Chapter 3
Parametric Statistical Models

Now that we have introduced probability models, we have the tools we need to put our statisticians' hats on. We want to make a probabilistic model that best describes the world around us. How is it that we can best move from our set of observations to a good model—a model that not only describes our samples, but the process that generated them? In this chapter, we start by making the assumption that the observed data follow a probability model, whose properties are described by a set of parameters. Now that we have this data, the statistical question is: how can we draw information from the samples about the parameters of the distributions that generated them? One basic assumption that aids this process enormously is that of independence.

Typically, statistical reasoning is built upon the assumption of **independence** of the observed measurements. In fact, most of the theoretical results in the subsequent chapter rely on this concept in one way or another. Therefore, we often assume that our observations are (or could be) replicated and that the outcome of one measurement does not depend on the outcome of any other. In practice, however, this requirement is not always met and, in fact, one often needs to think very carefully about whether it applies at all. To make this clear we look again at a very (!) simple and artificial example. Assume we are interested in the height difference between men and women and measure the heights of a single man and a woman once per day. Our database grows larger and larger. However, the data carry very little information about our research question because independence is violated. The data are taken from only two individuals! From the data we could, for example, extract information about height variation of an individual over different days, i.e. their height conditional upon the individual chosen. Either way, we still could not draw a conclusion, even with thousands of data-points and means that differed substantially. This is because our question relies on the individual chosen for each sample to be independent, which is clearly not the case here. We hope that this little example demonstrates the importance of sample independence. For the moment we will continue with this common assumption.

3.1 Likelihood and Bayes

Now is a good chance to introduce the concept of independence more formally. Let us begin with random variables $Y_1, \ldots, Y_n$ that come from the probability model

$$Y_i \sim F(y; \theta). \tag{3.1.1}$$

Here, $F(y; \theta)$ denotes a known distribution function, but the parameter θ is unknown. The parameter may contain multiple components, i.e. $\theta = (\theta_1, \ldots, \theta_p)$, but for simplicity's sake let us take θ to be univariate, i.e. $p = 1$. We let Θ be the set of possible parameter values, i.e. $\theta \in \Theta$. Note that the distribution function $F(y; \theta)$ is the same for all n observations. Finally, we postulate that the n observations are mutually independent, which gives

$$P(Y_1 \leq y_1, \ldots, Y_n \leq y_n) = P(Y_1 \leq y_1) \cdot \ldots \cdot P(Y_n \leq y_n) = \prod_{i=1}^{n} F(y_i; \theta).$$

This scenario is called the ***i.i.d.*** setting, where the abbreviation stands for *i*ndependently and *i*dentically *d*istributed random variables. We repeat that one should not accept this at face value, as in many real data situations, the $i.i.d.$ assumption may be violated. For now, however, we rely on the $i.i.d.$ assumption, defined as follows.

Definition 3.1 The data $y_1, \ldots, y_n$ are called **independently and identically distributed** ($i.i.d.$) if for $i = 1, \ldots, n$ we assume y_i to be the realisation of a random variable Y_i, with distribution $F(y; \theta)$ and that

$$P(Y_1 \leq y_1, \ldots, Y_n \leq y_n) = \prod_{i=1}^{n} F(y_i; \theta).$$

We now aim to use the observed data $y_1, \ldots, y_n$ to draw information about θ in the model. Given that the parameter θ is unknown, we may take the viewpoint of a subjective probability and assume that θ is random. Clearly this is a conceptional jump, but it has some attractive aspects that will become clear shortly. Bear in mind, a random variable is unknown but a realisation can be obtained by drawing and observing it.

Using the notion of subjective probability, we can formulate our uncertainty about θ in the form of a probability model. Following this path, we assume

$$\theta \sim F_\theta(\vartheta), \tag{3.1.2}$$

where $F_\theta(.)$ denotes a suitable distribution function for θ with $P(\theta \leq \vartheta) = F_\theta(\vartheta)$. This distribution is later called **prior distribution** (or sometimes *a priori* distribution) and it may depend on additional parameters, called hyperparameters,

which are omitted for now. To distinguish between the random variable θ from its possible realisations, we denote possible concrete values of θ with ϑ. One way to interpret the prior distribution (3.1.2) is as a representation of our existing knowledge about the parameter before observing any data. For instance, if we are sure that θ takes a particular value θ_0, then F_θ can be taken as a step function such that:

$$F_\theta(\vartheta) = \begin{cases} 0 & \vartheta < \theta_0 \\ 1 & \text{otherwise.} \end{cases}$$

On the other hand, if we are not at all sure about θ, then $F_\theta()$ can also demonstrate our lack of knowledge before observing the data. Let us take the example of a Bernoulli distribution, i.e. an experiment with two outcomes $Y \in \{0, 1\}$, where $P(Y = 1) = \theta$ and $P(Y = 0) = 1 - \theta$. The unknown parameter θ can hold values between 0 and 1. One strategy to express total uncertainty is to assume a flat prior, i.e. with density $f_\theta(\vartheta) = 1$ if $0 \leq \vartheta \leq 1$ and $f_\theta(\vartheta) = 0$ otherwise. That is, we think that all valid values for θ are equally likely. There are other ways to express uncertainty in our prior that will be covered in more detail in Chap. 5.

Now we observe $y_1, \ldots, y_n$ as realisations of (3.1.1) and obtain information about θ which we need to quantify. So let us use $f(y; \theta)$ as the density or probability function derived from (3.1.1). We can now use **Bayes rule** and calculate the conditional density

$$f_\theta(\vartheta | y_1, \ldots, y_n) = \frac{\left[\prod_{i=1}^{n} f(y_i; \vartheta)\right] f_\theta(\vartheta)}{\int_\Theta \left[\prod_{i=1}^{n} f(y_i; \tilde{\vartheta})\right] f_\theta(\tilde{\vartheta}) d\tilde{\vartheta}}. \tag{3.1.3}$$

Hence $f_\theta(\vartheta | y_1, \ldots, y_n)$ denotes the conditional density or probability function corresponding to the distribution function $F_\theta(\vartheta | y_1, \ldots, y_n)$. The distribution (3.1.3) is also called the **posterior distribution** as it is calculated after observing the data. Note that the denominator in (3.1.3) is a normalisation constant that does not depend on ϑ.

Looking at Eq. (3.1.3), we see that our knowledge of θ changes with data $y_1, \ldots, y_n$ being observed. We multiply the prior distribution with the density function given our data to obtain the posterior. The first component in the numerator of the ratio (3.1.3) essentially tells us how likely we are to observe our data realisations, given a particular parameter value. The prior expresses our prior knowledge on how likely we are to observe that parameter value at all. Taken together, these functions give us an updated distribution of the parameter that favours parameter values that are more likely to have produced our data. The function that describes the likelihood of producing our data given a particular parameter is defined as the **likelihood** of the data.

Definition 3.2 The **likelihood function** $L(\theta; y_1, \ldots, y_n)$ for an $i.i.d.$ sample $y_1, \ldots, y_n$ is defined by

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i; \theta) \tag{3.1.4}$$

with $f(y_i; \theta)$ as a density or probability function. Taking the logarithm defines the **log-likelihood function**

$$l(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i; \theta).$$

The likelihood function and likelihood theory play an essential role in statistics. This will be discussed in depth in Chap. 4. For now we conclude, because the denominator of (3.1.3) is simply a normalisation constant that does not depend upon ϑ, the posterior distribution is proportional to the product of the likelihood function and the prior, i.e.

$$f_\theta(\vartheta | y_1, \ldots, y_n) \propto L(\vartheta; y_1, \ldots, y_n) f_\theta(\vartheta),$$

where $\propto$ here means "is proportional to". This proportionality implies that the combination of the likelihood and the prior density provides the quantifiable information (under the given probability model) about θ that the data-points $y_1, \ldots, y_n$ provide. We will make use of this function in Chap. 5 when we discuss Bayesian inference. If there is no prior knowledge about θ, or if we do not want to include prior knowledge in our analysis, then a **non-informative prior** is assumed. If we use the previously mentioned flat prior for θ we find that all information about θ is contained in the likelihood function $L(\theta | y_1, \ldots, y_n)$, because our prior then gets absorbed into the denominator as a constant and we are left with

$$f(\vartheta | y_1, \ldots, y_n) \propto L(\vartheta; y_1, \ldots, y_n).$$

This in turn shows that the likelihood function is an essential tool for extracting information about θ.

3.2 Parameter Estimation

We have defined two central functions above for drawing information about the parameter θ in the model: the posterior distribution (3.1.3) and the likelihood function (3.1.4). In both cases, the parameter values that define models that better describe the data become themselves the more likely. The next step is to use these functions to find the best value for the unknown parameter given the data $y_1, \ldots, y_n$.

