Appendix. A Refresher

This appendix gives a condensed summary of the basic terminology used throughout the book. To the extent possible it is organised in the order of the chapters, indicating how they are connected. Not all the terms are standard or used in the same way as in the statistical literature. The appendix can be used as a glossary, although the principal definitions are accompanied by motivating examples.

A.1 Populations and Variables

We define *statistics* as the study of the values of variables on the members of populations. Any collection of units can be regarded as a population. Formally, a *population* is defined by a rule that arbitrates without any ambiguity, about any entity, as to whether it does or does not belong to (is a member of) the population. For instance, the population of the residents of a country is defined by a qualification stipulated by the relevant laws of the country. Such a population has to be associated with a date, to resolve the membership of those who were born or died, emigrated, immigrated, or qualified for residence by some other means around the designated date. The rule may be revised from time to time. A population need not comprise human subjects or other living organisms. Moments in time, repeated operations (e.g., in a production process), locations, or computer records may form a population, as can organisations defined by human subjects, such as companies, households, schools, and (local) administrative authorities.

A *variable* is defined on a population by its value for each member. Instead of these values, the variable may be defined by a procedure that establishes its value for each member. For instance, the income of a resident of a country is defined as the sum of all the payments received by the member in a given period of time. More details may be given to classify the payments into categories, such as income from employment, investments, pension, rents, sale of property, winnings in games of chance, and the like. The details of a definition

of a population or a variable that are essential to remove any ambiguity but are not listed every time we refer to the population or the variable are called the *small print*.

The *support* of a variable is defined as the set of all values that occur for the variable. The values of a variable may be counts (integers), numbers, categories, lists, or (unordered) sets of objects. They define the *type* of the variable. A variable is said to be *categorical* if its support comprises a finite number of values. These values may be associated with ordering, such as for the integers from one to six. Such a variable is called *ordered categorical*; its support consists of ordered categories. An unordered categorical variable is also called a *factor*. A variable is said to be *discrete* if its support comprises isolated values; around any value x in the support there is a neighbourhood that contains no value other than x. A variable is said to be *continuous* if its support contains no isolated values; any neighbourhood of any value x in the support contains at least one other value that belongs to the support. This definition is revised in Section A.3.

These definitions imply that the support is a subset of a space in which certain structures and operations are defined. For example, ordering is an operation; it assigns to each pair of values their comparison (the same, greater than, or smaller than). Whether two values are the same or not can also be regarded as a (trivial) operation. Neighbourhoods of points define a structure. Neighbourhoods are commonly defined by a *metric*. A metric is an operation that assigns to each pair of values (points) in the space their distance. The distance is nonnegative; $d(x_1\,,\,x_2) \geq 0$ for any pair of points x_1 and x_2 in the space. The only point in the distance of zero from any given point is the point itself; $d(x_1\,,\,x_2) = 0$ only when $x_1 = x_2$. The distance is symmetric, $d(x_1\,,\,x_2) = d(x_2\,,\,x_1)$, and satisfies the triangular inequality:

$$d(x_1\,,\,x_3) \leq d(x_1\,,\,x_2) + d(x_2\,,\,x_3)$$

for any three points $x_1\,,\,x_2\,,$ and $x_3\,.$ We can define the size of a value by its distance from a common reference point, called the *origin* and denoted by $0\,;$ that is, $s(x) = d(x, 0)$. The origin (its existence and location) is an element of the structure. A space is said to be *bounded* if there is a positive number M such that the distance between any two points in the space is shorter than M. The triangular inequality implies that a space is bounded only when there is an upper bound on the size of the values. If there is no such bound the space is said to be *unbounded*. Bounded and unbounded support are defined similarly.

Usually several variables are defined in a population. From one or several such variables, new variables can be defined by transformations, using operations that are well defined in the supports of the variables concerned. For instance, when the values of a variable are real numbers new variables can be defined by the usual arithmetic operations.

Clustering is a commonly occurring structure in populations of human subjects. For example, the members of a family, each of them also a member

of the population, form a cluster. Clusters may be nested, such as families (households) within streets, towns or villages, and districts, or cross-classified, such as families and birthplaces. Other structures can be defined by the values of one or several variables. For instance, the geographical location of a member of the population can be indicated by a categorical variable (place name) or even by its coordinates (latitude and longitude, both continuous variables).

Variables are defined because their values provide useful descriptions of the members of the population. For large populations, containing tens of thousands or even millions of members, a list of the values of a variable is not very useful for learning about the population; the values require some processing and summarising. This often takes the form of calculating certain *summaries* of the values. Examples of such summaries are the mean (average), range (the difference between the maximum and minimum value), the fraction of the members whose value exceeds a given threshold, and the proportion of the members who have a particular value. Such summaries are popularly referred to as *information*. A summary need not be a single number; it may comprise several numbers, although not many, because a summary is intended as an easy-to-digest, even if not comprehensive, description of the population.

Although it is usually derived from a single variable, a summary may involve several variables. For example, the proportion of members whose value of one variable exceeds the value of another variable is derived from two variables. However, this proportion depends only on the difference of the two variables. By defining this difference as another variable, the summary depends only on this new (constructed) variable.

An elementary task in statistics is associated with establishing the value of a population summary of a variable. This value could be determined by *enumeration*—by establishing the value of the variable on each member of the population and then evaluating the summary. We regard the task of evaluating a summary as elementary, requiring only minimum effort and expertise, *if* all the values are available. The principal difficulty is that enumeration requires resources, such as labour, equipment, services (including transport and telecommunications), and time, and therefore funding, and these are usually insufficient for an enumeration. Cooperation of the studied population, their goodwill, is another important resource.

Collecting information from every member of a large population is often a singularly unreasonable proposition, from the perspectives of both the member of the population (*respondent*) and the *consumer* of the information. The consumer associates the required information with a financial, ethical, professional, or some other benefit (value). They would be willing to finance, and assist by other means, the effort of collecting the information if the investment (expenditure) they make was recovered by the outcome—by valuable information that would facilitate the conduct of their business, such as governing the country (by adjusting policies and incentives), production and distribution for the retail trade, and location of service outlets. Instead of enumeration, the values of the variable of interest could be collected on only a subset of the

population. Such a subset is called a *sample*. The members of the population who belong to the sample are called *subjects*. The number of subjects in the sample is called the *sample size*. The value of the summary of interest, called the *target*, could not be established with precision but, hopefully, a summary of the sample would not be far off. Thus, the expense is reduced, but precision is sacrificed in the process.

The cost can be further reduced by establishing the value of the variable not precisely but subject to some approximation. For instance, instead of asking for a complete list of food and drink consumed in a given period of time (say, in a week), a questionnaire would inquire merely about the frequencies of eating certain kinds of food and consuming beverages in a short list of categories (types of food and drink). In this way, less detail is collected, but the exercise of eliciting information from the subjects is made easier and less intrusive.

In this description, we can readily identify two activities: selecting a sample (*sampling*) and eliciting the value (*measurement*). They are referred to as *processes*, because they are defined not by the selected sample and the recorded values, respectively, but by how they would be applied (methods) in any conceivable instance. Examples of these processes are all the adult human passengers on a selected list of rail services (date and number of the service) who are not employees of the railways, and requesting the subjects to complete a particular questionnaire that inquires about their experiences as railway passengers in the last few months. With this sampling process, members of the population who use rail services infrequently are less likely to be included in the sample than those who travel by rail frequently.

Ideally, we would like to draw (select) a sample in which the country's regions, age groups, occupational categories, and other attributes of the members of the studied population are represented in proportions that resemble their composition in the country. Similarly, a more elaborate process of measurement, with more detailed and clearly formulated questions, may be more useful than the responses to a single ambiguous question for which there is a limited set of response options, such as, at the extreme, only 'Agree' and 'Disagree'. More detailed questioning takes longer and detains the respondent for longer; it requires more preparation, instruction, and training of the interviewers and, as a result, a sample with fewer subjects (a smaller sample) can be afforded for the fixed resources available. Thus, higher quality of the measurement process may not serve well the primary purpose of the survey.

From the values of a variable recorded, possibly not precisely, on a sample of subjects, we cannot establish the value of the target; we can merely make a guess based on the available values and informed by the details of the sampling and measurement processes. Such a guess is called an *estimate*, and the process of deriving it is referred to as drawing (making) an *inference*. The process may be described by a mathematical formula, a verbal description, such as 'the proportion of subjects who responded with "Yes"', or it may be implemented in a computer program. The process (or procedure) by which the estimate is

evaluated (calculated) is called an *estimator*. A typical estimator is intended for a specific target. A desirable property of an estimator is that it is close to the target. The difference between estimates (numbers) and estimators (procedures) will become clearer in Section A.2.

Among the values and summaries defined so far, we can distinguish between population and sample quantities. A summary and a target are examples of population quantities; they can be established only when the values of the relevant variable are available for every member of the population. An estimate and the sample size are examples of sample quantities; they can be established from the values of the variable on the sample, after one application, or *realisation*, of the sampling and measurement processes.

With the terms defined so far, we can specify the role of statistics as making inferences about population quantities related to variables, when the resources available for these activities are limited. This entails specifying the processes of sampling, measurement, and estimation that yield the best inference that can be afforded with the available resources. To solve this problem, we have to agree first on what to regard as 'best' inference. Next, we require formulae for the cost of executing any considered sampling and measurement processes. The estimation process can also be associated with a cost, although it is usually fixed and trivial in comparison with the expenditure on the sampling and measurement processes.

A.2 Replications and Randomness

Replication is a key device for comparing alternative sampling, measurement, and estimation processes (*schemes*). Replication is the act of repeating (repeatedly applying) a set of processes, doing so each time without being affected in any way by the previous applications. Replications are *independent* applications of the same scheme. The outcome of a replication is called a *replicate*. Thus, we talk about replicate samples, replicate measurements, and replicate values of an estimator.

We assume that the estimator is perfectly replicable. That is, its application on the values of a variable for a given (*fixed*) set of subjects always yields the same estimate. In general, replicate estimates are not constant (are dispersed) because replications of the sampling process yield different sets of subjects, and they have different values of the observed variable. The replications of the measurement process might yield different values of the variable even if the same sample were drawn in the replications, or the measurement process were replicated on the entire population. The sampling and measurement processes involve randomness; we say that they are *stochastic*. Note that the sampling process cannot be replicated in practice; resources are usually available only for one application. Nevertheless, in some circumstances at least, we can discuss what results would be obtained in a long sequence

of replications. In Section A.8.1, we discuss a general method for generating replications on the computer.

Replicate measurements on the same subject are not constant because the measurement process is affected by the idiosyncrasies of the measurement instruments or agents (interviewers), momentary distractions influencing the respondents (subjects), and imperfect communication between the respondent and the interviewer. In some settings, measurements can be replicated (on the same set of subjects), especially when they leave no trace on the subject, are not costly to conduct, and have no ethical consequences. Such replications enable us to learn about the quality of the measurement.

Measurements are difficult to replicate when the subject or the interviewer can recall, even if only partially, the previous measurement. For instance, a school examination would be very difficult to replicate, especially if the same questions were presented in the second version of the exam. Nevertheless, we can speculate how different the results would be if a replication took place, with students unaffected by the experience of having taken the same exam in the past. We say that such a replication is *hypothetical*. When the idiosyncrasy of the measurement process is mainly due to the interviewer, his or her *assessment*, independence of the measurements can be ensured by engaging different interviewers who are not informed about each other's assessments or workloads (which subjects they assessed).

A.2.1 Efficiency

An estimator in a particular scheme is said to be *efficient* if its values obtained by replications (replicate estimates) are tightly concentrated around the target. To assess how close an estimate is to the target, we need a measure of its distance, the *deviation* of the estimate from the target. The difference of the estimate from the target is the obvious choice, although the sign of the difference is immaterial for the assessment of the size of the deviation. For an estimator in a scheme, represented by its replicate values, it is necessary to summarise its deviations from the target. Two important summaries, capturing two aspects of the deviations are *bias* and *dispersion*.

The bias is defined as the average deviation, with the sign of the deviation not ignored. The dispersion of an estimator is defined as the spread of its values around its mean. No bias, or being *unbiased*, and having small dispersion are desirable properties of an estimator. However, small bias is of little value if it is accompanied by large dispersion, and small dispersion is not useful if the bias is very large. Figure A.1 illustrates this with four examples of combinations of small and large bias with small and large dispersion. In each panel, the diagram, called a *histogram*, comprises bars. The height of each bar is proportional to the number of replicate estimates that fall into the (horizontal) range covered by the bar. Each histogram is based on 10 000 replications. The target is marked by a vertical line. The four panels have the same horizontal and vertical scales.

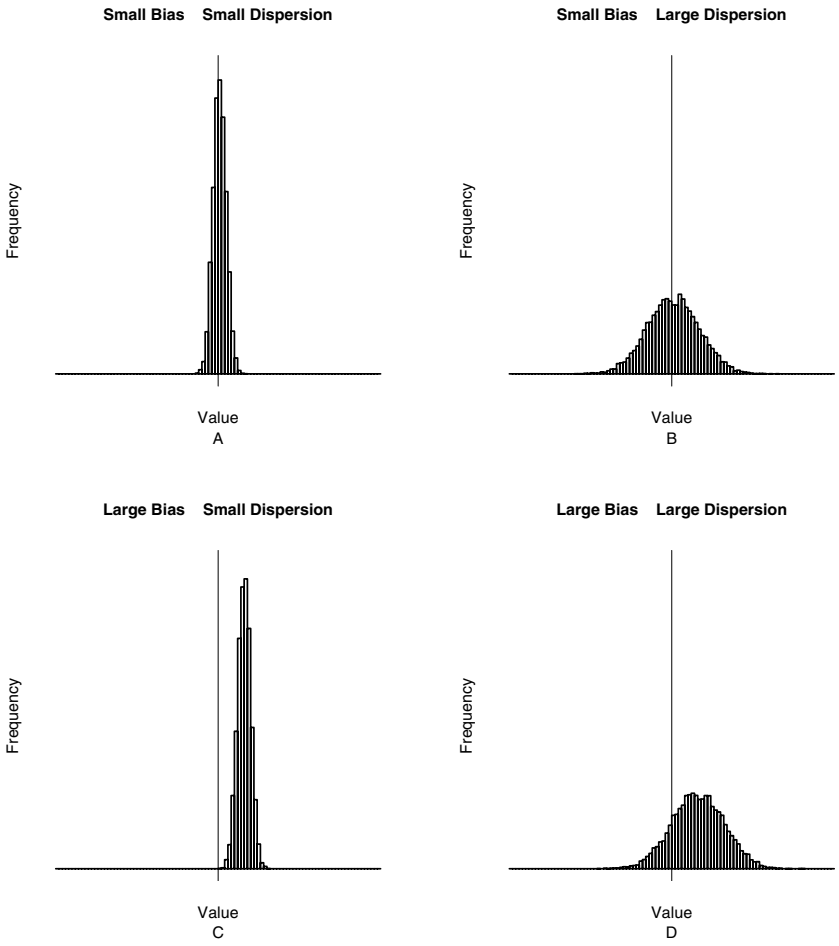

Fig. A.1. Histograms of replicate estimates for estimators with small and large biases and dispersions. The target is marked by a thin vertical line in each panel. The horizontal axes of the four histograms have the same scale.

The measurement process can be considered similarly, with the genuine value of the variable for a subject regarded as the target. The ideal measurement process recovers the target value in each replication. Otherwise, replicate measurements concentrated more tightly around the target are preferred. Note that the value of the variable for each member of the population is a potential target, and so the properties of the measurement process have to be considered for all members. The replicate measurements may be constant for some or all members of the population, and they may agree with the target for some members. The variable for which the value cannot be recovered with

precision is called *latent*. The variable that is recorded in its stead is called *manifest*. A latent variable may have several manifest versions, defined by different measurement instruments, or other circumstances (settings or small print) of the measurement process. Sets of replicate measurements can be regarded as separate variables.

Before defining a criterion for efficiency, of an estimator or a measurement process, which combines small bias and small dispersion, we introduce some notation.

A.3 Notation

The population is denoted by $\mathcal{P}$ and its members by integers $i = 1, 2, \ldots, N$. The number of members of the population, N, is called the *population size*. It need not be known but, to avoid some complications, we assume it to be finite, until specified otherwise. The values of a variable on the members of the population are denoted as $X_1, \ldots, X_N$, and the variable, or its value on an unspecified member, is denoted by X. It is practical to denote the collection of these values by a single symbol, $\mathbf{X}$; that is, $\mathbf{X} = (X_1, X_2, \ldots, X_N)^\top$. Any variable defined in a finite population is discrete because it cannot have more than N distinct values. However, when the number of unique values in $\mathbf{X}$ is large (then necessarily so is N), and any point on a continuum, such as a real interval, could, in principle, be a value of the variable, it is more appropriate to regard the variable as continuous. For example, income of the members of the labour force of a country is a continuous variable because any positive value, within a range, could be someone's income. Income is rounded to the smallest unit of currency, and so, strictly speaking, it is a discrete variable. However, it will turn out to be more constructive to regard it as a continuous variable.

The sampling and measurement processes are denoted by $\mathcal{S}$ and $\mathcal{M}$, respectively. The sample is denoted by $\mathbf{s}$, the number of its elements (subjects) by n, the subjects by $j = 1, \ldots, n$, and the values of the variable on the subjects by $x_1, \ldots, x_n$, or as $\mathbf{x}$. Note that (sample) subject $j = 1$ is distinct from (population) member $i = 1$, and their respective values x_1 and X_1 are not related in any way other than both being one of the N values in $\mathbf{X}$.

A population quantity, such as a target, is denoted by θ. For instance, θ may stand for the mean of a variable in a population. As the mean can be calculated for any numerical variable, a more complete notation includes the variable involved: $\theta(\mathbf{X})$. The 'same' variable may be defined in another population, and so the population may be added as another argument of θ, in addition to $\mathbf{X}$; $\theta(\mathbf{X}; \mathcal{P})$. As in most cases we work with a single population, this is not necessary. We regard two variables as different if they are defined in different populations, even if their descriptions are the same. In other words, the population is part of the small print in the definition of a variable.

An estimator of θ is denoted by $\hat{\theta}$ or $\hat{\theta}(\mathbf{x})$, although a more rigorous notation would include the sampling and measurement processes as arguments. The measurement process may be subsumed in the definition of the variable X. Then the values of one variable (on a sample of subjects) are used for making inferences about the summary of another variable. The sampling and measurement processes cannot be recognised from the sample values $\mathbf{x}$. That is, a particular sample $\mathbf{x}$, a set of n values, could be realised by several distinct pairs of sampling and measurement processes.

The replicate samples are denoted by $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(H)}$, where H is the number of replicates. Each of these samples is associated with an estimate, and these are denoted by $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(H)}$, or, more completely, as $\hat{\theta}^{(1)} = \hat{\theta}\left(\mathbf{x}^{(1)}\right), \ldots, \hat{\theta}^{(H)} = \hat{\theta}\left(\mathbf{x}^{(H)}\right)$, emphasising that we use the same estimator $\hat{\theta}$. The *expectation* of an estimator $\hat{\theta}$ is defined as the mean of the estimates in a large number of replications, that is, as

$$\frac{1}{H}\left(\hat{\theta}^{(1)} + \cdots + \hat{\theta}^{(H)}\right)$$

or, more precisely, as the limit of this expression with H diverging to infinity ($H \to +\infty$). The expectation of $\hat{\theta}$ is denoted as $\mathrm{E}(\hat{\theta})$. The expectation depends on the sampling process. We add the sampling process $\mathcal{S}$ to the notation, as $\mathrm{E}(\hat{\theta}; \mathcal{S})$ or $\mathrm{E}_{\mathcal{S}}(\hat{\theta})$, for emphasis or when we operate with several sampling processes.

The bias of $\hat{\theta}$ is denoted by $\mathrm{B}(\hat{\theta}; \theta)$:

$$\mathrm{B}\left(\hat{\theta}; \theta\right) = \mathrm{E}\left(\hat{\theta} - \theta\right).$$

It is essential to retain the target θ as an argument of B because an estimator may be used for more than one target; it may be unbiased for one target, and biased for another.

An obvious candidate for the estimator of a population quantity $\theta = \theta(\mathbf{X})$ is the same function of the sample values: $\hat{\theta} = \theta(\mathbf{x})$. For instance, the population mean may be estimated by the sample mean. Such estimators are called *naive*. (The term is not intended to be derogatory.) Note that θ has to be well defined for both N values in the population quantity $\theta(\mathbf{X})$ and n values in the estimator (sample quantity) $\theta(\mathbf{x})$. In fact, many estimators have to be similarly flexible, because the (replicate) samples need not have constant size n.

The *sampling variance* of an estimator $\hat{\theta}$ is defined as the expectation of the squared deviation of $\hat{\theta}$ from its expectation $\mathrm{E}(\hat{\theta})$:

$$\mathrm{var}\left(\hat{\theta}\right) = \mathrm{E}\left[\left\{\hat{\theta} - \mathrm{E}\left(\hat{\theta}\right)\right\}^2\right].$$

The mean squared error (MSE) of an estimator $\hat{\theta}$ is defined as the expectation of its squared deviation from the target θ:

$$\mathrm{MSE}(\hat{\theta};\, \theta) \;=\; \mathrm{E}\left\{ \left(\hat{\theta} - \theta \right)^2 \right\} .$$

The sampling variance and MSE depend on the sampling and measurement processes, and the MSE depends also on the target. The MSE, sampling variance, and bias are connected by the identity

$$\mathrm{MSE}\left(\hat{\theta};\, \theta\right) \;=\; \mathrm{var}\left(\hat{\theta}\right) + \left\{ \mathrm{B}\left(\hat{\theta};\, \theta\right) \right\}^2 . \tag{A.1}$$

Thus the sampling variance and the squared bias are two contributors to the MSE. An estimator with small MSE cannot have a large bias or a large sampling variance. We adopt the MSE as a measure of efficiency. Suppose $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$ are estimators intended for the same target θ. Then $\hat{\theta}_{\mathrm{A}}$ is said to be more efficient than $\hat{\theta}_{\mathrm{B}}$ for θ if $\mathrm{MSE}(\hat{\theta}_{\mathrm{A}}\,;\, \theta) < \mathrm{MSE}(\hat{\theta}_{\mathrm{B}}\,;\, \theta)$.

The MSE is an example of a *sampling-process quantity*. It characterises the sampling and estimation processes engaged. Except for some simple cases, it can be established only by replicating the sampling process many times. Usually, the MSE (of an estimator $\hat{\theta}$ for a target θ) depends on some population quantities, often the target itself, and so the MSE can itself be regarded as a target and estimated. As the MSE depends on some unknown (population) quantities, we may consider properties of the estimator $\hat{\theta}$ assuming specific values of these population quantities. One estimator of θ is said to be *uniformly more efficient* than another estimator of the same target if it is more efficient for any configuration of the population quantities on which their MSEs depend.

Estimators of a target may have strengths and weaknesses; they may be more efficient than their competitors for some configurations of population quantities and less efficient for others. When striving to choose an efficient estimator, information, however incomplete, about the relevant population quantities is sometimes invaluable; it can assist in discarding estimators that are inefficient for the particular setting.

A.4 Distributions

When studying the values of a variable in a population, we are usually not interested in the identities of the members; we say that the members are *anonymous*. Each member has a unique identifier. It is useful for tracing the various steps in the construction of the dataset and for connecting the values of two (or more) variables of a member. Given the values of all the defined variables for a member, the member's identifier has no information content and we treat it as a mere label.

We often wish to summarise a variable by how frequently certain values, and their ranges, arise. Examples of such summaries are:

- What proportion of the households in a country have income below a certain level?
- How many households comprise a single person each?
- How many students fail a particular examination?

To address the first of the listed questions, we define a new variable, U, equal to unity (or 'Yes') for members whose answer to the question

Is your household's income below £...?

is affirmative, and equal to zero (or 'No') if the answer is negative. The summary of interest, the proportion of 'Yes', is equal to the mean of the variable U. Such a *population proportion* is called a probability. We write $\mathrm{P}(U = 1)$ for this probability, but also as $\mathrm{P}(X < c^*)$, where X is income and c^* the value of the threshold income in the question.

