Chapter 2
Background in Probability

We begin our work with a summary of many essential concepts from probability theory. The chapter starts with different viewpoints on the very definition of probability and moves on to a number of foundational concepts in probability theory: expected values and variance, independence, conditional probability, Bayes theorem, random variables and vectors, univariate and multivariate distributions and limit theorems. We think a thorough understanding of these topics is absolutely necessary for the coming chapters and if they seem straightforward then simply skimming through this chapter would be appropriate. However, if some of these topics are not already clear then please keep in mind that this chapter is intended more as a reference than as a thorough explanation of the topics at hand. We refer to Heumann and Schomaker (2016) for a more basic introduction to the field of probability theory and statistics. Let us now begin with the deceptively difficult task of defining the meaning of probability itself.

2.1 Random Variables and Probability Models

2.1.1 Definitions of Probability

The foundation of statistical modelling is stochastic models. Data are assumed to be generated by a random process, which can be described by probabilities. In order to define probabilities, the notion of a random experiment is essential. A random experiment is a process for which the outcome is uncertain. The set of all possible outcomes is called sample space and denoted by Ω. For instance, when throwing a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$ and for the length of a person's life, we have $\Omega = [0, \infty)$. Events can be defined as subsets of the sample space Ω. Examples are $A = \{2, 4, 6\}$ ("even number"), $B = \{6\}$ ("six") or $C = (60, \infty)$ ("lifetime longer than 60"). Mathematically, probabilities are functions, which assign a number $P(A)$ between

0 and 1 to events A, where $A \subset \Omega$. If $P(A)$ is small, then the event A occurs seldomly and if $P(B)$ is close to 1, then B occurs rather often. Andrey Kolmogorov (1933) developed a system of axioms for probabilities, which still represent the basis of modern probability theory. We define these axioms as follows:

Definition 2.1 (Mathematical Probability Definition) The **probability** P is a function defined on the collection $\mathcal{A}$ of all subsets of a sampling space Ω, which fulfils the following properties.

$$(i)\ \ 0 \leq P(A) \leq 1 \text{ for all } A \in \mathcal{A} \tag{2.1.1}$$

$$(ii)\ \ P(\Omega) = 1 \tag{2.1.2}$$

$$(iii)\ \ P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{2.1.3}$$

for all mutually disjoint sequences of events, i.e. $A_i \in \mathcal{A}, i = 1, \ldots, \infty$ with $A_i \cap A_j = \emptyset$ for $i \neq j$.

This definition is purely formal and does not give any actual meaning to the term probability. However, it has proven to be both mathematically sound and practically useful. Despite the clear mathematical foundation, a long philosophical debate continues over the true definition of probability itself. We briefly discuss the two most common definitions, as we find them illustrative of the various strategies of statistical inference described in later chapters. The first interpretation of probabilities is based on infinite replications of the random experiment and was introduced by Richard Von Mises (1928).

Definition 2.2 (Frequentist Probability Definition) The **probability** of an event is defined as the limit of the relative frequency of its occurrence in a series of replications of the respective random experiment, that is

$$P(A) := \lim_{n \to \infty} \frac{n_A(n)}{n} \tag{2.1.4}$$

with n as the number of replications and $n_A(n)$ being the number of occurrences of A in n experiments.

The above definition has some shortcomings including, of course, that it is impossible to actually perform an infinite number of replications. Another important definition of probability was proposed by De Finetti (1974) and is based on the idea that uncertainty can also be seen as a lack of information.

Definition 2.3 (Subjective Probability Definition) The **probability** of an event A is defined as the degree of belief of an individual at a certain time with a certain amount of information. It is quantified by imaginary bets. That is, one's degree of

belief in A is P(A), if and only if P(A) units of utility is the price at which one would buy or sell a bet that pays 1 unit of utility if A occurs and 0 units if A does not occur.

Consider the example of a (fair) die roll. Before we throw a die, the number of dots it will show is unknown and once we have thrown the die and it lies on the table, we then know the number of dots showing. However, what happens if we throw the die by putting it in a cup? Shaking the cup and putting it upside down on the table fixes the number of dots showing on the die. Hence, the random experiment is realised. However, before we lift the cup, we still do not know what the die shows. With a closed cup on the table the random experiment is over, but the uncertainty about the number of dots still exists until the cup is lifted. Using De Finetti's approach we can still apply a probability model, i.e. the probability that the die under the cup shows six dots on the top is 1/6. This shows that uncertainty can be well expressed in terms of probability statements. Both definitions are essential in interpreting results using statistical models. An overview on defining probability with a more detailed discussion of the philosophical background can be found in Gillies (2000). No matter how probability is defined, Kolmogorov's axioms will always hold. Therefore, we proceed with some results in probability theory that result from these axioms.

2.1.2 *Independence, Conditional Probability, Bayes Theorem*

Definition 2.4 (Conditional Probability and Independence) Let A and B be two events. Assuming that $P(A) > 0$, the conditional probability of B given A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{2.1.5}$$

Two events A and B are called independent if and only if

$$P(A \cap B) = P(A) \cdot P(B). \tag{2.1.6}$$

The interpretation is that $P(B|A)$ is the probability of B if A is known to have occurred. In the case of rolling a fair die, let us take $B = \{6\}$ to be the event of rolling a 6, which has probability $1/6$. When we know that an even number occurred, i.e. $A = \{2, 4, 6\}$, then the probability of B is $1/3$, i.e. $P(B|A) = 1/3$. Note also that independence of A and B implies that $P(B|A) = P(B)$, i.e. knowing that A has occurred does not change the probability of B.

Property 2.1 (Law of Total Probability and Bayes Theorem) Let $\Omega = \bigcup_{i=1}^{\infty} A_i$ where $A_1, A_2, A_3, \ldots$ are countable partitions of Ω (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$.). Then, for each event B it holds that

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i) \tag{2.1.7}$$

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}. \tag{2.1.8}$$

Both equations can be directly deduced from Kolmogorov's axioms and Definition (2.1.5). The first equation describes the law of total probability and simply states that the probability of an event can be calculated as the sum of the probabilities of that event conditional on each event of a set of events, as long as that set encompasses all possible events of that partition. The second equation is also known as Bayes Theorem and is attributed to the English reverend and philosopher Thomas Bayes (1702–1761). The main idea of this theorem is to compute conditional probabilities "the other way round". The conditional probabilities $P(A_j|B)$ are calculated from conditional probabilities $P(B|A_j)$ and the (prior) probabilities $P(A_j)$. If the sets A_j characterise some unobservable events or some theories concerning the event B, then Bayes theorem gives us an understanding of what we learn about A_j when we observe B. This is the basis of what is called Bayesian inference, which is discussed later in Chap. 5.