This is to say, we are interested in a single value $\hat{\theta}$, which is a *plausible guess* for the unknown parameter θ. We call this guess an **estimate**. In what follows, we tackle the question of how to construct good and useful estimates from our data and how to assess the quality of these estimates. Note that our estimate $\hat{\theta}$ depends on the observed data, i.e.

$$\hat{\theta} = t(y_1, \ldots, y_n), \tag{3.2.1}$$

where $t : \mathbb{R}^n \mapsto \mathbb{R}$. We call any function of the data (and not the model parameters) $t(\cdot)$ a **statistic**. It is often helpful to bear in mind that $t(\cdot)$ depends on the sample size n, which we sometimes denote as $t_{(n)}(\cdot)$. The reader should also note that $t(y_1, \ldots, y_n)$ depends on the realisations $y_1, \ldots, y_n$ of random variables $Y_1, \ldots, Y_n$. However, taking the random variables instead of the realised values, leaves the statistic $t(Y_1, \ldots, Y_n)$ itself as a random variable. This allows us to model the properties of the statistic on average, independently of the random sample that is taken. This view will be essential for statistical reasoning, as we evaluate an estimate, not based on the concrete value that results from the sample $t(y_1, \ldots, y_n)$, but based on its stochastic counterpart $t(Y_1, \ldots, Y_n)$.

In the coming section, we examine a number of different approaches to the parameter estimation process. Maximum Likelihood chooses the model parameterisation that is most likely to have generated the given data, while Bayes estimation also includes the effect of our prior knowledge of the parameter. This is done by treating it as a random variable and finding its posterior probability after observing the data. Method of moments estimation attempts to find the parameter by matching the theoretical moments to the observed moments of the distribution as a system of equations. A different perspective is to take a traditional optimisation approach and minimise a loss between predicted and actual values, or even minimise the difference between the two entire distributions using the Kullback–Leibler divergence. We begin by looking at Bayes estimation.

3.2.1 Bayes Estimation

From the Bayesian perspective, the posterior distribution (3.1.3) given the data $y_1, \ldots, y_n$ contains all of the information that we have about θ. Therefore, it seems that the posterior mean value would be a natural candidate for estimating θ.

Definition 3.3 The **posterior mean estimate** $\hat{\theta}_{postmean}$ is defined by

$$\hat{\theta}_{postmean} = E_\theta(\vartheta | y_1, \ldots, y_n) = \int_{\vartheta \in \Theta} \vartheta f_\theta(\vartheta | y_1, \ldots, y_n) d\vartheta. \tag{3.2.2}$$

An alternative is to use the mode of the distribution:

Definition 3.4 The **posterior mode estimate** $\hat{\theta}_{postmode}$ is defined by

$$\hat{\theta}_{postmode} = \arg\max_{\vartheta} f_{\theta}(\vartheta | y_1, \ldots, y_n).$$

Taking the mode has a couple of advantages compared to the mean. Firstly, in order to calculate the posterior mean (3.2.2), integration is required, which can be cumbersome and numerically demanding. In contrast, taking the mode simply requires finding the maximum of the function, which is usually numerically much easier to solve or can even be calculated analytically. Secondly, the maximum of the posterior density function is, loosely speaking, the value with the highest probability, which also seems like a more intuitive candidate for θ.

3.2.2 Maximum Likelihood Estimation

If we assume a uniform prior, that is that $f_{\theta}(\vartheta)$ is constant, we can directly maximise the likelihood function itself. This gives the **Maximum Likelihood estimate**, commonly called the **ML estimate**.

Definition 3.5 The **Maximum Likelihood estimate** (or ML estimate) is defined by

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; y_1, \ldots, y_n).$$

The ML estimate can also be calculated by differentiating the log-likelihood function instead of the likelihood function. Not only is it simpler to differentiate a sum instead of a product, we will also learn in Chap. 4 that the log-likelihood has important properties. In fact, ML estimation is the most frequent estimation principle in statistics and we will cover it in the next chapter extensively. It is also important to note that, given that we are no longer considering our prior in this process, we can find the Maximum Likelihood estimate outside of the Bayesian framework. That is, we can estimate our parameter θ without assuming it to be random. The ML estimate and the posterior mode estimate obey the invariance property. This means that any transformation of the parameter directly yields the new estimate. Hence, if θ is the parameter and we are interested in the transformed parameter, $\gamma = g(\theta)$, where $g(.)$ is a bijective transformation function, i.e. there is exactly one g(x) for each x and vice versa, then $\hat{\gamma}_{ML} = g(\hat{\theta}_{ML})$ is the transformed ML estimate.

3.2.3 Method of Moments

We thought it remiss to introduce the concept of parameter estimation without introducing the first ever method for calculating parameter estimates. This method was introduced by Karl Pearson (1857–1936) and is based on relating the theoretical moments of random variables $Y_1, \ldots, Y_n$ to the empirical moments of the observed data $y_1, \ldots, y_n$. Assume that the expectation $E(Y^k) = \int y^k f(y; \theta) dy$ is a function of the unknown parameter θ. The empirical moments of the data are given by

$$m_k(y_1, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i^k.$$

For instance, in the case of a normal distribution, the parameters μ and σ can be written as $\mu = E(Y)$ and $\sigma^2 = E(Y^2) - (E(Y))^2$. An obvious choice for estimating the mean parameter μ is to take the mean

$$\hat{\mu} = m_1 = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

For the estimation of σ^2 one can use

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} y_i \right)^2 = m_2 - m_1^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The method of moments estimator relates the empirical and theoretical moments such that

$$E_{\hat{\theta}_{MM}}(Y^k) = m_k(y_1, \ldots, y_n),$$

where $\hat{\theta}_{MM}$ is the resulting method of moment estimator. Hence, the expectation derived from the probability model should match the observed value of the statistic. The calculation of $\hat{\theta}_{MM}$ is not always easy and sometimes requires numerical simulation routines. In exponential family distributions, however, the method of moments and Maximum Likelihood estimation coincide, which we show in the next chapter (see Example 9).

3.2.4 Loss Function Approach

So far, we have introduced probability models and then derived estimates for their parameters. An alternative approach is to make use of loss functions, which

formalise a penalty for deviances of our estimate from the true value. To do this, we again make use of our data $y_1, \ldots, y_n$ to derive an estimate for the unknown parameter θ. This estimate is denoted as $\hat{\theta}$ and is calculated from the data with the statistic $\hat{\theta} = t(y_1, \ldots, y_n)$. With this in mind, we now want to find out how close our estimate $\hat{\theta}$ is to the true but unknown parameter θ. Clearly, if $\hat{\theta}$ is equal to θ, we have estimated the true parameter perfectly and if $\hat{\theta} \neq \theta$, we have made some sort of error. The idea is now to quantify the extent of this error with a loss function defined as follows:

Definition 3.6 Let $\Theta \subset \mathbb{R}$ be the parameter space and let $t : \mathbb{R}^n \to \mathbb{R}$ be a statistic which is intended to estimate the parameter θ. With $\mathcal{T}$ we define the set of possible outcomes of $t(Y_1, \ldots, Y_n)$. A **loss function** $\mathcal{L}$ is defined as

$$\mathcal{L} : \mathcal{T} \times \Theta \to \mathbb{R}^+,$$

where the minimum value is equal to zero and occurs if both elements are equal, i.e. $\mathcal{L}(\theta, \theta) = 0$.

One example of a very common loss function is the squared loss defined by

$$\mathcal{L}(t, \theta) = (t - \theta)^2.$$

Another example makes use of absolute distances, i.e. $\mathcal{L}(t, \theta) = |t - \theta|$. Both make sense to describe different objectives and their use is problem-dependent. For example, the squared loss clearly penalises estimates far away from the true value θ more than the absolute loss. Historically, the squared loss was used in many settings because it was easily differentiable, but the absolute loss is now widely available if appropriate, even for large models. For estimation, the intention is that the loss $\mathcal{L}(t, \theta)$ should be as small as possible. The reader should take note, however, that because θ is unknown, this loss cannot be calculated and thus $\mathcal{L}(t, \theta)$ is a purely theoretical quantity. Our loss also depends on our sample, and hence is a scalar that is dependent upon the data that we have gathered.

Let us now move on to the next question. What if we no longer want to evaluate a single estimate, but instead want to determine if the *estimator* $t(y_1, \ldots, y_n)$ itself is effective? That is, we want to determine if our method for estimating θ works consistently well, no matter what sample we get. As usual, we want to model the properties of this function in general, so let us take our sample, and by extension our statistic, to be random. We can now take the expectation of our loss to find out how it reacts on average. To deal with this notationally, we denote concrete realisations of t from a real dataset as $t = t(y_1, \ldots, y_n)$ and let $t(.)$ describe the function itself. So, instead of looking at the loss, which evaluates the performance of our estimate for a single realised sample, we evaluate a stochastic sample that takes the probability model of $Y_1, \ldots, Y_n$ into account. This takes us from loss to risk and allows us to define the related risk function.

Definition 3.7 For a given loss function $\mathcal{L}(t, \theta)$ we define the **risk function** with

$$R(t(.), \theta) = E\Big(\mathcal{L}(t(Y_1, \ldots, Y_n), \theta)\Big) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \mathcal{L}(t(y_1, \ldots, y_n), \theta) \prod_{i=1}^{n} f(y_i; \theta) dy_i.$$

Note that this is an n-dimensional integral and due to independence the density results as product of marginal densities.

