SpringerBriefs in Statistics

A Concise Guide to Statistics

Bearbeitet von
Hans-Michael Kaltenbach

1. Auflage 2011. Taschenbuch. XIII, 111 S. Paperback
ISBN 978 3 642 23501 6
Format (B x L): 15,5 x 23,5 cm
Gewicht: 205 g

Chapter 2
Estimation

Abstract Estimation is the inference of properties of a distribution from an observed random sample. Estimators can be derived by various approaches. To quantify the quality of a given estimate, confidence intervals can be computed; the bootstrap is a general purpose method for this. Vulnerability of some estimators to sample contaminations leads to robust alternatives.

Keywords Maximum-likelihood · Confidence interval · Bootstrap

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay"

Sherlock Holmes

2.1 Introduction

We assume that n independent and identically distributed random samples $X_1, \ldots, X_n$ are drawn, whose realizations form an observation $x_1, \ldots, x_n$. Our goal is to infer one or more parameters θ of the distribution of the X_i. For this, we construct an *estimator* $\hat{\theta}_n$ by finding a function g, such that

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

is a "good guess" of the true value θ. Since $\hat{\theta}_n$ depends on the data, it is a random variable. Finding its distribution allows us to compute confidence intervals that quantify how likely it is that the true value θ is close to the estimate $\hat{\theta}_n$.

Example 10 Let us revisit the problem of sequence matching from Example 8 (p. 19) we already know that the number of matches in two random sequences is a random variable $M \sim \text{Binom}(n, p)$, but do not know the probability p, and want to infer it from given data. For this, let us assume we are given two sequences of length n each, and record the matches $m_1, \ldots, m_n$, where again $m_i = 1$ if position i is a match, and $m_i = 0$ if it is a mismatch, as well as the total number of matches $m = m_1 + \cdots + m_n$.

H.-M. Kaltenbach, *A Concise Guide to Statistics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-23502-3_2, © Hans-Michael Kaltenbach 2012

For any fixed value of p, we can compute the probability to see exactly the observed matches $m_1, \ldots, m_n$. The main new idea is to consider this probability as a function of the parameter p for given observations. This function is known as the *likelihood function*

$$L_n(p) = \mathbb{P}(M_1 = m_1, \ldots, M_n = m_n) = \prod_{i=1}^{n} \mathbb{P}(M_i = m_i) = p^m (1-p)^{n-m};$$

note that we can only write the joint probability as a product because we assume the positions (and therefore the individual matches) to be independent. We then seek the value $\hat{p}_n$ that maximizes this likelihood and gives the highest probability for the observed outcome. In this sense, it therefore "best" explains the observed data. Maximizing the likelihood is straightforward in this case: we differentiate the likelihood function with respect to p and find its roots by solving the equation

$$\frac{\partial L_n(p)}{\partial p} = 0.$$

Taking the derivative of $L_n(p)$ requires repeated application of the product-rule. It is therefore more convenient to use the *log-likelihood* for the maximization, given by

$$\ell_n(p) = \log L_n(p) = \sum_{i=1}^{n} \log \mathbb{P}(M_i = m_i) = m \log(p) + (n-m) \log(1-p).$$

Maximizing either $L_n(p)$ or $\ell_n(p)$ yields the exact same result, as the logarithm is a strictly increasing function, but we can conveniently differentiate each summand individually in the log-likelihood. In our case,

$$0 = \frac{\partial \ell_n(p)}{\partial p} = m \frac{1}{p} + (n-m) \left(-\frac{1}{1-p} \right),$$

which gives

$$\frac{m}{p} = \frac{n-m}{1-p} \iff p = \frac{m}{n}.$$

Thus, the desired estimate of the parameter value p is $\hat{p}_n = m/n$, the proportion of matches in the sequence.

It is important to understand the fundamental difference between the parameter p and its estimate $\hat{p}_n$: the parameter p is a fixed number, relating to the model describing the experiment. It is independent of the particular outcome m of the experiment. In contrast, its estimate $\hat{p}_n$ is a function of the data and takes different values for different samples. For studying general properties of this estimator, we will therefore consider $\hat{p}_n$ as the random variable M/n rather than its realization m/n. It then has

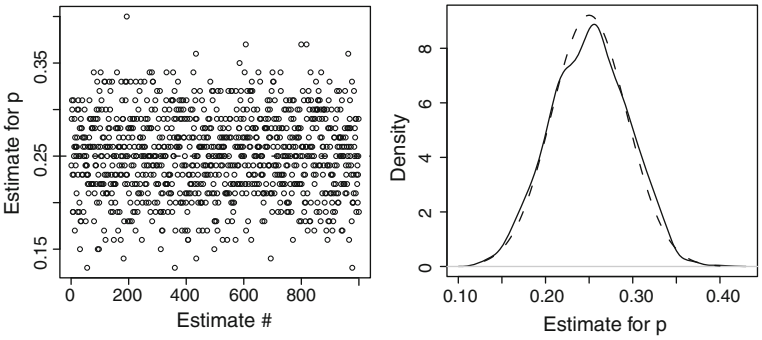

Fig. 2.1 *Left*: values of 1000 repetitions of estimating the matching probability of Binom(100,0.25) experiments. *Right*: density of estimate (*solid*) and Norm(0.25,0.001875) approximation (*dashed*)

a distribution and if we were to repeat the same experiment over and over, p would always be the same, but the estimate would yield a different realization of $\hat{p}_n$ each time.

Because an estimator is a random variable, it is helpful to either compute its entire distribution or some of its moments. For our example, we can easily work out the expectation and the variance of our estimator:

$$\mathbb{E}(\hat{p}_n) = \mathbb{E}\left(\frac{M}{n}\right) = \frac{1}{n}\mathbb{E}(M) = \frac{np}{n} = p,$$

which shows that the estimator is *unbiased* and thus—on average—yields the correct value for the parameter, and

$$\mathrm{Var}(\hat{p}_n) = \frac{1}{n^2}\mathrm{Var}(M) = \frac{p(1-p)}{n}.$$

The variance of the estimator decreases with increasing sample size n, which is intuitively plausible: by using more data, we are more confident about the correct value of the parameter p and expect the estimator to get closer to the true value with high probability. We also get a lower variance of the estimate if the variance of the data is smaller.

The estimator of a true parameter value $p = 0.25$ is studied in Fig. 2.1 on 1000 pairs of unrelated sequences of length 100. On the left, the values of $\hat{p}_n$ are given for each such pair. Most estimates lie reasonably close to the true value, but there are also some larger deviations. On the right, the empirical density function of $\hat{p}_n$ is given (solid line) together with a normal density with the same expectation and variance (dashed line). The values of the estimate closely follow the normal distribution and the mean nicely corresponds to the correct parameter value p.