2.1.3 Random Variables

Even though probabilities are formally defined on sets, it is often more convenient to make use of random variables which take real numbers. Formally, a random variable is defined as a mapping from the sample space Ω into the real numbers, that is

$$Y : \Omega \to \mathbb{R}.$$

The probability is now defined with relation to the values of Y instead of on subsets of Ω. As a simple example, consider rolling two dice and define the random variable Y as the sum of the two dice. Then Ω is $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ and $Y_{2D}((a, b)) = a + b$. In general, random variables are used to describe random phenomena. The subset of possible values of $\mathbb{R}$ is called the support of Y. If the support is finite or countably infinite, the random variable is called discrete. For instance, the above random variable Y_{2D} has the finite support $\{2, \ldots, 12\}$ and can be characterised by the probability function $P(Y_{2D} = y)$. Random variables that have a support consisting of all real numbers (or an interval of them) are called continuous random variables, e.g. the height of a randomly chosen child or the lifetime of an electronic device. The stochastic behaviour of a continuous random

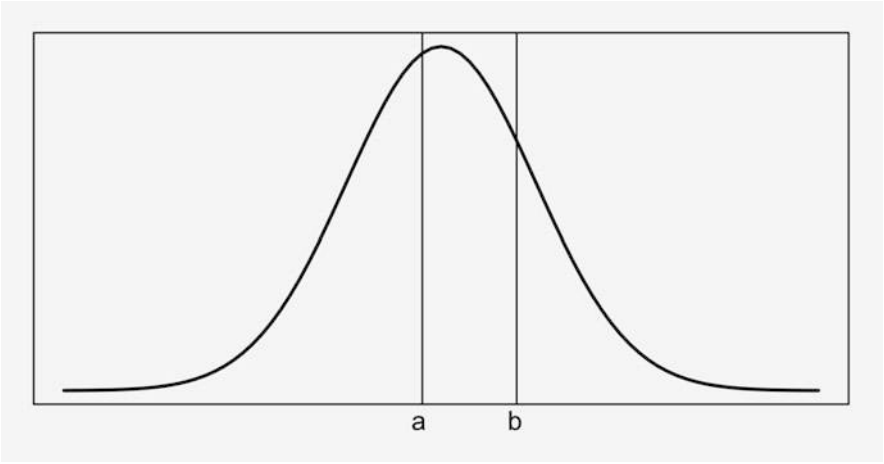

Fig. 2.1 Example of a density function

variable is described by its density function. The **density function** $f(.)$ is a positive function, such that

$$P(Y \in [a, b]) = \int_a^b f(y)dy,$$

as sketched in Fig. 2.1. From the second of Kolmogorov's axioms (2.1.2) we get $\int_{-\infty}^{\infty} f(t)dt = 1$.

For both discrete-valued and continuous random variables, the **distribution function** $F_Y(y) = P(Y \leq y)$ yields a unique description of the random variable Y, where y may take any possible real number, i.e. $y \in \mathbb{R}$. This is sometimes also called the **cumulative distribution function**. In the definition of $F_Y(y)$, we have introduced several notational concepts that are commonly used in statistics but are worth mentioning here explicitly. Firstly, random variables are commonly denoted with capital letters, while their realisations are denoted with lower case letters. Hence, whenever we write Y, it means that we do not know the variable's value, while $Y = y$ means that the random variable Y has taken the specific value y. In fact, this distinction is central to much of statistics, where models are generated with random variables, which are then put into practice by realising their values with real data. Secondly, when it is unclear, we put an index on the distribution or density function, meaning that this function (i.e. the probability model) belongs to the given random variable (e.g. F_Y). We will not be stringent with this index notation, because it is often obvious which random variable we are referring to. Note also that for

$y \to -\infty$ we have $F_Y(y) = 0$ and for $y \to \infty$ we get $F_Y(y) = 1$. Finally, we can relate $F_Y(.)$ to the density or probability function $f_Y(y)$ as follows:

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt. \tag{2.1.9}$$

It can be can notationally quite tedious, and not particularly informative, to thoroughly distinguish between discrete and continuous random variables. We therefore simplify the notation with the following convention, which will be used throughout the entire book. This allows us to have a general notation for both continuous and discrete random variables.

1. If Y is a discrete value random variable (e.g. with values $\{0, 1, 2, \ldots\}$), then the integral in Definition (2.1.9) refers to a sum, i.e.

$$\int_{-\infty}^{y} f_Y(\tilde{y})d\tilde{y} = \sum_{k:k\leq y} P(Y = k).$$

   Function $f_Y(y)$ in this case is equivalent to the probability function $P(Y = y)$.
2. If Y is continuous, then the integral in (2.1.9) is an integral in the usual sense (assuming integrability), such that

$$f_Y(y) = F_Y'(y)$$

   (at values y where $F_Y(y)$ is differentiable).
3. For mixed random variables that take discrete and continuous values, the integral in (2.1.9) is a mixture of sums and integrals.

The mathematically trained reader might not be satisfied with our "sloppyness" here. However, we do not want to get distracted from our main topic, which is statistics. For this reason, we content ourselves with being a little superficial with the notation. **Therefore, from now on, we denote $f_Y(y)$ as a density, even if Y is discrete valued**. This convention makes the notation much easier and we are convinced that the reader will quite easily understand when an integral refers to a sum or a real integral or a mixture of both. For a rigorous mathematical treatment of random variables, there are many textbooks on probability, such as Karr (1993).

Each random variable has some essential features, which we describe in the following definitions. The most prominent are the mean value, which we call the expectation, and the variability, which we call the variance.

Definition 2.5 (Expected Value, Variance and Moments)

1. The expected value or mean value of a random variable is defined as

$$E(Y) = \int y f_Y(y)dy.$$

2. The variance of a random variable is defined as

$$Var(Y) = E\Big(\{Y - E(Y)\}^2\Big) = \int \{y - E(Y)\}^2 f(y)dy,$$

which always results in $Var(Y) \geq 0$.
3. The k-th Moment is defined as $E(Y^k) = \int y^k f_Y(y)dy$ and the k-th central moment as $E(\{Y - E(Y)\}^k)$.

We often abbreviate the expectation with the parameter $\mu = E(Y)$ and the variance with the parameter $\sigma^2 = E(Y - E(Y)^2)$. Note also that the variance can be written in the following form:

$$\begin{aligned}\sigma^2 &= E\big(\{Y - \mu\}^2\big) = E\big(Y^2 - 2Y\mu + \mu^2\big) \\ &= E(Y^2) - 2\mu^2 + \mu^2 = E(Y^2) - \mu^2. \\ &= E(Y^2) - \{E(Y)\}^2.\end{aligned}$$

This follows from some properties of the expectation operator that we will define in the next section. We thereby assume that the corresponding integrals exist, i.e. are finite. We must also mention that for some exceptional distributions (e.g. the Cauchy-distribution) the expected value or the variance (or both) do not exist.

Property 2.2 (Expectation and Variance of a Sum of Random Variables) An important result is that the sum of the expectations of a set of random variables is equal to the expectation of their sum. Also important is that the variance of a sum of independent random variables is equal to the sum of their variances. For random variables $Y_1, \ldots, Y_n$ we get

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i), \tag{2.1.10}$$

and if $Y_1, \ldots, Y_n$ are independent, then

$$Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 Var(Y_i), \tag{2.1.11}$$

for arbitrary known values $a_1, \ldots, a_n$.