The parameter θ in the risk is the true unknown parameter. Taking the squared loss allows us to split the risk as calculated above into bias and variance components, an idea that is central to both statistics and machine learning. Let $\mathcal{L}(t, \theta) = (t - \theta)^2$ such that for $Y = (Y_1, \ldots, Y_n)$ we get

$$\begin{aligned} R(t(.), \theta) &= E\Big(\{t(Y) - \theta\}^2\Big) \\ &= E\Big(\{t(Y) \underbrace{-E(t(Y)) + E(t(Y))}_{=0} - \theta\}^2\Big) \\ &= \underbrace{E\Big(\{t(Y) - E(t(Y))\}^2\Big)}_{1} + \underbrace{E\Big(\{E(t(Y)) - \theta\}^2\Big)}_{2} + \\ &\quad \underbrace{2E\Big(\{t(Y) - E(t(Y))\}\{E(t(Y)) - \theta\}\Big)}_{3}. \end{aligned}$$

Note that Component 1 is given by the variance of $t(Y) = t(Y_1, \ldots, Y_n)$ while Component 2 defines the systematic error, which we will define as bias below. Finally, component 3 is by definition equal to 0, i.e. $E_\theta\Big(t(Y) - E_\theta(t(Y))\Big) \equiv 0$. So we can see that the squared loss decomposes the risk into the sum of the variance (Component 1) and the squared bias (Component 2). The variance describes how much variation we have in estimates sampled from the distribution with the given parameters, i.e. for a given θ, while the bias captures how much we systematically over or underestimate the correct θ. Together, these components define the **mean squared error** (MSE).

Definition 3.8 Given an estimate $t = t(Y_1, \ldots, Y_n)$, the **mean squared error** (MSE) is defined as

$$MSE(t(.), \theta) = E\Big(\{t(Y) - \theta\}^2\Big) = \mathrm{Var}_\theta\Big(t(Y_1, \ldots, Y_n)\Big) + \mathrm{Bias}\big(t(.); \theta\big)^2,$$

where $\mathrm{Var}_\theta\Big(t(Y_1, \ldots, Y_n)\Big)$ is the variance of estimate $t = t(Y_1, \ldots, Y_n)$ and $\mathrm{Bias}(t(.), \theta)$ is the systematic error, called **bias**, defined by

$$\mathrm{Bias}(t(.), \theta) = E\Big(t(Y_1, \ldots, Y_n)\Big) - \theta.$$

An estimate with vanishing bias, i.e. $\text{Bias}(t(.), \theta) = 0$ for all $\theta \in \Theta$, is also called an **unbiased** estimate.

If we now take an estimator that minimises the risk we can be sure that we have a good estimator, but only for our unknown true parameter θ. For example, if our estimator was simply a constant, i.e. $t(.) = c$ and our true parameter just so happened to be equal to be c, the risk would be 0. Of course, this would be a terrible estimator if θ took any other value, so somehow our best estimator is still dependent upon our true θ. Therefore, a useful strategy would be to estimate θ with the function $t(.)$ such that the risk is minimised for all $\theta \in \Theta$. In special cases, this distinction does not matter. With the mean squared error this happens if the estimate is unbiased and the variance does not depend on θ. In this case, we can simply select the estimator $t(.)$ such that the risk is minimal.

This approach will give the best estimator no matter what the true θ value is. In general, however, we need to be more precise when minimising the risk and need to take θ into account. A "cautious" strategy would be to apply the **minimax** approach. First we choose θ such that the risk is highest (maximal), then we select $t(.)$ such that the highest risk is smallest (minimal). In mathematical notation this means that we are looking for

$$\hat{\theta}_{minimax} = \arg\min_{t(.)} \Big(\max_{\theta \in \Theta} R(t(.), \theta) \Big), \tag{3.2.3}$$

where $t(.)$ is selected from all possible statistics, i.e. functions from $\mathbb{R}^n$ to $\mathbb{R}$. While the minimax approach may appear sound, it is certainly a very cautious strategy, as we aim to minimise the worst case error for our estimation and is therefore often not very practical.

The dependence of the risk $R(t(.), \theta)$ on the parameter θ can also be circumvented by following a Bayesian approach. In this case, we assume that the parameter θ has a prior probability $f_\theta(.)$, which allows us to find the expected value over all possible values of θ. This leads to the **Bayes risk**

$$R_{Bayes}\Big(t(.)\Big) = E_\theta\Big(R(t(.), \theta)\Big) = \int_\Theta R(t(.), \vartheta) f_\theta(\vartheta) d\vartheta,$$

where the risk of course depends on the prior distribution. The **Bayes-optimal** estimation is then found by minimising the Bayes risk, that is

$$\hat{\theta}_{Bayes} = \arg\min_{t(.)} R_{Bayes}\Big(t(.)\Big), \tag{3.2.4}$$

where again the minimum is taken over all possible statistics $t(.)$. Note that the risk is calculated with respect to the prior distribution. It seems, however, strategically more appropriate to use the posterior distribution instead of the prior. A more useful

alternative to the Bayes risk is therefore to use the posterior distribution yielding the **posterior Bayes risk**.

$$R_{post.Bayes}\Big(t(.) \mid y_1, \ldots, y_n\Big) = \int_\Theta \mathcal{L}\Big(t(y_1, \ldots, y_n), \vartheta\Big) f_\theta(\vartheta | y_1, \ldots, y_n) d\vartheta$$
$$= E_{\theta|y}\Big(\mathcal{L}(t(y), \theta) \mid y\Big).$$

We may again minimise $R_{post.Bayes}(.)$ with respect to $t(.)$ yielding a posterior Bayes risk estimate.

$$\hat{\theta}_{post.Bayes.risk} = \arg\min_{t(.)} R_{post.Bayes}\Big(t(.)|y_1, \ldots, y_n\Big).$$

3.2.5 Kullback–Leibler Loss

A loss function evaluates an estimate $\hat{\theta}$ by evaluating how far it is from the unknown parameter θ. An alternative would be to compare the two distributions directly, one parameterised by our estimate and the other by the true parameter. We therefore assume that θ is estimated by $\hat{\theta} = t(y_1, \ldots, y_n)$, such that $f(\tilde{y}; \hat{\theta})$ is the estimated probability function at some arbitrary value $\tilde{y} \in \mathbb{R}$. We now look at the log ratio

$$\log \frac{f(\tilde{y}; \theta)}{f(\tilde{y}; \hat{\theta})}, \tag{3.2.5}$$

which indicates the discrepancy between the true density $f(\tilde{y}; \theta)$ and the estimated density $f(\tilde{y}; \hat{\theta})$. Clearly, if $\hat{\theta} = \theta$ the ratio equals 1 and taking the log makes the term vanish. The log ratio (3.2.5) depends on the particular value of $\tilde{y}$. We can consider all possible values for $\tilde{y}$ by taking the expectation with respect to the true distribution. This suggests making use of the Kullback–Leibler divergence as introduced in Chap. 2.3.

With a slight change of notation we define

$$KL(\theta, \hat{\theta}) = KL(f(;\theta), f(;\hat{\theta})) = \int \log \frac{f(y;\theta)}{f(y;\hat{\theta})} f(y;\theta) dy.$$

Note that $KL(\theta, \hat{\theta}) > 0$ unless $\theta = \hat{\theta}$.

Note that the KL divergence is in fact a (non-symmetric) loss function, which may be explicitly defined as $\mathcal{L}_{KL}(t, \theta) = KL(\theta, t)$. With this in mind, we can construct a risk measure from the KL loss. Taking the statistic $t(.)$ this would be defined by

$$R_{KL}(t(.), \theta) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \mathcal{L}_{KL}\big(t(y_1, \ldots, y_n), \theta\big) \prod_{i=1}^{n} f(y_i; \theta) dy_i. \tag{3.2.6}$$

Note again that the above integral is n-dimensional, and due to independence, the density ends up as a product of marginal univariate densities. It seems like things are getting a bit complicated here, because we are integrating over the possible sample values $y_1, \ldots, y_n$, while in the Kullback–Leibler divergence itself we integrate over a possible observation $\tilde{y}$. This will be essential in Chap. 9, when we introduce model selection based on the Kullback–Leibler divergence. To simplify the calculations, we rely on a simple second-order Taylor series approximation and write

$$\log f(\tilde{y}, \hat{\theta}) \approx \log f(\tilde{y}, \theta) + \frac{\partial \log f(\tilde{y}, \theta)}{\partial \theta}(\hat{\theta} - \theta) + \frac{1}{2}\frac{\partial^2 \log f(\tilde{y}, \theta)}{(\partial \theta)^2}(\hat{\theta} - \theta)^2.$$