2.2 Constructing Estimators

To derive an estimator for a parameter θ, we need to construct the function $g(X_1, \ldots, X_n)$. There are multiple methods to do this and we will discuss the maximum-likelihood and the least-squares approach in more depth. Both methods rely on finding the parameter value that "best" explains the observed data, but there definition of "best" is different and requires finding the minimum or maximum of a certain function. A third approach, the minimax principle, will be presented in a more general framework in Sect. 2.5.

2.2.1 Maximum-Likelihood

To apply maximum-likelihood estimation in the general case, we need to specify a *family of distributions* that is parametrized by θ such that each value of θ selects one particular distribution from this family. In the previous example, this family was the set of all binomial distributions with fixed n, where each value for p selects one particular member of this family.

Here, we consider the density function $f(x; \theta)$, describing the family of distributions, and aim at estimating the parameter θ. The *likelihood function* for this parameter is

$$L_n(\theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

which in the discrete case corresponds to the joint probability that the underlying distribution generates the observed sample $x_1, \ldots, x_n$. For a family of continuous distributions, this product can no longer be directly interpreted as a probability, but the overall reasoning remains the same. The corresponding *log-likelihood function* is

$$\ell_n(\theta) = \log\left(L_n(\theta)\right) = \sum_{i=1}^{n} \log\left(f(x_i; \theta)\right).$$

The *maximum-likelihood estimator (MLE)* $\hat{\theta}_n$ of θ then corresponds to the value that maximizes the likelihood functions:

$$\hat{\theta}_n := \operatorname{argmax}_\theta L_n(\theta) = \operatorname{argmax}_\theta \ell_n(\theta).$$

Example 11 Let us suppose that we perform n measurements and have good reason to expect them to be normally distributed such that $X_1, \ldots, X_n \sim \text{Norm}(\mu, \sigma^2)$. The normal distribution can often be justified with the Central Limit Theorem. We want to estimate both parameters from the n observed values $x_1, \ldots, x_n$ using the

maximum-likelihood approach. Let us denote the parameters as $\theta = (\mu, \sigma)$ and start with setting up the likelihood function

$$L_n(\theta) = \frac{1}{\pi} \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2\right),$$

where we ignored constant factors in the second equation, as they do not contribute to the maximization (the symbol $\propto$ means "proportional to"). Abbreviating $\bar{X} = \frac{1}{n} \sum X_i$ and $S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$, we can eliminate the sum and simplify to

$$L_n(\theta) \propto \sigma^{-n} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{X} - \mu)^2}{2\sigma^2}\right),$$

from which we immediately derive the log-likelihood function

$$\ell_n(\theta) \propto -n \log(\sigma) - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}.$$

We maximize this function by taking the derivatives with respect to μ and σ, respectively. For deriving $\hat{\mu}_n$, the equation reads

$$\frac{\partial \ell_n(\theta)}{\partial \mu} = -\frac{n}{2\sigma^2} \left(-2\bar{X} + 2\mu\right),$$

and finding the roots yields the estimator

$$-2\bar{X} + 2\mu = 0 \iff \bar{X} = \mu.$$

Not surprising, the arithmetic mean is an estimator for the expectation. The maximum-likelihood estimators for the two parameters are then

$$\hat{\mu}_n = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}_n^2 = S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

where the derivation of $\hat{\sigma}_n$ follows the same ideas.

Calculating the distribution of an estimator will become crucial for establishing the bounds of a particular estimate. While this calculation is often difficult for general estimators, maximum-likelihood estimators have the convenient property of being *asymptotically normal*. Formally,

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \to \text{Norm}(0, 1) \text{ as } n \to \infty,$$

which simply means that if the sample size gets large enough, *any* maximum-likelihood estimator has a normal distribution. In retrospect, this explains the surprisingly good fit of the empirical and normal density in our introductory example (see Fig. 2.1 (right)).

2.2.2 Least-Squares

Instead of maximizing the likelihood of the observed outcome, we can also construct an estimator by looking at the distance of the observed outcome and the outcome that we would expect with a particular parameter value. The value that minimizes this distance is then an estimate for the parameter.

Let us denote by $h(\theta)$ the expected value of observations for parameter value θ. With measurements x_i, we then minimize the distance

$$d(\theta) = \sum_{i=1}^{n} (x_i - h(\theta))^2.$$

The *least-squares estimate (LSE)* $\hat{\theta}_n$ is then the value that minimized this squared difference between observed and expected data. This is a very common approach in regression (Chap. 4).

Example 12 Let us consider the sequence matching again, this time from a least-squares perspective, and compare the matches with their expected value. At each position i, the expected value of a match is $\mathbb{E}(M_i) = p$, while the observed match m_i is either zero or one. We minimize the sum of their squared differences:

$$\tilde{p}_n = \text{argmin}_p \sum_{i=1}^{n} (m_i - p)^2 .$$

Minimization is again done by finding the roots of the derivative. A quick calculation reveals

$$\frac{\partial}{\partial p} \sum_{i=1}^{n} (m_i - p)^2 = \frac{\partial}{\partial p} \left(\sum_{i=1}^{n} m_i^2 - 2p \sum_{i=1}^{n} m_i + \sum_{i=1}^{n} p^2 \right) = 0 - 2m + 2np,$$

which yields $\tilde{p}_n = \frac{m}{n}$. In this example, the least-squares and the maximum-likelihood estimator are identical, but this is not always the case.

2.2.3 Properties of Estimators

In principle, there is no reason why we should not define an estimator $\hat{\theta}_n = g(X_1, \ldots, X_n) = 0$, which completely ignores the data. It is a formally valid estimator, but quite useless in practice. The question therefore arises, how we can capture properties of an estimator and conclude that, for example the MLE is more useful than the proposed "zero-estimator"?

Consistency. The first useful property of an estimator is *consistency*, which means that with increasing sample size, the estimate approaches the true parameter value:

$$\hat{\theta}_n \to \theta \text{ as } n \to \infty.$$

While all three estimators $(\hat{p}_n, \bar{X}, S^2)$ in the binomial and normal examples are consistent, the above estimator $\hat{\theta}_n \equiv 0$ is obviously not, because the estimate does not get any closer to the true value, no matter how many samples we take.

Unbiasedness. Even if an estimator is consistent, it might still be that it systematically over- or underestimates the true value and introduces a *bias* in the estimate. The bias is given by the difference of expected and true value

$$\mathbb{E}(\hat{\theta}_n) - \theta,$$

and an estimator is called *unbiased* if this difference is zero. If we were to repeat the same sampling procedure multiple times, an unbiased estimator would on average neither over- nor underestimate the true parameter value.

The following example shows the bias in one of the estimators we constructed earlier.