The *distribution* of a variable X with real values is defined as any collection of probabilities from which the probability $\mathrm{P}(X < c)$ could be recovered for any real value c. Of course, such a collection is not unique. For instance, $\mathrm{P}(X < c)$ for every value c that occurs in the population is a distribution, but so is $\mathrm{P}(X > c)$ for every such c, or indeed $\mathrm{P}(X = c)$, so long as the number of distinct values c of X is finite. As every population we consider has a finite population size, the number of distinct values of the variable in the population is finite. By definition, it can be established whether two collections of probabilities correspond to the same set of probabilities $\mathrm{P}(X = c)$ for values c in their supports. If they do, it is practical to regard them as identical distributions. With this convention, the distribution is uniquely defined, and any conceivable probability involving X, such as $\mathrm{P}(X = c_1 \text{ or } X = c_2 \text{ or } \ldots \text{ or } X = c_K)$ can be derived by adding up the relevant probabilities $\mathrm{P}(X = c_k)$, $k = 1, 2, \ldots, K$, so long as the values $c_1, \ldots, c_K$ are distinct.

The distribution of a variable often comprises many probabilities, so it cannot be effectively presented in any tabular form. The distribution of a variable can be presented graphically by a *histogram*. An example is presented in Figure A.2 for a variable with 400 distinct values in a population of size 25 000. The vertical segments represent the distinct values and the height of each segment is equal to the *frequency*—how many times the value occurs in the population. Note the similarity in the layout with the histograms in Figure A.1. Part of the distribution is presented in tabular form in Table A.1, giving the frequencies of the 15 smallest values of the variable.

Some of the detail in Figure A.2 is unnecessary. The segments that are very close to one another could be represented by a single segment, or a bar, as in Figure A.1. Two examples of this are given in Figure A.3. Either histogram conveys much better that the most frequent values are around zero, all the values are nonnegative, very few values exceed 7.0, there are fewer values in the neighbourhood of 3.0 than elsewhere in the support, and so on. The vertical axis in Figure A.1 is in fractions (probabilities), whereas in Figure A.2 it is in counts (numbers of members). This has no impact on what we can

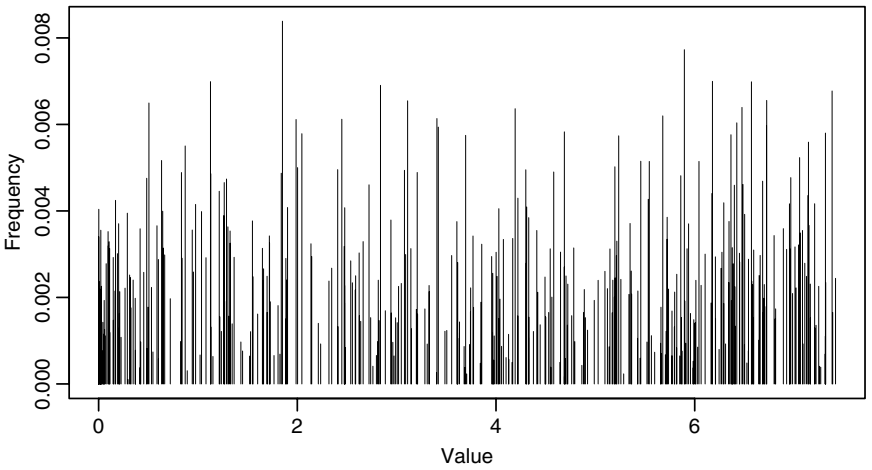

Fig. A.2. Histogram—graphical representation of the distribution of a variable.

Table A.1. The distribution of the variable in Figure A.2 (an extract).

Value	*Frequency*	*Value*	*Frequency*	*Value*	*Frequency*
0.0004	15	0.0035	52	0.0128	71
0.0008	52	0.0042	52	0.0160	155
0.0009	87	0.0048	43	0.0165	134
0.0022	62	0.0059	80	0.0231	23
0.0025	250	0.0115	18	0.0242	73

learn about the distribution; that is, the same information could be extracted from the diagrams with either layout.

The histogram in panel A is more detailed and the histogram in panel B somewhat coarser. The coarseness is given by the width of the bars or by the number of bars that cover the entire range of the values, in this example, from zero to 7.41. The more detailed histogram in panel A has 50 bars and the histogram in panel B 20 bars. Each histogram is associated with a table that lists the range of each bar with the corresponding frequency. Table A.2 presents such a table for the histogram in panel B.

A.4.1 Describing Distributions

Although we can reconstruct from the distribution most of the important facts about a variable, we do not always need to convey all the details of

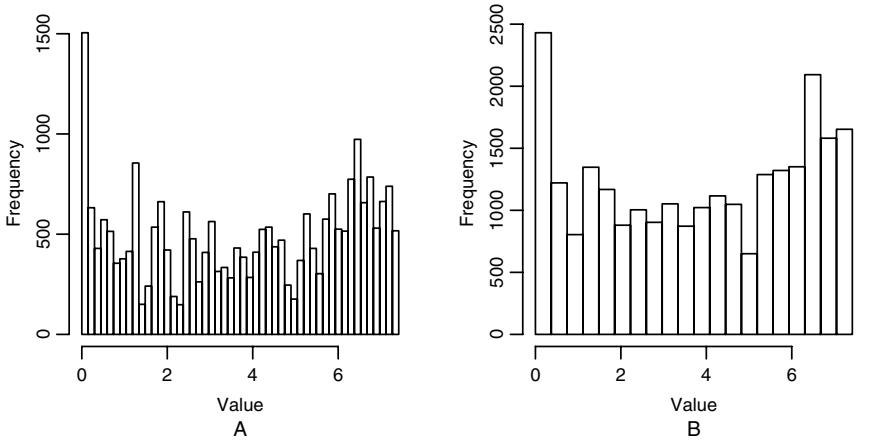

Fig. A.3. Coarse histograms of the variable in Figure A.2.

Table A.2. The ranges and frequencies associated with the histogram in panel B of Figure A.3.

Range	*Frequency*	*Range*	*Frequency*	*Range*	*Frequency*
0.000–0.370	3819	2.594–2.964	566	5.187–5.558	1218
0.370–0.741	811	2.964–3.334	914	5.558–5.928	1506
0.741–1.111	1412	3.334–3.705	1341	5.928–6.298	1503
1.111–1.482	616	3.705–4.075	709	6.298–6.669	1202
1.482–1.852	1046	4.075–4.446	1112	6.669–7.040	1478
1.852–2.223	345	4.446–4.816	1638	7.040–7.410	2066
2.223–2.594	721	4.816–5.187	977		

the distribution. Coarse histograms and the associated tables of frequencies condense the information about the distribution and present it in a form that is easy to digest. Often it useful to have a single-number or a succinct verbal description of a particular feature of the distribution. In this section, we define a few such features.

Any summary of the values of a variable that can be derived directly from the distribution is also a summary, or feature, of the distribution. For instance, the population mean of a variable can be expressed as

$$\mathrm{E}(X) \;=\; \frac{1}{N}\left(M_1\, C_1 + M_2\, C_2 + \cdots + M_K\, C_K\right) ,$$

where $M_1, \ldots, M_K$ are the frequencies (*multiplicities*) of the respective (unique) values $C_1, \ldots, C_K$ of X in the population, or as

$$\mathrm{E}(X) = C_1\,\mathrm{P}(X = C_1) + C_2\,\mathrm{P}(X = C_2) + \cdots + C_K\,\mathrm{P}(X = C_K)\,,$$

where $\mathrm{P}(X = C_k) = M_k/N$. Note that a more rigorous notation would use M_{C_k} instead of M_k, to associate the multiplicity with the value C_k, not with its order k.

Location Quantities

A population quantity is said to be a *location quantity* if it is a summary that involves one variable and adding a constant to or changing the scale of the variable corresponds to the same change of the quantity. That is, if d is a location quantity of X, then, for any given values (constants) a and b, $ad+b$ is the location quantity of the (linearly transformed) variable $aX+b$, formed by changing each value X_i to aX_i+b. This defining property of location quantities is also referred to as *invariance* with respect to linear transformations. Apart from the mean, the minimum and maximum are obvious location quantities.

The (population) *median* of a variable, or of a distribution, is defined as the value that is exceeded by exactly half the members of the population. For example, in a population that comprises $N = 41$ members, the median is equal to the 21st highest value of the variable. When the population size N is even, the median is not always unique. For example, any value between the 20th and 21st highest value is a median in a population of 40 members. If these two values coincide, then the median is unique. Otherwise, we may choose as the median the mean of these two values. If either of these values occurs more than once the weighted mean of the values may be used, with weights equal to the frequencies. The median of the distribution in Figure A.2 is 4.200; in this case it is a value that occurs in the population, for 131 members, so the median is unique, even though the population size N is even.

The upper quartile of a variable or distribution is defined as a value that is exceeded by exactly 25% of the values, and the lower quartile as a value exceeded by exactly 75% of the values. For the distribution in Figure A.2, these quartiles are 1.218 and 5.915. Both values occur in the population multiply, so both quartiles are unique.

More generally, for any number q between zero and unity, the q-quantile is defined as a value R_q for which $\mathrm{P}(X < R_q) = q$. In a more complete notation, we would write $R_q(X)$ instead of R_q, because the quantile depends on the values of the variable. We drop the argument X only when there is no ambiguity about the variable on which the quantile is evaluated. The p-percentile is defined as the $p/100$-quantile. For example, an upper quartile is a 0.75-quantile and a 75th percentile of the distribution. A particular quantile may not be unique. When it is not, any point in the interval between two consecutive values of the variable is this quantile, or a convention for averaging or weighting of the adjacent values may be adopted. With any convention that makes the quantiles unique we can refer to any particular quantile as *the* quantile. The quantiles and percentiles are location quantities. They have a

general invariance property that for any increasing function g, $R_q\{g(X)\} = g\{R_q(X)\}$; swapping the operations 'quantile' and 'function' does not alter the result.

A compact, though incomplete, description of a distribution is by the values of the minimum, lower quartile, median, upper quartile, and the maximum, possibly supplemented by the mean. For example, these values for the distribution in Figure A.2 are

$$(0.000,\ 1.218,\ 4.200,\ 5.915,\ 7.408)$$

and the mean is $\mathrm{E}(X) = 3.753$. The minimum can be regarded as the 0-quantile and the maximum as the 1-quantile of the distribution.

Dispersion Quantities

A population quantity, defined for a variable or a distribution, is called a *dispersion quantity* if it is unchanged when a constant is added to each value of the variable and is multiplied by $|\,b\,|$ when each value is multiplied by a constant b. The difference of any two quantiles (higher quantile – lower quantile) is a dispersion quantity, as is the *range*, the difference between the maximum and minimum. The difference between the two quartiles, $R_{0.75}(X) - R_{0.25}(X)$, is called the *interquantile range*.

The population variance is defined as the mean squared distance of the values from their mean:

$$\mathrm{var}(X) \;=\; \mathrm{E}\left[\{X - \mathrm{E}(X)\}^2\right] \;;$$

compare this with the definition of the sampling variance in Section A.3. When the context is insufficient to distinguish between the two kinds of variance the notation can be supplemented by subscripts to indicate whether a variance is over sampling or population, $\mathrm{var}_{\mathcal{S}}$ and $\mathrm{var}_{\mathcal{P}}$, respectively. The square root of the population variance, $\sqrt{\mathrm{var}_{\mathcal{P}}(X)}$, is called the *standard deviation*. The standard deviation is a dispersion quantity.

Symmetry and Unimodality

A distribution is said to be *symmetric* if it coincides with its reflection across the (suitably defined) median, that is, when the distributions of X and $2R_{0.50}(X) - X$ coincide. An example of a symmetric distribution is given in Figure A.4.

The mean and median of a symmetric distribution coincide; $\mathrm{E}(X) = R_{0.50}(X)$. Further, for any $0 \le q \le 1$, the q- and $(1-q)$-quantiles are equidistant from the median:

$$R_q(X) - R_{0.50}(X) \;=\; R_{0.50}(X) - R_{1-q}(X)\,.$$

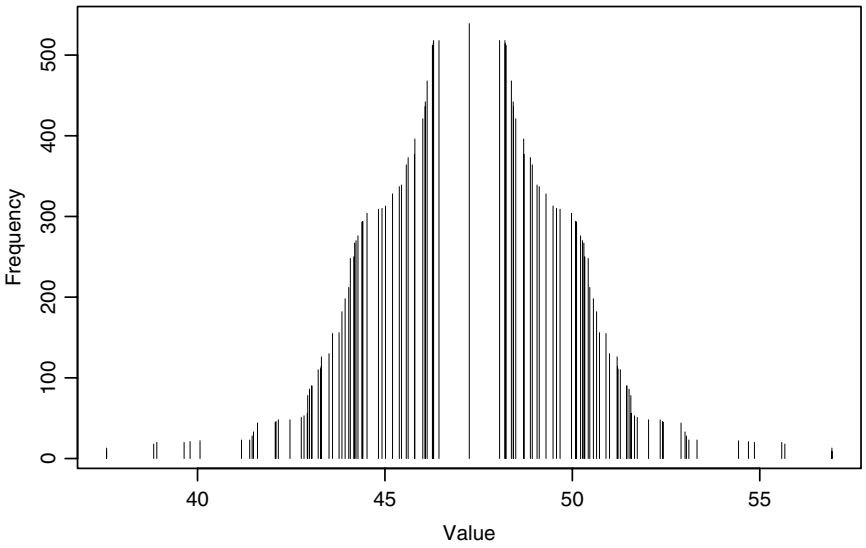

Fig. A.4. Example of a symmetric distribution.

A distribution is said to have a *mode* at a value X^* if both the nearest value smaller than X^* and the nearest value greater than X^* have smaller frequencies. Every distribution has a mode, but some distributions have several modes. For example, the distribution in Figure A.4 has one mode, at its median, but the distribution in Figure A.2 has numerous modes. The modes of a symmetric distribution are located symmetrically around the median. If X^* is a mode, then so is $2R_{0.50}(X) - X^*$. A distribution with a single mode is called *unimodal*, with two modes *bimodal* and, generally, with more than one mode as *multimodal*.

A.4.2 Approximating the Distribution by a Histogram

The graphs of the distributions in Figures A.2 and A.4 are rather unwieldy and contain too much detail that may not be relevant and would be better omitted from a compact summary. One way of achieving this is by *rounding* the values of the variable. The resulting variable can be regarded as a manifest, or *coarsened* version of the original variable. The distribution of the coarse variable is simpler because the variable has fewer possible values and these (the support) are located regularly.

In general, a coarsening is defined by a set of cut points $c_0 < c_1 < \ldots < c_K$ and values $d_1, \ldots, d_K$, such that $c_{k-1} \leq d_k \leq c_k$, $k = 1, \ldots, K$. If the original value of X is in the range $(c_{k-1}, c_k]$ the value of the coarse variable is set to d_k. Instead of the coarse variable we can define a variable $X^\dagger$ with values

equal to the category k into which the original value falls: if $c_{k-1} < X \leq c_k$, then $X^\dagger = k$. When the original variable is equal to one of the cut points c_k, coarsening can allocate it either to d_k (category k) or to d_{k+1} (category $k+1$). A coarsening with cut points $c_0 < c_1 < \ldots < c_K$ is said to be coarser than with cut points $c'_0 < c'_1 < \ldots < c'_{K'}$ if the set of values $(c_0, c_1, \ldots, c_K)$ is a subset of the set $(c'_0, c'_1, \ldots, c'_{K'})$. Necessarily, $K < K'$. A coarsening can be refined by introducing new cut points (and defining appropriate new values d_k), and made coarser by discarding one or several cut points c_k (and defining new values d_k in the affected intervals).

For a given variable, such as the annual income of a household, we may consider a few alternative ways of coarsening that are ordered according to their coarseness. For example, the income could be rounded to units, tens, hundreds, or thousands of £UK. The choice of the coarseness should be guided by the purpose of the analysis (summary) to which the variable (income) is to be subjected. For example, if the difference of several hundreds of £UK is not important, rounding to thousands is appropriate. By a coarser rounding we obtain a variable that is easier to handle, because is has fewer possible values, but we may lose some detail in the process. In contrast, less coarse (finer) rounding yields values that are closer to (or the same distance from) the original values but may contain too much detail for an effective presentation and study.

A (coarse) histogram of a variable can be identified with the distribution of a coarsened version of the variable. The cut points of the coarsening applied coincide with the limits of the bars. Figure A.3 gave an example of the impact on coarsening of a continuous variable.

A.5 Sampling Design

When we cannot afford to enumerate the population, we establish the values of the target variable only for a sample of subjects. Such an exercise is called a *survey*. Every survey involves a sampling process by which subjects are selected. Substantial advantages accrue when we can select the sampling process purposefully. Such a sampling process is said to be *controlled* or planned and is referred to as a *sampling design*.

A sampling design can be defined by its (unambiguous) description, such as

1. Select a member completely at random.
2. Select one member completely at random from those not yet selected.
3. Repeat step 2 until the specified number of subjects has been selected.

More formally, a sampling design is defined as a way of assigning to every subset of the population the probability that it would form the sample. That is, the collection of all subsets of the population's members, denoted by $\exp(\mathcal{P})$, is regarded as a new population, and a probability in this population

is interpreted by a reference to replications. This definition does not seem to be constructive, because most populations have very many subsets, equal to 2^N, where N is the population size. In a typical sampling design, most subsets have zero probability of forming a sample. For instance, when the sample size is set (prescribed or fixed) to be n the number of possible samples is $\binom{N}{n} = N!/\{n!(N-n)!\}$. The controlled nature of a sampling design rests not on which member is selected into the sample but on how the selection is conducted.

Sampling designs in which each member has the same probability p of being included in the sample, each pair of members has the same probability $p^{(2)}$, and so on, are called *simple random*. Of course, $p \neq p^{(2)}$. A member can be included in the sample several times. Sampling designs in which this is possible are called designs *with replacement*, and designs in which this is ruled out are called *without replacement*. The term *replacement* refers to the description of the sampling design by a mechanism of drawing subjects into the sample one by one. In designs with replacement, after being selected, a subject is retained in the pool of candidates for being selected in subsequent draws. Thus, in simple random sampling design with replacement, each member has the same probability of being drawn as the first subject, equal to $1/N$, but, irrespective of who was drawn first, the probability of being drawn as the second subject is also equal to $1/N$ for every member of the population. The number of times a member of the population is included in the sample is referred to as its *multiplicity*.

A sample is most conveniently specified as a list of its subjects. The order of the subjects in such a list is immaterial; (i_1, i_2, i_3, i_4) and (i_1, i_4, i_3, i_2) are identical samples, even when $i_2 \neq i_4$. However, multiplicity is an important feature; when $i_2 \neq i_3$, (i_1, i_2, i_3, i_3), (i_1, i_2, i_3, i_2), and $(i_1, i_2, i_2, i_3, i_2)$ are different samples, even though each of them comprises the same set of subjects, i_1, i_2, and i_3.

From the sampling design, we can derive the probability that a given member is included in the sample by adding up the probabilities of all the subsets that contain the member:

$$p_i = \sum_{\mathbf{s} \in \exp(\mathcal{P})} I(i \in \mathbf{s}) \mathrm{P}(\mathbf{s}; \mathcal{D}),$$

where P denotes the probability as a function of the set $\mathbf{s}$ and design $\mathcal{D}$ and I is the indicator function, equal to unity when its argument is true and to zero otherwise. A sampling design $\mathcal{D}$ is called *proper* if each member has a positive probability of being included in the sample. Members who have zero probability of being included in the sample are, in effect, excluded from the population that is studied because they are not considered in the sampling process.

A.5.1 Complex Sampling Designs

Stratification and clustering are two ways of defining a wide range of sampling designs. For stratification, the population is divided into subpopulations (groups) called *strata*, and a different sampling design is applied in each stratum. The sampling processes in the strata are independent; the *subsample* drawn in one stratum has no impact on the subsample drawn in another. Stratification has two important advantages. First, the unwieldy task of drawing a sample from one large population is simplified to a number of simpler tasks of drawing a sample from each of several smaller subpopulations, and second, the design can exercise tighter control over the within-stratum subsamples. For example, the within-stratum sampling designs may sample much more densely (with relatively greater subsample sizes) in some strata, at the expense of sparser sampling in other strata.

In most large populations of practical importance, the members are related; such populations are said to be *structured*. The most common element of the structure is clustering—members form small groups (such as families or households), the groups are further clustered (say, to clusters at level 2, such as neighbourhoods or classrooms), and these groups may be further clustered (clusters at level 3, such as districts or schools). In a *clustered sampling design*, a sampling design is applied to the clusters (at a particular level), and independent sampling designs are then applied in each selected cluster. The design applied in a cluster may itself be clustered. Such designs are called *multistage clustered*. The clusters involved in the first round of clustering (say, districts) are called *primary sampling units*, clusters in the next round (say, neighbourhoods) *secondary* sampling units, and so on. Subjects are the *elementary-level* sampling units (elements), unless all members of the selected clusters at level 2 (or at another level) are included in the sample.

An advantage of clustered sampling design is that it is focussed; the sampling in some (randomly selected) clusters is dense, at the expense of no sampling in some other clusters. Dense sampling may provide more information about the associations within the clusters, such as similarity of the members' values of the observed variables. In a clustered sampling design, we have a cluster-level design for each level of clustering and within-cluster designs. Clustered designs are in general easier to organise and manage. When the cost of accessing a cluster is substantial it is more economic to collect data in fewer clusters, but to do so from more (or all) subjects in the selected clusters.

A.5.2 Sampling Frame

A sampling design can be constructed with purpose and implemented effectively only when some basic information about the population is available. The *sampling frame* is a list of all the members of the population. In ideal circumstances, a sampling frame is *complete*, containing all members of the

population (without any omissions); *exclusive*, containing no objects that do not belong to the population; and *nonredundant*, containing no duplicates. A sampling frame that satisfies these three conditions is said to be perfect. For stratification and clustering, it would also identify the relevant strata and clusters into which the members belong.

Construction of a perfect sampling frame for a large population is rarely feasible. Commonly, a sampling frame of clusters is used, with some information about the composition of each cluster. For instance, a sampling frame may comprise the country's districts, or smaller administrative units, and the population size of each cluster may be available, usually subject to some approximation, for instance, because is it based on a population register that is a few months out of date. When a clustered sampling design is planned, construction of the sampling frame may be reduced to the selected clusters. For example, clustering by schools is a practical proposition in a survey of students. The school enrollments (sizes) may be available from a previous year, informing the sampling design for the schools as clusters, and within-school subsampling frames are obtained only from the schools that have been selected into the sample.

A.5.3 The Planned and Realised Sampling Processes

The sampling design reduces the study of a population to operations applied to a sample—eliciting information from the subjects and processing their responses (or information recorded about them) to make inferences about the population. A sampling design is essential to ensure good *representation* (representativeness) of the population by the sample. Good representation means that, in replications of the designed sampling process, samples would tend to have features similar to the population and, as a consequence, efficient inferences could be made about the population quantities related to the observed variables. The requirement of good representation has to be qualified by the targets—the population quantities for which inferences are sought.

Without a sampling design, the sampling process may conspire to yield samples that present a distorted image of the population. The image would be distorted in many replications. The image (a feature) may be distorted even in a sample drawn by a well-chosen sampling design because a distortion cannot be ruled out. As an example, consider a population that comprises $N = 100$ members and a binary variable, with values of zero and unity for 50 members each. A simple random sampling design without replacement and with fixed sample size $n = 10$ may yield a sample in which each subject's value of the variable is equal to zero. The probability of this event is $\binom{50}{40} / \binom{100}{90} = (50 \times \ldots \times 41)/(100 \times \ldots \times 91) \doteq 0.0006$. Thus, even an extreme distortion is possible, but its probability is very small; it would be present in only a small fraction of replicate samples.

Without a sampling design, such a 'protection' would not be available. Some distortion in the sample may be introduced by the sampling design.

For instance, by using a stratified sampling design, the smallest regions of the country may be overrepresented in the sample. If these regions tend to have high values of the observed variable the sample will tend to contain more high values than what might be regarded as an appropriate representation of the country. However, such a 'misrepresentation' can be taken into account when the sampling process is known. Without a sampling design we do not know how the sample is likely to have been distorted.