Writing $\hat{\theta} = t(y_1, \ldots, y_n)$ allows us to approximate the Kullback–Leibler risk (3.2.6) as

$$\begin{aligned}
&\int \ldots \int \left[\int \log \frac{f(\tilde{y}; \theta)}{f\big(\tilde{y}; t(y_1, \ldots, y_n)\big)} f(\tilde{y}; \theta) d\tilde{y} \right] \prod_{i=1}^{n} f(y_i; \theta) dy_i \\
&= \int \ldots \int \left[\int \log f(\tilde{y}; \theta) f(\tilde{y}, \theta) d\tilde{y} - \int \log f(\tilde{y}; t(y_1, \ldots, y_n)) f(\tilde{y}; \theta) d\tilde{y} \right] \prod_{i=1}^{n} f(y_i; \theta) dy_i \\
&\approx - \int \ldots \int \underbrace{\left(\int \frac{\partial \log f(\tilde{y}; \theta)}{\partial \theta} f(\tilde{y}; \theta) d\tilde{y} \right)}_{1} \Big(t(y_1, \ldots, y_n) - \theta \Big) \prod_{i=1}^{n} f(y_i; \theta) dy_i \\
&+ \frac{1}{2} \int \ldots \int \underbrace{\left(- \int \frac{\partial^2 \log f(\tilde{y}; \theta)}{(\partial \theta)^2} f(\tilde{y}; \theta) d\tilde{y} \right)}_{2} \Big(t(y_1, \ldots, y_n) - \theta \Big)^2 \prod_{i=1}^{n} f(y_i; \theta) dy_i .
\end{aligned}$$

We will see in Chap. 4 that component 1 is equal to zero and component 2 will later be defined as the Fisher information. Given that the Fisher information does not depend on the sample $y_1, \ldots, y_n$, we see that the Kullback–Leibler risk is roughly proportional to

$$\int \ldots \int \Big(t(y_1, \ldots, y_n) - \theta \Big)^2 \prod_{i=1}^{n} f(y_i; \theta) dy_i = E\Big((\hat{\theta} - \theta)^2 \Big),$$

which we defined as mean squared error risk above. The reverse also holds, meaning that if we are able to choose an estimate such that the mean squared error is small, then the estimated density will be close to the true density in the Kullback–Leibler sense. Minimising the mean squared error is therefore a useful strategy.

3.3 Sufficiency and Consistency, Efficiency

Now we have a number of different ways to construct estimators $t(.)$ for a sample $y_1, \ldots, y_n$, such as Maximum Likelihood estimation and posterior mean/mode estimation.

We also have a number of different ways of analysing an estimator's overall performance with various risk measures for a range of true parameter values, such as minimax and Bayesian risk. Let us now look at some further properties of our estimators that might also guide our choice, which will also introduce us to a number of important mathematical concepts. Instead of going too deep into theory here, we hope to simply motivate the key building blocks of and remain on a somewhat informal level.

3.3.1 Sufficiency

The first property we want to introduce is sufficiency. If we have a sample $y_1, \ldots, y_n$, we can try to condense all of the relevant information contained within this sample to the quantity $t(y_1, \ldots, y_n)$. Hence, from originally n different values we calculate a single value $t(y_1, \ldots, y_n)$, that should represent everything that we need to know. At first, this sounds like a tremendous loss of information. The question is now whether the information that we lose when taking our statistic instead of the entire sample is relevant or not. If $t(y_1, \ldots, y_n)$ has sufficient information about θ, it would mean that all information we get about θ from our sample is contained in $t(y_1, \ldots, y_n)$. This leads us to the concept of sufficiency. In fact, we can completely erase our data if we have calculated a sufficient estimate. This can be rather useful, particularly in the age of big data, as it states that one only needs to calculate and store $t(y_1, \ldots, y_n)$ in order to get all information from our sample relevant to our parameter of interest θ. To proceed, let us now define sufficiency more formally.

Definition 3.9 A statistic $t(Y_1, \ldots, Y_n)$ is called **sufficient** for θ if the conditional distribution $P(Y_1 = y_1, \ldots, Y_n = y_n | t(Y_1, \ldots, Y_n) = t_0; \theta)$ does not depend on θ.

This states, once again, that the distribution of the single values $y_1, \ldots, y_n$ is non-informative, if we know the value of the statistic $t(y_1, \ldots, y_n)$. This idea can be a little confusing at first glance, so an example might help to clarify.

Example 1 Let $Y_1, \ldots, Y_n$ be Bernoulli variables with values 0 or 1 and $P(Y_i = 1) = \pi$. Let our statistic be $t(.) = t(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} Y_i / n$, i.e. the mean. It holds that statistic $t(.)$ is sufficient for π. We can see that the statistic takes values

in set $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$ and we denote the resulting value with t_0. Correspondingly, $n_0 = nt_0$ is the sum of all of our variables. Then, using Bayes rule,

$$
\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n | t(Y_1, \ldots, Y_n) = t_0; \pi) \\
&= \frac{P\left(Y_1 = y_1, \ldots, Y_n = y_n, \sum_{i=1}^{n} Y_i = n_0; \pi\right)}{P\left(\sum_{i=1}^{n} Y_i = n_0; \pi\right)} \\
&= \begin{cases} \dfrac{\prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}}{\binom{n}{n_0}\pi^{n_0}(1-\pi)^{(n-n_0)}} & \text{for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{1}{\binom{n}{n_0}} & \text{for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Apparently, this distribution does not depend on π. ▷

It can be difficult to prove sufficiency in the above defined form, but the **Neyman-factorisation** makes it simple to find a sufficient statistic.

Property 3.1 A statistic $t(Y_1, \ldots, Y_n)$ is sufficient for θ if and only if the density or probability function decomposes to

$$
f(y_1, \ldots, y_n; \theta) = h(y_1, \ldots, y_n) g\Big(t(y_1, \ldots, y_n); \theta\Big), \tag{3.3.1}
$$

where $h(.)$ does not depend on θ and $g(.)$ depends on the data only through the statistic $t(y_1, \ldots, y_n)$.

The proof of this statement is rather simple and given at the end of this section. Note that sufficiency itself is a rather weak statement, as the original sample $(y_1, \ldots, y_n)$ itself is already sufficient. We therefore also require some concept of minimality defined as follows:

Definition 3.10 The statistic $t(y_1, \ldots, y_n)$ is minimal sufficient for θ if $t(.)$ is sufficient and for any other sufficient statistic $\tilde{t}(y_1, \ldots, y_n)$ there exists a function $m(.)$ such that $t(y_1, \ldots, y_n) = m(\tilde{t}(y_1, \ldots, y_n))$.

The definition states that if there exists a minimal sufficient statistic, then it can be calculated directly from any other sufficient statistic. Hence, we may reduce the

data $y_1, \ldots, y_n$ to the value of the minimal sufficient statistic, but we may not reduce it further without losing information about the parameter θ.

Sufficient statistics are also closely related to the exponential family distributions that we described in Sect. 2.1.5. We have written an exponential family distribution in the form

$$f(y; \boldsymbol{\theta}) = \exp\Big(t^T(y)\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})\Big)h(y).$$

This directly shows that, for a sample $y_1, \ldots, y_n$, one obtains $\sum_{i=1}^n t(y_i)$ as a minimal sufficient statistic.

Proof We here prove Neyman-factorisation. Assume that $t(.)$ is sufficient, then $f(y_1, \ldots, y_n | t(y) = t, \theta)$ does not depend on θ. Because $t(.)$ is calculated from $y_1, \ldots, y_n$, with the basic definition of conditional probabilities, we get

$$f(y_1, \ldots, y_n; \theta) = \underbrace{f(y_1, \ldots, y_n | t(y_1, \ldots, y_n) = t; \theta)}_{h(y_1, \ldots, y_n)} \underbrace{f_t(t | y_1, \ldots, y_n; \theta)}_{g(t(y_1, \ldots, y_n); \theta)},$$

where the first component does not depend on θ and the second component is the distribution of $t(.)$ which by construction depends only on $t(y_1, \ldots, y_n)$ and θ. Let us assume now that the density $f(y_1, \ldots, y_n)$ is factorised as in (3.3.1). The marginal density for $t(y_1, \ldots, y_n)$ is

$$\begin{aligned} f_t(t; \theta) &= \int_{t(y_1, \ldots, y_n) = t} f(y_1, \ldots, y_n; \theta) dy_1 \ldots dy_n \\ &= \int_{t(y_1, \ldots, y_n) = t} h(y_1, \ldots, y_n) g(t; \theta) dy_1 \ldots dy_n \\ &= g(t; \theta) \int_{t(y_1, \ldots, y_n) = t} h(y_1, \ldots, y_n) dy_1 \ldots dy_n. \end{aligned}$$

The conditional distribution can then be written as

$$\begin{aligned} f(y_1, \ldots, y_n | t(y_1, \ldots, y_n) = t; \theta) &= \frac{f(y_1, \ldots, y_n, t(y_1, \ldots, y_n) = t; \theta)}{f_t(t; \theta)} \\ &= \begin{cases} \frac{h(y_1, \ldots, y_n) g(t; \theta)}{g(t, \theta)} & \text{for } t(y_1, \ldots, y_n) = t \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

which does not depend on θ because $g(t, \theta)$ cancels out. □

3.3.2 Consistency

The principle of sufficiency describes how data can be condensed without losing information about the quantity of interest. The next question is how to quantify the gain of information for increasing sample size. We have seen that by taking the squared loss function, the risk equals the mean squared error (MSE). We have also seen that the MSE as loss measure approximates the Kullback–Leibler divergence. Let us look at this statement again, but explicitly focus on the role of the sample size n. In principle, we know that information increases with sample size, which means the estimate should approach the true value. This property can be formulated in mathematical terms, by once again making use of the mean squared error.