Example 13 Consider the estimates for the parameters μ and σ^2 of a normal distribution as given above. Are they unbiased? Let us start with $\bar{X}$; its unbiasedness is easily established by exploiting the linearity of the expectation:

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}n\mu = \mu.$$

The calculation for S^2 is slightly more elaborate and we skip some details:

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((X_i - \bar{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((X_i - \mu)(\bar{X} - \mu)\right) \\ &= \sigma^2 - \frac{2}{n}\sigma^2 + \frac{n\sigma^2}{n^2} = \sigma^2\left(1 - \frac{1}{n}\right) \neq \sigma^2.\end{aligned}$$

The MLE for the variance is therefore biased and systematically underestimates the true variance. It is nevertheless consistent, as the bias is proportional to $1/n$ and decreases rapidly to zero for increasing n.

The reason for this can presumably be best explained with the following argument: we use n sample points $X_1, \ldots, X_n$ for estimation and thus divide by n. However, we also use the *estimate* $\bar{x}$ instead of the true expectation μ. The value of any sample point is completely determined if we know $\bar{X}$ and the other remaining points. The *degrees of freedom* in the estimate are therefore $n - 1$ rather than n, as we already "used" one degree for estimating μ. Indeed,

$$S^2 = \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

is an unbiased estimator for the variance, but not a maximum-likelihood estimator.

Properties of ML-Estimators. Conveniently, maximum-likelihood estimators automatically have many desired properties. They are

- *consistent*: they approach the true parameter value with increasing sample size,
- *equivariant*: if r is a function, then $r(\hat{\theta}_n)$ is also the MLE of $r(\theta)$,
- *not* necessarily unbiased, so we need to take caution here,
- *asymptotically normal*: $\frac{\hat{\theta}_n - \theta}{\sqrt{\mathrm{Var}(\hat{\theta}_n)}} \to \text{Norm}\,(0, 1)$ as $n \to \infty$.

2.3 Confidence Intervals

The discussed properties of estimators provide valuable information for comparing and choosing an estimator, but they say rather little about the quality of a particular estimate. For example, consistency guarantees that the estimated value will approach the true value in the limit, but does not give information on how close it is to the true value, given some data with a certain number of samples.

For quantifying the quality of a particular estimate, we can compute a *confidence interval* (*CI*) around the estimate $\hat{\theta}_n$, such that this interval covers the true value θ with some high probability $1 - \alpha$. The narrower this interval, the closer we are to the true value, with high probability.

Let us go through the main ideas first, before we look into two concrete examples. For an estimator $\hat{\theta}_n$, the $(1 - \alpha)$-confidence interval is the interval

$$C = \left[\hat{\theta}_n - l, \hat{\theta}_n + u\right] ,$$

for a lower value l and an upper value u such that

$$\mathbb{P}(\theta \in C) = 1 - \alpha. \tag{2.1}$$

This interval C is random, because the value of $\hat{\theta}_n$ depends on the data. The location of the interval is determined by $\hat{\theta}_n$, and its width depends on the distribution of the estimator and in particular on the estimator's variance $\mathrm{Var}(\hat{\theta}_n)$. If the estimator's variance decreases, the confidence interval gets narrower. This allows us to conclude that the difference of true value and estimate gets smaller with decreasing variance, with high probability. We usually need to work with the square-root of the estimator's variance, which we call the *standard error*:

$$\mathrm{se}(\hat{\theta}_n) = \sqrt{\mathrm{Var}(\hat{\theta}_n)}.$$

For computing the confidence interval, we start by normalizing the estimator by shifting it by (the unknown) true parameter θ and scaling by $1/\mathrm{se}(\hat{\theta}_n)$. Provided the estimator is unbiased, this normalization simply shifts the estimator's distribution by its mean and scales by the standard error, which results in a new random variable with mean zero and standard error one. Equation 2.1 then becomes

$$\mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\mathrm{se}(\hat{\theta}_n)} \in \left[\frac{-l}{\mathrm{se}(\hat{\theta}_n)}, \frac{u}{\mathrm{se}(\hat{\theta}_n)}\right]\right) = 1 - \alpha. \tag{2.2}$$

Solving (2.2) requires that we find the two quantiles $q_{\alpha/2}$, $q_{1-\alpha/2}$ of the distribution of the normalized estimator, with $1-\alpha/2-\alpha/2 = 1-\alpha$. From these quantiles, we work out the upper value $u = q_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)$ and thus the interval bound $\hat{\theta}_n + q_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)$, and similar for the lower value l. For an unbiased estimator, the $(1-\alpha)$-confidence interval therefore takes the general form

$$C = \left[\hat{\theta}_n + q_{\alpha/2}\mathrm{se}(\hat{\theta}_n), \hat{\theta}_n + q_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)\right],$$

which simplifies by $q_{\alpha/2} = -q_{1-\alpha/2}$ if the estimator additionally has a symmetric distribution around its mean. The two main remaining problems are then to establish the distribution of $\hat{\theta}_n$ to calculate the quantiles and to estimate its variance.

If $\hat{\theta}_n$ is an unbiased maximum-likelihood estimator, we already know that the estimator has a normal distribution and the correct quantiles are $z_{\alpha/2}$ and $z_{1-\alpha/2}$. The shifted and scaled interval is then symmetric around zero and the confidence interval is immediately given by

$$C = \left[\hat{\theta}_n - z_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n),\ \hat{\theta}_n + z_{1-\alpha/2}\mathrm{se}(\hat{\theta}_n)\right].$$

Before we deal with the more general case of estimators that are not ML, let us first look into two concrete examples and work out confidence intervals for the sequence matching problem and the estimates for the normal parameters.

Example 14 We would like to compute an interval $[\hat{p}_n - l, \hat{p}_n + u]$ around the estimate $\hat{p}_n$ of the matching probability, such that the interval contains the true value p with given probability $1 - \alpha$:

$$\mathbb{P}(p \in [\hat{p}_n - l, \hat{p}_n + u]) = \mathbb{P}(\hat{p}_n - l \leq p \leq \hat{p}_n + u) = 1 - \alpha.$$

Because $\hat{p}_n$ is the maximum-likelihood estimator of p, its distribution approaches a normal distribution for large n. Its normalized form has a standard normal distribution:

$$\frac{p - \hat{p}_n}{\mathrm{se}(\hat{p}_n)} \sim \mathrm{Norm}\ (0, 1).$$

We can therefore immediately solve the following equation by using the corresponding quantiles z_α for u and l

$$\mathbb{P}\left(\frac{-l}{\mathrm{se}(\hat{p}_n)} \leq \frac{p - \hat{p}_n}{\mathrm{se}(\hat{p}_n)} \leq \frac{u}{\mathrm{se}(\hat{p}_n)}\right) = 1 - \alpha.$$

By exploiting the symmetry of the normal distribution, we derive

$$\frac{u}{\mathrm{se}(\hat{p}_n)} = z_{1-\alpha/2} \iff u = z_{1-\alpha/2}\mathrm{se}(\hat{p}_n) \text{ and } l = z_{\alpha/2}\mathrm{se}(\hat{p}_n),$$

The standard error of $\hat{p}_n$ is $\mathrm{se}(\hat{p}_n) = \sqrt{p(1-p)/n}$, leading to the requested confidence interval

$$C = \left[\hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right],$$

where we replaced the unknown true parameter value p by its estimate $\hat{p}_n$.