2.1.4 Common Distributions

There are some distributions that appear so often in statistics that familiarity with their parameters and features would greatly benefit the reader. The distributions

themselves depend on one or more parameters, which are conventionally denoted with Greek letters. We also follow the notational convention that the dependence of the distribution on the parameter is expressed with a semicolon, that is, the parameters of the distribution are listed after the semicolon. This convention is used throughout the rest of the book and is easily comprehended with the following standard distributions.

Definition 2.6 (Binomial Distribution $B(n, p)$)

- Application: Two outcomes: success | failure with a chance of success of $\pi \in [0, 1]$. We count the number of successes in n independent trials.
- Probability function: $P(Y = k; \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$, for $k = 0, \ldots, n$
- Expectation: $E(Y) = n\pi$
- Variance: $Var(Y) = n\pi(1 - \pi)$

Definition 2.7 (Poisson Distribution $Po(\lambda)$)

- Application: Count of events in a certain period of time with events occurring at an average rate of $\lambda > 0$.
- Probability function: $P(Y = k; \lambda) = \frac{\lambda^k}{k!} \exp(-\lambda)$ for $k = 0, 1, 2, \ldots$
- Expectation: $E(Y) = \lambda$
- Variance: $Var(Y) = \lambda$

Definition 2.8 (Normal Distribution $N(\mu, \sigma)$)

- Application: Metric symmetrically distributed random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.
- Density function: $f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$ for $y \in \mathbb{R}$
- Expectation: $E(Y) = \mu$
- Variance: $Var(Y) = \sigma^2$

Definition 2.9 (Exponential Distribution $Exp(\lambda)$)

- Application: Time between two events following a Poisson process, where the waiting time for the next event is on average $1/\lambda$, $\lambda > 0$.
- Density function: $f(y; \lambda) = \lambda \exp(-\lambda y)$ for $y \geq 0$
- Expectation: $E(Y) = \frac{1}{\lambda}$
- Variance: $Var(Y) = \frac{1}{\lambda^2}$

Definition 2.10 (t-Distribution $t(n)$)

- Application: Metric symmetrically distributed random variable with a considerable probability of extreme outcomes ("heavy tails"). Also used for statistical tests for the mean of normally distributed variables when the variance is unknown and estimated from the data and the degrees of freedom, n, is small.

- Density function: $f(y;d) = \frac{\Gamma(\frac{d+1}{2})}{\sqrt{d\pi}\Gamma(\frac{d}{2})}\left(1+\frac{y^2}{d}\right)^{-\frac{d+1}{2}}$
  where $\Gamma(.)$ is the gamma function and $d = 1, 2, \ldots$ is the degree of freedom (which in principle can take any positive value but we only work with positive discrete values here)
- Expectation: $E(Y) = 0$ for $d \geq 1$
- Variance: $Var(Y) = \frac{d}{d-2}$ for $d \geq 3$

Definition 2.11 (Chi-Squared Distribution $\mathcal{X}^2(n)$)

- Application: Squared quantities like sample variances, sum of n independent squared normal distributed random variables.
- Density function: $f(y;n) = \frac{y^{\frac{n}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$ where $n = 1, 2, 3, \ldots$
- Expectation: $E(Y) = n$
- Variance: $Var(Y) = 2n$

2.1.5 Exponential Family Distributions

Several of the above distributions allow for a mathematical generalisation, that in turn allows for the development of theories and models that apply to the entire model class. One such class of distributions that is central to statistics is the exponential family of distributions. This class consists of many of the most common distributions in statistics, for example, the normal, the Poisson, the binomial distribution and many more. Although these distributions are very different, as can be seen above, how they are constructed is the same.

Definition 2.12 A class of distributions for a random variable Y is called an **exponential family**, if the density (or probability function) can be written in the form

$$f_Y(y;\theta) = \exp\{t^T(y)\theta - \kappa(\theta)\}h(y), \tag{2.1.12}$$

where $h(y) \geq 0$ and $t(y) = (t_1(y), \ldots, t_p(y))^T$ is a vector of known functions and $\theta = (\theta_1, \ldots, \theta_p)^T$ is a parameter vector.

The density (or probability function) given in (2.1.12) is very general, and thus looks a little complicated, but we will see in the subsequent examples that the different terms simplify quite substantially in most real examples. Let us therefore consider the quantities in (2.1.12) in a little more depth. First of all, the function $t(y)$ is a function of the random variable and will later be called a statistic. The term θ is the parameter of the distribution, also called **natural parameter**, and quantity $\kappa(\theta)$ simply serves as normalisation constant, such that the density integrates out to

1. Hence, $\kappa(\theta)$ is defined by the following:

$$1 = \int \exp\{t^T(y)\theta\}h(y)dy \exp(-\kappa(\theta))$$
$$\Leftrightarrow \quad \kappa(\theta) = \log \int \exp\{t^T(y)\theta\}h(y)dy.$$

As a consequence, by differentiating $\kappa(\theta)$, we get

$$\frac{\partial \kappa(\theta)}{\partial \theta} = \frac{\int t^T(y)\exp\{t^T(y)\theta\}h(y)dy}{\int \exp\{t^T(y)\theta\}h(y)dy} = \int t^T(y) f_Y(y;\theta)dy = E\Big(t^T(Y)\Big).$$

Finally, the quantity $h(y)$ depends on the random variable and not on the parameter and we will see later that this quantity will not be of particular interest.

We will now demonstrate how some of the above introduced distributions can be written in the style of an exponential family. We thereby start with the normal distribution which can be rewritten as

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2}\frac{y^2 - 2y\mu + \mu^2}{\sigma^2} - \frac{1}{2}\log(\sigma^2)\right)\frac{1}{\sqrt{2\pi}} \\ &= \exp\left(\underbrace{(-\frac{y^2}{2}, y)}_{t^T(y)}\underbrace{\begin{pmatrix}\frac{1}{\sigma^2}\\ \frac{\mu}{\sigma^2}\end{pmatrix}}_{\theta} - \underbrace{\frac{1}{2}(-\log\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^2})}_{\kappa(\theta)}\right)\underbrace{\frac{1}{\sqrt{2\pi}}}_{h(y)}, \end{aligned}$$

where $\theta_1 = \frac{1}{\sigma^2}$ and $\theta_2 = \frac{\mu}{\sigma^2}$ such that

$$\kappa(\theta) = \frac{1}{2}\left(-\log(\theta_1) + \frac{\theta_2^2}{\theta_1}\right) = -\frac{1}{2}\left(\log(\theta_1) - \frac{\theta_2^2}{\theta_1}\right).$$

Note that:

$$\begin{aligned} \frac{\partial \kappa(\theta)}{\partial \theta_1} &= -\frac{1}{2}\left(\frac{1}{\theta_1} + \frac{\theta_2^2}{\theta_1^2}\right) = -\frac{1}{2}(\sigma^2 + \mu^2) = E\left(-\frac{Y^2}{2}\right) = E\Big(t_1(Y)\Big) \\ \frac{\partial \kappa(\theta)}{\partial \theta_2} &= \frac{\theta_2}{\theta_1} = \mu = E(Y) = E\Big(t_2(Y)\Big). \end{aligned}$$

Hence, we see that the normal distribution is an exponential family distribution. The same holds for the Binomial distribution, because

$$\begin{aligned} f(y) = P(Y = y) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= \left(\frac{\pi}{1-\pi}\right)^y (1-\pi)^n \binom{n}{y} \\ &= \exp\left(\underbrace{y}_{t(y)} \underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\theta} - \underbrace{n\log\left(\frac{1}{1-\pi}\right)}_{\kappa(\theta)}\right) \underbrace{\binom{n}{y}}_{h(y)}, \end{aligned}$$

where $\theta = \log(\frac{\pi}{(1-\pi)})$ is also known as log odds ratio and $\kappa(\theta)$ results in

$$\kappa(\theta) = n\log(1+\exp(\theta)).$$

As $\pi = \exp(\theta)/(1+\exp(\theta))$ one gets

$$\frac{\partial\kappa(\theta)}{\partial\theta} = n\frac{\exp(\theta)}{1+\exp(\theta)} = n\pi = E(Y) = E\Big(t(Y)\Big).$$

Among many others the Poisson distribution is also part of the exponential family, the proof of which we leave as an exercise for the reader.