Suppose $\theta(\mathbf{X})$ is a population quantity of interest (a target). Here, θ can be interpreted as a mathematical formula or a computer program. Good representation can be interpreted that θ applied on the sample $\mathbf{x}$, $\theta(\mathbf{x})$, would be close to $\theta(\mathbf{X})$. Of course, an adjustment is necessary when θ is a total, but this can be 'built in' to the definition of θ. At the outset, when the sampling design is specified, the sample is not yet available; only the process by which it is formed is specified. We say that at that point the sample is *random*. Any sample quantity is also random at that point; its value is not known, but its distribution could, in principle, be established, by replications. In particular, it is meaningful to discuss how a population quantity would be estimated.

As a result of applying the sampling design, a sample is drawn. It is referred to as the *realised* sample. With it or, more precisely, with the values of the relevant variable on the subjects, the selected estimator can be evaluated and an estimate obtained. If the sampling design is implemented as prescribed the survey is concluded by reporting the estimate. In practice, the analysis (calculations made on the realised sample) is more extensive—several estimators are evaluated and each estimator is associated with its estimated MSE. Other forms of inference may be conducted, such as evaluating confidence intervals; these are dealt with in Section A.20.

The good properties of an estimator are usually contingent on the sampling design. In large populations, most sampling designs are impossible to implement exactly as planned. The sampling frame is usually imperfect and some of the selected subjects may not be available for an interview or may exercise their right not to cooperate with the survey. As a result, the probabilities of the samples that could be realised are altered. The sampling design as a process is contaminated by an *imperfection* process. This 'contaminated' version of the sampling design is called the *realised* sampling process. We cannot refer to it as a sampling design because it is not under our control. Without a detailed description of the imperfection process, the probabilities of the possible samples for a realised sampling process are not known.

The estimator selected at the planning stage may be efficient when the planned and realised sampling processes coincide, but with the imperfections its properties may have been altered somewhat. The estimator does not have the properties it would have had had the planned sampling design been implemented perfectly. When the realised process deviates from the plan only slightly we can expect the 'realised' properties of the estimators to deviate from the 'planned' properties also only slightly. Hence the strong incentive to

reduce the difference between the planned and realised processes, even if it cannot be eliminated altogether.

A.6 Measurement Processes

This section deals with describing measurement processes. In general, we intend to make inferences about a variable X, but we can obtain or record (measure, elicit, or the like) only the values of a related variable Y. The measurement process can be motivated as the way in which a particular value of the latent (underlying) variable X is 'converted' (distorted) to the value of the manifest (observed or recorded) variable Y. The measurement process can either be described by the mechanism that distorts the value of X in the process of its measurement, or by the distribution of the differences $Y - X$. Instead of these differences, the ratios Y/X, their logarithms, $\log(Y/X)$, or the differences after some other monotone transformation, $f(Y) - f(X)$, may be considered. As an alternative, the distributions of Y may be described in each of the subpopulations defined by the unique values of X. Of course, this is not practical when X is continuous, unless these distributions have some features in common.

When X is a categorical variable and Y is an attempt to recover the value of X, we refer to a *misclassification* process. Such a process can be described by the table of the probabilities $\mathrm{P}(X = x, Y = y)$ for each pair of possible values x and y. A desirable property of a misclassification process is that the probability of agreement, $\mathrm{P}(X = Y)$, equal to the total of the probabilities $\mathrm{P}(X = x\,, Y = x)$ over the possible values x, is close to unity. When X is an ordinal categorical variable another desirable property is that when disagreement occurs, $X \neq Y$, it is frequently by only one point on the scale. For instance, when the possible values of X are $1, 2, \ldots, K$, we prefer a manifest variable Y for which $\mathrm{P}(\mid X - Y \mid \geq 2)$ is small.

Apart from the latent value X_i, the manifest value Y_i may depend on the values of some other variables. For instance, if the task of measurement is assigned to several judges, the identity of the judge assigned to assess a particular subject is a relevant (categorical) variable. The manifest value Y_i may be influenced even by the values of some variables, including X, for other subjects. For instance, a judge's assessment may be influenced by the other assessments made (recently) by the same judge, or by instructions received (or made aware of) halfway through the assessment process. Of course, it is wise to avoid such influences (by training and appropriately instructing the judges), but the process of measurement is not always under our control and training and instruction entail costs drawn from the same budget as the other survey tasks.

If the value of X can be established we can learn about the distortion $Y - X$ directly, by applying the measurement process on a sample of subjects. Otherwise, when the act of measurement alters the state of the subject at

most temporarily (in particular, it does not destroy it) and is not costly, we can learn about the measurement process by replicating it on subjects. Thus, we observe two (or more) versions of the variable Y, $Y^{(1)}$ and $Y^{(2)}$; the pairs may be observed on the entire sample of subjects, a subsample of the subjects, or an entirely different sample drawn from the same population. These two variables have identical distributions. Observations from other populations have to be considered with care because the properties of the measurement process may be specific to the population. By the same token, if repeated observations are made on a subsample of subjects, this subsample (just like the sample) should be representative of the population.

The variables $Y^{(1)}$ and $Y^{(2)}$ differ because they are affected by different settings, such as the assigned judge, observed circumstances that are beyond our control, such as the temperature and the environment in which the interview is conducted, and other inexplicable influences (circumstances) that defy our understanding. We may consider versions of Y associated with each conceivable set of circumstances (moments or *contexts*). These contexts can themselves be regarded as a population. Unlike populations considered so far, they may be *infinite*. To draw a clearer distinction, we refer to the population that is the original target of our inferences, as the *target population*, and to the contexts as an *incidental* or *nuisance* population. This qualifier reflects our position—if the context had no impact on the measurement, or indeed, if Y coincided with X, our task of making inferences about X would be simpler.

The properties of a measurement process are described by the distribution of the measurements on a member of the target population. The measurements are taken in the population of contexts. For each member i of the target population, we denote by $Y_i^{(\mathrm{m})}$ the variable defined as the manifest value in the population of contexts. The bias of the measurement $\mathrm{B}_i^{(\mathrm{m})}$ is defined as the expectation of the measurement deviations,

$$\mathrm{B}_i^{(\mathrm{m})} = \mathrm{E}\left(Y_i^{(\mathrm{m})} \,|\, i\right) - X_i\,,$$

taken over the population of contexts. We write i behind the bar $|$ to indicate that the expectation is taken with the member i fixed; the expectation is *conditional* on and relates solely to member i. The expression for bias requires a definition of the expectation, because E has so far been defined only for finite populations. The expectation for an infinite population is defined as the limit over sequences of increasing subpopulations, such that each member is eventually included in a subpopulation. The details are postponed to Section A.7. We considered similar limits in the context of replications of a sampling process in Section A.3.

The measurement variance and mean squared error (MSE) are defined similarly to the expectation:

$$\mathrm{var}\left(Y_i^{(\mathrm{m})} \,\middle|\, i\right) = \mathrm{E}\left[\left\{Y_i^{(\mathrm{m})} - \mathrm{E}\left(Y_i^{(\mathrm{m})}\right)\right\}^2 \,\middle|\, i\right],$$

$$\mathrm{MSE}\left(Y_i^{(\mathrm{m})};\, X_i \,\middle|\, i\right) = \mathrm{E}\left\{ \left(Y_i^{(\mathrm{m})} - X_i\right)^2 \,\middle|\, i \right\};$$

they coincide for unbiased measurement processes, when $\mathrm{B}_i^{(\mathrm{m})} = 0$ for each member i.

Conditioning on member i in these equations is essential. Without it the measurement process would be replicated on a different member each time and $\mathrm{var}(Y_i^{(1)})$ would depend also on the dispersion of the values of X and on the process used for selecting the member to be observed.

Just like the expectation, variance, and MSE, other features and properties defined for a (finite) target population can be defined also for a population of measurements. These features include symmetry, the median, and quantiles (percentiles), except for minimum and maximum (the zero- and unity-quantiles) that need not exist. The definition of the mode is also problematic.

Every feature defined for one subject or member of the population has an obvious equivalent for every other member; after all, the ordering (labelling) of the members of the population is immaterial. Description of the measurement process by one or a few quantities for each member is impractical for a population of moderate or large size. The ultimate simplification is attained when the measurement process has the same properties for every member. Of course, this is a very special case. For example, suppose X has possible values 0, 1, 2, ..., 10, and its manifest version Y deviates by at most one unit in either direction, with probability 0.1:

$$\mathrm{P}(Y = X - 1 \,|\, 1 < X < 10) \;=\; \mathrm{P}(Y = X + 1 \,|\, 1 < X < 10) \;=\; 0.1\,,$$

unless $X = 0$ or $X = 10$. When $X = 0$ or $X = 10$, only one kind of deviation is possible: observing $Y = 1$ instead of $X = 0$ and observing $Y = 9$ instead of $X = 10$. This measurement process is symmetric and unbiased, so long as $X \neq 0$ and $X \neq 10$.

We prefer measurement processes that have smaller MSEs, and among those with identical MSEs, those with smaller absolute bias $|\,\mathrm{B}\,|$. Of course, it may be difficult to compare measurement processes when their properties are specific to the members of the population. The distribution of the deviations $Y - X$ may be the same within each subpopulation defined by a categorical variable (such as men and women, or occupational categories in human populations), or by the value of X itself.

An appealing property of a measurement process is that its distribution depends on the observed subject only through the values of a limited set of variables, and the identity of the observed member is irrelevant otherwise. A measurement process is said to be *impartial* if the distribution of Y depends only on the value of X. A measurement process is said to be *additive* if the deviations $Y - X$ have the same distribution for every member of the population. Such a process can be described as

$$Y \;=\; X + \varepsilon\,,$$

where the distribution of ε is independent of the observed members' values of X. Additive processes for which ε is independent of the background variables are impartial. Section 6.2 contains more details on impartiality and additivity.

Properties of a measurement process can be changed dramatically by a transformation. By way of an example, suppose X is a monetary value, such as the total value of a company's liabilities. For many companies, their total liabilities are not defined with precision because guesses have to be made about some of its elements, and other elements may depend on the prices in the near future. A plausible model for a particular measurement (assessment or audit) process is that

$$Y_i \;=\; X_i\,\delta_i\,,$$

where the distribution of δ_i does not depend on the company (i). Such a measurement process is called *multiplicative*. The logarithms of the assessed and 'true' liabilities satisfy an additive measurement model. For instance, suppose a typical deviation from X is by 1%, and deviations in access of 2.5% are very rare. A deviation of 1% corresponds to £10 000 for a large company with liabilities of £1 million, but only £100 for a small company with liabilities of £10 000. After taking logarithms, such deviations correspond to log-deviations of $\log(0.01) = 0.0095$, irrespective of the underlying value of the liabilities.

A.7 Infinite Populations

The distribution of a variable in an infinite population cannot be established by counting the number of members with each specific value because these counts may be infinite for some or all of the possible values. Even when the counts are infinite it is meaningful to consider how much more frequently one value occurs than another. For example, the distribution of the outcomes of the single toss of a fair coin is given by the probabilities:

$$\mathrm{P}(Y = \text{head}) \;=\; \mathrm{P}(Y = \text{tail}) \;=\; \tfrac{1}{2}\,.$$

We could verify this by replicating the toss many times and observing that about half the outcomes are heads. (The number of replications has to be large and specified up front.)

The distribution of a general variable in an infinite population is defined similarly. The distributions are considered for a sequence of samples $\mathbf{s}_h\,,\ h = 1, 2, \ldots$, such that each sample is a subsample of the following sample, that is, $\mathbf{s}_h \subset \mathbf{s}_{h+1}\,$, and the union of all the samples coincides with the population; every member i belongs to all samples $\mathbf{s}_h$ for $h \geq h_i\,$; the index h_i is specific to member i. The distribution of the variable is defined as the limiting distribution as h increases above all bounds.

This definition requires two qualifications: how a limiting distribution is defined and by what process the sequence of samples is constructed. For simplicity, we consider first variables that have only a finite number of possible

Table A.3. The observed counts of the outcomes in replicate draws from the distribution with probabilities 0.08, 0.15, 0.27, 0.32, and 0.18 of the respective categories 1–5. The corresponding frequencies are plotted in Figure A.5.

	Outcome				
Replications	1	2	3	4	5
100	10	10	25	38	17
1000	79	155	250	334	182
10 000	798	1489	2694	3184	1837
100 000	7835	15 058	27 005	32 295	17 807

values. With the increasing sample size, the segments or bars of the histogram become taller, even if their relative sizes are not changed radically. The effect of the sample size can be removed by plotting the proportions of subjects in each value category, while keeping the total length of the segments constant, equal to unity. With such a *standardisation*, the limiting distribution is defined by the limits of the proportions for each category.

An example of convergence in distribution is given in Figure A.5. The numbers of replicates (sample sizes) on which the plotted distributions are based are 100, 1000, 10 000 and 100 000, given in the subtitle of each panel. The probabilities, 0.08, 0.15, 0.27, 0.32, and 0.18, of the respective categories 1–5 are marked in each panel by thin horizontal bars. On the scale used for plotting, the five deviations of the sample proportions from the corresponding probabilities are substantial for $n = 100$ in panel A and minute for $n = 100\,000$ in panel D. Table A.3 gives the four distributions in tabular form, expressed as counts for each category. In contrast to the proportions, the counts tend to differ from their expectations by wider margins with more replications; for instance, in 100 replications, outcome 1 was observed ten times, in two more cases than expected, whereas in 100 000 replications, outcome 1 was observed 165 fewer times than the expected count of 8000. Convergence occurs for the proportions, not for the counts.

Control over the sampling process is essential to avoid distortions such as overrepresentation of a category. To simplify matters, we define simple random sampling from an infinite population by a sequence of replications of drawing a single subject without any prejudice for or against any of the members' attributes. This sounds like a circular definition, but we cannot define a sampling process by probabilities because the probability of any one member being drawn is equal to zero.

The distributions drawn in Figure A.5 are called *sampling distributions* because they depend on the sample (of occasions) drawn. Their limit is called the *population distribution*. The adjectives *sampling* and *population* are used

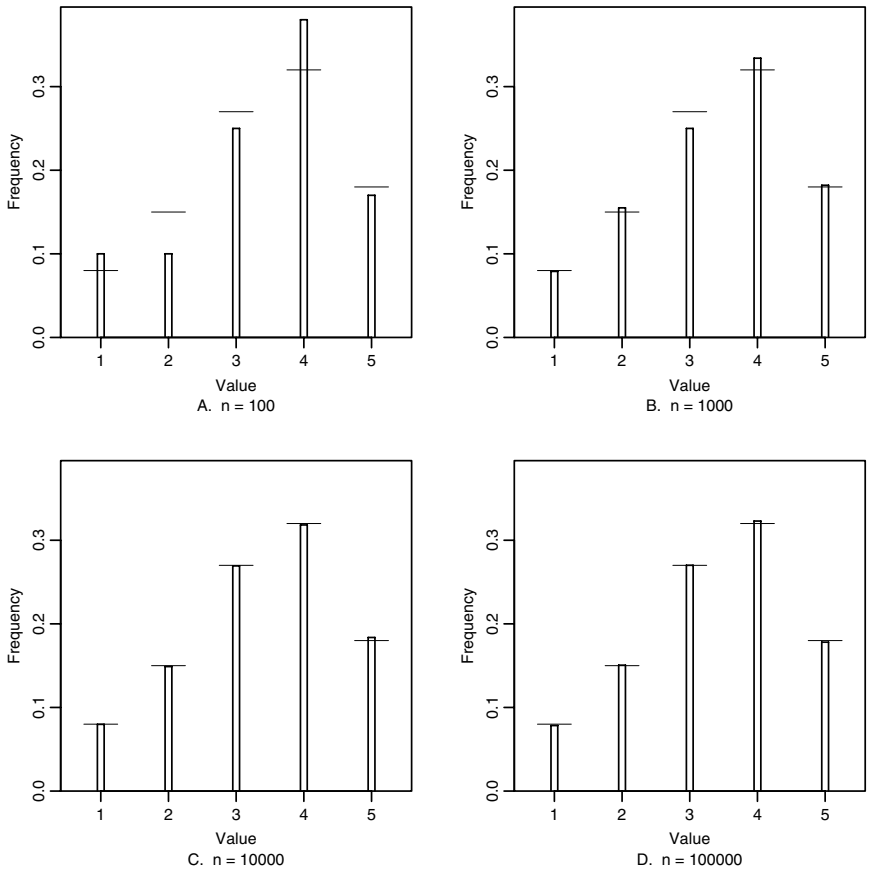

Fig. A.5. Illustration of convergence in distribution. The sample size for each distribution is given in the subtitle of each panel. The limiting frequencies are marked by horizontal bars.

in the same way as for quantities or sets of quantities derived from them. A distribution can be regarded as a collection of quantities.

A.7.1 Continuous Distributions

The possible values of some variables cover the entire continuum or an interval of real numbers, and so, without rounding, each value may be unique. The distribution of such a variable cannot be described as a limit of sampling distributions, because each (finite) sampling distribution is full of spikes corresponding to the values of the individual subjects. Yet the density of the spikes informs us about the ranges in which values are more or less frequent. Such features are more succinctly depicted by a coarsened histogram. It is practical

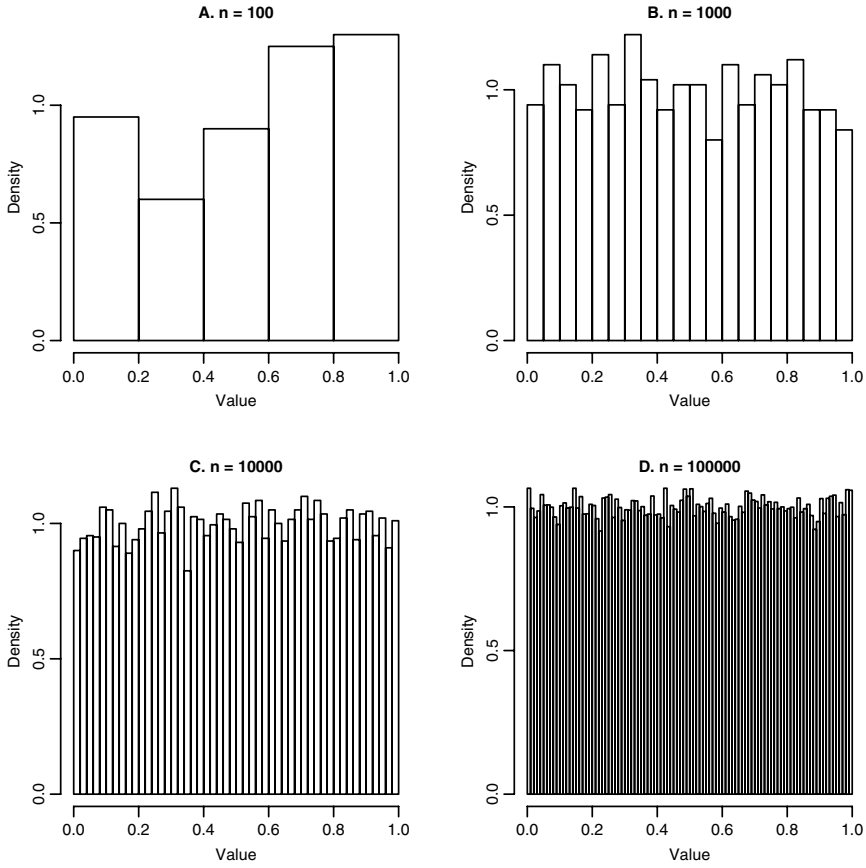

Fig. A.6. Illustration of the convergence in distribution for a continuous variable. The sample size for each distribution is given in the title of each panel.

to plot the standardised histogram, in which the area covered by the bars is equal to a set value, such as unity. The distribution of a continuous variable in an infinite population is defined as the limit of the standardised histograms for random samples with sample sizes increasing beyond all bounds, while the bars of the histograms get narrower (the coarsening is refined) as the sample size increases.

An illustration paralleling Figure A.5 is given in Figure A.6. In the limiting distribution, each bar has the same (unit) height. For the sample size $n = 100$, this could not be anticipated, but for $n = 100\,000$ it is obvious, although one may argue that the limit could still have an irregular pattern. The limiting histogram, with the bar widths converging to zero, is called the *density* of the distribution, if the limit is well defined. A distribution that has a density is called *absolutely continuous*. We drop the qualifier 'absolutely' because we

very rarely come across variables or distributions that are continuous but not absolutely continuous.

A distribution with a constant density on its support is called *uniform*. The *standard uniform* distribution is the uniform distribution with the support on $(0, 1)$. Its density is $f(x) = 1$ for $x \in (0, 1)$ and $f(x) = 0$ otherwise. The limiting distribution in Figure A.6 is standard uniform.

For any continuous distribution, the area under the density is equal to unity:

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

A density defines a unique distribution by the identity

$$\mathrm{P}(c < X < d) \;=\; \int_{c}^{d} f(x)\,\mathrm{d}x$$

for any pair of real numbers $c < d$.

For a variable X or its distribution, the distribution function is defined as

$$F(x) = \mathrm{P}(X \le x)\,,$$

a function in $(-\infty, +\infty)$. It is nondecreasing, with limits of zero and unity at $-\infty$ and $+\infty$, respectively. The distribution function and the density of a continuous distribution are related by the identity

$$f(x) \;=\; F'(x)$$

(the derivative of F at x). Therefore, the distribution function of a continuous distribution is differentiable.

Strictly speaking, it cannot be proven that the limiting distribution in Figure A.6 is the uniform or any other distribution. To justify the uniform as the limit, we have to supplement the evidence based from sampling with the conjecture, or appeal to 'good reason', that the density is smooth. In the case of Figure A.6, it may be difficult to argue why the density should deviate from a constant (unity) according to no apparent pattern. In practice, it is often much more difficult to integrate the information extracted from the data and obtained from other sources, such as descriptions of the studied setting and relevant findings made by other parties, and conclude with a simple description of the sought distribution and its properties.

A.7.2 Superpopulations: Models

Although finite, many human populations are large, with several million members, and the distributions of variables defined for them are often very close to continuous distributions with densities that contain no sharp edges or sudden changes. Such densities are called *smooth*. Formally, a density $f(x)$ is said to

be smooth in an interval if it is differentiable at each point of the interval and its derivative, denoted by $f'(x)$, is a continuous function. Note that a smooth density f corresponds to an 'even smoother' distribution function F; since $f(x) = F'(x)$, F is twice continuously differentiable.

Using a continuous distribution has several advantages; continuous distributions tend to be easier to describe, by a mathematical formula or graph, and various operations with them are easier to execute. The use of such a distribution may be justified by a reference to an infinite-size *superpopulation*. This is a hypothetical (nonexistent) population from which the studied population is assumed to have been drawn as a random sample. For instance, in a different context, such as the same survey conducted at a different time point and using slightly different questions, the survey would be conducted on a different population, but the population would have been realised by the same sampling process applied to the same superpopulation. We may even make inferences about the superpopulation, regarding its features as more stable (less transient) than the features of the population. After all, the population, regarded as a simple random sample, is, by definition, a faithful miniature of the superpopulation.

Superpopulation and its description (by a distribution) are an example of a *model*, an analyst's construct. A model is a stylised (simplified) description of a studied population using one or several variables defined on it. For instance, a model may provide a description of how the values of several variables are related in a population. The price of the simplification is a loss of precision and detail, but it may well be worth it if the result is greater insight and better understanding of the studied population.

A.8 Distributions

Any sequence of values $v_1, v_2, \ldots$, of finite or infinite length, and a corresponding sequence of positive numbers $p_1, p_2, \ldots$ that add up to unity, forms a discrete distribution, by prescribing for a random variable X that

$$\mathrm{P}(X = v_k) \;=\; p_k\,.$$

Any nonnegative function $f^\dagger$ that has a finite area underneath it,

$$\int f^\dagger(u)\,\mathrm{d}u \;<\; +\infty\,,$$

defines a continuous distribution by the density obtained by standardising $f^\dagger$, that is, by defining

$$f(x) \;=\; \frac{f^\dagger(x)}{\int f^\dagger(u)\,\mathrm{d}u}\,.$$

For example, the exponential distribution is defined by the density

$$f(x) \;=\; \theta \exp(-\theta x)\,, \tag{A.2}$$

for $x > 0$ and $f(x) = 0$ otherwise; θ is a positive constant. The distribution function is obtained from the density by integration:

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

and the density is obtained from the distribution by differentiation: $f(x) = F'(x)$. The distribution function of the exponential distribution given by (A.2) is

$$F(x) \;=\; \mathrm{P}(X \le x) \;=\; 1 - \exp(-\theta x)\,,$$

for $x > 0$. We use the singular *distribution* for the density $f(x)$ in (A.2) with a specific *parameter* θ, and the plural *distributions* for the collection of distributions, for θ in the entire range $(0, +\infty)$ or its subset.