An immediate caveat of the approximation of the true distribution of the estimator $\hat{p}_n$ by its asymptotic normal distribution is that this confidence interval is only valid for large sample sizes n and parameter values not too close to zero or one. For small p, for example, the confidence interval would also consider the case that $\hat{p}_n$ takes on a negative value, which is not possible. Hence, the approximations for this confidence interval are not always valid and more sophisticated intervals exist.

Example 15 Let us consider the estimator for the expectation of normally distributed data, i.e., $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ with $X_i \sim \text{ Norm } (\mu, \sigma^2)$. Being the ML-estimator, this random variable has a normal distribution. We already checked that it is unbiased, and we easily compute its variance $\mathrm{Var}(\bar{X})$ as

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},$$

where we could take the sum outside the variance because we assumed the X_i to be independent. Thus, the normalized distribution of the difference in true and estimated mean is

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \text{ Norm } (0, 1).$$

Again, we do not know the true variance and need to estimate is using the unbiased estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, which leads to the normalized random variable

$$\frac{\bar{X}-\mu}{S/\sqrt{n}},$$

which does *not* have a standard normal distribution. We can derive its correct distribution by looking at the estimated variance in more detail. In particular, let us consider the quotient of the true and estimated variance:

$$(n-1)\frac{S^2}{\sigma^2} = (n-1)\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2.$$

Each summand is the square of a standard normal variable and there are $(n-1)$ independent such variables. Thus, from Sect. 1.4, we know that the sum has a χ^2-distribution with $(n-1)$ degrees of freedom. Replacing the true variance by its estimate, we derive the distribution

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma/\sqrt{n}}{S/\sqrt{n}} \sim \frac{\text{Norm}(0,1)}{\sqrt{\frac{1}{n-1}\chi^2(n-1)}},$$

which from Sect. 1.4 we recognize as a t-distribution with $(n-1)$ degrees of freedom. We therefore derive the correct $(1-\alpha)$-confidence interval

$$C = \left[\bar{X} - t_{1-\alpha/2}(n-1\frac{S}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$$

for the estimator $\bar{X}$ of the expected value. Again, this interval gets narrower if we increase the sample size n or decrease the variance σ^2 of the data.

As an example, let us repeatedly take 10 samples from a Norm(5,16) distribution and compute the corresponding 0.9-confidence interval for the estimated mean $\bar{X}$. For each such computation, we derive a slightly different interval, both in terms of the center of the interval (due to the estimated mean $\bar{X}$) and the length of the interval (due to the estimated variance of $\bar{X}$). For 25 repetitions, the confidence intervals are plotted next to each other in Fig. 2.2. Some intervals, such as the 5th and the 24th, do not cover the true value. To demonstrate the effect of estimating the variance, we compute the correct t-based and the incorrect normal confidence intervals, both using the estimated variance, for the 5th sample (which is too far away from the true mean) as

$$C^t = [1.396, 4.717] \quad \text{and} \quad C^{\text{norm}} = [0.631, 5.482].$$

The normal quantiles overestimate the width of the interval, such that the normal interval contains the true value, while the t-based does not.

2.3.1 The Bootstrap

For computing the confidence interval for a given estimate, we frequently encounter two problems: finding the variance of an estimator, and working out the distribution of an estimator that is not an MLE. In addition, the theory leading to normal (or t-based) confidence intervals is based on the asymptotic distribution of the estimator, which might be quite different than the distribution for small sample sizes. A very popular way for solving these problems is by using the *bootstrap* method, which aims at estimating all necessary quantities directly from the data themselves. While mainly used for computing the estimator's variance, the bootstrap method

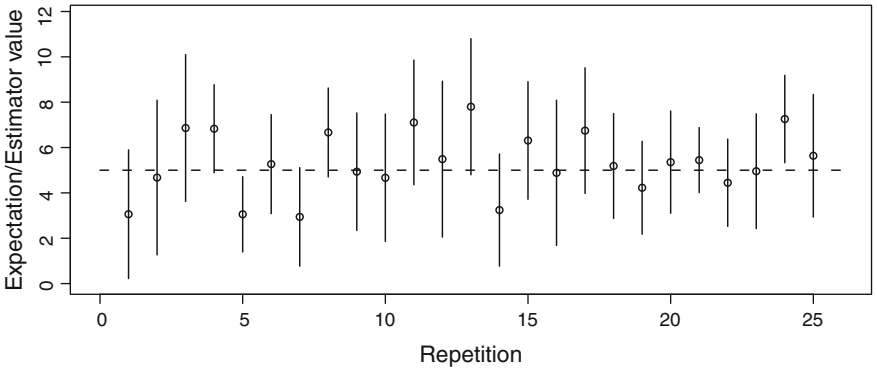

Fig. 2.2 Result of 25 repetitions of estimating the mean of a Norm(5,16) distribution, each with 10 samples. *Dashed line*: True value of mean $\boldsymbol{\mu} = 5$, *solid lines*: 0.9-confidence intervals for estimators, points: estimated value of mean in *i*th repetition

also allows to compute higher moments, and even allows computation of confidence intervals for estimators with non-normal distribution.

Let us suppose we take b independent samples $Y_1, \ldots, Y_b$ from a distribution. Then, by the laws of large numbers, the sample mean approaches the true expectation for increasing b. The same argument still holds if we apply a function h on mean and expectation:

$$\frac{1}{b}\sum_{i=1}^{b} h(Y_i) \to \mathbb{E}(h(Y_1)).$$

For example, we recover the variance estimator by choosing $h = (Y_i - \bar{Y})^2$.

The key idea on how this helps is the following: let us consider any estimator $\hat{\theta}_n$ and denote by $F(x) = \mathbb{P}(\hat{\theta}_n \leq x)$ its cumulative distribution function. For the beginning, we are interested in calculating $\mathrm{Var}(\hat{\theta}_n)$. This is comparatively easy if we know the distribution F of the estimator. If we do not, we can try to estimate this distribution by $\hat{F}$, and subsequently estimate the variance using this estimated distribution as an approximation. Let us assume we are given a set of data $x_1, \ldots, x_n$. For estimating $\hat{F}$, we re-sample new data from this given set, uniformly and with replacement, such that each x_i has the same probability to be re-sampled, and can also be re-sampled several times. We repeat this re-sampling b times, where $x^\star_{j,1}, \ldots, x^\star_{j,n}$ is the jth new sample. From each such sample, we compute the estimator $\hat{\theta}^\star_{n,j}$, leading to a total of b estimates. Each sample is prone to be different from the others, and so are the estimated values.