2.1.6 Random Vectors and Multivariate Distributions

We must also introduce the concept of multivariate random variables. A sequence of random variables $Y_1, Y_2, \ldots, Y_q$ is often represented as a random vector $(Y_1, Y_2, \ldots, Y_q)$. To allow for dependence between the elements of the random vector, we need to extend our previously introduced concept of the probability distribution towards a multivariate distribution function. The corresponding (multivariate) cumulative distribution function is given by

$$F(y_1, \ldots, y_q) = P(Y_1 \leq y_1, \ldots, Y_q \leq y_q).$$

If $F(.)$ is continuous and differentiable, we can relate the cumulative distribution function to the density function $f(y_1, \ldots y_q)$ with

$$P(a_1 \leq Y_1 \leq b_1, \ldots a_q \leq Y_q \leq b_q) = \int_{a_1}^{b_1} \cdots \int_{a_q}^{b_q} f(y_1, \ldots, y_q) dy_1 \ldots dy_q. \tag{2.1.13}$$

Integrating out all variables except for one gives the **marginal distribution**. We can also extend the concept of conditional probability to multivariate random variables.

Definition 2.13 (Marginal and Conditional Distribution) The marginal density with respect to Y_1 is given by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1, \ldots, y_q) dy_2 \ldots dy_q. \tag{2.1.14}$$

The conditional density is analogue to the conditional probability. We give the definition for the two dimensional case as

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f(y_1, y_2)}{f(y_2)} \text{ for } f(y_2) > 0. \tag{2.1.15}$$

From (2.1.15) we can calculate conditional expectation and variance, e.g.

$$\begin{aligned} E(Y_1|Y_2 = y_2) &= \int y_1 f_{Y_1|Y_2}(y_1|y_2) dy_1 \\ Var(Y_1|Y_2 = y_2) &= \int \{y_1 - E(Y_1|Y_2 = y_2)\}^2 f_{Y_1|Y_2}(y_1|y_2) dy_1 \\ &= E(Y_1^2|Y_2 = y_2) - \{E(Y_1|Y_2 = y_2)\}^2. \end{aligned}$$

Of particular interest is the dependence structure among the components of the random vector. This can be expressed in the covariance matrix, which is defined as follows:

$$Cov(Y) = \begin{pmatrix} Cov(Y_1, Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_q) \\ Cov(Y_2, Y_1) & \ddots & & \vdots \\ \vdots & & \ddots & \\ Cov(Y_q, Y_1) & \cdots & & Cov(Y_q, Y_q) \end{pmatrix},$$

where

$$Cov(Y_j, Y_k) = E\left[\left\{Y_j - E(Y_j)\right\}\left\{Y_k - E(Y_k)\right\}\right] = E(Y_j Y_k) - E(Y_j)E(Y_k)$$

for $1 \leq j, k \leq q$. It is clear that $Cov(Y_j, Y_j) = Var(Y_j)$ and that if Y_j and Y_k are independent, then $Cov(Y_j, Y_k) = 0$. This follows easily, as under independence we have

$$f(y_j, y_k) = f_{Y_j}(y_j) f_{Y_k}(y_k),$$

such that $E(Y_jY_k) = E(Y_j)E(Y_k)$. As well as the covariance, we will occasionally focus on the correlation, which is defined as follows:

$$Cor(Y_j, Y_k) = \frac{Cov(Y_j, Y_k)}{\sqrt{Var(Y_j)Var(Y_k)}}.$$

In Chap. 10 we will discuss more details and advanced statistical models for multivariate random vectors. For now, we will simply present most commonly encountered multivariate distribution, namely the multivariate normal distribution. It is defined as follows:

Definition 2.14 The random vector $(Y_1, \ldots, Y_q)^T$ follows a **multivariate normal distribution** if the density takes the form

$$\begin{aligned} f(y_1, \ldots, y_q) &= \frac{1}{(2\pi)^{q/2}} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} \sum_{j=1}^{q} \sum_{k=1}^{q} (y_j - \mu_j)(y_k - \mu_k)\Omega_{jk} \right) \\ &= \frac{1}{(2\pi)^{q/2}} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} (y - \mu)^T \Omega (y - \mu) \right), \end{aligned}$$

where $\Omega = \Sigma^{-1}$ and $y = (y_1, \ldots, y_q)^T$. We denote this with $Y \sim N(\mu, \Sigma)$. The mean vector $\mu = (\mu_1, \ldots, \mu_q)^T$ is the vector of expected values, i.e. $\mu_j = E(Y_j)$ and the covariance matrix Σ has entries

$$\Sigma_{jk} = Cov(Y_j, Y_k).$$

A conditional normal distribution is derived as follows. Let vector $(Y_1, \ldots, Y_q)^T$ be divided into two subcomponents, labelled (Y_A^T, Y_B^T). Accordingly, let Σ be divided into the four submatrices

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix},$$

where $A, B \subset \{1, \ldots, q\}$ with $A \cap B = \emptyset$ and $A \cup B = \{1, \ldots, q\}$. Then

$$Y_A | y_B \sim N\left(\mu_A - \Sigma_{AB}\Sigma_{BB}^{-1}(y_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \right),$$

where μ_A and μ_B are the corresponding subvectors of μ. Hence, the conditional distribution of a multivariate normal distribution is again normal. This also holds for the marginal distribution, i.e.

$$Y_A \sim N(\mu_A, \Sigma_{AA}).$$

While the normal distribution is central for modelling multivariate continuous variables, the multinomial distribution is central for modelling discrete-valued random variables.