The exponential distributions are a special case of the gamma distributions given by the densities

$$f(x) \;=\; \frac{1}{\Gamma(b)} \theta^b x^{b-1} \exp(-\theta x)\,,$$

for $x > 0$ and $f(x) = 0$ otherwise; b and θ are positive constants and Γ is the gamma function; see Section A.12 for more details.

An important role in statistics is played by the *normal* distributions. They are defined by the densities

$$f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \tag{A.3}$$

where μ is a real and σ^2 a positive constant. We denote this distribution by $\mathcal{N}(\mu, \sigma^2)$. The normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the *standard normal* distribution.

For continuous distributions, we can define various features similarly to their counterparts for discrete distributions, with probabilities $\mathrm{P}(X = x_k)$ replaced by the values of the density $f(x)$. A continuous distribution is said to be symmetric around a value c if $f(c - \Delta) = f(c + \Delta)$ for every constant Δ. For example, the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is symmetric around μ. A continuous distribution is said to have a mode at value c if its density has a local maximum at c. The distribution is said to be unimodal (bimodal, trimodal, and so on), if it has one (two, three, or more) modes.

The expectation of a continuous distribution with density $f(x)$ is defined as

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

if the integral is well defined. Equivalently, the expectation can be defined as the limit of the expectations of a sequence of discrete distributions that

converge to the distribution with density f, so long as the limit exists and the integral is well defined. The variance of a continuous distribution with density $f(x)$ is defined as

$$\mathrm{var}(X) \;=\; \mathrm{E}\left[\{X - \mathrm{E}(X)\}^2\right]$$

if the integrals involved are well defined. Equivalently, we have

$$\mathrm{var}(X) \;=\; \mathrm{E}\left(X^2\right) - \{\mathrm{E}(X)\}^2\,,$$

so long as $\mathrm{E}\left(X^2\right)$ is well defined. If it is, then so is $\mathrm{E}(X)$. For example, the mean of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is μ, and its variance is equal to σ^2.

The q-quantile of a continuous distribution with density $f(x)$ is the value u for which

$$\mathrm{P}(X < u) \;=\; \int_{-\infty}^{u} f(x)\,\mathrm{d}x \;=\; q\,. \tag{A.4}$$

This probability, the distribution function, in fact, is a continuous nondecreasing function of u, and so the equation in (A.4) always has a solution, for every $q \in (0, 1)$; the solution either is unique or is any point in an interval. In the latter case, a sensible convention is to declare the centre of the interval as the quantile.

Earlier we defined sampling processes for finite populations. Sampling from a continuous distribution or a related superpopulation requires a new definition, because the probability of any particular real value is equal to zero. We define a random draw from the standard uniform distribution by the process of drawing a single value in the range $(0, 1)$ without any prejudice. A practical implementation of this can rely on a sequence of independent draws from the discrete uniform distribution on (0, 1, . . . , 9). Let these draws be $(l_1\,, l_2\,, \ldots)$. Then the draw from the standard uniform distribution is

$$\sum_{h=1}^{\infty} 10^{-h} l_h\,,$$

that is, l_h is the digit in the decimal place h. A random draw from a continuous distribution is defined as the q-quantile of this distribution, where q is a random draw from the standard uniform distribution. A random sample of size n is defined as a sequence of n replicate (independent) random draws from a distribution. Note that this definition confers a pivotal role on the standard uniform distribution. From a random sample $\mathbf{X} = (X_1\,, X_2\,, \ldots, X_n)$ from the standard uniform distribution we obtain a random sample from a continuous distribution with distribution function F by the elementwise quantile transformation

$$F^{-1}(\mathbf{X}) \;=\; \left\{F^{-1}(X_1), F^{-1}(X_2), \ldots, F^{-1}(X_n)\right\}\,.$$

A distribution is given by a set of probabilities, a density, a distribution function, or the like. Frequently we consider a *class of distributions*; they are a

finite or infinite set of distributions that have a common (or similar) functional form and are distinguished by the values of one or several parameters.

For example, the set of all exponential distributions, given by the density $f(x) = \theta \exp(-\theta x)$, where θ is in the range $(0, +\infty)$, is a class of distributions. The distributions in this class are characterised by the value of one parameter, θ. Such a class is said to be single-parameter. The class of all the normal distributions $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in (-\infty, +\infty)$ and $\sigma^2 \in (0, +\infty)$, is a two-parameter class of distributions. In principle, any collection of distributions can be regarded as a class, and they need not have a description in terms of one or a few parameters.

A.8.1 Simulations

With a considerable simplification, a typical problem in statistics can be described as follows. The values of a variable are available for a random sample drawn from an infinite population, and the population distribution of the variable is known to belong to a given class of distributions. The task is to estimate this distribution or its summary. Ideally, we would like to identify it, but that is rarely possible. For a one-parameter class, the quantity of interest may be the value of the characterising parameter, such as θ for the exponential distributions given by (A.2). The key assumption made is that the process that generates the values of the studied variable is well described by one of the distributions in the posited class. We can *simulate* (mimic) the process of generating a random sample from an infinite population (distribution) on the computer.

Figure A.7 displays the histogram of a computer-generated random sample of size 50 000 drawn from the standard normal distribution. The density of the normal distribution, suitably scaled, is superimposed. The distribution of the computer-generated values is called *empirical*. Any summary of the empirical distribution, such as its mean and variance and, in relation to an estimator, bias and MSE, are also called empirical.

The histogram shows that with a large sample size the empirical and population distributions differ only slightly and have essentially the same features, except for a modicum of roughness of the empirical distribution. For instance, unlike the population distribution, the empirical distribution may have more than one mode. The symmetry is not reproduced, but the empirical distribution is very close to symmetry.

The term (computer) *simulation* refers to replications of the assumed data-generating process (on a computer). In principle, any device could be used for simulation, but the modern computer has no practical competitor, especially when a large number of replicates and a nontrivial amount of computing are required. By simulation, we can generate replicates fast and at a fraction of the cost of replicates generated by the studied processes ('real life') and can assess the properties and, more generally, learn about the posited distributions. To this process, we can attach estimation, using several (candidate) estimators,

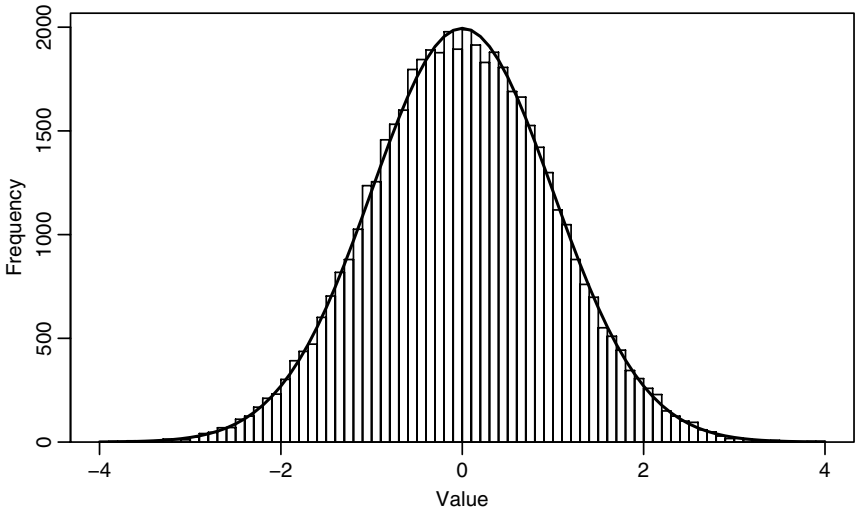

Fig. A.7. Histogram of a computer-generated random sample of size 50 000 from the standard normal distribution, with the density of $\mathcal{N}(0, 1)$ superimposed.

and compare their properties, efficiency in particular. In this way, we can engage in an informed planning of a study in which observations are expensive (as regards finance, labour, ethics or any form of undesirable destruction), and decide how to strike a balance between the conflicting goals of high precision in estimating a target and low expenditure.

A.9 Classes of Distributions and Models

Simulations can use the computing power as a replacement for the analytical ability to derive properties of estimators. Sampling and measurement processes can be explored similarly. In practice, we face a task much more difficult than simulation because only one replication of a process, governed by a distribution that is not known to us, is available. When we know, or assume, that the sought distribution belongs to a particular class, we might look for the member of the class that resembles the observed values more closely than any other. For this, we have to define a metric for 'resemblance' but also develop approaches to identifying suitable classes of distributions based on the information about the studied processes.

If we had a perfect understanding of how a particular process operates, we could anticipate what kind of values it would produce. In a typical setting, our understanding is far from complete but is not totally hollow. We study a process to enhance or supplement our partial understanding of it. We can entice the process to run its course and yield values of one or several key

variables on a sample of subjects or occasions (observational units). From this output, commonly referred to as *data*, we want to estimate certain population quantities related to the process. We can interpret this problem as an *inverse* task to simulations. While in simulations we can implement a process and obtain output (data), the task in practice is to infer from the data obtained some properties (details) of the underlying process.

A powerful general approach to addressing this problem is by specifying a model for the studied (target) process. This model is a collection of processes, and we assume that one of them is the studied process. For instance, if we believe that the process generates values drawn at random from one of a class of distributions, then this class forms the model. Suppose the class comprises all the normal distributions, $\mathcal{N}(\mu, \sigma^2)$, with unknown values of μ and σ^2, and it would be valuable to know the values of these *parameters*, μ and σ^2. The target process may be more complex than one of the model distributions, but the simplicity in the model specification may be rewarded by a better choice of an estimator or, more importantly, a better understanding of the studied process.

For models or classes of distributions in general, we can define a partial ordering. Model A is said to be narrower than model B if every distribution in model A is also contained in model B. A model with a narrower class of distributions has the advantage that we have fewer candidates for the target process, so the search among them is, in principle, easier. On the other hand, if we choose a wider class of distributions we do not reduce the possibility that this class contains the target distribution or contains a distribution that is close to the target distribution. This balancing act between *specificity* (narrowness) and *validity* (containing the 'true' distribution) is one of the principal unresolved problems in statistics. A model that contains the target distribution is called *valid*. In later chapters we will also find that estimation based on narrower models is more efficient, *if* the model is valid.

A.10 Normal Distributions

The class of normal distributions appears in many theoretical derivations as well as in practical applications. It has some very useful properties which make it a popular choice for a model. From its density given by (A.3), we can easily deduce that the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is symmetric and has a single mode at μ. These properties imply that its mean, if it exists, is equal to μ. This can be verified by evaluating the integral

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\ &= \mu \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y \;=\; \mu\,,\end{aligned}$$

after the transformation $y = (x - \mu)/\sigma$, and realising that the first integrand in the second line is the density of the standard normal distribution. The second integrand is symmetric and has the primitive function $-\exp(-y^2/2)$, so its integrals over $(-\infty, 0)$ and $(0, +\infty)$ are well defined and add up to zero. Similarly, it can be shown that the variance of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is equal to σ^2. A more elegant proof of this is provided after deriving another property of the class of normal distributions.

If X has the distribution $\mathcal{N}(\mu, \sigma^2)$, then $Y = (X - \mu)/\sigma$ has the standard normal distribution $\mathcal{N}(0, 1)$. This can be derived by expressing the distribution of Y by the probabilities

$$\mathrm{P}(Y < c) \,=\, \mathrm{P}(X < \mu + c\sigma)\,,$$

and expressing this as

$$\begin{aligned}\mathrm{P}(X < \mu + c\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\mu + c\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} \exp\left(-\tfrac{1}{2}y^2\right) \,\mathrm{d}y\,,\end{aligned}$$

which is the distribution function of the standard normal distribution. The variance of the standard normal distribution is equal to

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \exp\left(-\tfrac{1}{2}y^2\right) \,\mathrm{d}y \,=\, 1\,,$$

derived by integrating by parts, using the primitive function $-\exp(-y^2/2)$ for $y \exp(-y^2/2)$. As the standard deviation is a dispersion quantity, the standard deviation of $\mathcal{N}(\mu, \sigma^2)$ is equal to σ and the variance to σ^2.

We can declare each constant μ the (degenerate) normal distribution $\mathcal{N}(\mu, 0)$. With this convention, any linear transformation of a normally distributed variable is normally distributed. We say that the normal distributions are closed with respect to linear transformations. Any normal distribution can be formed by a linear transformation of the standard normal distribution, and any normal distribution $\mathcal{N}(\mu, \sigma^2)$ with positive variance σ^2 can be transformed linearly, as from X to $(X - \mu)/\sigma$, to become the standard normal. The application of this transformation, $g(X) = (X - \mu)/\sigma$, is referred to as *standardisation*.

Owing to the closure of the normal distribution with respect to linear transformations, an arbitrary quantile of any normal distribution can be derived straightforwardly from the corresponding quantile of the standard normal distribution. Denote by Φ the distribution function of $\mathcal{N}(0, 1)$, so that its inverse $\Phi^{-1}(q)$ is the quantile as a function of the probability q. As every quantile is a location quantity, the q-quantile of $\mathcal{N}(\mu, \sigma^2)$ is equal to $\mu + \sigma\Phi^{-1}(q)$. Many estimators encountered in practice are approximately normally distributed with the approximation being very close when the sample size is large.

A.10.1 Log-Normal Distributions

Many variables are expressed in physical units, such as degrees Fahrenheit, miles, or degrees of latitude, which have alternatives, such as degrees Celsius, kilometres, and radians, respectively. A class of distributions would have a strong appeal if, among other conditions, it would be equally well suited for either of the scales defined by the alternative units. When the units are linearly related, as in the listed examples, the normal distribution has the obvious advantage because it is closed with respect to linear transformations.

In practice, we often encounter variables for which classes that are closed with respect to multiplication would be suitable. For example, a natural operation for values defined in monetary units is multiplication, often expressed in terms of percentages. By taking logarithms, multiplication converts to addition, a linear transformation. Some laws of physics involve multiplication (or division) and units that imply multiplication. Area and volume are cases in point, and speed and shape involve division. By taking logarithms, length, area, and volume are expressed in identical units, such as 'log-meter'.

These examples motivate the *log-normal* distribution. A variable X is said to have a log-normal distribution if its logarithm has a normal distribution. If a variable X is distributed according to $\mathcal{N}(\mu, \sigma^2)$, then its exponential (the inverse of logarithm) has mean $\exp(\mu + \frac{1}{2}\sigma^2)$ and variance $\exp(2\mu)\exp(\sigma^2)\left\{\exp(\sigma^2) - 1\right\}$. To prove this, we evaluate $\mathrm{E}\{\exp(kX)\}$ for $k = 1$ and $k = 2$. By reorganising the terms in the arguments of the exponentials we obtain

$$\begin{aligned}
\mathrm{E}\left\{\exp(kX)\right\} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp(kx) \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2} + \frac{(\mu + k\sigma^2)^2}{2\sigma^2}\right\} \int_{-\infty}^{+\infty} \exp\left[-\frac{\left\{x - (\mu + k\sigma^2)\right\}^2}{2\sigma^2}\right] \\
&= \exp\left(k\mu + \tfrac{1}{2}k^2\sigma^2\right) ,
\end{aligned}$$

exploiting the fact that the density of $\mathcal{N}\left(\mu + k^2\sigma^2, \sigma^2\right)$ integrates to unity. The expression for the mean follows directly ($k = 1$), and the variance is derived from the identity

$$\mathrm{var}\{\exp(X)\} = \mathrm{E}\{\exp(2X)\} - [\mathrm{E}\{\exp(X)\}]^2.$$

The normal and log-normal distributions provide an effective illustration that nonlinear transformations and expectation E do not commute. We have

$$\mathrm{E}\{\exp(X)\} \geq \exp\{\mathrm{E}(X)\} ,$$

with equality only when $\sigma^2 = 0$. The finite-sample version of this inequality is equivalent to the statement that the geometric mean of a set of positive numbers never exceeds the arithmetic mean:

$$\sqrt[n]{\prod_{i=1}^{n} x_i} \;\leq\; \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$

with equality only when all n values x_i coincide.

A.11 Uniform Distributions

The uniform distribution was introduced in Section A.7.1. In this section, we explore it in greater detail. The class of continuous uniform distributions is given by the densities $f(x) = 1/(\theta_2 - \theta_1)$ for $x \in (\theta_1\,, \theta_2)$ and $f(x) = 0$ elsewhere; $\theta_1 < \theta_2$ are the two parameters that define a distribution. We denote the uniform distribution on $(\theta_1\,, \theta_2)$ by $\mathcal{U}(\theta_1\,, \theta_2)$. Although θ_1 has to be smaller than $\theta_2\,$, it is expedient to regard the constant θ as the (degenerate) uniform distribution $\mathcal{U}(\theta, \theta)$. For $\theta_1 < \theta_2\,$, the distribution function of $\mathcal{U}(\theta_1\,, \theta_2)$ is piecewise linear: $F(x) = 0$ for $x < \theta_1\,$, $F(x) = (x - \theta_1)/(\theta_2 - \theta_1)$ for $x \in [\theta_1\,, \theta_2]$ and $F(x) = 1$ for $x > \theta_2\,$.

Similarly to the standard normal distribution, the standard uniform distribution is obtained from an arbitrary nondegenerate distribution $\mathcal{U}(\theta_1\,, \theta_2)$, with $\theta_1 < \theta_2\,$, by the linear transformation $g(X) = (X - \theta_1)/(\theta_2 - \theta_1)$. Conversely, any uniform distribution is obtained from the standard uniform by the transformation $\theta_1 + (\theta_2 - \theta_1)X$. The class of uniform distributions is closed with respect to linear transformations.

The distribution $\mathcal{U}(\theta_1\,, \theta_2)$ is symmetric, with mean $(\theta_1 - \theta_2)/2$ and variance $(\theta_1 - \theta_2)^2/12$. The latter expression is obtained by integration (for the standard uniform distribution),

$$\int_0^1 \left(x - \tfrac{1}{2}\right)^2 \mathrm{d}x \;=\; \tfrac{1}{12}\,,$$

and using the fact that the standard deviation is a dispersion quantity. The class of uniform distributions defines a model for randomness in a variety of settings. By a number randomly drawn from a given range, such as $(0, 100)$, we mean a random draw from the distribution $\mathcal{U}(0, 100)$. The condition of 'no prejudice' for or against any particular value in the support is interpreted as a constant density of the distribution from which a draw (selection) is to be made. It implies that the probability of a draw falling to any particular interval depends solely on the length of the interval; $\mathrm{P}(a < X < b) = (b - a)/(\theta_2 - \theta_1)$, so long as $\theta_1 \leq a \leq b \leq \theta_2\,$.

The standard uniform distribution has the role of a pivot among all the continuous distributions. Suppose variable X has a continuous distribution with distribution function $F(x)$, strictly increasing throughout its support $(\xi_1\,, \xi_2)$; either bound ξ_1 or ξ_2 may be infinite. Then $F(X)$, the distribution function applied as a transformation, has the standard uniform distribution. This follows immediately from the identity

$$\mathrm{P}\{F(X) < u\} = u$$

for $u = F(x) = \mathrm{P}(X < x)$.

Suppose continuous variable X has distribution function $F(x)$ and $G(x)$ is another distribution function with a density. Then the variable $G^{-1}\{F(X)\}$ has the distribution function $G(x)$. Thus, a continuous variable can be transformed to have any other continuous distribution. In particular, we can construct (by simulations) a variable with any conceivable continuous distribution.

Polynomial and Other Distributions Derived from Uniform

A polynomial distribution is derived by taking a power of a variable with uniform distribution. We focus here on polynomial distributions with support on $(0, 1)$; distributions with supports on other intervals are derived straightforwardly. Suppose variable X has uniform distribution on $(0, 1)$, $X \sim \mathcal{U}(0, 1)$, so that $\mathrm{P}(X < x) = x$ for any $x \in (0, 1)$. Then the variable $Y = X^k$ has the distribution function $\mathrm{P}(Y < x) = \sqrt[k]{x}$, density $k^{-1}x^{1/k-1}$, expectation $1/(k + 1)$ and standard deviation $k/(k + 1)/\sqrt{2k + 1}$.

The exponential distribution is obtained as the negative logarithm of the uniform distribution. If $X \sim \mathcal{U}(0, 1)$, then the distribution function of $-\log(X)$ is $\mathrm{P}\{-\log(X) < x\} = \mathrm{P}\{X > \exp(-x)\} = 1 - \exp(-x)$. Note that while $\mathrm{E}(X) = \frac{1}{2}$, $\mathrm{E}\{-\log(X)\} = 1$, different from $\exp(\frac{1}{2})$; exponentiation and expectation cannot be exchanged. An arbitrary exponential distribution is obtained by the transformation $-\theta^{-1}\log(X)$ from $X \sim \mathcal{U}(0, 1)$ for a positive θ.

A.12 Beta and Gamma Distributions

The class of beta distributions is defined by the densities

$$f(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1} \tag{A.5}$$

for $x \in (0, 1)$ and positive constants a and b. The gamma function $\Gamma(a)$ is defined for positive arguments a as

$$\Gamma(a) = \int_0^{+\infty} x^{a-1}\mathrm{e}^{-x}\,\mathrm{d}x\,.$$

For integers a, $\Gamma(a) = (a - 1)! = 2 \times 3 \times \ldots \times (a - 1)$. The beta density with parameters a and b is denoted as $B(a, b)$. Its expectation is $a/(a + b)$ and its variance is $ab/\left\{(a + b)^2(a + b + 1)\right\}$. The expectations of the beta distributions are in the entire range $(0, 1)$ and their variances in the range

$\{0, \mu(1-\mu)\}$ where μ is the expectation. Beta distributions with mean equal to $\frac{1}{2}$, when $a = b$, are symmetric.

The gamma distributions are defined by the densities

$$f(x) \;=\; \frac{1}{\Gamma(\alpha)}\theta^{\alpha}x^{\alpha-1}\exp\left(-\theta x\right)\;,$$

where α and θ are arbitrary positive constants and $x > 0$. The mean and variance of a gamma distribution are α/θ and α/θ^2, respectively. The parameter α is referred to as the shape and θ as the scale. Exponential distributions are a special case (a *subclass*) of gamma, with shape $\alpha = 1$. A gamma distribution with shape parameter equal to integer α can be derived as the distribution of the sum of α independent variables each with the exponential distribution with the same parameter θ; see Section 2.2.

A.13 Classes of Discrete Distributions

The simplest nontrivial discrete distribution is the binary distribution, also known as the Bernoulli distribution. It has probabilities $1-p$ and p, $0 < p < 1$, on the two points of its support, 0 and 1, respectively. Its expectation is p and variance $p(1-p)$, obtained directly from the corresponding definitions for a general discrete distribution.

A *binomial distribution* is defined as the number of successes in a sequence of n independent trials, each of which has the same probability p of yielding a successful outcome. Thus, it is a sum of n independent and identically distributed binary variables. Its distribution is given by the probabilities

$$\mathrm{P}(X = k) \;=\; \binom{n}{k}\, p^k(1-p)^{n-k}. \tag{A.6}$$

Its expectation and variance are np and $np(1-p)$, respectively. They are both n-multiples of the Bernoulli distribution from which the binomial is generated. This is not a coincidence; see Section A.17. Every binomial distribution is either unimodal or its highest probability is attained at two consecutive points. The latter is the case when $k = p(n+1)$ is an integer, and then the highest probabilities are attained at $k-1$ and k. Otherwise it is unimodal, with mode either at the integer part of $p(n+1)-1$, denoted by $[p(n+1)-1]$, or at the following integer, $[p(n+1)]$. Every binomial distribution with $p = \frac{1}{2}$ is symmetric, but no other binomials are, unless we define the degenerate binomial distributions with $p = 0$ and $p = 1$.