If the data are a representative sample, this re-sampling gets us all the information needed: there are only few sample points with extreme values. These therefore get re-sampled rarely and are only present in a few bootstrap samples. On the other hand, "typical" values are sampled often, possibly even multiple times, into one bootstrap

sample. The estimation is then performed more often on sets of "typical" data values than on more extreme, more unlikely combinations of values. This in turn gives a correct impression of how the estimator varies with varying data.

Overall, there are n^n different ways draw new samples of size n; this number is very large even for moderate n which ensures that we do get a different sample each time. The name "bootstrap" refers to the seemingly impossible task to lift ourselves out of the unknown variance problem by using the straps of our own boots, namely the data we have.

The algorithm. We can write the general bootstrap procedure for estimating the variance in a more algorithmic form as

- Draw $X_1^\star, \ldots, X_n^\star$ uniformly with replacement from $\{x_1, \ldots, x_n\}$.
- Compute $\hat{\theta}_{n,i}^\star = g(X_1^\star, \ldots, X_n^\star)$ from this bootstrap sample.
- Repeat the two steps b times to get the estimates $\hat{\theta}_{n,1}^\star, \ldots, \hat{\theta}_{n,b}^\star$.
- Compute the estimator's bootstrap variance estimate

$$\mathrm{Var}(\hat{\theta}_n) \approx v_{\mathrm{boot}} = \frac{1}{b}\sum_{i=1}^{b}\left(\hat{\theta}_{n,i}^\star - \frac{1}{b}\sum_{j=1}^{b}\hat{\theta}_{n,j}^\star\right)^2.$$

In R, the package `boot` offers a function `boot()` that simplifies the computation of statistics by the bootstrap method. Once the b bootstrapped values for $\hat{\theta}_n$ are computed, there are several ways to form a confidence interval.

The normal CI. If the estimator has a normal distribution, we may simply replace the variance $\mathrm{Var}(\hat{\theta}_n)$ by its bootstrap estimate v_{boot} and form the usual normal (or t-based) confidence interval

$$C^{\mathrm{norm}} = \hat{\theta}_n \pm z_{1-\alpha/2}\sqrt{v_{\mathrm{boot}}}.$$

The percentile CI. The second method relies on using the bootstrap samples of estimator values to compute the empirical quantile of its distribution and works for all unbiased estimators. For this, we sort the bootstrap estimates such that $\hat{\theta}_{n,(1)}^\star \le \cdots \le \hat{\theta}_{n,(b)}^\star$; again, the estimate with index *(i)* is the ith smallest one. Then for $k = b\alpha$ (suitably rounded to the next integer), $\hat{\theta}_\alpha^\star := \hat{\theta}_{n,(k)}^\star$ is the empirical α-quantile of the distribution of $\hat{\theta}_n$ and we form the empirical percentile confidence interval

$$C^{\mathrm{percentile}} = \left[\hat{\theta}_{\alpha/2}^\star, \hat{\theta}_{1-\alpha/2}^\star\right].$$

Example 16 Sometimes, the data are not normally distributed, but their logarithms are. The *log-normal distribution* with parameters μ and σ^2 is the distribution of $X = \exp(Y)$, with $Y \sim \mathrm{Norm}\,(\mu, \sigma^2)$. These parameters are the mean and variance of Y, but not of X. The distribution of X is asymmetric and we want to quantify this asymmetry using the skewness measure from Sect. 1.6.5. An unbiased estimator for the skewness θ of a sample is

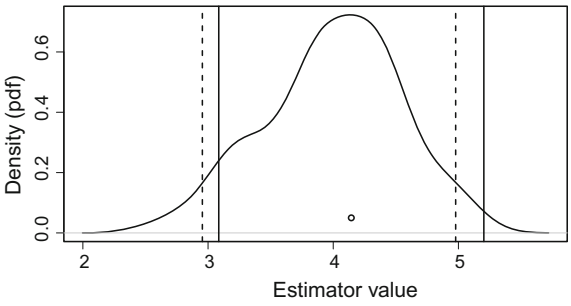

Fig. 2.3 Density of skewness estimates for log-normal sample of size 500. Normal (*solid*) and pivot (*dashed*) 0.95-confidence intervals are given, computed from $b = 100$ bootstrap samples. The circle is the original estimate of skewness

$$\hat{\theta}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}}.$$

Instead of working out the distribution of this estimator, we apply the bootstrap method to derive the percentile and normal confidence interval for a given estimate.

For illustration, we generate $n = 500$ samples $X_1, \ldots, X_n$ of log-normal random variables with parameters $\mu = 0,\ \sigma = 1$. We then compute the skewness estimate $\hat{\theta}_n$, followed by $b = 100$ bootstrap samples from the X_i. Normal and pivot confidence intervals are finally computed for $\alpha = 0.05$.

The results are given in Fig. 2.3, where the estimate $\hat{\theta}_n$ is indicated by the small circle, and the normal and percentile 0.95-CIs are given by the solid and dashed lines, respectively. The solid black line gives the empirical density function of the estimator $\hat{\theta}_n$ estimated from the bootstrap samples. Interestingly, the percentile confidence interval is not symmetric around the estimate, because of a skewed estimator distribution.

2.4 Robust Estimation

There is one major problem with the estimators that we discussed so far: they all assume that each sample point is taken from the same underlying distribution, and there are thus no contaminations in the sample. A contamination can be an "outlier" that, by eye, can clearly be identified as an incorrect measurement, for example. Many estimators are very sensitive to such contaminations.

Example 17 A sample of $n = 20$ points from a Norm(5,4) is taken. In addition, the sample data is contaminated by only $n' = 2$ outliers, leading to the empirical density given in Fig. 2.4. Estimating the mean of the sample distribution gives the location indicated by the solid vertical line; it is substantially shifted to the right from the correct expected value which would be somewhere near the maximum of the density in this case.

To quantify how sensitive an estimator is to contaminations in the sample, the *robustness* of an estimator is measured by its *breakdown point*. It refers to the proportion of contaminations an estimator can handle before it gives arbitrarily large values. For the arithmetic mean, even one contamination has the capacity to make the estimate arbitrarily large, as even a single outlier very far away from the rest of the sample "pulls" the whole estimate away from the correct mean of the uncontaminated sample. Its breakdown point is therefore zero. The same argument holds for the two estimators for the variance, which both also have a breakdown point of zero.

While this might in practice not be as dramatic as theory suggests, simply because contaminations far away from the sample are often unlikely, it is nevertheless a reason to be uncomfortable, as it means that even a small amount of contamination can potentially yield very misleading results. For a small number of sample points, we might try a visual inspection to see if there are any unusual values in the sample, but this is clearly not a good strategy if we want to investigate large amounts of data. We will therefore investigate *robust estimators* with high breakdown points as alternatives for common estimators. Here, we will discuss robust alternatives for estimating the location and scale. They all rely on *order statistics* of the sorted sample, again denoted $x_{(1)} \leq \cdots \leq x_{(n)}$, and typically estimate empirical quantiles.