Definition 3.11 An estimate $\hat{\theta} = t(y_1, \ldots, y_n)$ is called **consistent** if

$$MSE(\hat{\theta}, \theta) = \mathrm{Var}_\theta(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2 \longrightarrow 0 \text{ as } n \to \infty.$$

Hence, if $\hat{\theta}$ is consistent, it approaches the unknown value θ with increasing sample size n. Using Chebyshev's inequality (Dudewicz and Mishra 1988) we can show that for any $\delta > 0$,

$$P\Big(|\hat{\theta} - E(\hat{\theta})| \geq \delta\Big) \leq \frac{\mathrm{Var}_\theta(\hat{\theta})}{\delta^2}.$$

Hence, if both, $\mathrm{Bias}(\hat{\theta}, \theta)^2 \to 0$ and $\mathrm{Var}_\theta(\hat{\theta}) \to 0$ for increasing sample size n, it follows that $P(|\hat{\theta} - \theta| \geq \delta) \to 0$. In other words, the estimate is getting closer to and more peaked around the true unknown parameter θ as n grows. This property is usually called weak consistency. There are many other consistency properties in use in statistics, see e.g. Dudewicz and Mishra (1988) or Lehmann and Casella (1998). For the purpose of this book, however, we will stick with the (MSE) consistency defined above. Consistency is thereby an essential property which any reasonable estimate should fulfil. It simply means that the larger the data, the more precise the conclusions are that can be drawn about the parameter of interest.

3.3.3 Cramer-Rao Bound

As the MSE is a key property of an estimator, the intention should therefore be to have the MSE as small as possible. This poses the question whether there is a lower threshold for the MSE. That is to say, for a given sample size n, the mean squared error cannot be smaller than a lower limit. Logically, it follows that some error will be introduced by the sampling process that cannot be overcome, no matter how good our estimator is. In fact, such a lower limit exists and is known as the **Cramer-Rao bound**. This bound holds for a wide class of distributions which are **Fisher-regular**.

Definition 3.12 The distribution $f(y;\theta)$ is **Fisher-regular** if the following properties hold:

1. The support of Y is not dependent upon θ, i.e. the set $\{y : f(y;\theta) > 0\}$ does not depend on θ.
2. The possible parameter space Θ is open, i.e. if θ is univariate, it has the form $\Theta = (a, b)$ with $a < b$.
3. The probability function $f(y;\theta)$ can be differentiated twice with respect to θ.
4. Integration and differentiation are exchangeable, i.e.

$$\int \frac{\partial}{\partial\theta} f(y;\theta)dy = \frac{\partial}{\partial\theta}\int f(y;\theta)dy.$$

It should be noted that Fisher-regularity is not a strong assumption. Still, it can be violated. One standard counterexample is the uniform distribution on $[0, \theta]$, i.e.

$$f(y;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } y \in [0,\theta] \\ 0 & \text{otherwise} \end{cases}.$$

Here, requirement 1 is violated as the support of Y depends on θ.

We will deal with Fisher-regular distributions in depth in Chap. 4. For now we use the idea of Fisher-regularity to find the lower bound for the mean squared error. This requires a further definition, namely that of the Fisher information. The Fisher information can be understood as the information available about θ in the data $y_1, \ldots, y_n$. It plays a central role in statistics and will also be treated in depth in Chap. 4. We will briefly provide a definition here as we need the Fisher information to describe the lower bound of the MSE.

Definition 3.13 Assume that $f(y;\theta)$ is Fisher-regular. We define with

$$I(\theta) = E\Big[-\frac{\partial^2 \log f(Y;\theta)}{\partial\theta\partial\theta}\Big] \tag{3.3.2}$$

the **Fisher information**. The component

$$J(\theta) = -\frac{\partial^2 \log f(y;\theta)}{\partial\theta\partial\theta} \tag{3.3.3}$$

is also known as **observed Fisher information**. If θ is multivariate, we define $I(\theta) = E\left[\frac{\partial^2 \log f(Y;\theta)}{\partial\theta\partial\theta^T}\right]$ as the Fisher information matrix and similarly $J(\theta)$ as the observed Fisher information matrix.

We will show in Chap. 4 that the Fisher-matrix can equivalently be written as

$$I(\theta) = E\Big[\left(\frac{\partial \log f(Y;\theta)}{\partial\theta}\right)\left(\frac{\partial \log f(Y;\theta)}{\partial\theta}\right)^T\Big]. \tag{3.3.4}$$

With these definitions in place, we now have all the tools we need to derive the Cramer-Rao bound.

Property 3.2 Let $\hat{\theta} = t(Y_1, \ldots, Y_n)$ be an estimate for θ, $Y_1, \ldots, Y_n$ *i.i.d.* and $Y_i \sim f(y; \theta)$ is drawn from a Fisher-regular distribution. The mean squared error of $t(.)$ is always larger than the lower **Cramer-Rao bound**

$$MSE(\hat{\theta}, \theta) \geq \text{Bias}^2(\hat{\theta}, \theta) + \frac{\left(1 + \frac{\partial \text{Bias}(\hat{\theta}, \theta)}{\partial \theta}\right)^2}{I(\theta)}.$$

In particular, if the estimate is unbiased one has

$$MSE(\hat{\theta}, \theta) \geq \frac{1}{I(\theta)}. \tag{3.3.5}$$

The Cramer-Rao bound demonstrates why the Fisher information plays a central role. In fact, if we are able to find an unbiased estimate with variance equal to the inverse Fisher information, then we have found one of the best possible estimates. We will show in the next chapter that such a class of estimators can be calculated with the Maximum Likelihood approach.

Example 2 Let us demonstrate the Cramer-Rao bound with a simple example. Assume we have normally distributed random variables $Y_i \sim N(\mu, \sigma^2)$, which are *i.i.d.* We are interested in estimating the mean and propose the estimate

$$t(y) = \sum_{i=1}^{n} w_i y_i$$

for some weights w_i. To obtain unbiasedness it is easy to see that we need to postulate $\sum_{i=1}^{n} w_i = 1$, because

$$E(t(Y)) = \sum_{i=1}^{n} w_i \mu.$$

We now question how to choose the weights such that the variance of $t(y)$ is minimised. Note that

$$Var(t(Y)) = \sum_{i=1}^{n} w_i^2 \sigma^2.$$

If we set $w_i = 1/n + d_i$ we get $\sum_{i=1}^n d_i = 0$ and hence

$$Var(t(Y)) = \sum_{i=1}^{n}(\frac{1}{n} + d_i)^2\sigma^2 = \sum_{i=1}^{n}(\frac{1}{n^2} + d_i^2)\sigma^2$$
$$= \sigma^2/n + \sum_{i=1}^{n} d_i^2\sigma^2.$$

Because $\sum_{i=1}^n d_i^2 \geq 0$, unless $d_i = 0$, we see that the arithmetic mean, which results when we set the weights w_i equal to $1/n$, in fact has the smallest variance, just as the Cramer-Rao bound states. ▷

Proof Here we prove the Cramer-Rao bound. We emphasise that the proof requires properties which will be derived in the next chapter. It is given here for completeness and may be skipped on a first reading. Note that for unbiased estimates we have

$$\theta = E(\hat{\theta}) = \int t(y) f(y;\theta) dy.$$

Differentiating both sides with respect to θ yields

$$1 = \int t(y) \frac{\partial f(y;\theta)}{\partial \theta} dy$$
$$= \int t(y) \frac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta) dy$$
$$= \int t(y) s(y;\theta) f(y;\theta) dy,$$

where $s(\theta; y) = \partial \log f(y;\theta)/\partial\theta$. We will later call $s(\theta; y)$ the score function and prove in Chap. 4 that

$$E\Big(s(\theta; y)\Big) = \int s(\theta; y) f(y;\theta) dy = 0.$$

This implies that

$$1 = \int t(y) s(\theta; y) f(y;\theta) dy = \int \Big(t(y) - \theta\Big)\Big(s(\theta; y) - 0\Big) f(y;\theta) dy$$
$$= \mathrm{Cov}\Big(t(Y); s(\theta; Y)\Big).$$

With the **Cauchy-Schwarz** inequality one obtains

$$1 = \mathrm{Cov}\Big(t(Y); s(\theta; Y)\Big) \leq \sqrt{\mathrm{Var}_\theta(t(Y))}\sqrt{\mathrm{Var}_\theta(s(\theta; Y))}.$$

Note that

$$\begin{aligned}\mathrm{Var}_\theta\Big(s(\theta;Y)\Big) &= \int \Big(s(\theta;y)-0\Big)^2 f(y,\theta)dy \\ &= \int s^2(\theta;y) f(y;\theta)dy \\ &= \int \frac{\partial \log f(y;\theta)}{\partial\theta}\frac{\partial \log f(y;\theta)}{\partial\theta} f(y;\theta)dy \\ &= I(\theta)\end{aligned}$$

using (3.3.4). As for unbiased estimates we have the equality $\mathrm{Var}_\theta(t(Y)) = MSE(t(Y),\theta)$, we obtain

$$\begin{aligned}1 &\le \sqrt{\mathrm{Var}_\theta(t(Y))}\sqrt{\mathrm{Var}_\theta(s(\theta;Y))} \\ \frac{1}{\sqrt{\mathrm{Var}_\theta(s(\theta;Y))}} &\le \sqrt{\mathrm{Var}_\theta(t(Y))} \\ \frac{1}{I(\theta)} &\le MSE(t(y),\theta).\end{aligned}$$