Definition 2.15 The random vector $Y = (Y_1, \ldots, Y_K)^T$ follows a multinomial distribution if

- $Y_k \in \mathbb{N}$ for all $k = 1, \ldots, K$
- $\sum_{k=1}^{K} Y_k = n$
- $P(Y_1 = y_1, \ldots, Y_K = y_K) = \frac{n!}{y_1! \cdots y_K!} \prod_{k=1}^{K} \pi_k^{y_k}$ where $\pi_k \in [0, 1]$ and $\sum_{k=1}^{K} \pi_k = 1$

We notate this as $Y \sim Multi(n, \boldsymbol{\pi})$ with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$. The random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_K)^T$ has the expectation

$$E(\boldsymbol{Y}) = \boldsymbol{\pi}$$

and the covariance matrix is given by

$$\begin{aligned} Var(\boldsymbol{Y}) &= diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T \\ &= \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \ldots & -\pi_1\pi_K \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \ldots & \\ \vdots & & & \vdots \\ -\pi_K\pi_1 & & \ldots & \pi_K(1-\pi_K) \end{pmatrix}. \end{aligned}$$

Multivariate distributions can also be used to derive moments for subsets of the random vector. We exemplify this for a bivariate setting, where (Y, X) are two random variables. The mean and variance of Y can be derived from the conditional mean and variance of Y given $X = x$ as follows:

Property 2.3 (Iterated Expectation) For two random variables X and Y we find

$$\begin{aligned} E_Y(Y) &= E_X(E_Y(Y|X)) \\ Var_Y(Y) &= E_X(Var_Y(Y|X)) + Var_X(E_Y(Y|X)), \end{aligned}$$

where E_X and Var_X indicate the expectation and variance with respect to random variable X, while E_Y and Var_Y denote the corresponding conditional quantities for random variable Y.

We will need this statement later in the book and fortunately, the proof is rather simple as can be seen below.

Proof Note that

$$E_Y(Y) = \int y f(y) dy = \int \underbrace{\int y f(y|x) dy}\, f_X(x) dx$$

$$= \int E_Y(Y|x) f_X(x) dx = E_X\Big(E_Y(Y|X)\Big).$$

Moreover with $\mu = E_Y(Y)$ and $\mu(x) = E_Y(Y|x)$ we get

$$\begin{aligned} Var_Y(Y) &= E_Y(Y^2) - \{E_Y(Y)\}^2 \\ &= E_X[\underbrace{E_Y(Y^2|X)}_{=}] - \{E_X[E_Y(Y|X)]\}^2 \\ &= E_X[\overbrace{Var_Y(Y|X) + \{E(Y|X)\}^2}] - \{E_X[E_Y(Y|X)]\}^2 \\ &= E_X[Var_Y(Y|X)] + \underbrace{E_X[E_Y(Y|X)^2] - \{E_X[E_Y(Y|X)]\}^2}_{=} \\ &= E_X[Var_Y(Y|X)] + \overbrace{Var_X[E_Y(Y|X)]}. \end{aligned}$$

□

2.2 Limit Theorems

Numerous concepts in statistical reasoning rely on asymptotic properties. In its most simple case this is, for instance, the behaviour of the arithmetic mean of n independent observations. For a sample of n individuals, we assume that the observations from one individual do not depend on the observations from another individual, which we call independence. This concept of independence has important properties and justifies nearly all asymptotic arguments in statistical reasoning. "Asymptotic" thereby means that one explores how a quantity derived from a sample behaves if the sample size increases, i.e. that is the number of (independent) observations n tends to infinity. Asymptotic arguments will help to quantify uncertainty and we will see that asymptotic normality plays a fundamental role in statistical reasoning. Before motivating this in more depth, let us visualise this with a simple example. A random walk is recursively defined as $Y_t = Y_{t-1} + X_t$ for $t = 1, 2, \ldots$, where we assume

$$X_t = \begin{cases} 1 & \text{with probability } \pi \\ -1 & \text{with probability } 1 - \pi. \end{cases}$$

It is natural to start with $Y_0 = 0$, so that we can rewrite $Y_n = \sum_{i=1}^{n} X_i$. We assume that the X_i are independent and set $\pi = 1/2$, which gives $E(X_i) = \pi - (1 - \pi) = 0$ and $Var(X_i) = E(X_i^2) = 1$. We run 100.000 simulations of this random walk and stop at $n = 10.000$. In Fig. 2.2 (top plot) we plot the resulting values $y_1, y_2, \ldots, y_{10000}$ for 3 of these simulations. The shaded region of Fig. 2.2 shows the range in which 90% of the observed values lie. The perceptive reader might note that the boundary of the shaded area appears similar to $\sqrt{n}$. We therefore conclude that, even though Y_n is a sum of n random variables, its spread does not grow with order n but "only" with order $\sqrt{n}$. This is a fundamental property, which in fact justifies most asymptotic reasoning in statistics. It implies that if we divide Y_n by $\sqrt{n}$, we get the behaviour shown on the middle plot of Fig. 2.2. Because $Var(Y_n) = n$, the variance of $\frac{Y_n}{\sqrt{n}}$ is 1. The reader can see that $Z_n := Y_n/\sqrt{n}$ shows a stabilised behaviour and 90% of the simulations remain in the shaded region. We say that Y_n has the asymptotic order of $\sqrt{n}$, meaning that Y_n is (stochastically) proportional to $\sqrt{n}$. Without giving an exact definition here it means that the variance of $Y_n/\sqrt{n}$ does not depend on n (if n is large). In the bottom panel of Fig. 2.2, we show the arithmetic mean $Y_n/n = \sum_{i=1}^{n} X_i/n$ and observe that the mean of the X-values gets closer to 0 with increasing n. We say that $\frac{1}{n}\sum_{i=1}^{n} x_i$ converges in probability. This property is known as the **law of large numbers**. Its central statement is that the arithmetic mean of independent variables which have all the same distribution converges to its expected value. An exact definition is given later.

Now let us explore whether we can say something about the distribution of $Z_n = Y_n/\sqrt{n}$. In Fig. 2.3, we show the empirical distribution of the 100.000 simulations for $n = 10$ and $n = 100$ as a histogram. We overlay the plot with a normal distribution. We see that for $n = 100$ a striking concordance with a normal distribution is already apparent. Hence, if we divide the sum of n independent random variables by $\sqrt{n}$ we obtain a stable behaviour, where the distribution of $Z_n = Y_n/\sqrt{n}$ approximately follows a normal distribution. This property is central to statistics and is expressed in the central limit theorem. All in all, there are three theorems we want to highlight. All rely on the assumption of independent and identically distributed random variables, which we explicitly define in the next section. Formally, it means that in a sequence of random variables $X_1, X_2, \ldots, X_n$

1. X_i and X_j are mutually independent,
2. X_i and X_j have the same distribution

for all $i \neq j$.