2.4.1 Location: Median and k-Trimmed Mean

Median. In addition to the expectation, the *median* is another measure for the location of a distribution. It corresponds to the 0.5-quantile $q_{0.5}$ of a distribution such that $\mathbb{P}(X \leq q_{0.5}) = 0.5$. We can estimate any α-quantile from the sorted sample simply by finding the correct index from $k = n\alpha$. If k is an integer, we simply select the kth smallest value, i.e., $\hat{q}(\alpha) = x_{(k)}$. If k is not an integer, we compute the two nearest integer k', k'' and interpolate the corresponding values $x_{(k')}$ and $x_{(k'')}$. Various ways for interpolation exist, many of which are implemented in the `quantile()` function in `R`. The estimate $\hat{q}(0.5)$ for the median is therefore simply the sample point in the middle (or the average of the two surrounding ones). As such, it does not use any information about the actual values of the sample points, but only uses information about the *rank*, i.e., their indices in the sorted sample.

Example 18 For $n = 8$ given sample points

6.39, 0.887, 1.521, 8.635, 7.742, 7.462, 6.631, 5.511,

the sorted values are

0.887, 1.521, 5.511, 6.39, 6.631, 7.462, 7.742, 8.635.

The median is estimated as $\hat{q}(0.5) = \frac{1}{2} \times (6.39 + 6.631) = 6.5105$ by interpolating between the two sample points with ranks 4 and 5.

Changing the largest value from 8.635 to 108.635 changes the mean substantially from $\hat{\mu} = 5.5974$ to $\hat{\mu}' = 18.0974$, but leaves the median unchanged at $\hat{q}'(0.5) = 6.5105$. In `R`, the median is computed by `median()`.

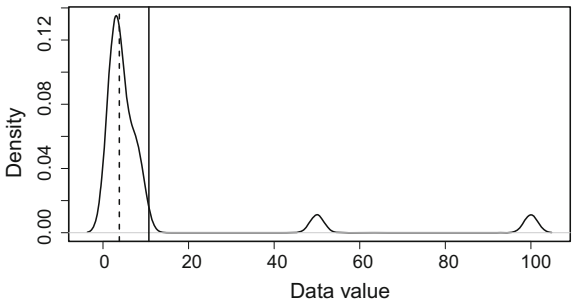

Fig. 2.4 Estimation of the location of a sample of 20 normally distributed points contaminated with 2 "outliers" with values 50 and 100. The true mean is $\boldsymbol{\mu} = 5$. The two outliers "pull" the estimation of the mean to the right, leading to $\bar{x} = 10.72$ (*solid line*). The robust estimate of the median is $\hat{m}_{n+n'} = 3.82$ (*dashed line*), with a true median of 5

For the contaminated normal sample of Example 17, the estimated median $\hat{q}_{0.5} = 3.821$ is given by the dashed vertical line in Fig. 2.4. It is reasonably close to the true median $q_{0.5} = 5$ and largely unaffected by the two contaminations.

Let us look at the robustness of the median. We note that because it only considers ranks and not values, we can increase all points larger than the median arbitrarily without changing it. The same holds for decreasing smaller values and we conclude that the median has breakdown point 50%.

Both the mean and the median are measures for the location of the distribution, trying to give a single number to describe where "most" of the values are. The mean gives the expectation or average of a sample, wheres the median indicates the point such that half of the data is smaller (resp. larger). If the distribution is symmetric, the two values are identical, but they differ for skewed distributions. Because of its large breakdown point, the median is often the better choice for estimating the location. However, it cannot always be interpreted as the expected value of the distribution.

Trimmed means. If we still want to specifically estimate the expectation robustly, the k-trimmed mean is a good alternative to the median. It also uses the sorted values of the sample, but drops the k lowest and highest sample points of the data. The rationale is that contaminations are likely to be much smaller or much larger than the uncontaminated sample points. Formally,

$$\hat{\mu}(k) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}.$$

For $k = 0$, we recover the ordinary arithmetic mean again, for $k = n/2$ (taken to the next suitable integer), we recover the median. The k-trimmed mean is thus a generalization of both estimators. The choice for k is somewhat arbitrary of course, and should always be stated if this estimator is used; common choices are $k = 0.05 \times n$ and $k = 0.25 \times n$.

Example 19 For the 8 sample points of the previous example and $k = 1$, the k-trimmed mean reads

$$\hat{\mu}(1) = \frac{1}{8}(x_{(2)} + \cdots + x_{(7)}) = 5.876,$$

so the smallest and largest value are ignored and the ordinary arithmetic mean is computed from the remaining data. Again, changing the largest sample value from 8.635 to 108.635 does not change the estimate, as this point is ignored in the computation. In `R`, the k-trimmed mean can be accessed by `mean(..., trim=...)`.

2.4.2 Scale: MAD and IQR

Similar considerations lead to two robust alternatives for measuring the scale of a distribution: the *median absolute deviation* (MAD) and the *inter-quartile-range* (IQR).

Median absolute deviation. The MAD follows the same ideas as the variance, but measures the median of the absolute distance to the median:

$$\text{MAD} = \text{median}_i(|x_i - \text{median}_j(x_j)|).$$

Inter-quartile range. The IQR computes the difference between the 0.25- and 0.75-quantile, and is given by the rectangle in a boxplot (see Sect. 1.8.4) that contains the medium 50% of the data:

$$\text{IQR} = q_{\frac{3}{4}} - q_{\frac{1}{4}}.$$

Comparison to the variance. Both the MAD and the IQR give different measures for the scale compared to the variance. Similar to the median, they both are more based on the ranks and not the absolute values of the particular data, and have high breakdown points. For the normal distribution, $\sigma \approx 1.48 \times \text{MAD}$, so variance and MAD are scaled versions of each other. In `R`, both MAD and IQR are easily accessible via the functions `mad()` and `IQR()`.

Example 20 For $n = 20$ samples contaminated with $n' = 2$ "outliers" of Example 17, the various estimators are summarized in the following table. The second column gives the values estimated on the contaminated sample, the third column gives the values computed on the uncontaminated subset of the sample.

Estimator	Value	True value
$\bar{x} = \hat{\mu}$	11.19	4.81
$\hat{m} = q(0.5)$	4.39	4.28
$\hat{\mu}(\alpha = 0.1)$	5.11	4.62
$\text{s} = \hat{\sigma}$	22.12	1.95
IQR	2.46	1.52
MAD	1.23	1.07

As we would expect, the mean and the standard deviation give very different values on the contaminated and uncontaminated sample. In contrast, their robust counterparts all give estimates on the contaminated sample that are reasonably close to the uncontaminated values.

2.5 Minimax Estimation and Missing Observations

In addition to maximum-likelihood and least-squares, *minimax estimation* is a third principle to construct estimators.