A *Poisson distribution* can be derived as the limit of binomial distributions as the probability p converges to zero and the number of trials diverges to infinity at such a speed that np converges to a finite constant λ. The Poisson distributions are given by the probabilities

$$\mathrm{P}(X = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\;,$$

for $k = 0, 1, \ldots$ and parameter $\lambda > 0$. The expectation and variance of this distribution are both equal to λ. Every Poisson distribution is either unimodal, with mode at one of the integers next to λ, or, when λ is an integer, the highest probability is shared by the points $\lambda - 1$ and λ.

A geometric distribution is derived as the number of failures prior to the first success in a sequence of independent binary trials. Let $q = 1 - p$ be the probability of failure; $q \in (0, 1)$. Then

$$\mathrm{P}(X = k) \,=\, pq^k,$$

for a geometrically distributed variable X. The expectation of this variable is q/p and variance q/p^2. To prove this, we differentiate both sides of the identity

$$\sum_{k=0}^{\infty} p(1-p)^k = 1\,;$$

$$\sum_{k=0}^{\infty}(1-p)^k - \sum_{k=1}^{\infty} kp(1-p)^{k-1} = 0\,. \tag{A.7}$$

The first summation, the total of a geometric sequence, is equal to $1/p$ and second to the $1/q$-multiple of the expectation; hence $\mathrm{E}(X) = q/p$. Another differentiation of (A.7) yields the identity

$$-2\sum_{k=1}^{\infty} k(1-p)^{k-1} + \sum_{k=2}^{\infty} k(k-1)p(1-p)^{k-2} \,=\, 0\,,$$

and by relating the two summations to $\mathrm{E}(X)$ and $\mathrm{E}\{X(X-1)\}$, respectively, we obtain the identity

$$\mathrm{E}\,\{X(X-1)\} \,=\, \frac{2q^2}{p^2}\,,$$

from which the result for $\mathrm{var}(X) = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \{\mathrm{E}(X)\}^2$ follows immediately.

A.13.1 Discrete Uniform Distributions

The class of discrete uniform distributions is defined by equal probabilities on each possible outcome. For example, the toss of a fair coin is represented by the uniform distribution on the set (H,T), head and tail of the coin. Casting a die corresponds to the uniform distribution on its six faces, or on the digits $1, \ldots, 6$. Similarly, by drawing a random digit, we mean drawing one member from the population $(0, 1, \ldots, 9)$ with probability equal to 0.1 for each digit. Lottery and games of chance provide further examples (and applications) of discrete uniform distributions.

The distributions supported on a finite number of values are called *multinomial*. The discrete uniform and binomial are their subclasses. A discrete

distribution that is supported on infinitely many values can be approximated with arbitrary precision by a distribution that has the same probabilities on a suitably selected finite subset of the support, and the remaining probability is gathered in a single point. This sequence of subsets can be set to the n values that have the highest probabilities, with an arbitrary way of resolving ties, and letting n diverge to infinity.

A.14 Discrete Bivariate Distributions

So far, we have considered in detail only distributions defined on the real numbers, $(-\infty, +\infty)$, or its subsets, such as finite-length intervals and integers. We refer to such distributions as *univariate*. However, distributions can be defined in any space. In this section, we consider distributions on $\mathcal{I}^2 = \mathcal{I} \times \mathcal{I}$, where $\mathcal{I}$ is the set of integers 0, 1, ... They are essential for dealing with pairs of univariate variables and for studying how they are associated.

A discrete bivariate distribution is derived from a pair of random variables, X_1 and X_2; it is given by the probabilities

$$\mathrm{P}(X_1 = x_1 \text{ and } X_2 = x_2)\,,$$

for all integers x_1 and x_2. The two variables are said to be *independent* if

$$\mathrm{P}(X_1 = x_1 \text{ and } X_2 = x_2) \;=\; \mathrm{P}(X_1 = x_1)\,\mathrm{P}(X_2 = x_2) \tag{A.8}$$

for all x_1 and x_2. Independence is a special case of association. A trivial case of dependence (the negation of independence) is the association of the variable with itself; the identity

$$\mathrm{P}(X = x) \;=\; \{\mathrm{P}(X = x)\}^2\,,$$

required for independence, holds only when $\mathrm{P}(X = x)$ is equal to zero or unity, so a variable is independent with itself only when it is degenerate (supported by a single value x). As a nontrivial example of dependence, consider a single cast of a die and denote by X_1 and X_2 the dichotomous variables that indicate whether the outcome is even (2, 4 or 6), and whether it exceeds 3. Both variables have binary distributions on (Yes, No), with identical distributions given by $\mathrm{P}(X_1 = \text{Yes}) = \mathrm{P}(X_2 = \text{Yes}) = \frac{1}{2}$. However,

$$\mathrm{P}(X_1 = \text{Yes and } X_2 = \text{Yes}) = \tfrac{1}{3}\,,$$

different from $\mathrm{P}(X_1 = \text{Yes}) \times \mathrm{P}(X_2 = \text{Yes}) = \frac{1}{4}$.

A probability that involves both variables X_1 and X_2 is called *joint*, and a probability that involves only one of them is called *marginal*. These terms are motivated by the table of the (joint) probabilities in Table A.4.

In general, we have the identity

Table A.4. Joint and marginal probabilities (an example).

	X_2		
X_1	Yes	No	*Margin*
Yes	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$
No	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
Margin	$\frac{1}{2}$	$\frac{1}{2}$	

$$\mathrm{P}(X_1 = x_1) \;=\; \sum_x \mathrm{P}(X_1 = x_1 \text{ and } X_2 = x)\,,$$

where the summation is over the support of X_2. From the joint distribution (probabilities) we can derive the marginal distributions, but these are not sufficient for recovering the joint distribution of a pair of variables. If we know that X_1 and X_2 are independent, then their joint distribution is easily reconstructed from the marginal distributions according to (A.8).

A.14.1 Conditional Distributions

A conditional distribution is defined for a variable and a condition. A simple example is drawn from Table A.4. The conditional distribution of X_1, given that the outcome of casting a die exceeds 3 (X_2 is equal to 'Yes'), is defined by

$$\mathrm{P}(X_1 = \text{Yes} \,|\, X_2 = \text{Yes}) \;=\; \tfrac{2}{3}\,.$$

This is derived by reducing our attention to the outcomes that satisfy the condition stated behind the vertical bar $|$. The probability is derived from the first column of the table as $\frac{1}{3} \,/\, \left(\frac{1}{3} + \frac{1}{6}\right)$ or, in general, as

$$\mathrm{P}(X_1 = x_1 \,|\, X_2 = x_2) \;=\; \frac{\mathrm{P}(X_1 = x_1 \text{ and } X_2 = x_2)}{\mathrm{P}(X_2 = x_2)}\,.$$

This may be easier to motivate by replacing the probabilities in the table by counts. It corresponds to multiplying each entry in the table by a large number, but it has no impact on the ratio.

The roles of the two variables in the probability $\mathrm{P}(X_1 = x_1 \,|\, X_2 = x_2)$ differ substantially. In general,

$$\mathrm{P}(X_1 = x_1 \,|\, X_2 = x_2) \;\neq\; \mathrm{P}(X_2 = x_2 \,|\, X_1 = x_1)$$

even though $\mathrm{P}(X_1 = \text{Yes} \,|\, X_2 = \text{Yes}) \;=\; \mathrm{P}(X_2 = \text{Yes} \,|\, X_1 = \text{Yes}) \;=\; \frac{2}{3}$ in Table A.4. The following example confirms this rule. Let X_1 be the higher

outcome of the two casts of a die and X_2 the sum of the two outcomes. The conditional probability that $X_1 = 6$, given that the total is $X_2 = 8$, is $\mathrm{P}(X_1 = 6 \,|\, X_2 = 8) = \frac{2}{5}$, whereas the conditional probability that $X_2 = 8$ given that six has come up at least once is $\mathrm{P}(X_2 = 8 \,|\, X_1 = 6) = \frac{2}{11}$. A conditional probability is well defined only when the condition itself has a positive probability.

The two sets of probabilities are connected by the identity

$$\mathrm{P}(X_1 = x_1 \,|\, X_2 = x_2) \;=\; \frac{\mathrm{P}(X_2 = x_2 \,|\, X_1 = x_1)\,\mathrm{P}(X_1 = x_1)}{\mathrm{P}(X_2 = x_2)} \tag{A.9}$$

(assuming that $\mathrm{P}(X_2 = x_2)$ and $\mathrm{P}(X_1 = x_1)$ are both positive), called the *Bayes theorem*. Further, the denominator in (A.9) can be expressed as

$$\mathrm{P}(X_2 = x_2) \;=\; \sum_x \mathrm{P}(X_2 = x_2 \,|\, X_1 = x)\,\mathrm{P}(X_1 = x)\,,$$

where the summation is over the (finite) support of X_1.

Let $p_{kh} = \mathrm{P}(X_1 = k \text{ and } X_2 = h)$, $k = 1, \ldots, K$ and $h = 1, \ldots, H$, be the two-way table of the joint probabilities (rows k and columns h) associated with discrete variables X_1 and X_2. Then the conditional distribution of X_1 given $X_2 = h$ is derived by standardising column h of the table, that is, dividing its entries by the column total (so that they add up to unity):

$$\mathrm{P}(X_1 = k \,|\, X_2 = h) \;=\; \frac{p_{kh}}{p_{1h} + p_{2h} + \cdots + p_{Kh}}\,.$$

It is useful to introduce the notation p_{k+} and p_{+h} for the marginal probabilities, as they are derived by summing up over the index that is replaced by the summation sign '+'. The conditional distribution of X_2 given $X_1 = k$ is derived by interchanging the roles of the rows and columns:

$$\mathrm{P}(X_2 = h \,|\, X_1 = k) \;=\; \frac{p_{kh}}{p_{k+}}\,.$$

Of course, we assume that none of the margins (row or column totals) vanish; otherwise the row or the column concerned could be deleted.

Variables X_1 and X_2 are independent when none of the conditional probabilities $\mathrm{P}(X_1 = k \,|\, X_2 = h)$ depend on the condition and all are equal to $\mathrm{P}(X_1 = k)$. This follows immediately from the definition of independence. When the conditional probabilities $\mathrm{P}(X_1 = k \,|\, X_2 = h)$ do not depend on h for any h in the support of X_2, the conditional probabilities $\mathrm{P}(X_2 = h \,|\, X_1 = k)$ do not depend on the condition either. Independence is a symmetric property, but conditional probabilities are not symmetric.

A.15 Bivariate Continuous Distributions

Most of the definitions and theory of bivariate discrete distributions carry directly over to bivariate continuous distributions, with the various probabil-

ities replaced by densities. A pair of continuous random variables X_1 and X_2 is said to have a continuous joint distribution when the limit

$$\lim_{\delta \to 0+} \frac{\mathrm{P}(|X_1 - x_1| < \delta \text{ and } |X_2 - x_2| < \delta)}{\delta^2} \tag{A.10}$$

exists for every x_1 and x_2 in the respective supports of X_1 and X_2. The limit is called their *joint density* and is denoted by $f(x_1, x_2)$. When any ambiguity might arise, we indicate the variables as subscripts of the density function, such as $f_{\mathrm{X}_1, \mathrm{X}_2}$ for the density of (X_1, X_2).

Even though every (univariate) continuous distribution has a density, not every pair of continuous distributions has a (bivariate) joint density. As an example, suppose X_1 has the standard uniform distribution and let $X_2 = 1 - X_1$. It is easy to show that the limit in (A.10) is either equal to zero (when $x_1 + x_2 \neq 1$) or diverges to $+\infty$. Therefore the joint distribution of the pair (X_1, X_2) is not continuous.

Two continuous distributions are independent when they have a joint density $f(x_1, x_2)$ and it is equal to the product of the (marginal) densities of the (univariate) component variables: $f(x_1, x_2) = f_{\mathrm{X}_1}(x_1) f_{\mathrm{X}_2}(x_2)$. This definition is equivalent to the natural definition of independence, requiring that

$$\mathrm{P}(X_1 \in U_1 \text{ and } X_2 \in U_2) = \mathrm{P}(X_1 \in U_1)\,\mathrm{P}(X_2 \in U_2)$$

for any pair of intervals U_1 and U_2.

The marginal densities are obtained from the joint density of a pair of variables by integration:

$$f_{\mathrm{X}_1}(x_1) = \int_{-\infty}^{+\infty} f(x_1, x)\,\mathrm{d}x$$

and similarly for X_2.

The conditional density of X_1, given a value of X_2, is defined as

$$f_{\mathrm{X}_1}(x_1 \,|\, X_2 = x_2) = \frac{f(x_1, x_2)}{f_{\mathrm{X}_2}(x_2)},$$

so long as the denominator is positive. In parallel with conditional probabilities, in general, $f_{\mathrm{X}_1}(x_1 \,|\, X_2 = x_2) \neq f_{\mathrm{X}_2}(x_2 \,|\, X_1 = x_1)$. The two sets of conditional distributions are connected by the Bayes theorem for conditional densities,

$$f_{\mathrm{X}_1}(x_1 \,|\, X_2 = x_2) = \frac{f_{\mathrm{X}_2}(x_2 \,|\, X_1 = x_1)\, f_{\mathrm{X}_1}(x_1)}{f_{\mathrm{X}_2}(x_2)},$$

and $f_{\mathrm{X}_2}(x_2) = \int f_{\mathrm{X}_2}(x_2 \,|\, X_1 = x)\, f_{\mathrm{X}_1}(x)\,\mathrm{d}x$; compare with (A.9).

Two continuous random variables are independent when the conditional distribution of one, given a value of the other, does not depend on the value, that is, when $f_{\mathrm{X}_1}(x_1 \,|\, X_2 = x_2) = f_{\mathrm{X}_1}(x_1)$ for all x_1 and x_2 in the respective supports of X_1 and X_2.

A bivariate distribution is defined for *any* two random variables, and these may be of different types, such as one continuous and one discrete. A practical way of defining their joint distribution is by the set of (continuous) conditional distributions $f_{\mathrm{X}_1}(x_1 \,|\, X_2 = x_2)$ given category x_2 of the discrete variable X_2. The marginal density of X_1 is

$$f_{\mathrm{X}_1}(x_1) = \sum_x f_{\mathrm{X}_1}(x_1 \,|\, X_2 = x)\, \mathrm{P}(X_2 = x)\,,$$

where the summation is over the (discrete) support of X_2. The distribution of X_1 is called a *discrete mixture* (of the conditional distributions given the categories of X_2). When X_2 is supported by a finite set of values, the mixture is said to be finite. A natural mechanism (process) for generating a draw from such a distribution is by drawing first a value of X_2, which determines the distribution from which X_1 is to be drawn next. A finite mixture of variables can be expressed as

$$X \;=\; I_1 X_1 + I_2 X_2 + \cdots + I_K X_K\,,$$

where I_k is the *indicator* of category k: $I_k = 1$ if category k is realised and variable X_k used, so that $X = X_k$, and $I_k = 0$ otherwise. It is assumed that $I_1, \ldots, I_K$ are independent of all the *constituent* variables $X_1, \ldots, X_K$. However, they are correlated among themselves, as $I_1 + \cdots + I_K = 1$.

A.16 Operating with Bivariate Distributions

The expectations, medians, quantiles, variances, and the like, are defined for bivariate distributions componentwise, that is, separately for each variable (component), so they entail no new definitions. An important quantity that describes the association of two variables X_1 and X_2 is the *covariance*, denoted by $\mathrm{cov}(X_1, X_2)$. It is defined as

$$\mathrm{cov}(X_1, X_2) \;=\; \mathrm{E}\left[\{X_1 - \mathrm{E}(X_1)\}\,\{X_2 - \mathrm{E}(X_2)\}\right]$$

so long as all three expectations (including those of X_1 and X_2) are well defined. Simple operations yield the identity

$$\mathrm{cov}\,(X_1, X_2) \;=\; \mathrm{E}\,(X_1\, X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2)\,, \tag{A.11}$$

if the expectations are well defined. Variance is a special case of covariance—it is the covariance of a variable with itself: $\mathrm{var}(X) = \mathrm{cov}(X, X)$.

It is expedient to use a single symbol for the pair of variables; $\mathbf{X} = (X_1, X_2)^\top$, so that $\mathbf{X}$ is a 2×1 column vector. A clash with the notation introduced in the context of sample surveys in Section A.3 is unavoidable. Later we will use boldface symbols for vectors of arbitrary (unspecified) finite

length. We write $\mathrm{E}(\mathbf{X}) = \{\mathrm{E}(X_1), \mathrm{E}(X_2)\}^\top$, and define the variance matrix $\mathrm{var}(\mathbf{X})$ as the matrix

$$\mathrm{var}(\mathbf{X}) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) \end{pmatrix}.$$

For an arbitrary 2×1 vector $\mathbf{a}$, $\mathrm{E}\left(\mathbf{a}^\top \mathbf{X}\right) = \mathbf{a}^\top \mathrm{E}(\mathbf{X})$ and $\mathrm{var}\left(\mathbf{a}^\top \mathbf{X}\right) = \mathbf{a}^\top \mathrm{var}\left(\mathbf{X}\right) \mathbf{a}$, so long as each element of $\mathrm{E}(\mathbf{X})$ and $\mathrm{var}(\mathbf{X})$ is well defined. The latter identity implies that $\mathrm{var}(\mathbf{X})$ is a nonnegative definite matrix, and hence

$$\{\mathrm{cov}(X_1, X_2)\}^2 \leq \mathrm{var}(X_1)\, \mathrm{var}(X_2), \tag{A.12}$$

with equality only when X_1 and X_2 are linearly dependent (and when $\mathbf{a}^\top \mathbf{X}$ is equal to a constant for a nonzero vector $\mathbf{a}$).

The *correlation* of two variables X_1 and X_2 that have well-defined (finite) positive variances is defined as

$$\mathrm{cor}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}(X_1)}\, \sqrt{\mathrm{var}(X_2)}}.$$

The inequality in (A.12) implies that $-1 \leq \mathrm{cor}(X_1, X_2) \leq 1$. Further, $\mathrm{cor}(X_1, X_2) = \pm 1$ only when X_1 and X_2 are linearly dependent. In such a case, X_1 and X_2 are said to be *perfectly correlated*; $\mathrm{cor}(X_1, X_2) = 1$ when $X_1 - cX_2$ is constant for a positive constant c, and $\mathrm{cor}(X_1, X_2) = -1$ when $X_1 - cX_2$ is constant for a negative constant c.

Variables X_1 and X_2 are said to be uncorrelated when $\mathrm{cov}(X_1, X_2) = 0$. When X_1 and X_2 are independent they are also uncorrelated. To prove this, we evaluate first $\mathrm{E}(X_1 X_2)$. The distribution of the product $Y = X_1 X_2$ is in general given by the density

$$f_{\mathrm{Y}}(y) = \int f_{\mathrm{X}_1}\left(\frac{y}{x_2} \,\middle|\, X_2 = x_2\right) f_{\mathrm{X}_2}(x_2)\, \mathrm{d}x_2,$$

obtained by conditioning on X_2. Hence, discarding the condition (owing to independence),

$$\mathrm{E}(X_1 X_2) = \int y \int f_{\mathrm{X}_1}\left(\frac{y}{x_2}\right) f_{\mathrm{X}_2}(x_2)\, \mathrm{d}x_2\, \mathrm{d}y,$$

and the change of variables $x_1 = y/x_2$ yields the identity

$$\mathrm{E}(X_1 X_2) = \int x_1 f(x_1)\, \mathrm{d}x_1 \int x_2 f(x_2)\, \mathrm{d}x_2,$$

that is, $\mathrm{E}(X_1)\, \mathrm{E}(X_2)$. Now, $\mathrm{cov}(X_1, X_2) = 0$, according to (A.11), and so also $\mathrm{cor}(X_1, X_2) = 0$. The proof carries over to discrete variables, by replacing each integral with the corresponding summation.

Table A.5. Example of a pair of variables that are uncorrelated but dependent.

	X_1			
X_2	-2	-1	1	2
1	0	$\frac{1}{4}$	$\frac{1}{4}$	0
2	$\frac{1}{4}$	0	0	$\frac{1}{4}$

Note however, that absence of correlation does not imply independence. The following is a simple example of two dependent uncorrelated discrete variables. Variable X_1 is uniformly distributed on $(-2, -1, 1, 2)$, that is, each value in its support has probability $\frac{1}{4}$. Variable X_2 is equal to the absolute value of X_1, so it is uniformly distributed on $(1, 2)$; see Table A.5. The two variables are dependent because $\mathrm{P}(X_1 = 1 \text{ and } X_2 = 2) = 0$ whereas $\mathrm{P}(X_1 = 1) = \frac{1}{4}$ and $\mathrm{P}(X_2 = 2) = \frac{1}{2}$, yet they are uncorrelated because $\mathrm{E}(X_1\, X_2) = \mathrm{E}(X_1) = 0$.

Independence is maintained by transformations. If X_1 and X_2 are independent variables, then so are their transformations $g_1(X_1)$ and $g_2(X_2)$. Of course, the transformed variables $g_1(X_1)$ and $g_2(X_2)$ may be independent when the original variables X_1 and X_2 are not. Further, if X_1 and X_2 are both independent of X_3, then any function of X_1 and X_2 is also independent of X_3.

The covariance has the following 'quadratic' property. If $\mathrm{cov}(X_1, X_2)$ is well defined for a pair of variables X_1 and X_2, then

$$\mathrm{cov}(a_1 X_1 + c_1\,, a_2 X_2 + c_2) \;=\; a_1 a_2\, \mathrm{cov}(X_1\,, X_2)$$

for arbitrary constants a_1, a_2, c_1, and c_2. Hence

$$\mathrm{cor}(a_1 X_1 + c_1\,, a_2 X_2 + c_2) \;=\; \mathrm{sign}(a_1\, a_2)\, \mathrm{cor}(X_1\,, X_2)\,,$$

where sign is the function equal to $+1$ for positive, -1 for negative arguments, and $\mathrm{sign}(0) = 0$; the absolute value of the correlation is unaffected by (non-trivial) linear transformations. Recall that $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$, since standard deviation is a dispersion quantity.

For any two variables with finite variances,

$$\mathrm{var}(X_1 + X_2) \;=\; \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\mathrm{cov}(X_1\,, X_2)\,;$$

the covariance is well defined, so long as any two variances in this expression are. For independent variables X_1 and X_2

$$\mathrm{var}(X_1 + X_2) \;=\; \mathrm{var}(X_1) + \mathrm{var}(X_2)\,. \tag{A.13}$$

This is a key identity for working with random samples, sequences of independent and identically distributed random variables.

For expectations, we have an identity similar to (A.13), except that it holds even when X_1 and X_2 are correlated;

$$\mathrm{E}(X_1 + X_2) \;=\; \mathrm{E}(X_1) + \mathrm{E}(X_2)\,, \tag{A.14}$$

so long as two of the expectations are well defined. This result is derived similarly to its counterpart for the product of two independent variables:

$$\begin{aligned}
\mathrm{E}\,(X_1 + X_2) &= \int\int y f_{\mathrm{X}_1}(y - x_2 \,|\, X_2 = x_2)\, f_{\mathrm{X}_2}(x_2)\,\mathrm{d}x_2\,\mathrm{d}y \\
&= \int x f_{\mathrm{X}_1}(x)\,\mathrm{d}x + \int x_2 f_{\mathrm{X}_2}(x_2)\,\mathrm{d}x_2 \\
&= \mathrm{E}(X_1) + \mathrm{E}(X_2)\,,
\end{aligned}$$

after the change of variables $x = y - x_2$.

A.17 Random Samples

A single observation of a process, yielding a random draw from a distribution, is rarely of much use because it conveys little information. Much more commonly we work with a sequence of independent realisations of the studied process and the resulting random sample from a distribution. This section summarises the properties of random samples.

Suppose X_1, X_2, …, X_n is a random sample from a distribution with finite mean μ and finite variance σ^2. Then the mean of the sample, $\bar{X} = (X_1 + \cdots + X_n)/n$ has the expectation μ and variance σ^2/n. This result, derived from (A.14), supports the estimation of the (finite) expectation of a distribution. It can be rephrased as follows: if the population mean μ is finite, then the sample mean of a random sample is an unbiased estimator of μ and, if the population variance is finite, its sampling variance converges to zero as the sample size increases above all bounds. That is, greater sample size is rewarded by greater precision, confirming the intuition that more data amounts to more information and yields better inference (about μ).