Property 2.4 (Central Limit Theorem) Let $Y_1, Y_2, Y_3, \ldots$ be independent and identically distributed random variables, with mean zero and variance σ^2. The distribution of $Z_n = \sum_{i=1}^{n} Y_i/\sqrt{n}$ converges to a normal distribution with mean zero and variance σ^2, which we denote as

$$Z_n \xrightarrow{d} N\left(0, \sigma^2\right).$$

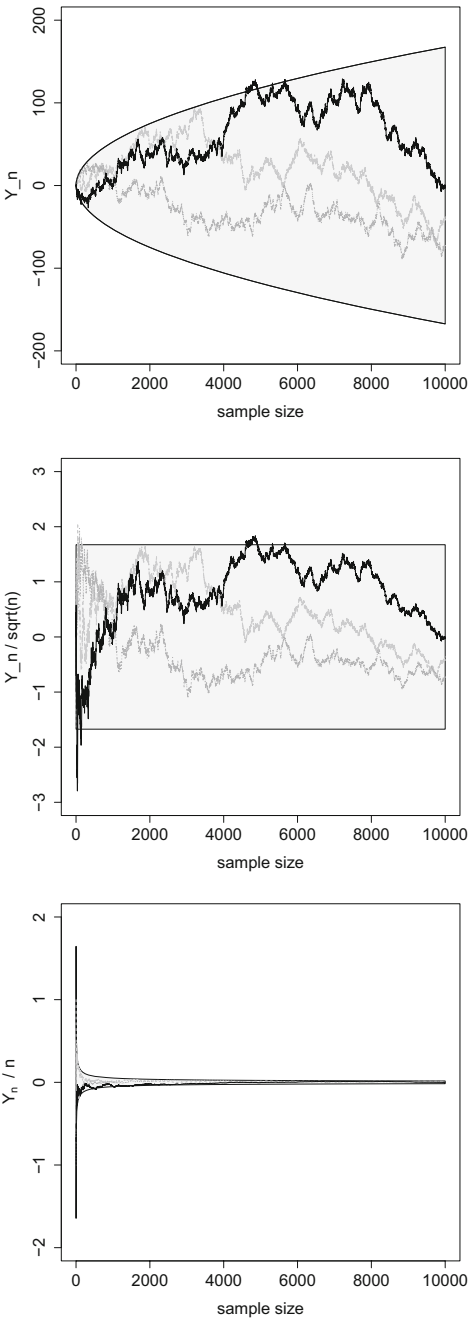

Fig. 2.2 Top plot: Random Walks with $n = 10.000$ steps. Middle plot: Random Walks divided by $\sqrt{n}$. Bottom plot: divided by n

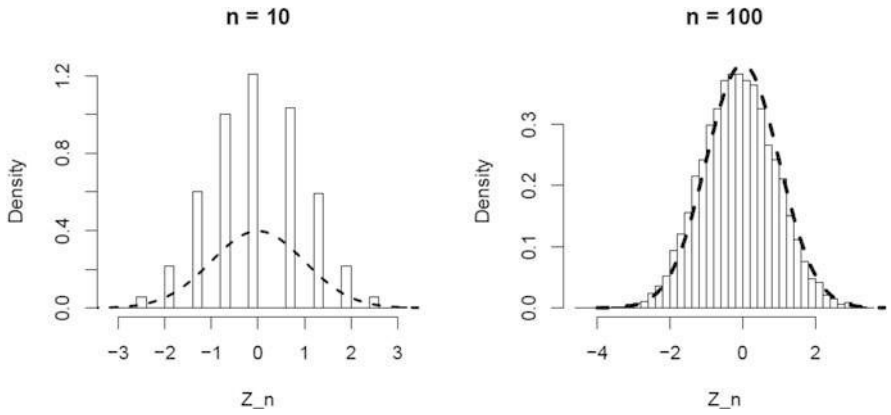

Fig. 2.3 Distribution of $Y_n/\sqrt{n}$ at $n = 10$ and $n = 100$ given 100.000 simulations

The word "central" has a double meaning here, as the central limit theorem truly is a core component of statistics. Note that the only assumption we make is that random variables $Y_1, \ldots, Y_n$ have a finite variance, but they can otherwise follow arbitrary distributions. Because the central limit theorem is so important, it seems worthwhile to sketch a proof.

Proof We show that the **moment generating function** of Z_n converges to the moment generating function of a normal distribution. The moment generating function of a random variable Y thereby is defined as

$$M_Y(t) = E\left(e^{tY}\right).$$

Note that

$$\left.\frac{\partial^k M_Y(t)}{(\partial t)^k}\right|_{t=0} = E(Y^k),$$

which justifies the name of the function as its k-th derivative generates the k-th moment of random variable Y, which, as a reminder, is $E(Y^k)$. A companion of the moment generating function is the cumulant generating function, defined as the logarithm of the moment generating function, i.e.

$$K_Y(t) = \log M_Y(t).$$

For a normally distributed random variable $Z \sim N(\mu, \sigma^2)$, the moment generating function is given by

$$M_Z(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

which gives

$$K_Z(t) = \mu t + \frac{1}{2}\sigma^2 t^2.$$

This implies that the first two derivatives $\partial^k K_U(t)/\partial t\big|_{t=0}$ are equal to the mean μ and the variance σ^2 and all higher order derivatives are equal to zero.

Moreover, note that for any constant $c > 0$ one has

$$M_{cY}(t) = E\left(e^{tcY}\right) = M_Y(ct).$$

The last component needed for our proof is that a random variable is uniquely defined by its cumulant (or moment) generating function and vice versa (as long as the moments and cumulants are finite). Hence, if we prove convergence of the cumulant generating function, it implies convergence of the corresponding random variables as well. Assume now that $Y_1, \ldots, Y_n$ are independent and identically distributed random variables (but not necessarily normal), each drawn from the distribution $F_y()$ with mean value μ and variance σ^2. Then, the moment generating function of $Z_n = (Y_1 + \ldots + Y_n)/\sqrt{n}$ is given by

$$\begin{aligned} M_{Z_n}(t) &= E\left(e^{t(Y_1+Y_2+\ldots+Y_n)/\sqrt{n}}\right) \\ &= E\left(e^{tY_1/\sqrt{n}} \cdot e^{tY_2/\sqrt{n}} \cdot \ldots \cdot e^{tY_n/\sqrt{n}}\right) \\ &= E\left(e^{tY_1/\sqrt{n}}\right) \cdot E\left(e^{tY_2/\sqrt{n}}\right) \cdot \ldots \cdot E\left(e^{tY_n/\sqrt{n}}\right) \\ &= M_Y\left(t/\sqrt{n}\right)^n. \end{aligned}$$

Correspondingly, the cumulant generating function for Z_n is given by

$$K_{Z_n}(t) = nK_Y\left(t/\sqrt{n}\right).$$

Taking derivatives gives

$$\left.\frac{\partial K_{Z_n}(t)}{\partial t}\right|_{t=0} = \frac{n}{\sqrt{n}}\left.\frac{\partial K_Y(t)}{\partial t}\right|_{t=0} = \sqrt{n}\mu$$

$$\left.\frac{\partial^2 K_{Z_n}(t)}{(\partial t)^2}\right|_{t=0} = \frac{n}{n}\left.\frac{\partial^2 K_Y(t)}{(\partial t)^2}\right|_{t=0} = \sigma^2$$

and all higher order derivatives tend to zero as n goes to infinity. Using Taylor series expansion we can derive $K_{Z_n}(t)$ around $K_Z(0)$ with

$$K_{Z_n}(t) = K_{Z_n}(0) + \frac{\partial K_{Z_n}(t)}{\partial t}|_{t=0} t + \frac{1}{2}\frac{\partial^2 K_{Z_n}(t)}{(\partial t)^2} t^2 + \ldots$$

$$= 0 + \sqrt{n}\mu t + \frac{1}{2}\sigma^2 t^2 + \ldots,$$

where the terms collected in $\ldots$ are of order $1/\sqrt{n}$ or smaller. Hence, for $n \to \infty$ one has

$$K_{Z_n}(t) \to K_Z(t),$$

where $Z \sim N(\sqrt{n}\mu, \sigma^2)$. This in turn proves the central limit theorem. □

We should finally note that in fact the central limit theorem has been proven to apply under a much weaker set of assumptions, where the random variables need not be identically distributed or may be dependent.