2.5.1 Loss and Risk

Before introducing minimax estimation, let us briefly look into a theoretical framework that allows us to compare the performance of various estimators and derive new principles for their construction.

The *loss-function* $\mathscr{L}(\theta, \hat{\theta})$ measures the distance from the true parameter value and its estimate. Two popular choices for loss functions are the squared loss $\mathscr{L}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ and the absolute loss $\mathscr{L}(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$.

The loss depends on the actual value of the estimator, and thus on the specific sample. To get a more general measure, we therefore look at the expected loss, known as the *risk* of the estimator

$$\mathscr{R}(\theta, \hat{\theta}) = \mathbb{E}\big(\mathscr{L}(\theta, \hat{\theta})\big).$$

For example, the risk of an unbiased estimator $\hat{\theta}$ with respect to the squared loss function is simply its variance:

$$\mathscr{R}(\theta, \hat{\theta}) = \mathbb{E}\big(\mathscr{L}(\theta, \hat{\theta})\big) = \mathbb{E}\big((\theta - \hat{\theta})^2\big) = \mathbb{E}\big((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\big).$$

A small risk indicates that on average, for all possible true parameter values, the estimator is not too far off.

Example 21 Let us calculate the risk for the maximum-likelihood estimator $\hat{p}_n$ of the matching probability p in the sequence matching example with respect to squared loss of the MLE $\hat{p}_n = M/n$:

$$\mathscr{R}(p, \hat{p}) = \mathbb{E}\big((\hat{p}_n - p)^2\big) = \mathrm{Var}(\hat{p}_n) = \frac{p(1-p)}{n}.$$

As shown in Fig. 2.5 (solid line), the risk is highest for $p \approx 1/2$ and lowest for values near the boundary. Intuitively, for $p = 0$, we will not observe any matches, and always estimate correctly.

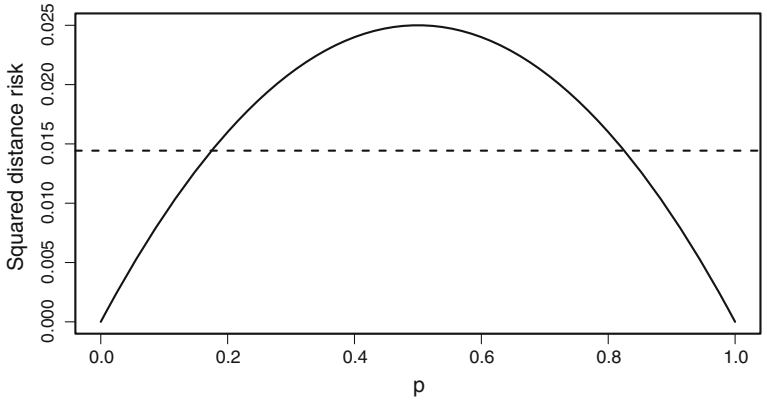

Fig. 2.5 Risk function for MLE $\hat{p}$ (*solid line*) and minimax estimator $\tilde{p}$ (*dashed line*) of the matching probability p

2.5.2 Minimax Estimators

For some applications, we might be interested in having a guarantee that the risk is not too high for any possible true value of the parameter. For this, we construct a *minimax estimator* $\tilde{\theta}$ such that the maximal risk

$$\max_{\theta} \mathscr{R}(\theta, \tilde{\theta})$$

is minimal. This means that while we allow the loss of this estimator to be larger for some values of θ, it will stay lower *on average* than any other estimator.

Example 22 Let us consider a different estimator for the matching example:

$$\tilde{p}_n = \frac{M + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}}$$

has squared loss risk

$$\mathscr{R}(p, \tilde{p}_n) = \frac{n}{4(n + \sqrt{n})^2},$$

which is constant for all parameter values p and is smaller than the maximal risk of the MLE $\hat{p}_n$. Indeed, $\max_p \mathscr{R}(p, \hat{p}_n) = 1/4n$ but $\mathscr{R}(p, \tilde{p}_n) < 1/4n$ for all p, as one easily verifies. As we see in Fig. 2.5, however, the risk is not always lower than the MLE-risk, but is lower for mid-range parameter values and higher in the extremes.

Minimax estimators are also useful if some potential outcomes of an experiment have small probabilities and may not be observed due to a small sample size.

Example 23 Let us consider the following problem: the possible outcome of an experiment is one of s different types, such as `A/C/G/T` in the former random

sequence examples, with each sequence position being one of the four possible nucleotides.

Such experiments are described by a *multinomial distribution*, which is a generalization of the binomial distribution to more than two outcomes. The probability for an observation of category i is p_i, and $\sum_{i=1}^{s} p_i = 1$. A result of such an experiment is a vector $(X_1, \ldots, X_s)$ containing the number of samples of category i in X_i, so that $\sum_{i=1}^{s} X_i = n$ is the total number of sample points.[1]

The probability mass function of $(X_1, \ldots, X_s)$ is given by

$$\mathbb{P}(X_1 = k_1, \ldots, X_s = k_s) = \frac{n!}{k_1! \cdots k_s!} p_1^{k_1} \cdots p_s^{k_s},$$

and the binomial distribution is recovered by setting $s = 2$, in which case $p_2 = 1 - p_1$ and $k_2 = n - k_1$, leading to the binomial coefficient. The expected number of observations in the ith category is $\mathbb{E}(X_i) = np_i$. In the sequence example, we would thus expect to see np_A `A` and np_G `G` in a sequence of length n. The probability of category i can be estimated by the maximum-likelihood estimator

$$\hat{p}_{n,i} = \frac{k_i}{n}.$$

A common rule-of-thumb suggests choosing a sample size n such that at least five observations are expected in each category for estimating the various probabilities with some confidence:

$$n \min_i(p_{n,i}) \geq 5 \Rightarrow n \geq \frac{5}{\min_i(p_{n,i})}.$$

For the possible nucleotides with probabilities

$$(p_A, p_C, p_G, p_T) = (1/2, 1/4, 1/8, 1/8),$$

we would thus need at least $5/0.125 = 40$ samples to reliably estimate all probabilities.

In practice, we usually do not know these probabilities, of course, and sometimes have no control over the possible sample size n. Imagine that we only have a sequence of 20 nucleotides. If it happens not to have any `G`, we consequently estimate $\hat{p}_G = 0$. This will have undesired consequences if we use these values for a model to describe the sequence matching probabilities, because the model would assume that `G` can never occur and will thus incorrectly predict the possible number of matchings.

One way of dealing with this problem is to introduce *pseudo-counts* by pretending that there is a certain number of observations in each category to begin with. For example, let us put a observations in each category before conducting the actual

[1] Note that in contrast to the previous notation, all X_i together describe *one* experiment (or sample point).

experiment, and therefore see $x_i + a$ observations in category i after the experiment. Then, we can use the estimate

$$\tilde{p}_{n,i} = \frac{x_i + a}{sa + n}$$

for the categories' probabilities, which is simply the MLE for the modified data. With $a > 0$, each estimate is strictly larger than (but potentially very close to) zero.