□

3.4 Interval Estimates

3.4.1 Confidence Intervals

We have now defined and discussed a few properties relevant to the evaluation of point estimates. Our estimate $\hat{\theta} = t(Y_1,\ldots,Y_n)$ itself, however, is just a single value and no information is given yet about how close our estimate is to the true parameter θ. Even though we know that for consistent estimates we have

$$P\left(\mid \hat{\theta}-\theta \mid > \delta\right) \to 0 \tag{3.4.1}$$

for $n \to \infty$ and for any $\delta > 0$. While these results are useful and informative, they do not tell us anything about what occurs in real life. In real life, despite our best efforts, we can never take infinitely sized samples. As a consequence, some knowledge about the range of likely values would definitely be of value. Let us therefore move to constructing **interval estimates** instead of single point estimates. Such intervals will show us how close our estimate is to the true value. Interval estimates are defined by a left boundary $t_l(Y_1,\ldots,Y_n)$ and a right boundary $t_r(Y_1,\ldots,Y_n)$, which give an interval of $\big[t_l(Y_1,\ldots,Y_n), t_r(Y_1,\ldots,Y_n)\big]$. Clearly,

to achieve a useful interval $t_l(Y_1, \ldots, Y_n) < t_r(Y_1, \ldots, Y_n)$ for all values of $Y_1, \ldots, Y_n$. These statistics can be defined as follows:

Definition 3.14 The interval $CI = \left[t_l(Y_1, \ldots, Y_n), t_r(Y_1, \ldots, Y_n)\right]$ is called a **confidence interval** for θ with confidence level $1 - \alpha$ if

$$P_\theta\Big(t_l(Y_1, \ldots, Y_n) \leq \theta \leq t_r(Y_1, \ldots, Y_n)\Big) \geq 1 - \alpha \tag{3.4.2}$$

for all θ. The value $(1 - \alpha)$ is called the confidence level and α is chosen as a small value, e.g. $\alpha = 0.01$ or $\alpha = 0.05$.

The probability statement (3.4.2) can be reformulated as

$$P_\theta\Big(\theta \in [t_l(Y_1, \ldots, Y_n), t_r(Y_1, \ldots, Y_n)]\Big) \geq 1 - \alpha.$$

It is important to bear in mind that $t_l(Y_1, \ldots, Y_n)$ and $t_r(Y_1, \ldots, Y_n)$ are both random variables. However, it is also clear that defining $t_l(Y_1, \ldots, Y_n) = -\infty$ and $t_r(Y_1, \ldots, Y_n) = \infty$ gives a $1-\alpha$ confidence interval, as $P(-\infty \leq \theta \leq \infty) = 1$ for any $\theta \in \mathbb{R}$. This is, of course, a useless interval. However, it demonstrates that our intention should be to choose the smallest possible interval and that means choosing "$\geq 1 - \alpha$" in (3.4.2) in fact means "$= 1 - \alpha$".

The construction of such a confidence interval is easy if a pivotal statistic exists.

Definition 3.15 A quantity $g\big(Y_1, \ldots, Y_n; \theta\big)$ is called a **pivotal statistic** if its distribution does not depend on θ. The distribution of $g(Y_1, \ldots, Y_n; \theta)$ is also called a pivotal distribution.

Exact pivotal quantities are rare. However, approximate pivotal distributions are quite common due to the central limit theorem. In fact, if the sample size n is large, the estimate $\hat{\theta} = t(Y_1, \ldots, Y_n)$ in many cases follows approximately a normal distribution. We denote this as

$$\hat{\theta} = t(Y_1, \ldots, Y_n) \overset{a}{\sim} N\left(\theta, Var(\hat{\theta})\right), \tag{3.4.3}$$

where the letter a in the formula above stands for asymptotic. We then can construct an approximate pivotal statistic with:

$$g\left(t((Y_1, \ldots, Y_n); \theta)\right) = \frac{t(Y_1, \ldots, Y_n) - \theta}{\sqrt{Var(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \overset{a}{\sim} N(0, 1), \tag{3.4.4}$$

which asymptotically follows a standard normal $N(0,1)$ distribution. As $N(0,1)$ does not depend on θ, the statistic is pivotal. The pivotal distribution can now be used to construct a confidence interval in the following way. With (3.4.4) we have

$$1-\alpha \approx P\left(z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \leq z_{1-\alpha/2}\right)$$

$$\Leftrightarrow 1-\alpha \approx P\left(\hat{\theta}+z_{\alpha/2}\sqrt{Var(\hat{\theta})} \leq \theta \leq \hat{\theta}+z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}\right),$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of a $N(0,1)$ distribution and accordingly, $z_{1-\alpha/2}$ the $1-\alpha/2$ quantile. Because $z_{\alpha/2} = -z_{1-\alpha/2}$, we obtain the confidence interval as

$$CI = \left[\underbrace{\hat{\theta}-z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}}_{t_l(Y_1,\ldots,Y_n)}, \underbrace{\hat{\theta}+z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}}_{t_r(Y_1,\ldots,Y_n)}\right]. \tag{3.4.5}$$

One should note that $Var(\hat{\theta})$ may itself depend on θ, the unknown parameter. It may also depend on some other parameters, which are unknown. In the first case it is reasonable to replace θ with its estimate $\hat{\theta}$. If the variance $Var(\hat{\theta})$ depends on other parameters, one needs to estimate them. In fact, by doing so we estimate the variance $Var(\hat{\theta})$ and denote this with $\widehat{Var(\hat{\theta})}$. If the estimator $\widehat{Var(\hat{\theta})}$ is consistent, then the confidence interval is still asymptotically valid and (3.4.5) now changes to

$$CI = \left[\underbrace{\hat{\theta}-z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\theta})}}}_{t_l(Y_1,\ldots,Y_n)}, \underbrace{\hat{\theta}+z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{\theta})}}}_{t_r(Y_1,\ldots,Y_n)}\right]. \tag{3.4.6}$$

The construction of a confidence interval based on asymptotic normality (3.4.3) is easy and convenient and works in a large number of practical applications. It fails, however, if the observed data are extreme in the sense of the underlying distribution. A typical and quite common situation occurs if we are interested in rare events. Assume, for instance, we are interested in the failure of a technical component and the available database consists of n observations (e.g. trials) out of which in Y cases a technical component failed. Assuming that Y is the realisation of the binomial distribution

$$Y \sim B(n,\pi),$$

we are interested in a confidence interval for the failure probability π. If π is small then Y is small and arguments relying on the asymptotic normality are questionable. In this case, we need to look more closely into the distributional model. Bear in

mind that if we observe $y = 0$, for instance, it does not mean that $\pi = 0$. Technical systems do fail and just because we have not observed a failure yet does not mean that we can be certain that we will never observe a failure in the future. We therefore need to construct a confidence interval for π which also includes values with $\pi > 0$. In this case, we can use (3.4.2) and restructure it by allowing the error probability α to be allocated in equal size on both sides of the interval. That is, we construct the confidence interval $\left[t_l(Y), t_r(Y)\right]$ with $t_l(Y) < t_r(Y)$ as

$$P\Big(t_l(Y) > \pi; \pi\Big) \leq \alpha/2 \text{ and } P\Big(t_r(Y) < \pi; \pi\Big) \leq \alpha/2,$$

such that

$$P\Big(t_l(Y) \leq \pi \leq t_r(Y); \pi\Big) \geq 1 - \alpha.$$

Let us first look at the left-hand side. Note that $Y \in \{0, 1, 2, \ldots, n\}$ and thus $t_l : \{0, \ldots, n\} \to [0, 1]$. If we could "invert" t_l we could derive the following:

$$\begin{aligned} & P\Big(t_l(Y) > \pi; \pi\Big) \leq \alpha/2 \qquad (3.4.7) \\ \Leftrightarrow & P\Big(Y > t_l^{-1}(\pi); \pi\Big) \leq \alpha/2 \\ \Leftrightarrow & P\Big(Y \leq t_l^{-1}(\pi); \pi\Big) \geq 1 - \alpha/2. \end{aligned}$$

Let $t_l^{-1}(\pi) \in \{0, \ldots, n\}$, where clearly for value $t_l^{-1}(\pi) = n$ the inequality is fulfilled for all values of π. We therefore look at values $k \in 0, \ldots, n-1$ and consider

$$P(Y \leq k; \pi)$$

as a function of π. For every k there exists a unique left sided π_{lk}, such that

$$P\left(Y \leq k; \pi_{lk}\right) = \alpha/2,$$

where index l refers to the left side. Setting $\pi_{ln} = 1$, we have thereby defined the inverse function $t_l^{-1} : \{\pi_{l0}, \ldots, \pi_{ln}\} \to \{0, \ldots, n\}$ that fulfils (3.4.7). Accordingly, for $Y \in \{0, \ldots, n\}$, we can now set

$$t_l(y) = \pi_{ly}.$$

In the same way, we can define the right-hand statistic $t_r(y) = \pi_{ry}$. This construction yields confidence intervals $[t_l(y), t_r(y)]$ which are shown for various values of n and y/n in Fig. 3.1. Note that for $y \equiv 0$ and hence $y/n = 0$, we always