The central limit theorem is visible in the middle plot of Fig. 2.2, but even clearer in Fig. 2.3. The bottom plot of Fig. 2.2, on the other hand, demonstrates what is known as the law of large numbers, which we define as follows.

Property 2.5 (Law of Large Numbers) . Let $Y_1, Y_2, Y_3, \ldots$ be independent and identically distributed random variables with mean μ. Then for every $\varepsilon > 0$ we have

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu| > \epsilon) \to 0$$

The law of large numbers states that the arithmetic mean (or any other quantity which is expressible as an arithmetic mean) converges to its mean value. In particular, the difference between the mean value μ and the arithmetic mean $\sum_i Y_i/n$ becomes infinitely small for the sample size n increasing. One important special case results when Y_i is a Bernoulli variable with $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$ and $E(Y_i) = p$. Then $\frac{1}{n}\sum_{i=1}^{n} Y_i \to p$. Since $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is the relative frequency of the outcome 1, this result can be interpreted as follows. When we independently repeat an experiment with a 0/1 outcome very often, then the relative frequency of successes converges to its success probability. This mirrors in an interesting way the Frequentist definition of probability (see Definition 2.2).

Let us go a step further and consider *i.i.d.* continuous random variables. As an example we use the exponential distribution with parameter $\lambda = 1$, i.e.

$$Y_i \sim Exp(\lambda), i = 1, \ldots, n.$$

In Fig. 2.4, a histogram for sample sizes $n = 30$ and $n = 1000$ is given. The histogram comes to closely resemble the density function (solid line) as the sample

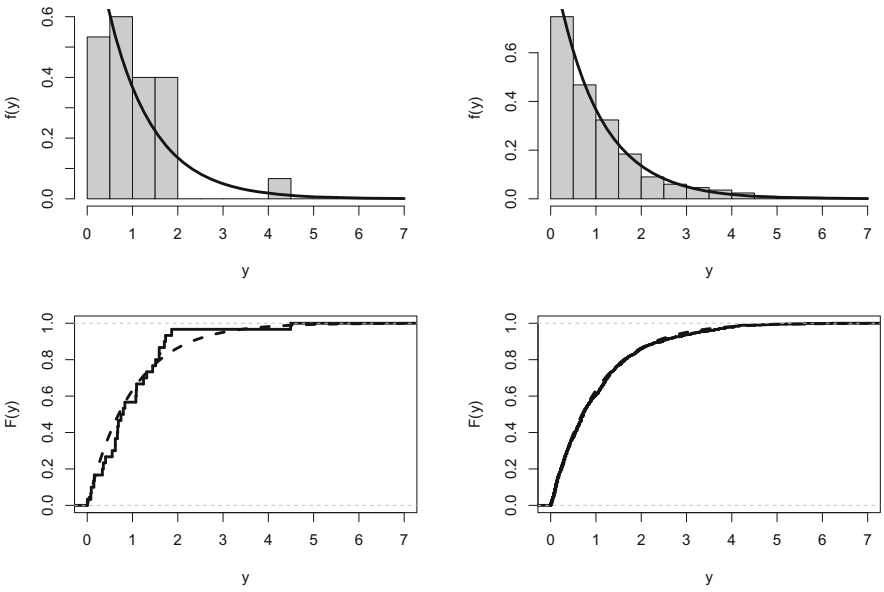

Fig. 2.4 Histograms (top row) and empirical distribution functions (bottom row) for n = 30 (left column) and n = 1000 (right column)

size increases. Alternatively, we can look at the empirical distribution function of the data. This is defined by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \leq y\}}$$

in other words, the proportion of data-points in the dataset with values less than y. For $n \to \infty$, the empirical distribution converges to the (theoretical) distribution $F(y)$, which is illustrated in Fig. 2.4.

Property 2.6 (Glivenko-Cantelli Theorem) Let $y_1, \ldots, y_n$ be a random sample from a distribution with distribution function $F(.)$. Then

$$lim_{n \to \infty} F_n(y) = F(y) \quad \text{for all } y \in \mathbb{R}. \tag{2.2.1}$$

$F_n(y)$ is the empirical distribution function for the sample $y_1, \ldots, y_n$.

The statement goes back to Glivenko (1933) and Cantelli (1933). The theorem is of central importance in data analysis as it states that the empirical distribution function is a good estimate for the true distribution function. We will extensively exploit the theorem in Chap. 9, when we discuss resampling, specifically with bootstrapping. For the proof we refer to the standard probability literature, e.g. Karr (1993), page 206.

2.3 Kullback–Leibler Divergence

In statistics, we are often faced with comparing two distributions. Assume that $f(y)$ is a density (or probability function) and so is $g(y)$. Our intention is to measure how far apart $f(.)$ and $g(.)$ are. Consider Fig. 2.5 (top plot), where we show two densities $f(.)$ and $g(.)$. One approach would be to look at the log ratio of the two distribution