Let us assume we observed $(x_A, x_C, x_G, x_T) = (13, 6, 0, 1)$. How large should we choose a? If we choose it too large, it would spoil the whole estimation and assign almost identical probabilities everywhere, independent of the data. For example, with $a = 1000$ the estimates are

$$(\tilde{p}_A, \tilde{p}_C, \tilde{p}_G, \tilde{p}_T) = (0.252, 0.25, 0.249, 0.249).$$

If we choose a too small, it might not have an effect and we end up with non-zero, but extremely low probabilities. Indeed, for $a = 0.1$,

$$(\tilde{p}_A, \tilde{p}_C, \tilde{p}_G, \tilde{p}_T) = (0.642, 0.299, 0.005, 0.054).$$

We can calculate a reasonable compromise by selecting a such that we minimize the maximal risk of the corresponding estimator. For parameters of the multinomial distribution, this minimax estimator is achieved by choosing

$$a = \frac{\sqrt{n}}{s}.$$

For the example, $a = 1.118$ and we estimate

$$(\tilde{p}_A, \tilde{p}_C, \tilde{p}_G, \tilde{p}_T) = (0.577, 0.291, 0.046, 0.087),$$

which is fairly close to the correct values, taking into account that we do not have many data available.

The seemingly ad-hoc estimator in Sect. 2.5.2 for the binomial case was derived in this way.

2.6 Fisher-Information and Cramér-Rao Bound

We conclude the chapter by a brief discussion of the idea of Fisher-information, from which we can derive a theoretical lower bound for the variance of an estimator. This bound tells us how precise we can actually estimate a given parameter with a fixed number of samples.

Recall the definition $\ell_n(\theta) = \sum_i \log(f(x_i; \theta))$ of the log-likelihood function. The *Fisher-score* is simply the derivative of this function with respect to the parameter(s),

$$\text{Fisher-score} = \frac{\partial \ell_n(\theta)}{\partial \theta},$$

and we calculate a maximum-likelihood estimator by finding its roots. In addition, the *Fisher-information* describes the curvature of the likelihood function around a parameter value θ. It is given by

$$I_n(\theta) = \sum_{i=1}^{n} \text{Var}\left(\frac{\partial \ell_n(\theta)}{\partial \theta}\right) = -n\mathbb{E}\left(\frac{\partial^2 \ell_n(\theta)}{\partial \theta^2}\right) = nI(\theta).$$

Loosely speaking, a large information indicates that the likelihood function will change noticeably when moving from θ to a nearby value θ'; the parameter value can then be estimated more reliably. A small information indicates a shallow "valley" in the likelihood function, where substantially different parameter values lead to almost identical values of $\ell_n(\theta)$.

Example 24 Let us again consider the matching example with log-likelihood function $\ell_n(p) = M \log(p) + (n - M) \log(1 - p)$ and

$$\frac{\partial \ell_n(p)}{\partial p} = \frac{M}{p} - \frac{n - M}{1 - p}.$$

The Fisher-information is

$$\begin{aligned} I_{n(p)} &= -n\mathbb{E}\left(\frac{\partial^2 \ell_n(p)}{\partial p^2}\right) = -n\mathbb{E}\left(-\frac{M}{p^2} - \frac{n - M}{(1 - p)^2}\right) \\ &= \frac{n}{p^2}\mathbb{E}(M) + \frac{n}{(1 - p)^2}\mathbb{E}(n - M) = \frac{n}{p(1 - p)}. \end{aligned}$$

For sequences of length $n = 20$ nucleotides and $m = 12$ observed matches, the log-likelihood function and its Fisher-information are given in Fig. 2.6. As expected, the log-likelihood is highest at $p = m/n$. The Fisher-information does not take into consideration the actual observed matches and shows that for parameters p in the mid-range, the information carried by a sample is much lower than for more extreme parameter values near zero or one. This tells us that true values near the boundaries are much easier to estimate, as they lead to more dramatic expected changes in the likelihood function. These properties of the likelihood and information functions become more pronounced if we increase the number of samples from $n = 20$ to $n = 60$.

Cramér-Rao bound. The main importance of the Fisher-information is that it allows us to calculate the smallest possible variance that can be achieved with a given estimator and a given sample size. This *Cramér-Rao bound* states that

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)},$$

and we cannot decrease the variance of an estimator $\hat{\theta}_n$ below the reciprocal of its information. For getting estimates with lower variance and thus, for example,

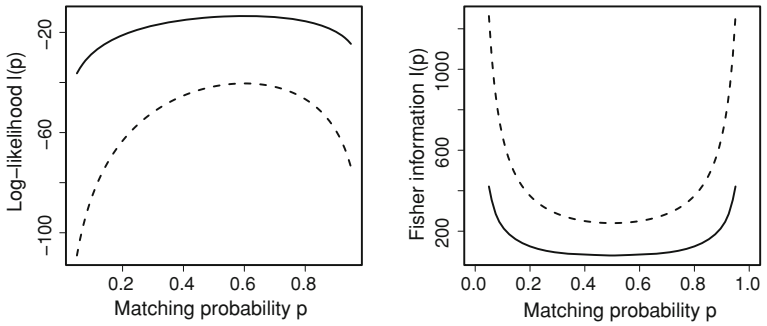

Fig. 2.6 Log-likelihood function $\ell_n(p)$ (*left*) and Fisher information $I_n(p)$ (*right*) for the sequence matching example with $n = 20$ and $m = 12$ observed matches (*solid lines*) and $n = 60$, $m = 36$ (*dashed lines*)

narrower confidence intervals, we either need to increase the sample size (because $I_n(\theta) \approx nI(\theta)$) or choose another estimator. Indeed, some estimators can be shown to have lowest variance among all other estimators for the same parameters.

2.7 Summary

Estimation allows us to infer the value of various properties of a distribution, such as its location, from data. We can construct corresponding estimators by the maximum-likelihood, the least-squares, and the minimax approach. Estimators are random variables because their realization depends on a given random sample. Their properties such as consistency and unbiasedness allow us to compare different estimators.

For a concrete estimation, we can compute confidence intervals around the estimated value to quantify how good the estimate is. These intervals contain the true value with high probability. For their computation, we need to know the distribution of the estimator to find the corresponding quantiles, and its standard error to scale correctly. The bootstrap offers a practical method to establish confidence intervals by resampling the data and computing empirical quantiles from the corresponding estimated values.

The breakdown point describes the sensitivity of an estimator to contaminations in the data. Because many classical estimators have a very low breakdown point, we should usually try to use robust alternatives, such as the median. Many robust estimators are based on the ranks of sample points rather than their values.

Missing observations and small sample sizes can cause major problems when estimating multinomial probabilities. Using minimax estimation to calculate pseudo-counts enables us to partly circumvent these problems.