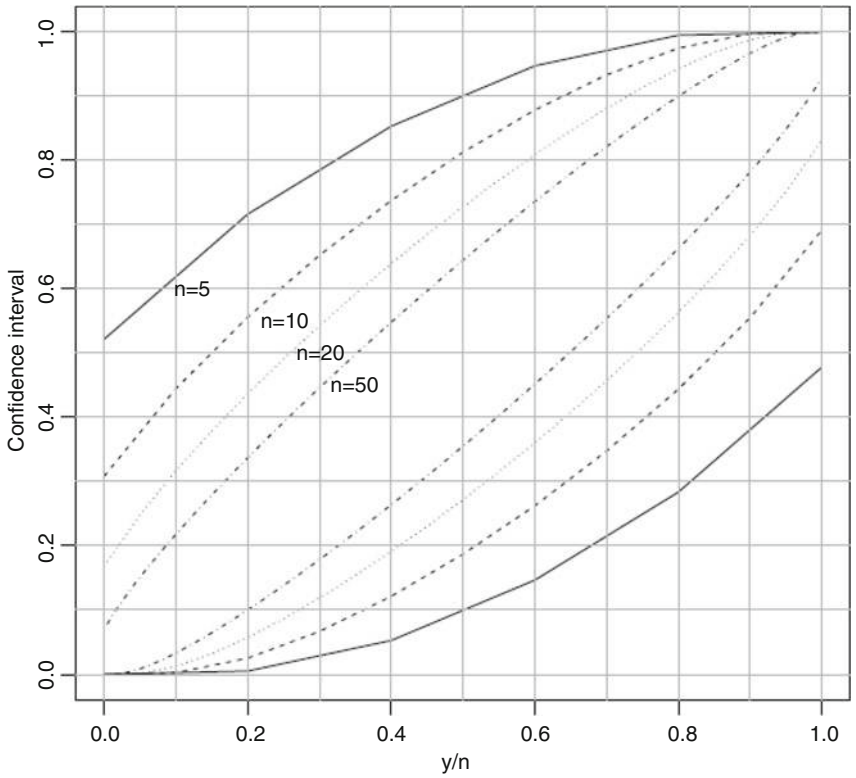

Fig. 3.1 Confidence intervals for π for binomial distribution

obtain $t_l(0) = 0$. For instance, if $n = 5$ and $y = 0$ then $t_r(0)$ is even larger than 0.5. The same holds symmetrically for $t_l(n)$.
This construction for confidence intervals was devised by Clopper and Pearson (1934) and is therefore sometimes called the Clopper-Pearson interval. Another name for it is the exact binomial confidence interval, to distinguish it from the asymptotic confidence interval, which is only adequate for large n.

Example 3 When the nuclear disaster happened in Fukushima in 2011, it was the second serious nuclear accident after Chernobyl in 1986. These are just two events, but it raised doubts as to whether the safety of nuclear power plants corresponds to the stated risk of one accident per 250,000 years. In Kauermann and Kuechenhoff (2011) the question is tackled with statistical tools. We give the main reasoning and conclusion here again. Setting the number of nuclear power plants to 442 (which was the number of actively running nuclear power plants in 2011), we consider this number as valid for the years 1981–2011 (which is a good proxy, as there is little variation in this number over the 30 years). We consider the binary events $Y_{it} \in \{0, 1\}$, where Y_{it} indicates whether the i-th power plant had a serious disaster

in year t. Assuming independence between the years and between the power plants we have

$$Y = \sum_i \sum_t Y_{it}$$

as Binomial distributed variable with $n = 30 \cdot 442$ and π being the yearly risk for a nuclear disaster of a power plant. We have two accidents observed, so our data to fit the model is $Y = 2$. The Maximum Likelihood estimate results as

$$\hat{\pi} = \frac{2}{30 \cdot 442} \approx \frac{1.5}{10000} \approx \frac{1}{6607}.$$

In other words, given the data we estimate the probability of an accident in the order of every 6667 years for each reactor. This is apparently much larger than the one every 250,000 years, which is the reported safety risk. More important, however, is to assess the confidence in our estimate. This can be done with the methods just described leading to the following exact confidence interval for π:

$$\left[\frac{1}{54000}, \frac{1}{1800}\right].$$

We see that the safety level $1/250{,}000$ is not within the confidence interval meaning that the fact that two nuclear accidents have been observed indicates that the proposed safety level is not valid. We do not question the model here nor do we want to interpret the result in depth. The example is given to demonstrate that statistical reasoning can also be used if events are rare. ▷

In practice, confidence intervals are a very important tool to handle uncertainty of estimators. However, the correct interpretation of confidence intervals is very important. Given a concrete sample $y_1, \ldots, y_n$ one can estimate a corresponding confidence interval with:

$$\big[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)\big].$$

Note that if $y_1, \ldots, y_n$ are realised values, then the interval boundaries $t_l(y_1, \ldots, y_n)$ and $t_r(y_1, \ldots, y_n)$ are realised values as well—they are concrete numbers. This implies that probability statements like "What is the probability that parameter θ lies within the interval $\big[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)\big]$?" are in principle not valid using the probability model for Y. As the values $y_1, \ldots, y_n$ are observed, there is no randomness with respect to random variables. As a consequence, our only technical answer would be that we simply do not know whether θ lies in the interval or not. This is not a very satisfactory conclusion. A way out of this problem could, of course, be to use a Bayesian viewpoint and formulate our uncertainty about θ with a probability function. We will demonstrate this later. For now we still want to consider θ as a fixed but unknown parameter

for which a (realised) confidence interval is given. Even though we cannot quantify with a formal probability statement whether θ lies in the interval, we may formulate this with subjective probabilities using DeFinetti's approach (see Definition 2.3) leading to a confidence statement. In fact, our *confidence* that θ lies in the interval $\left[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)\right]$ can be quantified to be $1-\alpha$. That is also to say that in the long run, if we repeated the data collection process, we expect in $(1-\alpha) \cdot 100$ % of the cases that the parameter lies in the interval.

3.4.2 Credibility Interval

Let us conclude this chapter by exploring confidence intervals from a Bayesian perspective. Our knowledge about the parameter θ is given by the posterior distribution $f_\theta(\vartheta \mid y_1, \ldots, y_n)$. The posterior distribution can directly be used to construct an interval $\left[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)\right]$ such that

$$P_\theta\left(\vartheta \in \left[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)\right] | (y_1, \ldots, y_n)\right) = \int_{t_l(y_1,\ldots,y_n)}^{t_r(y_1,\ldots,y_n)} f_\theta(\vartheta \mid y_1, \ldots, y_n) d\vartheta \geq 1-\alpha.$$

In the Bayesian terminology, such an interval is called a **credibility interval**. A natural choice is to set $t_l()$ and $t_r()$ such that

$$\int_{-\infty}^{t_l(y_1,\ldots,y_n)} f_\theta(\vartheta \mid y_1, \ldots, y_n) d\vartheta = \int_{t_r(y_1,\ldots,y_n)}^{\infty} f_\theta(\vartheta \mid y_1, \ldots, y_n) d\vartheta = \frac{\alpha}{2},$$

that is, we cut off a probability mass of $\alpha/2$ on the left and right of the posterior probability. This choice is not optimal, as it may occur that for $\vartheta_1 \notin [t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)]$ and for $\vartheta_2 \in [t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)]$ one has

$$f_\theta(\vartheta_1 \mid y_1, \ldots, y_n) > f_\theta(\vartheta_2 \mid y_1, \ldots, y_n).$$

Hence, the density may be larger for values outside of the credibility region compared to values within the credibility region. This drawback can be avoided by using a **highest posteriori density** credibility interval or, in short, the highest density interval

$$HDI(y_1, \ldots, y_n) = \{\theta;\, f_\theta(\theta \mid y_1, \ldots, y_n) \geq c\},$$

where c is chosen such that

$$\int_{\vartheta \in HDI(y_1,\ldots,y_n)} f_\theta\big(\vartheta \mid (y_1,\ldots,y_n)\big)d\vartheta = 1-\alpha.$$

That is, we choose an interval where values are greater than a threshold density c, such that the region integrates to $1-\alpha$. To demonstrate the idea let us use the binomial example from above again. Assume that $Y_1, .., Y_n \in \{0, 1\}$ are independent Bernoulli variables, such that $Y = \sum_{i=1}^{n} Y_i$ is binomial with parameter n and π. We assume a flat prior for π that the posterior is

$$f_\pi(\pi \mid y) \propto \pi^y(1-\pi)^{n-y},$$

where $\propto$ means "is proportional to". As will be shown in Chap. 5, this gives a beta distribution, such that

$$\pi \mid y \sim Beta(1+y, 1+n-y).$$

The resulting highest density credibility intervals are shown in Fig. 3.2 for the setting $n = 5$ and $y = 0$ (left plot) and $n = 5$ and $y = 2$ (right plot). The generic credibility intervals for different values of n and y/n are shown in Fig. 3.3.
Clearly, the plot resembles the one in Fig. 3.1, although they are built upon different reasoning. While Fig. 3.1 gives what is called exact confidence intervals, Fig. 3.3 relies on Bayesian reasoning and provides credibility intervals. Interestingly, the coincidence of the confidence intervals and credibility intervals occurs in many places and we will see it later in the book.