The sample mean is unbiased for the population mean even when the observations are correlated. However, independence is essential for the result about the variance, $\mathrm{var}\left(\bar{X}\right) = \sigma^2/n$. In fact, absence of any correlation among X_j would suffice, but the additional generality is of little practical importance. When any pair of observations X_j have the same positive correlation ρ, then

$$\mathrm{var}\left(\bar{X}\right) \;=\; \frac{\sigma^2}{n}\left(1 + \frac{n-1}{n}\rho\right)\,,$$

so inferences are profoundly handicapped vis-à-vis independent observations with the same sample size.

It might seem that the variance σ^2 would also be estimated without bias naively, as

$$\hat{\sigma}^2_\dagger \;=\; \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2 .$$

That is not the case;

$$\begin{aligned}\mathrm{E}\left(\hat{\sigma}^2_\dagger\right) &= \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left\{\left(X_j-\mu\right)^2\right\} - \mathrm{E}\left\{\left(\bar{X}-\mu\right)^2\right\} \\ &= \sigma^2 - \mathrm{var}\left(\bar{X}\right) \;=\; \frac{n-1}{n}\sigma^2 ,\end{aligned}$$

and so

$$\hat{\sigma}^2 \;=\; \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2$$

is unbiased for σ^2. The unity subtracted from the divisor for $\hat{\sigma}^2$ is interpreted as a *degree of freedom* lost due to estimating a parameter, in this case μ. Indeed, if μ were known, the estimator

$$\hat{\sigma}^2 \;=\; \frac{1}{n}\sum_{j=1}^{n}\left(X_j-\mu\right)^2$$

would be unbiased.

In many settings, the sample mean $\bar{X}$ is efficient, or nearly so, but that is not always the case. For example, for a uniform distribution $\mathcal{U}(\theta_1\,,\theta_2)$, the sample mean is a very inefficient estimator of the population mean $\frac{1}{2}\left(\theta_1+\theta_2\right)$; the average of the extremes, $\frac{1}{2}\left(\max x_j + \min x_j\right)$, is much more efficient. In contrast, the sample mean of a normally distributed random sample is efficient for the population mean, and the average of its extremes is very inefficient.

A.18 Regression

Regression of one variable, Y, on another, X, is the term used for the conditional expectation of Y given a value of X, $\mathrm{E}\left(Y \mid X = x\right)$, treated as a function of the value x. It provides a description, alternative or additional to covariance and correlation, for the association of the two variables. For example, when X and Y are independent, $\mathrm{E}\left(Y \mid X = x\right) = \mathrm{E}(Y)$, and regression of Y on X is constant. The conditional variance $\mathrm{var}\left(Y \mid X = x\right)$, as a function of x, is called the *residual variance*.

The regression can be interpreted as the contribution made by X to the information about Y. When we know the distribution of Y but we do not observe Y, its expectation $\mathrm{E}(Y)$ is a reasonable estimate (prediction) of the value of Y that would or might be observed in the future, especially when

the distribution of Y is symmetric. By definition, $\mathrm{E}(Y)$ is unbiased for the realisation of Y, and its variance is $\mathrm{var}(Y)$. If we know the value of X that accompanies Y, we should be able to predict the value of Y more efficiently (with smaller MSE)—the additional information, in the form of the value of X, should not be detrimental to our efforts at predicting the value of Y. The logic of this statement does not hold up all the time, certainly not when the expectation or the variance of Y is not defined, but a weaker result is obtained from the identity

$$\mathrm{var}(Y) \;=\; \mathrm{E}_{\mathrm{X}}\left\{\mathrm{var}(Y \,|\, X = x)\right\} \;+\; \mathrm{var}_{\mathrm{X}}\left\{\mathrm{E}(Y \,|\, X = x)\right\} \;,$$

in which the 'outer' expectation and variance, with the subscript X, are over the distribution of X, that is, over all possible conditions $X = x$. This identity implies that $\mathrm{var}(Y) \geq \mathrm{E}_{\mathrm{X}}\left\{\mathrm{var}(Y \,|\, X = x)\right\}$, and equality occurs only when the regression $\mathrm{E}\,(Y \,|\, X = x)$ is constant, that is, when X and Y are uncorrelated. Therefore, by using the regression X helps us in predicting Y *on average*.

A.19 Multivariate Distributions

For a vector of more than two variables, some but not all of the definitions for bivariate distributions are extended straightforwardly. Let $\mathbf{X} = (X_1\,, \ldots, X_K)^\top$ be a vector of K variables. Its expectation is defined componentwise,

$$\mathrm{E}(\mathbf{X}) \;=\; \left\{\mathrm{E}(X_1), \ldots, \mathrm{E}(X_K)\right\}^\top \;,$$

and its variance matrix is defined as the matrix of the variances and covariances of its components,

$$\mathrm{var}(\mathbf{X}) \;=\; \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1\,, X_2) & \ldots & \mathrm{cov}(X_1\,, X_K) \\ \mathrm{cov}(X_2\,, X_1) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2\,, X_K) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_K\,, X_1) & \mathrm{cov}(X_K\,, X_2) & \ldots & \mathrm{var}(X_K) \end{pmatrix} .$$

Commonly, the following notation is used: $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$, with diagonal elements $\sigma_k^2 = \mathrm{var}(X_k)$, and $\sigma_{kh} = \mathrm{cov}(X_k\,, X_h)$, with the convention that $\sigma_{kk} = \sigma_k^2$.

The correlation matrix of $\mathbf{X}$ is defined as the matrix of the pairwise correlations $\mathrm{cor}(X_k\,, X_h)$, with unities on the diagonal. Let $\boldsymbol{\sigma} = \mathrm{diag}_{(k)}(\sigma_k)$ be the diagonal matrix with the standard deviations of $\mathbf{X}$ on its diagonal; then $\mathrm{cor}(\mathbf{X}) = \boldsymbol{\sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\sigma}^{-1}$. The variables in $\mathbf{X}$ are said to be uncorrelated if they are pairwise uncorrelated, so that the correlation and variance matrices of $\mathbf{X}$ are diagonal. Recall that the correlation of two variables is not defined if one of them has zero variance or its variance is not defined.

The variables in $\mathbf{X}$ are said to be *mutually independent* if each variable X_k is independent of any transformation $g(\mathbf{X}_{-k})$ that involves the variables in $\mathbf{X}$ except for X_k. Mutual independence is a stricter condition than pairwise independence.

A multivariate distribution, say of a vector $\mathbf{X}$, has univariate marginals, the distributions of $X_1, X_2, \ldots, X_K$, bivariate marginals, the joint distributions of any pair of components X_k and X_h of $\mathbf{X}$ ($k \neq h$), and so on, $(K-1)$-variate marginals, the distributions of $\mathbf{X}_{-k}$, $k = 1, \ldots, K$, obtained by dropping one of the components of $\mathbf{X}$.

A.19.1 Multivariate Normal Distributions

A random vector $\mathbf{X}$ is said to have multivariate normal distribution if any linear combination of its components, $\mathbf{a}^\top \mathbf{X}$, has a univariate normal distribution. This definition may appear a bit awkward, but a vector comprising normally distributed variables as its components may not satisfy this definition, and it is desirable to exclude such (joint) distributions. The joint density of the multivariate normal distribution is

$$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^K} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are a vector of length K and a $K \times K$ symmetric positive definite matrix, respectively. It turns out that $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma}$; this can be verified directly by integration. We denote a K-variate normal distribution by $\mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; the subscript K will be dropped whenever the dimension K is immaterial or is obvious from the context (e.g., when the length of $\boldsymbol{\mu}$ is specified).

Any marginal distribution of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is also (multi- or univariate) normal with its mean vector and variance matrix obtained as the corresponding subvector of $\boldsymbol{\mu}$ and submatrix of $\boldsymbol{\Sigma}$. The proof of this is immediate from the definition of $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, since any linear combination of a subvector is also a linear combination of the original vector. More generally, let $\mathbf{A}$ be a $H \times K$ matrix of constants of full rank H ($H \leq K$). Then $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ implies that $\mathbf{AX} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$. For a choice of $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^\top = \boldsymbol{\Sigma}^{-1}$, $\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ is the vector of zeros (of length K) and $\mathbf{I}$ the $K \times K$ identity matrix. A vector with the distribution $\mathcal{N}_K(\mathbf{0}, \mathbf{I})$ can be constructed from K independent standard normal variates, variables distributed identically according to $\mathcal{N}(0, 1)$.

The class of multivariate normal distributions is complete in the sense that a distribution can be constructed for any vector of means $\boldsymbol{\mu}$ and any positive definite matrix $\boldsymbol{\Sigma}$ of compatible dimensions. The class of multivariate normal distributions can be extended by attaching to a vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ linear combinations of the components of $\mathbf{X}$, and permuting the resulting vector. The variance matrix of this vector is not positive definite, but it has no negative eigenvalues—it is nonnegative definite.

The class of multivariate log-normal distributions is formed by componentwise log-transformations of vectors with multivariate normal distributions. Other classes of distributions, with prescribed univariate marginals can be formed by componentwise transformations, using the standardized uniform distribution as a pivot.

A.19.2 Regression with Normally Distributed Variables

A common task in statistics is concerned with the (joint) distribution of a vector or variable constructed by mathematical operations. In most cases, the solution involves expressions that are not analytic and can at best be evaluated only approximately. Some exceptions from this 'rule' involve the class of normal distributions.

Suppose (X, Y) have bivariate normal distribution with the vector of expectations $(\mu_{\mathrm{X}}\,,\,\mu_{\mathrm{Y}})$, variances σ_{X}^2 and σ_{Y}^2 and covariance σ_{XY}. We derive the regression of Y on X. First, we require the conditional distribution of Y given X, which we denote by $(Y \,|\, X)$. Its density is given by the ratio $f(x, y)/f_{\mathrm{X}}(x)$, and this is equal to

$$\frac{\sigma_{\mathrm{X}}}{\sqrt{2\pi \det(\boldsymbol{\Sigma})}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} x-\mu_{\mathrm{X}} \\ y-\mu_{\mathrm{Y}} \end{pmatrix}^{\top} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} x-\mu_{\mathrm{X}} \\ y-\mu_{\mathrm{Y}} \end{pmatrix} + \frac{(x-\mu_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2} \right\},$$

where $\boldsymbol{\Sigma}$ is the variance matrix of (X, Y). Its inverse is

$$\boldsymbol{\Sigma}^{-1} \;=\; \frac{1}{\sigma_{\mathrm{X}}^2\,\sigma_{\mathrm{Y}}^2 - \sigma_{\mathrm{XY}}^2} \begin{pmatrix} \sigma_{\mathrm{Y}}^2 & -\sigma_{\mathrm{XY}} \\ -\sigma_{\mathrm{XY}} & \sigma_{\mathrm{Y}}^2 \end{pmatrix}.$$

Hence, after simplifying the argument of the exponential, we obtain

$$(Y \,|\, X = x) \;\sim\; \mathcal{N}\left\{ \mu_{\mathrm{Y}} + \frac{\sigma_{\mathrm{XY}}}{\sigma_{\mathrm{X}}^2}(x - \mu_{\mathrm{X}}),\; \sigma_{\mathrm{Y}}^2 - \frac{\sigma_{\mathrm{XY}}^2}{\sigma_{\mathrm{X}}^2} \right\}.$$

Several aspects of this result are remarkable. First, the normal distribution is 'closed' with respect to conditioning; if (X, Y) is (bivariate) normally distributed, then $(Y \,|\, X = x)$ is also normally distributed. Next, the regression of Y on X is linear, with slope equal to $\sigma_{\mathrm{XY}}/\sigma_{\mathrm{X}}^2$. And finally, the residual variance is constant, not depending on the value of X in the condition. The ratio $\sigma_{\mathrm{XY}}^2/\sigma_{\mathrm{X}}^2$ is the reduction of the variance due to knowing the value of X. The squared correlation

$$\rho^2 \;=\; \frac{\sigma_{\mathrm{XY}}^2}{\sigma_{\mathrm{X}}^2\,\sigma_{\mathrm{Y}}^2}\,, \tag{A.15}$$

is the fraction of the variance of Y by which the variance of the prediction of the value of Y is reduced when we know the value of X. Prediction with smaller MSE and, when the prediction is unbiased, with smaller variance is preferred, so variables X for which ρ^2 is high are particularly valuable,

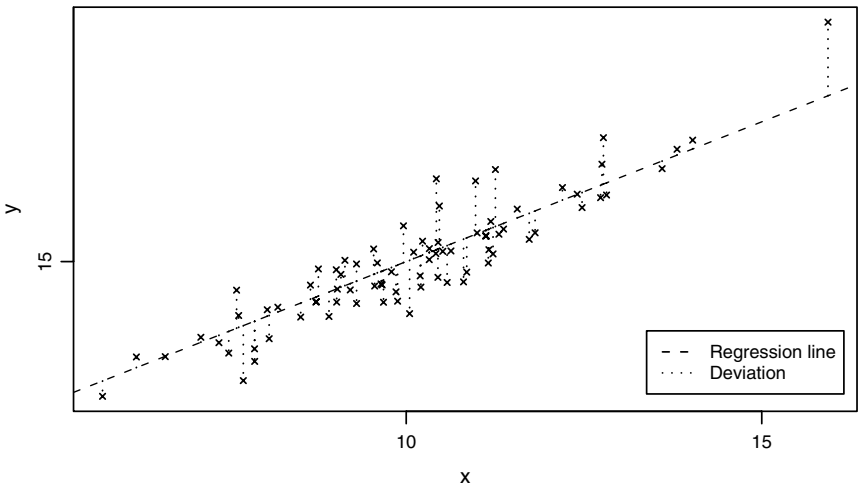

Fig. A.8. Example of ordinary regression.

especially when their observation is easy and inexpensive. When X and Y are strongly associated and observing X is much cheaper than observing Y, we may observe only X and pay a small 'penalty' of uncertainty due to the positive residual variance.

The regression of Y on X can be obtained also as the linear transformation $a + bX$ that differs from Y least, that is, by minimising $\mathrm{E}\left\{(Y - a - bX)^2\right\}$. This expectation is equal to

$$\mathrm{var}(Y - bX) + \{\mathrm{E}(Y - a - bX)\}^2 \;=\; \sigma_{\mathrm{Y}}^2 - 2b\sigma_{\mathrm{XY}} + b^2\sigma_{\mathrm{X}}^2 + (\mu_{\mathrm{Y}} - a - b\mu_{\mathrm{X}})^2\,,$$

a quadratic function of a and b. Its minimum is attained when $a = \mu_{\mathrm{Y}} + b\mu_{\mathrm{X}}$ and $b\sigma_{\mathrm{X}}^2 - \sigma_{\mathrm{XY}} = 0$. Hence $b = \sigma_{\mathrm{XY}}/\sigma_{\mathrm{X}}^2$.

A geometric interpretation of regression of Y on X is that in the plot of the values of X against Y, the points are scattered around the straight (regression) line with intercept $a = \mu_{\mathrm{Y}} + b\mu_{\mathrm{X}}$ and slope $\sigma_{\mathrm{XY}}/\sigma_{\mathrm{X}}^2$; hence the terms *regression slope* for $b = \sigma_{\mathrm{XY}}/\sigma_{\mathrm{X}}^2$ and regression line for $a + bx$. The dispersion of the points around the regression line, measured vertically, is equal to the residual variance. See Figure A.8 for an illustration.

A.20 Formulating Inferences

The starting point of a statistical analysis is to specify the target—the population quantity that we would like, in ideal circumstances, to determine. This goal is reduced to estimation, based on the values of the key (and maybe some other) variables on a sample of subjects. The outcome of the analysis is an

estimate of the target. It is an established practice in statistics to indicate the precision of the estimator that was applied, to inform how much faith can be placed in the estimate. The MSE is commonly adopted as a measure of the precision. It is equal to the sampling variance when the estimator is unbiased. Usually the MSE cannot be determined and has to be estimated. Thus, the estimation task comprises two parts:

- estimation of the target θ, by evaluating an estimator $\hat{\theta}$;
- estimation of $\mathrm{MSE}(\hat{\theta};\theta)$, by evaluating an estimator $\widehat{\mathrm{MSE}}\left(\hat{\theta};\theta\right)$.

Since efficient estimation is valued, the analyst's reward should be inversely proportional to $\mathrm{MSE}(\hat{\theta};\theta)$ and the analyst has an obvious incentive to present the estimate in as good a light as possible. One unfair means of doing this is by estimating $\mathrm{MSE}(\hat{\theta};\theta)$ with a negative bias (underestimating the MSE). In contrast, overestimating the MSE is comparable to underselling the product of the analyst's effort—understating the quality of the analysis.

Commonly, the term *estimation* is used for generating a statement of the form $\{\hat{\theta},\ \widehat{\mathrm{MSE}}(\hat{\theta};\theta)\}$ for a target θ. We say that such an estimation is *dishonest* if the MSE is underestimated. Note that underestimation does not mean that $\widehat{\mathrm{MSE}}(\hat{\theta};\theta) < \mathrm{MSE}(\hat{\theta};\theta)$, because the realised value of $\widehat{\mathrm{MSE}}$ may exceed the MSE even when it does not do so in expectation (on average in replications). A more appropriate, although rather cumbersome, term for underestimation might be 'dishonest in the long run', standing for $\mathrm{E}\{\widehat{\mathrm{MSE}}(\hat{\theta};\theta)\} \leq \mathrm{MSE}(\hat{\theta};\theta)$.

A.20.1 Confidence Intervals

The 'honestly' estimated root-MSE indicates how far we can expect the estimate to be from the target on average, if we replicated the sampling and estimation processes many times. An alternative formulation of the inference is by a *confidence interval*. The confidence interval is defined as an interval $(C_{\mathrm{L}}\,,\,C_{\mathrm{H}})$ delimited by sample quantities C_{L} and C_{H} that satisfy the inequality

$$\mathrm{P}(C_{\mathrm{L}} < \theta \text{ and } C_{\mathrm{H}} > \theta) \;\geq\; \alpha\,, \tag{A.16}$$

where α, called the *level of confidence*, is a prescribed (a priori set) value in the range $(0, 1)$. Practical choices for α are values close to unity, so that the interval with the data-dependent (random) bounds C_{L} and C_{H} is very likely to contain the target, an unknown constant. As a convention, $\alpha = 0.95$ and $\alpha = 0.99$ are often used, allowing an error rate (probability of not covering the target by the confidence interval) of not more than 5% and 1%, respectively. We may set out to obtain a confidence interval with a particular level of confidence, but as a result of errors, incorrect assumptions, or approximations, we end up with a level of confidence that does not satisfy the condition in (A.16). The probability on the left-hand side of (A.16) is called the *coverage rate*. It may depend on some parameters, even on the target itself.

In analogy with honest estimation, we say that a confidence interval is honest if condition (A.16) is satisfied; that is, if the coverage rate does not fall short of the (intended) level of confidence. For the confidence intervals for a specified target and level of confidence, we can define a partial ordering. Confidence interval A is said to be narrower than confidence interval B, if A is a subset of B. Note that both A and B are data-dependent, so the lengths of A and B, as random variables, may overlap even when A is narrower than B.

For a given target and confidence level, there may be several alternative confidence intervals. We should discard all dishonest intervals and compare the honest ones by their length or expected length. Shorter (narrower) confidence intervals are preferred because they narrow down the range of plausible values of the target. Suppose confidence interval A has coverage rate 95%, equal to the level of confidence, and has constant length 2.7. Suppose another confidence interval, B, has coverage rate 96.5% and length constant 2.5. Confidence interval B is preferred because it is shorter. The fact that it could, in principle, be improved by reducing it so that its coverage rate would match the level of confidence is beside the point; there may be a confidence interval better than B, but it is not A. The length of a confidence interval is a sample quantity, so it may be random. This makes the comparison of confidence intervals more difficult than this example may suggest. Comparison of the realisations of the two confidence intervals is not sufficient, although it may happen that one confidence interval is longer than another for every replication (with probability equal to one).

In some settings, only confidence intervals of the form $(-\infty, C_{\mathrm{H}})$ are of interest. Such intervals, as well as intervals of the form $(C_{\mathrm{L}}, +\infty)$ are called *one-sided*. We can adopt any function of the pair $(C_{\mathrm{L}}, C_{\mathrm{H}})$ as the criterion for what we regard an optimal confidence interval among the intervals with the prescribed level of confidence. However, only three criteria are of any practical relevance: minimum width of the interval, preferring intervals that are symmetric around the estimate ($C_{\mathrm{L}} = \hat{\theta} - \xi$ and $C_{\mathrm{C}} = \hat{\theta} + \xi$ for a suitably defined sample quantity ξ), and using only one-sided intervals (either $C_{\mathrm{L}} = -\infty$ and preferring small C_{H} or $C_{\mathrm{H}} = +\infty$ and preferring large C_{L}). And finally, instead of confidence intervals, we may consider confidence *regions*, which may be any subsets of real numbers. We need to do this very rarely, most often to consider a pair of intervals symmetric around zero, such as $(-a, -b) \cup (a, b)$ for some positive numbers $a < b$, when $\left(a^2, b^2\right)$ would be a confidence interval for the square of the parameter.

A confidence interval (or region) is interpreted as *a* range of plausible values of a parameter. Naturally, it is often corrupted to *the* range of possible values. We should bear in mind that the confidence limits are sample quantities and a 'surprise', in the form of $\theta \notin (C_{\mathrm{L}}, C_{\mathrm{H}})$, has a positive probability. When several confidence intervals are considered, the probability of such a surprise in connection with at least one of the intervals may be much greater than the error rate associated with a single interval.

Problems and Exercises

A.1. Formulate for your country the current definitions of

(a) the citizen;
(b) the resident.

Consult the appropriate sources for the legal definitions. Discuss elements or clauses in these definitions that might be ambiguous or contentious.

A.2. Describe the population of all applications to study at a university in your country in a recent year. Relate it to the population of all applicants in the year, to all functioning universities and their departments, and to the population of all eligible persons. Define some variables and structures (divisions into clusters) for these populations. Describe the variables by type and specify their supports.

A.3. On the web site of a national statistical institute, such as the Office for National Statistics (ONS, `www.ons.gov.uk`) or the Catalan Statistical Institute (IDESCAT, `www.idescat.es`), find summaries of basic sociodemographic and economic indicators for a recent year and determine which of them were established by enumeration and which by a survey. Identify some consumers of such information and how they would draw benefit from them. For any sample quantity you come across, identify its target and look for any information about its precision. Are these sample quantities estimates or estimators? What about their precision?

A.4. In the training of assessors (graders) of essays in a particular academic subject, the trainees are presented essays as typical examples that should be marked with a particular score. The trainees are not informed about these scores. In one such exercise, copies of an essay that is supposed to be marked 75 are distributed and the twelve trainees mark it, without conferring, as 72, 70, 83, 79, 76, 76, 75, 69, 73, 80, 70, and 75. In another scheme, the trainees discuss an essay that is supposed to be marked 62 and collectively come to the conclusion that it should be marked 65. In a third scheme, the trainees are presented an essay that is supposed to be marked 95, are informed about this score, and are asked to agree or raise objections to the score. Discuss the merits and drawbacks of each scheme for essay marking and for assessment of the accuracy of the marking in the case when each essay is marked only by one person assigned at random from the pool of qualified assessors.

A.5. In a particular context, the sampling variance of an estimator is 2.6 and its bias for a given target θ is 1.2. An alternative estimator of the same target θ is unbiased, with the sampling variance 3.5. Which estimator of θ is more efficient? Could your choice, when applied to a dataset, be more distant from the target than the other estimate?