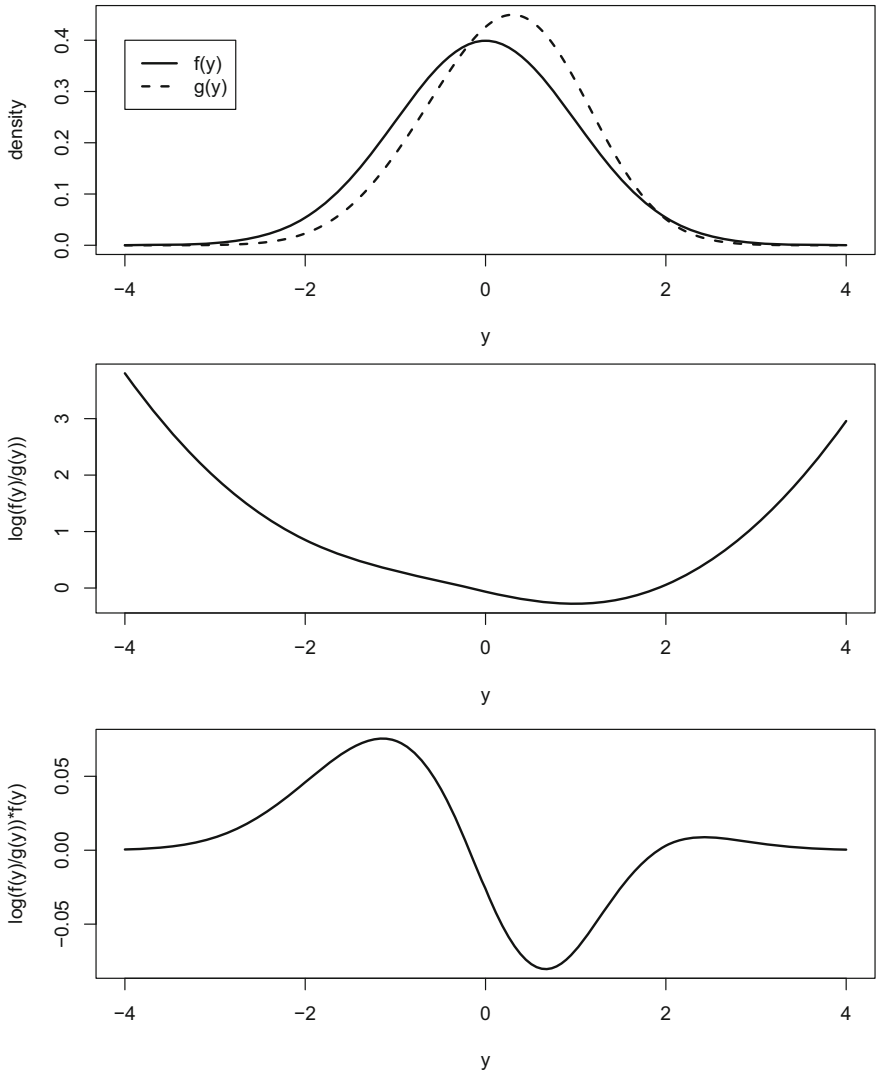

Fig. 2.5 Two different densities (top plot), their log ratio (middle plot), and log ratio weighted with $f(y)$ (bottom plot)

functions $\log(f(y)/g(y))$, which is visualised in the middle plot of Fig. 2.5. Clearly, this is largest in absolute terms at the boundary where both densities are small. To compensate for this, we can weight the ratio with respect to one of the two densities. This is shown in the bottom plot of Fig. 2.5, where we plot $\log(f(y)/g(y))f(y)$. After applying this change, the function is most pronounced around -1 and 1, where the difference (and hence distance) between the densities $f(y)$ and $g(y)$ matters most. We can now integrate this function, which leads us to the definition of the Kullback–Leibler divergence.

Definition 2.16 Let $f(y)$ and $g(y)$ be two densities or probability functions with the same support, i.e.$\{y : f(y) > 0\} = \{y : g(y) > 0\}$. The **Kullback–Leibler divergence** (KL divergence) is defined by

$$KL(f(.), g(.)) = \int_{-\infty}^{\infty} \log \frac{f(y)}{g(y)} f(y)dy = E_f\{\log \frac{f(Y)}{g(Y)}\}. \tag{2.3.1}$$

The Kullback–Leibler divergence will be used in many places throughout this book and thus it might be useful to discuss the measure in more depth. First of all, we need to emphasise that the Kullback–Leibler measure is a divergence and not a distance. The difference is that a distance measure needs to be symmetric, i.e. the distance from $f(.)$ to $g(.)$ is the same as the distance from $g(.)$ to $f(.)$. This simple property does not hold for the KL divergence and in fact $K(f(.), g(.)) \neq K(g(.), f(.))$ unless $g(.) = f(.)$. Nonetheless, it holds that:

$$KL(f(.), g(.)) = \begin{cases} 0 & \text{if } f(y) = g(y) \text{ for all } y \in \mathbb{R} \\ > 0 & \text{otherwise.} \end{cases}$$

This property can be easily seen as for the (natural) logarithm it holds $\log(x) \leq x-1$ for all $x \geq 0$ and $\log(x) = x - 1$ if and only if $x = 1$. This implies for two densities $f(y)$ and $g(y)$ that

$$\begin{aligned} KL(f(.), g(.)) &= \int \log \frac{f(y)}{g(y)} f(y)dy = -\int \log \frac{g(y)}{f(y)} f(y)dy \\ &\geq \int \big(1 - \frac{g(y)}{f(y)}\big) f(y)dy = 1 - 1 = 0. \end{aligned}$$

Note that the Kullback–Leibler divergence decomposes to

$$\begin{aligned} KL(f(.), g(.)) &= \int \log(f(y)) f(y)dy - \int \log(g(y)) f(y)dy \\ &= E_f\{\log\ f(Y)\} - E_f\{\log\ g(Y)\}, \end{aligned}$$

where the first component is also defined as the **entropy** of $f(y)$. The Kullback–Leibler divergence will be used at several places throughout the book. Indeed, it turns out that this quantity is of central importance in statistical reasoning.

2.4 Exercises

Exercise 1
A random variable Y with values in the interval $(0, 1)$ is sometimes modelled with a Beta distribution. The probability density function of a Beta(α, β) distributed random variable Y is given by

$$f(y) = \frac{1}{\mathrm{B}(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} ,$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta function

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} ,$$

and $\Gamma(\cdot)$ denotes the Gamma function. Show that the Beta distribution is a member of the exponential family and determine the (natural) parameters.

Exercise 2
A family of distributions with two parameters, the distribution function

$$F(y|a, b) = F_0\left(\frac{y-a}{b}\right) , a \in \mathbb{R}, b > 0$$

and $F_0(y)$ defined for $a = 0, b = 1$ is called a location scale family. a is the location parameter and b is the scale parameter.

1. Show that the density f in the continuous case is

$$f(y|a, b) = \frac{1}{b} f_0\left(\frac{y-a}{b}\right) .$$

2. The density of a generalised t-distributed random variable $Y \sim t_\nu(\mu, \sigma^2)$ with location parameter μ, scale parameter $\sigma \in \mathbb{R}_+$ and $\nu > 2$ degrees of freedom is

$$f(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\sigma}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} .$$

The moments are $E(Y) = \mu$ and $Var(Y) = \frac{\nu}{\nu-1}\sigma^2$. Show that the family of generalised t-distributions is an exponential family for fixed ν.

Exercise 3 (Use R Statistical Software)
Random numbers are often used to design simulation studies for evaluating the performance of statistical procedures. Generate a sample of $n = 200$ trivariate normal random numbers with mean $\mu = (3.0, 10.0, 50.0)$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1.0 & 0.6 & 0.6 \\ 0.6 & 1.0 & 0.6 \\ 0.6 & 0.6 & 1.0 \end{pmatrix} .$$

1. As a first method, use the fact that, if X is a random vector of independent standard normal $N(0, 1)$ variables, then

$$Y = \mu + \Sigma^{1/2} X$$

has the desired mean and covariance matrix. Use e.g. the Cholesky decomposition in `R` to compute the matrix root $\Sigma^{1/2}$. By repeating the experiment $S = 1000$ times, check empirically that your random numbers have the desired expected values, variances and covariances.
2. A more comfortable way is to use the function `rmvnorm` in the package `mvtnorm`. Repeat your experiment using this function.
3. Calculate the (conditional) variance of the conditional distribution of $(Y_1|Y_2, Y_3)$.

Exercise 4 (Use R Statistical Software)
Write a simulation program using random numbers and plots to show that the central limit theorem holds for independent Poisson distributed random variables. Evaluate, whether the speed of convergence depends on the choice of the parameter $\lambda : \lambda = 0.1, \lambda = 0.5, \lambda = 1.0, \lambda = 5, \lambda = 10, \lambda = 50$. Discuss your results.