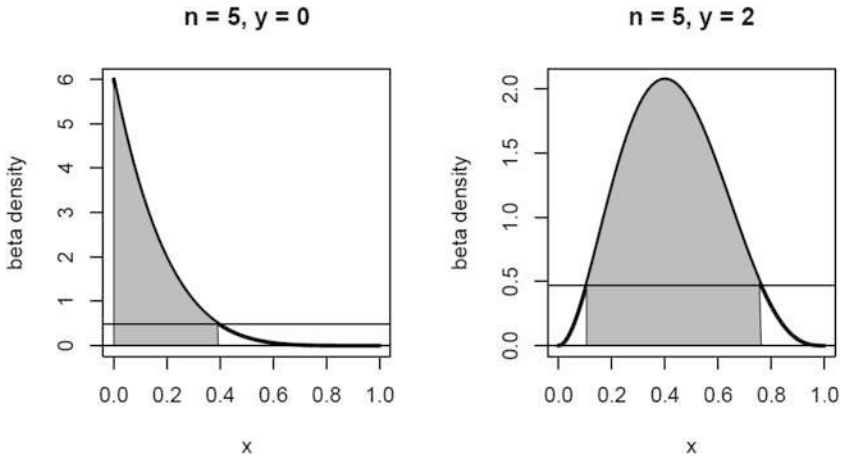

Fig. 3.2 Highest density credibility interval for binomial data with $n = 5$ and $y = 0$ (left) and $n = 5$ and $y = 2$ (right)

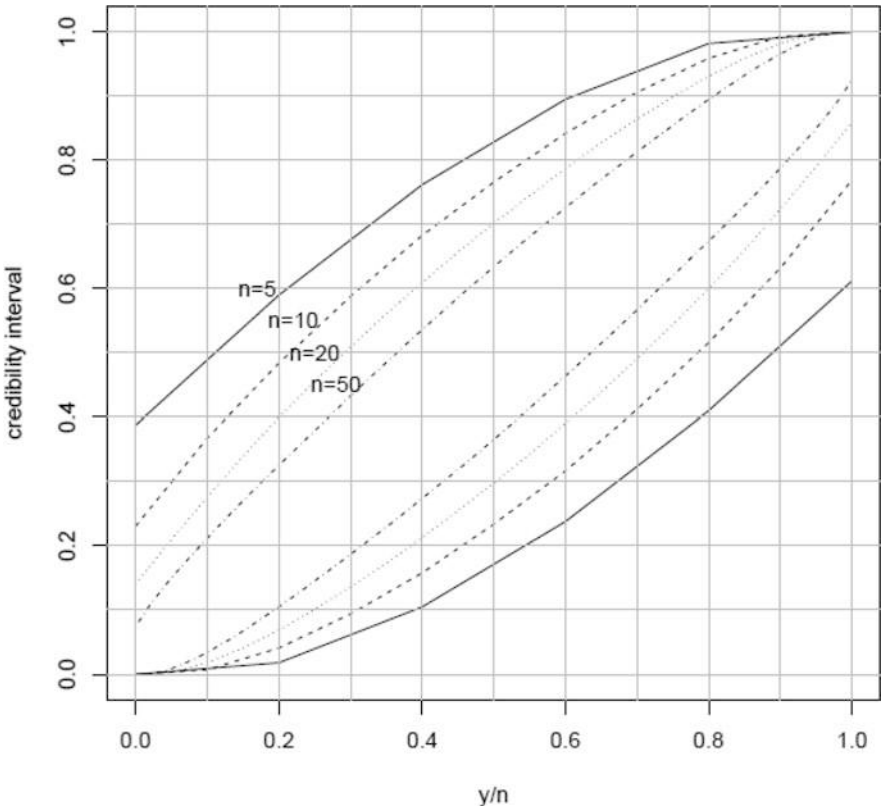

Fig. 3.3 Highest posterior density credibility intervals (given a flat prior) for different values of n and y/n

Example 4 Let us illustrate the important role of confidence and credibility intervals in some real applications. In surveys on voting behaviour, randomly sampled people are asked about their potential vote. In order to determine the share of population that votes for party A, we can define the random variable:

$$Y_i^{(A)} \text{ with } Y_i^{(A)} = \begin{cases} 1 & \text{if the i-th person votes for A} \\ 0 & \text{otherwise.} \end{cases}$$

If we define the unknown true share as π_A, the distribution of Y_i is $B(1, \pi_A)$ assuming a simple random sample from a large population. Furthermore, independence of the Y_i can be assumed and $Y_1, \ldots, Y_n$ are $i.i.d.$. Then $\sum_{i=1}^{n} Y_i \sim B(n, \pi_A)$ and the estimator is $\hat{\pi}_A = \frac{1}{n} \sum_{i=1}^{n} Y_i$, which is the relative frequency of voters for party A in our sample. For the construction of a confidence interval, we can

Table 3.1 95% Confidence intervals for the proportion of voting intention for different parties based on a survey conducted by Infratest dimap on May, 2nd, 2019

Party	No. of votes in sample	$\hat{p}$ in %	Asymptotic CI in %	Exact CI
CDU/CSU	421	28	[25.7; 30.2]	[25.7;30.0]
SPD	271	18	[16.1; 19.9]	[16.1; 20.0]
AFD	181	12	[10.4; 13.7]	[10.4; 13.8]
FDP	120	8	[6.6; 9.3]	[6.7; 9.5]
Linke	135	9	[7.5; 10.4]	[7.6; 10.5]
Gruene	301	20	[18; 22]	[18.0; 22.1]

use (3.4.3) for a large sample size n applying the asymptotic normality of $\hat{\pi}_A$. As $Var(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n}$, the interval is

$$\left[\hat{\pi}_A - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n}}, \hat{\pi}_A + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n}}\right].$$

For a survey conducted by Infratest dimap on May 2nd, 2019, $n = 1505$ were sampled and asked for their voting intention. The results including asymptotic and Clopper-Pearson intervals are given in Table 3.1.

Note that the intervals are relatively large in spite of the rather high sample size. ▷

Example 5 (Validation of Machine Learning Algorithms) Let us examine how confidence intervals can be calculated for machine learning methods. In an analysis of professional soccer players, Rossi et al. (2018) developed an algorithm to predict injuries of players with GPS training data. The authors use a test sample to check the predictive power of their algorithm. In their approach, 9 out of 14 injuries could be predicted, while in the second best approach 6 out of 14 injuries could be successfully predicted. The rate is 64% (43%). The respective confidence intervals for the rate of successful prediction are [0.39, 0.83] (Method 1) and [0.17, 0.71] (second method). The intervals show a high uncertainty in the prediction performance of the model, which is due to the small sample size of the validation data. ▷

3.5 Exercises

Exercise 1

Let $Y_i \in \{0, 1\}$ be independent Bernoulli variables, such that $Y = \sum_{i=1}^{n} Y_i \sim B(n, \pi)$. Given the data y we want to estimate π.

1. Derive the ML estimate and the method of moment estimate.

2. We now look at estimates of the form

$$t(y) = \frac{y + a}{a + b + n},$$

where a and b need to be chosen appropriately. Derive the MSE(t, π).
3. Taking the squared risk $\mathcal{L}(t, \pi) = (t - \pi)^2$, we obtain (with differentiation) the maximum risk given a and b. Plot the risk for different values of a and b, including $a = 0$ and $b = 0$. Given your results choose the minimax estimate.

Exercise 2 (Use R Statistical Software)
We consider a sample $Y_1, \ldots, Y_n$ from a uniform distribution on the interval $[0, \theta]$ with density function $f(y|\theta) = 1/\theta$ for $y \in [0, \theta]$ and $f(y|\theta) = 0$ otherwise. We estimate θ with the maximum value in the sample, i.e. $\hat{\theta} = Y_{(n)}$.

1. Illustrate why $\hat{\theta}$ is a biased estimate.
2. Show that $\hat{\theta}$ is the Maximum Likelihood estimate.
3. Show that $\theta^* = 2\bar{Y} = \frac{2}{n}\sum_{i=1}^{n} Y_i$ is an unbiased estimate for θ.
4. Check your results empirically by generating uniform random numbers in the interval $[0, 5]$. Try different sample sizes n: $n = 5$, $n = 10$, $n = 50$, $n = 100$, $n = 500$ and discuss your findings.

Exercise 3 (Use R Statistical Software)
In the data file `injured.csv`, the weekly numbers of pedestrians severely injured in traffic in a city are recorded for 10 years ($n = 520$). We assume that the numbers are independent realisations of a Poisson distributed random variable Y with constant intensity parameter λ.

1. Derive and calculate the Maximum Likelihood estimate $\hat{\lambda}_{\text{ML}}$ of λ given the available data.
2. Calculate the method of moments estimate for λ.
3. Calculate a 95% confidence interval, assuming asymptotic normality of the Maximum Likelihood estimate.