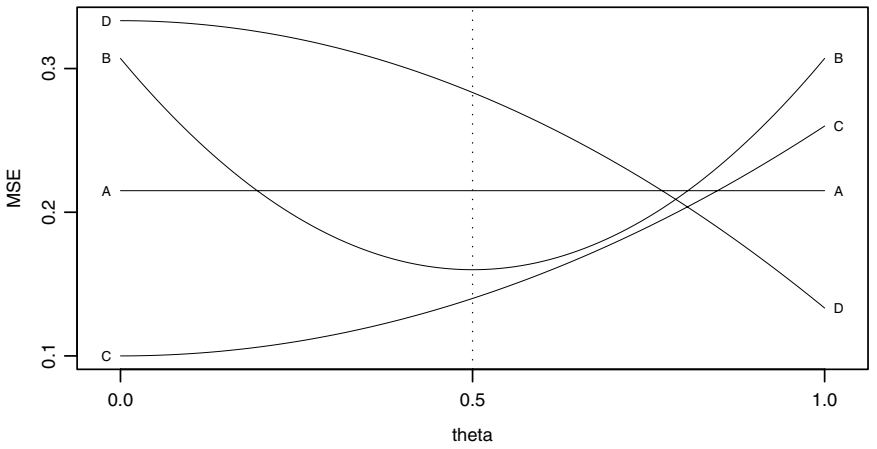

Fig. A.9. Plot of the MSEs of four estimators of θ as functions of θ. For Exercise A.9.

A.6. In a population of households in a region, the percentage of those with per capita income below a certain level is of interest. A simple random sample of households is selected and their size (number of members of the household) and income in the last year, rounded to hundreds of $US, are established. Define the naive estimator of the sought percentage. Consider how different it would be if the income were established with greater or smaller precision.

A.7. Derive the identity in (A.1)

(a) from the definitions of expectation and variance;
(b) from the definition of replications.

Suppose the square root of the MSE (root-MSE) is equal to the bias. What can be said about the sampling variance?

A.8. Calculate the mean and variance of the variable defined on a population of 22 members by their values:

$$3,\ 7,\ 11,\ 11,\ 13,\ \ 4,\ 7,\ 8,\ 11,\ 12,\ \ 2,\ 7,\ 9,\ 15,\ 18,\ \ 6,\ 6,\ 9,\ 12,\ 12,\ \ 10,\ 15\,.$$

Find the interquantile range of this variable. Present these values in a way that is more informative about the distribution of the variable.

A.9. Figure A.9 is a plot of the MSEs of four estimators of the same target θ as functions of the parameter θ in the interval $(0, 1)$. Is any of the estimators uniformly more efficient than one of the others? Describe the strengths and weaknesses of the four estimators. Suppose we can reduce our attention to θ in the interval $\left(0, \frac{1}{2}\right)$. Would your answer be different?

A.10. Which of the following summaries of a variable X are location quantities?

median; $\max_i X_i$; the standard deviation of X; the value for member 1, X_1; the interquantile range; sum of squares of the values of X; the number of positive values of X; the distribution of X.

Which are dispersion quantities?

A.11. What is the mode of a distribution with support on the digits 0, 1, ..., 9, if the value 0 is attained for more than 60% of the population? Is there a distribution with the same support that has mode at 5 and the frequency of this value is 15%? Give an example of a symmetric trimodal distribution with this support.

A.12. Construct a without-replacement sampling design that is different from the simple random sampling design and in which every member has the same probability of being included in the sample. On a population of small size, say, $N = 4$, construct a sampling design in which each pair has the same probability of being included in the sample but each member has a different probability.

A.13. Summarise the advantages of designs with stratification and clustering in a national survey of individuals or households. Find in the literature or on the Web an example of a national survey with a stratified clustered sampling design and discuss its details and what information about the design is not given.

A.14. A market research company engages interviewers who are assigned to be at particular locations in the centres of some cities and large shopping areas of the UK or United States at selected times of the day (lunchtime, the afternoon rush hour, Saturday morning, and the like). They collect responses to a questionnaire about a car brand from a given number (quota) of adult English-speaking passersby. Comment on the difficulties in making population inferences based on the collected data.

A.15. Construct a small population with the values of one variable and design a sampling scheme for this population. Replicate the sampling design in this population and record the values of an estimator based on the sample drawn. Compare the replicate values of the estimator with the target and estimate the bias and MSE of the estimator.

This exercise can be conducted on paper, with a device for random selection, e.g., based on a table of random numbers, but it is much more effective when executed on a computer using a more extensive population.

A.16. Generate at least 100 values of the discrete manifest variable Y for a latent variable X according to the matrix of probabilities $\mathrm{P}(Y = k, X = h)$ for $k = 1, 2, 3, 4$ (rows) and $h = 1, 2, 3, 4$ (columns)

$$\begin{pmatrix} \frac{1}{10} & \frac{1}{25} & \frac{1}{100} & 0 \\ \frac{1}{20} & \frac{1}{5} & \frac{1}{20} & \frac{1}{100} \\ \frac{1}{100} & \frac{1}{10} & \frac{1}{5} & \frac{1}{20} \\ \frac{1}{50} & \frac{1}{50} & \frac{1}{25} & \frac{1}{10} \end{pmatrix}.$$

Find the marginal distributions of X and Y from this matrix and estimate the latter from the generated values of Y.

For each value $k = 1, 2, 3$, and 4 of X, find the bias, measurement variance, and MSE of the misclassification process given by this matrix. How can it be verified using computer-generated values of X and Y?

A.17. Relate what is commonly understood by the term 'impartial jury' in the criminal justice system, 'impartial assessment' in educational testing, and 'impartial referee' in a sport event to the definition of impartiality in Section A.6. How is impartiality related to the frequency of incorrect decisions in these examples?

A.18. Replicate the example in Figure A.5 and Table A.3 in your own computational environment with probabilities and numbers of replications of your choice.

A.19. Revise the rules for integration of continuous functions on finite intervals and how integration and differentiation are associated. When can the order of integration and differentiation be exchanged, that is,

$$\frac{\partial}{\partial x} \int_0^1 f(u, x)\, \mathrm{d}u \;=\; \int_0^1 \frac{\partial f(u, x)}{\partial x}\, \mathrm{d}u\,?$$

Under what conditions do we have the identity

$$\frac{\partial}{\partial x} \int_0^x f(u)\, \mathrm{d}u \;=\; f(x)$$

for a continuous function f?

A.20. Find the distribution function $F(x) = \mathrm{P}(X < x)$ of a continuous variable with support on $(-1, 1)$ and density $f(x) = C(x+1)(x+2)$ for a suitable constant C. Sketch the density and describe the properties of this distribution. What is the probability of a positive value of X?

A.21. For a random sample from the uniform distribution on $(\theta_1\,, \theta_2)$, consider the following estimators of the population mean:

(a) the sample mean;
(b) the sample median;
(c) the average of the first and third sample quartiles;
(d) the average of the maximum and minimum of the sample.

Compare these estimators by simulation. That is, set a sample size n (say, $n = 150$) and the limits $\theta_1 < \theta_2$, and replicate many (say, $K = 1000$) times the following steps:

(A) draw a random sample of size n from $\mathcal{U}(\theta_1, \theta_2)$;
(B) evaluate the estimators a,–d, on this sample.

Finally, calculate the empirical biases, variances, and MSEs of these estimators. Explain why the study can focus on the standard uniform distribution, setting $\theta_1 = 0$ and $\theta_2 = 1$, so that the target is $\frac{1}{2}$.
If you have difficulties with this question, repeat the study for several settings of θ_1 and θ_2 and compare the results. As an alternative, you could conduct the separate studies within a single simulation study by linearly transforming each sample from $\mathcal{U}(0, 1)$ to be a sample from $\mathcal{U}(\theta_1, \theta_2)$.

A.22. Repeat the study in the previous exercise, with the normal distribution in place of the uniform. Explain why this study can focus on the standard normal distribution.

A.23. For revision. How is the integral of a function over an infinite interval defined? Revise the calculus of infinite sequences and sums, in particular the principal results about their convergence.
Let $p_k = 1/k - 1/(k+1)$ for $k = 1, 2, \ldots$. Do these values define a distribution on the integers? If not, do Cp_k for a positive constant C?

A.24. Show that when the expectation of the square of a distribution, $\mathrm{E}\left(X^2\right)$, is well defined, then so is the expectation $\mathrm{E}(X)$. Find a distribution for which $\mathrm{E}(X)$ is well defined, but $\mathrm{var}(X)$ is not.
Hint: Consider the functions $f(x) = x^{-k}$ on $(0, 1)$ and $(1, +\infty)$ for $k > 0$.

A.25. Show that when a discrete distribution on the integers $0, 1, \ldots$ has an expectation, then the summation $\sum_{k=1}^{\infty} F_k$, where $F_k = p_0 + \cdots + p_k$, converges and its limit is equal to the expectation.

A.26. Generate the empirical distribution of the continuous uniform distribution on $(0, 1)$ and the negative logarithm, $-\log(X)$, of the values. Compare the latter with the empirical distribution of the exponential distribution with parameter $\theta = 1$.

A.27. Carry out a simulation study to compare estimators of your choice for the expectation of the distribution with density $f(x) = 2x$ for $x \in (0, 1)$. Include the trimmed mean among the estimators. The p%-trimmed mean is defined as the mean of the subsample formed by discarding $\frac{1}{2}p\%$ of the smallest and $\frac{1}{2}p\%$ of the largest observations.

A.28. Derive $\mathrm{E}\left(X^k\right)$ for a centred normally distributed random variable X and integer k. Derive $\mathrm{var}\{(X - \mu)^2\}$ for an arbitrary distribution $\mathcal{N}(\mu, \sigma^2)$.
Hint: Apply integration by parts.

A.29. *Asymptotic normality.* For an estimator of your choice, show empirically that its distribution approaches normality as the sample size increases. Hint: Conduct separate simulations of the estimator for samples of sizes 10, 30, 100, 300, 1000, or similar, and draw the empirical distributions. Each simulation should be based on the same number of replications.

A.30. Derive the moments $\mathrm{E}\left(X^k\right)$ for X with distribution $\mathcal{U}(0,1)$. Confirm that

$$-\log\left\{\mathrm{E}(X^k)\right\} \,<\, \mathrm{E}\left\{-\log(X^k)\right\}$$

both analytically and by simulations. Check that

$$\mathrm{var}\left(X^k\right) \,=\, \frac{k^2}{(k+1)^2(2k+1)}\,.$$

Derive an expression for $\mathrm{E}\{X^k - 1/(k+1)\}^3$.

A.31. Check the expressions for the expectations and variances of the beta and gamma distributions given in Section A.12. Prove the identity

$$\mathrm{E}\left\{(X-\mu)^3\right\} \,=\, \mathrm{E}\left(X^3\right) - 3\mu\,\mathrm{var}(X) - \mu^3$$

for any distribution with expectation μ and finite third moment $\mathrm{E}(X^3)$.

A.32. Derive the probability in (A.6) and prove the statements about the modes of binomial distributions made in Section A.13.
Hint: Consider the ratios of $\mathrm{P}(X = k)/\mathrm{P}(X = k+1)$ for integers k and variable X with binomial distribution.

A.33. Prove that for every fixed integer k the probability $\mathrm{P}(X = k\,;\, n, p)$ for a binomial variable $\mathcal{B}(n,p)$ converges to the probability $\mathrm{P}(Y = k; \lambda)$ for a Poisson-distributed variable when n and p are such that $np_n \to \lambda$ as $n \to \infty$.

A.34. Devise a way of drawing a random sample from a binomial distribution using a source of independent draws from a uniform distribution.

Describe how a random sample from a discrete uniform distribution could be generated using a source of independent draws from the standard uniform distribution. How about a random sample from a multinomial distribution?

A.35. Show that for an exponentially distributed variable X, its conditional distribution given that $X > t$ is also exponential for every real constant t. Show that no binomial distribution $\mathcal{B}(n,p)$ with $n > 2$ has this property for any $k < n$. What about the geometric distributions?

A.36. Table A.6 gives the conditional probabilities $\mathrm{P}(X_1 = k \,|\, X_2 = h)$ for $h = 0, 1, \ldots, 9$. Marginally, X_2 has the binomial distribution $\mathcal{B}(9, 0.4)$. Find the marginal distribution of X_1. Construct the table of conditional probabilities $\mathrm{P}(X_2 = h \,|\, X_1 = k)$.
Note: This exercise is intended not for pencil and paper but for computer programming.

Table A.6. The conditional distributions of X_1 given values of X_2. The entries of the table are $\mathrm{P}(X_1 = k \,|\, X_2 = h)$, $h = 0, 1, \ldots, 9$ and $k = 0, 1, 2$.

	X_2									
X_1	0	1	2	3	4	5	6	7	8	9
0	0.70	0.60	0.55	0.50	0.40	0.40	0.35	0.30	0.25	0.15
1	0.20	0.20	0.20	0.20	0.25	0.30	0.35	0.40	0.40	0.45
2	0.10	0.20	0.25	0.30	0.35	0.30	0.30	0.30	0.35	0.40

A.37. Show that independence is maintained by transformations. That is, if X_1 and X_2 are independent, then so are $g_1(X_1)$ and $g_2(X_2)$ for any two functions g_1 and g_2. Show by example that the converse does not apply. That is, find transformations g_1 and g_2 and variables X_1 and X_2 such that X_1 and X_2 are not independent, but $g_1(X_1)$ are $g_2(X_2)$ are.
Hint: Focus on discrete variables with few categories.

A.38. Explore the variety of distributions that can be generated as mixtures of two or three univariate normal distributions. Write a programme for plotting the densities of such distributions and execute it for a range of trinomial probabilities (p_1, p_2, p_3), expectations (μ_1, μ_2, μ_3), and variances $\left(\sigma_1^2, \sigma_2^2, \sigma_3^2\right)$. Explain why no generality is lost by setting $\mu_1 = 0$ and $\sigma_1^2 = 1$. Present the conditional and the mixture densities in a suitable graph.

A.39. For a discrete mixture of a set of continuous distributions with densities $f_1, f_2, \ldots, f_K$, derive the conditional distribution of the category given the realised value of the mixture.

A.40. Find a class of distributions that are closed with respect to mixing. That is, if K distributions belong to this class, then so does their finite mixture with any multinomial distribution.

A.41. Revise the following topics: matrix calculus, including inverse and determinant; properties of (symmetric) positive and nonnegative definite matrices; and eigenvalue and other decompositions.

A.42. Show that if variable X is a mixture of variables $X_1, \ldots, X_K$, each of them with finite variance, then

$$\begin{aligned} \mathrm{E}(X) &= \sum_{k=1}^{K} p_k \,\mathrm{E}\left(X_k\right) , \\ \mathrm{cov}\left(X, X_k\right) &= p_k \,\mathrm{var}\left(X_k\right) , \\ \mathrm{var}(X) &= \sum_{k=1}^{K} p_k \left[\mathrm{var}(X_k) + \{\mathrm{E}(X_k)\}^2\right] - \{\mathrm{E}(X)\}^2, \end{aligned}$$

where p_k is the probability of category k.

A.43. Derive the covariance and correlation of a pair of indicators I_k and I_h in a multinomial distribution. The indicator for category k is defined as $I_k = 1$ if component k is realised and $I_k = 0$ otherwise.

A.44. Derive the distributions of the sum and product of two independent uniformly distributed variables. Check your results by simulations.

A.45. Suppose a set of identically distributed variables $X_1, X_2, \ldots, X_K$ is such that each pair of them has the same correlation $\rho = \text{cor}(X_{k_1}, X_{k_2})$. Find the highest lower bound for ρ.
Hint: Consider the variance of the sample mean.
Relate this result to the correlation of the K-nomial distribution with equal probabilities $1/K$.

A.46. Construct a symmetric matrix with unities on its diagonal and all other entries in the range $(-1, 1)$ that is not a correlation matrix. Interpret this matrix as an example that its symmetry and all correlations in the range $[-1, 1]$ are not sufficient for it to be a variance matrix.

A.47. Compare the regressions $\text{E}(X \mid Y)$ and $\text{E}(Y \mid X)$ for a vector (X, Y) with bivariate normal distribution and relate them to the correlation $\text{cor}(X, Y)$. Explain why the product of the regression slopes is not equal to unity when the two variables are not perfectly correlated.

A.48. Explore how the regression $\text{E}(Y \mid X)$ for a vector (X, Y) with bivariate normal distribution is altered when Y is replaced by $Y+\xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$ is independent of of both X and Y. What change is brought about by replacing X with $X + \xi$?

A.49. Suppose (C_L, C_H) is an honest confidence interval for a population quantity μ. Find honest confidence intervals for $\exp(\mu)$ and μ^2. Discuss the advantages and drawbacks of constructing confidence intervals for σ^2, $\log(\sigma)$, and $\sqrt{\sigma}$ for a population variance σ^2.

References

1. Aitkin, M., Anderson, D. A., Hinde, J. P., and Francis, B. J.: *Statistical Modelling in GLIM*, 2nd ed. Oxford University Press, Oxford (2003)
2. Aitkin, M., and Longford, N. T.: Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society* Series A **149**, 1–43 (1986)
3. Akaike, H.: A new look at statistical model identification. *IEEE Transactions on Automatic Control* AU–**19**, 716–722 (1974)
4. Armitage, P.: *Statistical Methods in Medical Research*. Blackwell, Oxford (1971)
5. Bayes, T.: An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 330–418 (1763)
6. Benjamini, Y., and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* Series B **57**, 289–300 (1995)
7. Berger, J. O.: *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer-Verlag, New York (1985)
8. Bernardo, J. M., and Smith, A. F. M.: *Bayesian Theory*. Wiley, New York (1994)
9. Bird, S. M. (Chair of the Working Party), Cox, D. R., Farewell, V. T., Goldstein, H., Holt, D., Smith, P. C.: Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society* Series A **168**, 1–27 (2005)
10. Box, G. E. P., and Tiao, G. C.: *Bayesian Inference in Statistical Analysis*. Wiley, New York (1973)
11. Breslow, N.: Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44 (1984)
12. Breslow, N., and Clayton, D. G.: Approximate inference in generalized linear models. *Journal of the American Statistical Association* **88**, 9–25 (1993)
13. Bryk, A. S., and Raudenbush, S. W.: *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA (1992)
14. van Buuren, S., Boshuizen, H. C., and Knook, D. L.: Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694 (1999)
15. Carroll, R. J., Ruppert, D., and Stefanski, L. A.: *Measurement Error in Nonlinear Models*. Chapman and Hall, London (1995)

16. Cassella, G., and George, E. I.: Explaining the Gibbs sampler. *The American Statistician* **46**, 167–174 (1992)
17. Chatfield, C.: *The Analysis of Time Series: An Introduction*, 6th ed. Chapman and Hall/CRC, Boca Raton (2004)
18. Chow, S. C., and Liu, J. P.: *Design and Analysis of Bioequivalence and Bioavailability Studies*. Marcel Dekker, New York (1992)
19. Cochran, W. G.: The planning of observational studies of human populations. *Journal of the Royal Statistical Society* Series A **128**, 134–155 (1965)
20. Cochran, W. G.: *Sampling Techniques*, 3rd ed. Wiley, New York (1977)
21. Cochran, W. G., and Cox, G.: *Experimental Designs*. Wiley, New York (1957)
22. Cochran, W. G., and Rubin, D. B.: Controlling bias in observational studies: a review. *Sankya* A **35**, 417–446 (1973)
23. Cook, J. R., and Stefanski, L. A.: A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314–1328 (1995)
24. Copas, J. B., and Shi, J. Q.: A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* **10**, 1–15 (2001)
25. Cox, D. R.: *Planning of Experiments*. Wiley, New York (1958)
26. Cox, D. R.: Regression models and life tables. *Journal of the Royal Statistical Society* Series B **34**, 187–220 (1972)
27. Cox, D. R., and Hinkley, D. V.: *Theoretical Statistics*. Chapman and Hall, London (1974)
28. Cox, D. R., and Oakes, D.: *Analysis of Survival Data*. Chapman and Hall, London (1984)
29. Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N.: *Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*. Wiley, New York (1972)
30. Crowder, M. J., and Hand, D. J.: *Analysis of Repeated Measures*. Chapman and Hall, London (1990)
31. Davidian, M., and Giltinan, D. M.: *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London (1995)
32. DeGroot, M. H.: *Optimal Statistical Decisions*. McGraw-Hill, New York (1970)
33. Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* Series B **39**, 1–38 (1977)
34. Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K.: Estimation in covariance component models. *Journal of the American Statistical Association* **76**, 341–353 (1981)
35. Dennis, J. E. Jr., and Schnabel, R. B.: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ (1983)
36. DerSimonian, R., and Laird, N. M.: Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188 (1986)
37. Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L.: *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press, Oxford (2002)
38. Dobson, A.: *An Introduction to Generalized Linear Models*, 2nd ed. Chapman and Hall/CRC, London (2001)

39. Draper, D., and Gittoes, M.: Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society* Series A **167**, 449–474 (2004)
40. DuMouchel, W. H., and Harris, J. E.: Bayes methods for combining results of cancer studies in humans and other species. *Journal of the American Statistical Association* **78**, 293–315 (1983)
41. Duval, S., and Tweedie, R.: Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56**, 455–463 (2000)
42. Efron, B., and Morris, C.: Limiting the risk of Bayes and empirical Bayes estimators—part I: the Bayes case. *Journal of the American Statistical Association* **66**, 807–815 (1971)
43. Efron, B., and Morris, C.: Limiting the risk of Bayes and empirical Bayes estimators—part II: the empirical Bayes case. *Journal of the American Statistical Association* **67**, 130–139 (1972)
44. Efron, B., and Morris, C.: Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117–130 (1973)
45. Efron, B., and Morris, C.: Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**, 311–319 (1975)
46. Efron, B., and Tibshirani, R.: *An Introduction to the Bootstrap*. Chapman and Hall, London (1993)
47. Eisenhart, C.: Effect of rounding or grouping data. In: Eisenhart, C., Hastay, M. W., and Wallis, W. A. (eds.) *Selected Techniques of Statistical Analysis*. McGraw-Hill, New York (1947)
48. Fay, R. A., and Herriot, R. E.: Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277 (1979)
49. Firth, D., and Harris, I. R.: Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545–555 (1991)
50. Fisher, R. A.: *The Design of Experiments*, 2nd ed. Oliver and Boyd, London (1949)
51. Fleiss, J. L.: *The Design and Analysis of Clinical Experiments*. Wiley, New York (1986)
52. Frangakis, C., and Rubin, D. B.: Principal stratification in causal inference. *Biometrics* **58**, 21–29 (2002)
53. Freeman, P. R.: The performance of the two-stage analysis of two-treatment cross-over designs. *Statistics in Medicine* **8**, 1421–1432 (1989)
54. Fuller, W. A.: *Measurement Error Models*. Wiley, New York (1987)
55. Gelfand, A. E., and Smith, A. F. M.: Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409 (1990)
56. Gelman, A., Carlin, J. B., Stern, H., and Rubin, D. B.: *Bayesian Data Analysis*, 2nd ed. Chapman and Hall/CRC, New York (2003)
57. Gilks, W. R., Richardson, S., and Spiegelhalter, D. (eds.): *Practical Markov Chain Monte Carlo*. Chapman and Hall, New York (1996)
58. Gill, P. E., Murray, W., and Wright, M. H.: *Practical Optimization*. Academic Press, New York (1981)
59. Glass, G.: Primary, secondary, and meta-analysis of research. *Educational Researcher* **5**, 3–8 (1976)

60. Goldstein, H.: *Multilevel Statistical Models*, 3rd ed. Edward Arnold, London (2003)
61. Goldstein, H., and Healy, M. J. R.: The graphical presentation of a collection of means. *Journal of the Royal Statistical Society* Series A **158**, 175–177 (1995)
62. Goldstein, H., and Spiegelhalter, D. J.: League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society* Series A **159**, 385–443 (1996)
63. Good, I. J.: *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, MA (1965)
64. Greenland, S.: Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology* **135**, 1301–1309 (1994)
65. Greenland, S.: Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society* Series A **168**, 267–306 (2005)
66. Hartigan, J. A.: *Bayes Theory*. Springer-Verlag, New York (1983)
67. Hartley, H. O., and Rao, J. N. K.: Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108 (1967)
68. Harvey, A. C.: *Time Series Models*, 2nd ed. Harvester Wheatsheaf, London (1993)
69. Harville, D. A.: Bayesian inference for variance components using only error contrasts. *Journal of the American Statistical Association* **69**, 383–385 (1974)
70. Harville, D. A.: *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York (1997)
71. Healy, M. J., and Westmacott, M.: Missing values in experiments analyzed on automatic computers. *Applied Statistics* **5**, 203–206 (1956)
72. Hedges, L. V., and Olkin, I.: *Statistical Methods for Meta-Analysis*. Academic Press, Orlando (1985)
73. Heitjan, D. F.: Inference from grouped continuous data: a review. *Statistical Science* **4**, 164–183 (1989)
74. Heitjan, D. F., and Little, R. J. A.: Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics* **40**, 13–29 (1991)
75. Heitjan, D. F., and Rubin, D. B.: Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association* **85**, 304–314 (1990)
76. Henderson, C. R.: Estimation of variance and covariance components. *Biometrics* **9**, 226–252 (1953)
77. Henderson, C. R.: *Applications of Linear Models in Animal Breeding*. University of Guelph, Canada (1984)
78. Hochberg, Y., and Tamhane, A.: *Multiple Comparison Procedures*. Wiley, New York (1987)
79. Hoeting, J., Madigan, D., Raftery, A. E., and Volinsky, C. T.: Bayesian model averaging: a tutorial. *Statistical Science* **14**, 381–417 (1998)
80. Holt, D., and Smith, T. F. M.: Post-stratification. *Journal of the Royal Statistical Society* Series A **142**, 33–46 (1979)
81. Holland, P. W.: Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–970 (1986)
82. Horton, N. J., and Kleinman, K. P.: Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* **61** 79–90 (2007)

83. Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685 (1952)
84. Hosmer, D. W., and Lemeshow, S.: *Applied Logistic Regression*. Wiley, New York (1989)
85. Huggins, R. M., and Loesch, D. Z.: On the analysis of mixed longitudinal growth data. *Biometrics* **54**, 583–595 (1998)
86. Jackson, D., Copas, J., and Sutton, A. J.: Modelling reporting bias: the operative mortality rate for ruptured abdominal aortic aneurysm repair. *Journal of the Royal Statistical Society* Series A **168**, 737–752 (2005)
87. Jamshidian, M., and Jennrich, R. I.: Acceleration of the EM algorithm by using quasi-Newton methods. *Journal of the Royal Statistical Society* Series B **59**, 569–587 (1997)
88. Jeffreys, H.: *Theory of Probability*, 3rd ed. Oxford University Press, New York (1961)
89. Jennrich, R. I., and Schluchter, M. D.: Unbalanced repeated measures models with structured covariance matrices. *Biometrics* **42**, 805–820 (1986)
90. Jones, B., and Kenward, M. G.: *Design and Analysis of Cross-over Trials*. Chapman and Hall, London (1989)
91. Kalbfleisch, J. D., and Prentice, R. L.: *The Statistical Analysis of Failure Time Data*. Wiley, New York (1980)
92. Kass, R. E., and Raftery, A. E.: Bayes factors and model uncertainty. *Journal of the American Statistical Association* **90**, 773–795 (1995)
93. Kass, R. E., and Steffey, D.: Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84**, 717–726 (1989)
94. Kennedy, W. J. Jr., and Gentle, J. E.: *Statistical Computing*. Marcel Dekker, New York (1980)
95. Kish, L.: *Survey Sampling*. Wiley, New York (1965)
96. Kish, L.: *Statistical Design for Research*. Wiley, New York (1987)
97. Laird, N. M., and Ware, J. H.: Random-effects models for longitudinal data. *Biometrics* **38**, 963–974 (1982)
98. Lange, K.: *Numerical Analysis for Statisticians*. Springer-Verlag, New York (1999)
99. Lee, Y., and Nelder, J. A.: Hierarchical generalized linear models. *Journal of the Royal Statistical Society* Series B **58**, 619–678 (1996)
100. Lee, Y., and Nelder, J. A.: Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987–1006 (2001)
101. Liang, K.-Y., and Zeger, S. L.: Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986)
102. Lindley, D. V.: *Introduction to Probability and Statistics from a Bayesian Viewpoint*, two volumes. Cambridge University Press, New York (1965)
103. Lindley, D. V.: Decision analysis and bioequivalence trials. *Statistical Science* **13**, 136–141 (1998)
104. Lindsey, J. K.: *Models for Repeated Measurements*. Oxford University Press, Oxford (1993)
105. Lindstrom, M. J., and Bates, D. M.: Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–687 (1988)

106. Linn, R. L., and Hastings, C. N.: A meta-analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement* **21**, 245–259 (1984)
107. Little, R. J. A.: Regression with missing X's: a review. *Journal of the American Statistical Association* **87**, 1227–1237 (1992)
108. Little, R. J. A.: Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134 (1993)
109. Little, R. J. A.: Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121 (1995)
110. Little, R. J. A., and Rubin, D. B.: *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York (2002)
111. Little, R. J. A., and Yau, L.: Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* **52**, 1324–1333 (1996)
112. Longford, N. T.: A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817–827 (1987)
113. Longford, N. T.: *Random Coefficient Models*. Oxford University Press, Oxford (1993)
114. Longford, N. T.: Logistic regression with random coefficients. *Computational Statistics and Data Analysis* **17**, 1–15 (1994)
115. Longford, N. T.: *Models for Uncertainty in Educational Testing*. Springer-Verlag, New York (1995)
116. Longford, N. T.: Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society* Series A **162**, 227–245 (1999)
117. Longford, N. T.: Selection bias and treatment heterogeneity in clinical trials. *Statistics in Medicine* **18**, 1467–1474 (1999)
118. Longford, N. T.: An alternative definition of individual bioequivalence. *Statistica Neerlandica* **54**, 14–36 (2000)
119. Longford, N. T.: Synthetic estimators with moderating influence: carryover in crossover trials revisited. *Statistics in Medicine* **20**, 3189–3203 (2001)
120. Longford, N. T.: An alternative to model selection in ordinary regression. *Statistics and Computing* **13**, 67–80 (2003)
121. Longford, N. T.: Editorial: Model selection and efficiency: is 'Which model . . . ?' the right question? *Journal of the Royal Statistical Society* Series A **168**, 469–472 (2005)
122. Longford, N. T.: *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York (2005)
123. Longford, N. T.: Correspondence: a comment on Jackson, Copas and Sutton (2005). *Journal of the Royal Statistical Society* Series A **169**, 647–648 (2006)
124. Longford, N. T.: On standard errors of model-based small-area estimators. *Survey Methodology* **33**, 69–79 (2007)
125. Longford, N. T., Ely, M., Hardy, R., and Wadsworth, M. E. J.: Handling missing data in diaries of alcohol consumption. *Journal of the Royal Statistical Society* Series A **163**, 381–402 (2000)
126. Longford, N. T., McCarthy, I., and Dowse, G.: Patterns of house price inflation in New Zealand. In: van Montfort, K., Oud, J., and Satorra, A.: *Longitudinal Models in the Behavioral and Related Sciences*, chapter 17, pp. 403–433. L. Erlbaum Associates, Mahwah, NJ (2007)

127. Longford, N. T., and Pittau, M. G.: Stability of household income in European countries in the 1990's. *Computational Statistics and Data Analysis* **51**, 1364–1383 (2006)
128. Longford, N. T., and Rubin, D. B.: Performance assessment and league tables: comparing like with like. Working paper No. 994, Department of Economics and Business Studies, University Pompeu Fabra, Barcelona, Spain (2006)
129. Longford, N. T., Tyrer, P., Nur, U. A. M., and Seivewright, H.: Analysis of a long-term study of neurotic disorder, with insight into the process of non-response. *Journal of the Royal Statistical Society* Series A **169**, 507–523 (2006)
130. Lord, F. M., and Novick, M.: *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA (1968)
131. Louis, T. A.: Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society* Series B **44**, 226–233 (1982)
132. McCullagh, P., and Nelder, J. A.: *Generalized Linear Models*, 2nd ed. Chapman and Hall, London (1989)
133. McCulloch, C. E.: Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170 (1997)
134. McLachlan, G. J., and Krishnan, T.: *The EM Algorithm and Extensions*. Wiley, New York (1997)
135. Meijlijson, I.: A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society* Series B **51**, 127–138 (1989)
136. Meng, X.-L., and van Dyk, D.: The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society* Series B **59**, 511–567 (1997)
137. Nelder, J. A., and Pregibon, D.: An extended quasi-likelihood function. *Biometrika* **74**, 221–232 (1987)
138. Nelder, J. A., and Wedderburn, R. W. M.: Generalized linear models. *Journal of the Royal Statistical Society* Series A **135**, 370–384 (1972)
139. O'Hagan, A: Fractional Bayes factors and model comparison. *Journal of the Royal Statistical Society* Series B **57**, 99–138 (1995)
140. Olkin, I.: Meta-analysis: current issues in research synthesis. *Statistics in Medicine* **15**, 1253–1257 (1996)
141. Orchard, T., and Woodbury, M. A.: A missing information principle: theory and applications. In: *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1., pp. 697–715. University of California, Berkeley (1972)
142. Patterson, H. D., and Thompson, R.: Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971)
143. Pawitan, Y.: *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford (2001)
144. Pearl, J.: *Causality*. Cambridge University Press, New York (2000)
145. Piantadosi, S.: *Clinical Trials: A Methodological Perspective*. Wiley, New York (1997)
146. Pinheiro, J. C., and Bates, D. M.: *Mixed-Effects Models in* `S` *and* `Splus`. Springer-Verlag, New York (2000)
147. Pocock, S. J.: *Clinical Trials: A Practical Approach*. Wiley, Chichester, UK (1983)

148. Potthoff, R. F., Woodbury, M. A., and Manton, K. G.: 'Equivalent sample size' and 'equivalent degrees of freedom' refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association* **87**, 383–396 (1992)
149. Prentice, R. L.: Surrogate end points in clinical trials: definitions and operational criteria. *Statistics in Medicine* **8**, 431–440 (1989)
150. Press, S. J.: *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, 2nd ed. Wiley, New York (2003)
151. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006
152. Ramsay, J. O., and Silverman, B. W.: *Functional Data Analysis*. Springer-Verlag, New York (1997)
153. Rao, C. R.: *Linear Statistical Inference and Its Applications*. Wiley, New York (1973)
154. Rao, J. N. K.: *Small Area Estimation*. Wiley, New York (2003)
155. Raudenbush, S. W., and Bryk, A. S.: Empirical Bayes meta-analysis. *Journal of Educational Statistics* **10**, 75–98 (1985)
156. Rosenbaum, P. R.: *Observational Studies*, 2nd ed. Springer-Verlag, New York (2002)
157. Rubin, D. B.: Estimation of causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701 (1974)
158. Rubin, D. B.: Inference and missing data. *Biometrika* **63**, 581–592 (1976)
159. Rubin, D. B.: Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association* **75**, 373–380 (1980)
160. Rubin, D. B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172 (1984)
161. Rubin, D. B.: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480 (1990)
162. Rubin, D. B.: EM and beyond. *Psychometrika* **56**, 241–254 (1991)
163. Rubin, D. B.: Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **46**, 1213–1234 (1991)
164. Rubin, D. B.: Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489 (1996)
165. Rubin, D. B.: *Multiple Imputation for Nonresponse in Surveys*, 2nd ed. Wiley, New York (2002)
166. Rubin, D. B.: Causal inference using potential outcomes: design, modelling, decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association* **100**, 322–331 (2005)
167. Rubin, D. B., and Schenker, N.: Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* **10**, 585–598 (1991)
168. Särndal, C.-E., Swensson, B., and Wretman, J.: *Model Assisted Survey Sampling*. Springer-Verlag, New York (1992)
169. Schafer, J. L.: *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London (1996)
170. Schafer, J. L.: Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3–15 (1999)
171. Schall, R., and Luus, H. G.: On population and individual bioequivalence. *Statistics in Medicine* **12**, 1109–1124 (1993)

172. Scharfstein, D. O., Rotnitzky, A., and Robins, J. M.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1146 (1999)
173. Schenker, N., Treiman, D. J., and Weidmann, L.: Analyses of public use Decennial Census data with multiply imputed industry and occupation codes. *Applied Statistics* **42**, 545–556 (1990)
174. Schuirmann, D., J.: A comparison of two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680 (1987)
175. Schwartz, G.: Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978)
176. Searle, S. R.: *Linear Models for Unbalanced Data*. Wiley, New York (1987)
177. Searle, S. R., Cassella, G., and McCulloch, C. E.: *Variance Components*. Wiley, New York (1992)
178. Seber, G. A. F.: *Linear Regression Analysis*. Wiley, New York (1977)
179. Senn, S. J.: *Cross-over Trials in Clinical Research*. Wiley, Chichester, UK (1993)
180. Senn, S. J.: *Statistical Issues in Drug Development*. Wiley, Chichester, UK (1997)
181. Sheiner, L. B.: Bioequivalence revisited. *Statistics in Medicine* **11**, 1777–1788 (1992)
182. Sheppard, W. F.: On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proceedings of the London Mathematical Society* **29**, 353–380 (1898)
183. Smith, A. F. M., and Roberts, G. O.: Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* Series B **55**, 3–102 (1993)
184. Spiegelhalter, D. J.: Surgical audit: statistical lessons from Nightingale and Codman. *Journal of the Royal Statistical Society* Series A **162**, 45–58 (1999)
185. Spiegelhalter, D. J., Aylin, P., Best, N. G., Evans, S. J. W., and Murray, G. D.: Commissioned analysis of surgical performance by using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society* Series A **165**, 1–31 (2002)
186. Strenio, J. F., Weisberg, H. I., and Bryk, A. S.: Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics* **39**, 71–86 (1983)
187. Sutton, A. J., Abrams, K. R., Jones, D. R., and Sheldon, T. A.: *Methods for Meta-Analysis in Medical Research*. Wiley, Chichester, UK (2000)
188. Thall, P. F., and Vail, S. C.: Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–671 (1990)
189. Tanner, M. A.: *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd ed. Springer-Verlag, New York (1993)
190. Tanner, M. A., and Wong, W. H.: The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association* **82**, 528–550 (1987)
191. Venables, W. N., and Ripley, B. D.: *Modern Applied Statistics with* `S-plus`, 4th ed. Springer-Verlag, New York, 2002
192. Verbeke, G., and Molenberghs, G.: *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York (2000)

193. Wedderburn, R. W. M.: Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447 (1974)
194. Whitehead, J.: *The Design and Analysis of Sequential Clinical Trials*, 2nd ed. Wiley, Chichester, UK (1997)
195. Williams, D. A.: Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144–148 (1982)
196. Wolter, K. M.: *Introduction to Variance Estimation*. Springer-Verlag, New York (1985)
197. Wu, C. F. J.: On convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103 (1983)
198. Zeger, S. L., Liang, K.-Y., and Albert, P.: Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060 (1988)

Index

χ^2 distibution, 251
t-distribution, 238
t-statistic, 238

Akaike information criterion (AIC), 58
analysis of variance (ANOVA), 1
approximate Bayes bootstrap (ABB), 151
assignment, 202
  at random (AR), 214
  completely at random (CAR), 204
  nonignorable, 214
  not at random (NAR), 214
asymptotic, 38
  efficiency, 41
  normality, 41
  variance, 41
attenuation, 173
auxiliary
  information, 87
  variable, 87

Bayes
  factors, 107
  theorem, 104, 438
Bayesian information criterion (BIC), 58
between-imputation variance, 136, 145
bias, 53, 400
bioequivalence, 247
blinding, 28
block-randomisation, 209
bootstrap, 190
  approximate Bayesian, 223
borrowing strength, 278

caliper matching, 224
carryover effect, 240
cause, 24, 202
chained equations, 148
Cholesky decomposition, 281
cluster, 265, 397
  longitudinal data, 355
  randomisation, 208
clustering, 413
coarse data, 186
coarsening, 186, 410
cohort, 344
comparable institutions, 223
complementary log-log link, 302
complete data, 130
  plausible, 145
complete record, 132
complete-case analysis, 135
complete-data
  design, 139
  statistics, 131
completed data, 145
composition, 49
concentration matrix, 340
confidence
  interval, 449
    one-sided, 450
  level, 449
  region, 450
conjugate distribution, 109, 326
consistency, 40

context, 417
  of a study, 372
  target, 387
context-level variance, 376
convergence
  of distributions, 41
  weak, 41
correlation, 441
covariance, 440
coverage rate, 449
Cramér–Rao inequality, 42
critical
  region, 52
  value, 51
crossover design, 239

data, VII
  completion, 131
  editing, 155
  reduction, 131, 135
data-generating
  distribution, 104
  process, 427
degrees of freedom, 12, 50, 444
density, 422
  smooth, 423
design
  effect, 79
  sampling, 411
    clustered, 413
    multistage clustered, 413
    proper, 412
    semi-systematic, 100
    stratified, 413
    systematic, 70
deviation, 400
difference estimator, 88
direct estimator, 92
discrimination, 229
dispersion, 400
  quantity, 409
distribution, 405, 424
  χ^2, 33, 50
  t, 238
  Bernoulli, 434
  beta, 433
  binomial, 434
  conditional, 437
  data-generating, 104
  empirical, 427
  exponential, 424
  gamma, 306, 425
  geometric, 435
  log-normal, 431
  multinomial, 435
  normal, 425
    multivariate, 446
    standard, 425
  Poisson, 65, 306, 434
  population, 420
  posterior, 103
  prior, 104, 258
  sampling, 420
  uniform, 423, 432
donor, 150
drop-in, 133
drop-out, 133

ecological fallacy, 297
effect, 24, 202
  counterfactual, 28
efficiency, VII, 400
  asymptotic, 41
element of context, 372
EM algorithm, 141, 177, 279, 311
empirical Bayes models, 92
enumeration, 67, 397
envelope of models, 49
error contrasts, 276
estimate, 398
estimation, VII
  dishonest, 449
  honest, 449
estimator, VII, 399
  constituent, 50
  Horvitz–Thompson (HT), 69
  maximum likelihood (ML), 38
  moment-matching, 63
  multiple-imputation (MI), 145
  naive, 403
  ratio, 88
  shrinkage, 42
  single-model-based, 50
  synthetic, 4, 48
    ideal, 5
exclusion restriction, 221
expectation, 403
expected loss, 253

exploiting similarity, 279
exponential family, 304
extrapolation, 25

factor, 396
feature, 21, 123
finite-population correction, 73
Fisher scoring algorithm, 272, 307
frequency, 405
functional data, 356
funnel plot, 381

Gaussian quadrature, 323
generalised linear model
  (GLM), 301
generalised linear models
  mixed (GLMM), 319
Gibbs sampler, 113
grouped data, 194

heteroscedasticity, 19
histogram, 400, 405
homoscedasticity, 8
Horvitz–Thompson estimator, 69
hot deck, 150, 223
hypothesis testing, 54

idempotent matrix, 11
ignorable, 141
impartiality, 168
importance sampling, 111
improper prior, 106
imputation, 131, 135
  deterministic, 136
  hot-deck, 150, 223
  multiple (MI), 144, 177
  nearest-neighbour, 150
  single (SI), 146
  stochastic, 136
incomplete data, 130
  statistic, 131
independence, 399, 436
  mutual, 446
indicator of selection, 67
inference, 398
information matrix
  expected, 39, 143
  observed, 39
informative, 139
intermediate variable, 203
interpolation, 25
interquantile range, 409
invariance, 408
iterative reweighting, 309

Laplace approximation, 324
latent variable, 153, 163, 402
level of nesting, 287
leverage, 18
likelihood, 37
  maximum
    (ML), 37, 270, 303, 307
    restricted (REML), 276
  ratio, 57, 311
    test, 57
Likert scale, 187
linear model, 7
  generalised, 301
linear predictor, 302
link function, 302
  canonical, 304
listwise deletion, 135
location quantity, 408
logit link, 302
longitudinal analysis, 293

manifest variable, 153, 163, 402
Markov property, 340
match, 30
matching, 213, 218
  caliper, 224
maximum likelihood (ML), 37, 270, 303,
    307
mean squared error (MSE), VII, 403
measurement
  error, 165
  impartial, 168, 418
  process, 163
    additive, 418
    biased, 165
    multiplicative, 419
median, 408
meta-analysis, 371
  multivariate, 387
meta-design, 377
metric, 396
MI estimator, 145
minimax estimation, 243

misclassification, 163, 175, 416
missing
  at random (MAR), 138
  completely at random (MCAR), 138
  information, fraction of, 142
  not at random (NMAR), 139
  value, 131
mixture, 152, 153, 310, 356, 440
mode, 410
model, 424, 429
  selection, 51
    multistage, 52
  uncertainty, 3
  valid, 429
moments, method of, 63, 94, 190
monotone response patterns, 134, 148
Monte Carlo Markov chain (MCMC), 113
moving average (MA), 343
multifeature, 21
multilevel model, 266, 287
multiple imputation (MI), 144, 177
multiplicative deviations, 327
multiplicity, 412

naive estimator, 6, 403
Newton method, 255
Newton–Raphson algorithm, 307
nonignorable, 139
nonresponse
  item, 133
  mechanism, 130, 138
  section-level, 133
  unit, 133
nuisance population, 417

ordinary least squares (OLS), 10, 40
origin, 396
outcome, 2, 202
  potential, 202, 235
outlier, 17, 155
overdispersion, 309

panel, 348
  rotating, 348
parameter, 425
partial record, 132
pattern-mixture model, 140
Pearson residuals, 315
percentile, 408
permutation test, 66
personalisation, 291
placebo, 28
plausible
  distribution, 146
  estimate, 145
  parameter value, 137
  value, 137, 146
pointwise unbiased, 165
Poisson distribution, 65, 306, 434
population, 67, 395
  distribution, 420
  nuisance, 417
  quantity, 399
  size, 402
  target, 417
posterior
  distribution, 103
    predictive, 122
poststratification, 81, 160
potential outcome, 202, 235
power
  of a test, 235
  of selection, 53
precision, 374
prediction, 12
prior
  distribution, 104, 258
    predictive, 122
  improper, 106
  noninformative, 105
probability, 405
  joint, 436
  marginal, 436
  proportional to size, 76
probit link, 302
process, data-generating, 427
projection matrix, 11
promotion process, 291
pseudo-observation, 106
publication bias, 380

quantile, 408
quantity
  dispersion, 409
  location, 408
  population, 399
  sample, 399

sampling-process, 404
quartile, 408
quota sampling, 98

raking, 82
random
draw, 426
sample, 426
start, 74
random coefficients, 265
GLMs with (GLMrc), 318
randomisation, 27, 204, 234
range, 409
Rao–Blackwell theorem, 45
ratio estimator, 88
recipient, 150
regression, 444
logistic, 303
ordinary, 8, 39
regression part, 267
regularity conditions, 41
rejection sampling, 112
replication, 399
hypothetical, 400
representativeness, 414
residual, 11
deviance, 316
Pearson, 315
response
distribution, 132
indicator, 131
stability, 132

sample, 67, 398
quantity, 399
sampling
clustered, 77
multistage, 77
design, 411
planned, 80
proper, 69
realised, 80
distribution, 420
frame, 71, 413
mechanism, 68
process, 67
quantity, 404
realised, 415
simple random (SRS), 68, 72
stratified, 76
weights, 69
score, 39
test, 57
selection
indicator of, 67
mechanism, 140
model, 140
semi-systematic sampling design, 100
sensitivity analysis, 149
shrinkage, 42, 51, 82
simulation, 427
simulation–extrapolation (SimEx), 184
size
population, 402
sample, 398
small print, 396
small-area estimation, 92, 292
specificity, 429
standard deviation, 409
standardisation, 430
stem-and-leaf, 22
step length, 74
stochastic process, 399
stratification, 76, 413
subject, 398
submodel, 49
sufficient statistics, 44
linear, 45, 141
minimal, 45
supermodel, 49
superpopulation, 424
support, 396
survey, 411
survival analysis, 194
SUTVA, 207
symmetry plot, 21
systematic review, 372

target, VII, 398
population, 417
time series, 342
autoregressive (AR), 343
moving average (ARMA), 344
moving average (MA), 343
stationary, 343
tolerance interval, 104
treatment, 26, 202
heterogeneity, 29

trimming, 81
two-level data, 289
  balanced, 296
two-level model, 266

utility, 253

validity, 169, 429
variable, 395
variance
  components, 267
  function, 306
  population, 409
  residual, 444
  sampling, 403
variation
  part, 267
  pattern of, 267

wave, 348
white noise, 165