Chapter 3
Estimation Theory

This chapter gives an introduction to decision and estimation theory. This introduction is based on the books of Lehmann [243, 244], the lecture notes of Künsch [229] and the book of Van der Vaart [363]. This chapter presents classical statistical estimation theory, it embeds estimation into a historical context, and it provides important aspects and intuition for modern data science and predictive modeling. For further reading we recommend the books of Barndorff-Nielsen [23], Berger [31], Bickel–Doksum [33] and Efron–Hastie [117].

3.1 Introduction to Decision Theory

We start from an observation vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ taking values in a measurable space $\mathbb{Y} \subset \mathbb{R}^n$, where $n \in \mathbb{N}$ denotes the number of components Y_i, $1 \le i \le n$, in $\boldsymbol{Y}$. Assume that this observation vector $\boldsymbol{Y}$ has been generated by a distribution belonging to the family $\mathcal{P} = \{P(\cdot; \theta); \theta \in \boldsymbol{\Theta}\}$ being parametrized by a parameter set $\boldsymbol{\Theta}$.

Remarks 3.1 There are some subtle points in the notation that we are going to use. We use $P(\cdot; \theta)$ for the distribution of the observation vector $\boldsymbol{Y}$, and if we consider a specific component Y_i of $\boldsymbol{Y}$ we will use the notation $Y_i \sim F(\cdot; \theta)$. We make this distinction as in estimation theory one often considers i.i.d. observations $Y_i \sim F(\cdot; \theta)$, $1 \le i \le n$, with (in this case) joint product distribution $\boldsymbol{Y} \sim P(\cdot; \theta)$. This latter distribution is then used for purposes of maximum likelihood estimation, etc. The family $\mathcal{P}$ is parametrized by $\theta \in \boldsymbol{\Theta}$, and if we want to emphasize that this parameter is a k-dimensional vector we use boldface notation $\boldsymbol{\theta}$, this is similar to the EFs introduced in Chap. 2, but in this chapter we do not restrict to EFs. Finally, we assume identifiability meaning that different parameters θ give different distributions $P(\cdot; \theta) \in \mathcal{P}$.

M. V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer Actuarial, https://doi.org/10.1007/978-3-031-12409-9_3

To fix ideas, assume we want to determine $\gamma(\theta)$ of a given functional $\gamma(\cdot)$ on $\boldsymbol{\Theta}$. Typically, the true value $\theta \in \boldsymbol{\Theta}$ is not known, and we are not able to determine $\gamma(\theta)$ explicitly. Therefore, we try to *estimate* $\gamma(\theta)$ from data $\boldsymbol{Y} \sim P(\cdot; \theta)$ that belongs to the same $\theta \in \boldsymbol{\Theta}$. As an example we may think of working in the EDF of Chap. 2, and we are interested in the mean $\mu = \mathbb{E}_\theta[Y] = \kappa'(\theta)$ of Y. Thus, we aim at determining $\gamma(\theta) = \kappa'(\theta)$. If the true θ is unknown, and if we have an observation Y from this model, we can try to estimate $\gamma(\theta) = \kappa'(\theta)$ from Y. This motivation is based on estimation of $\gamma(\theta)$, but the following framework of decision making is more general, for instance, it may also be used for statistical hypothesis testing.

Denote the *action space* of possible decisions (actions) by $\mathbb{A}$. In decision theory we are looking for a *decision rule* (*action rule*)

$$A : \mathbb{Y} \to \mathbb{A}, \qquad \boldsymbol{Y} \mapsto A(\boldsymbol{Y}), \tag{3.1}$$

which should be understood as an educated guess for $\gamma(\theta)$ based on observation $\boldsymbol{Y}$. A decision rule is evaluated in terms of a (given) *loss function*

$$L : \boldsymbol{\Theta} \times \mathbb{A} \to \mathbb{R}_+, \qquad (\theta, a) \mapsto L(\theta, a) \geq 0. \tag{3.2}$$

$L(\theta, a)$ describes the loss of an action $a \in \mathbb{A}$ w.r.t. a true parameter choice $\theta \in \boldsymbol{\Theta}$. The *risk function* of decision rule A for data generated by $\boldsymbol{Y} \sim P(\cdot; \theta)$ is defined by

$$\theta \ \mapsto \ \mathcal{R}(\theta, A) = \mathbb{E}_\theta[L(\theta, A(\boldsymbol{Y}))] = \int_{\mathbb{Y}} L\left(\theta, A(\boldsymbol{y})\right) dP(\boldsymbol{y}; \theta), \tag{3.3}$$

where $\mathbb{E}_\theta$ is the expectation w.r.t. the probability distribution $P(\cdot; \theta)$. Risk function (3.3) describes the long-term average loss of using decision rule A. As an example we may think of estimating $\gamma(\theta)$ for unknown (true) parameter θ by a decision rule $\boldsymbol{Y} \mapsto A(\boldsymbol{Y})$. Then, the loss function $L(\theta, A(\boldsymbol{Y}))$ should describe the *estimation loss* if we consider the discrepancy between $\gamma(\theta)$ and its estimate $A(\boldsymbol{Y})$, and the risk function $\mathcal{R}(\theta, A)$ is the *average estimation loss* in that case.

Good decision rules A should provide a small risk $\mathcal{R}(\theta, A)$. Unfortunately, this statement is of rather theoretical nature because, in general, the true data generating parameter θ is not known and the goodness of a decision rule for the true parameter cannot be evaluated explicitly, but the risk can only be estimated (for instance, using a bootstrap approach). Moreover, typically, there does not exist a uniformly best decision rule A over all $\theta \in \boldsymbol{\Theta}$. For these reasons we may (just) try to eliminate decision rules that are obviously not good. We give two introductory examples.

Example 3.2 (Minimax Decision Rule) Decision rule A is called minimax if for all alternative decision rules $\widetilde{A} : \mathbb{Y} \to \mathbb{A}$ we have

$$\sup_{\theta \in \boldsymbol{\Theta}} \mathcal{R}(\theta, A) \leq \sup_{\theta \in \boldsymbol{\Theta}} \mathcal{R}(\theta, \widetilde{A}).$$

A minimax decision rule is the best choice in the worst case of the true θ, i.e., it minimizes the worst case risk. ■

Example 3.3 (Bayesian Decision Rule) Assume we are given a distribution π on $\boldsymbol{\Theta}$. Decision rule A is called Bayesian w.r.t. π if it satisfies

$$A \;=\; \arg\min_{\widetilde{A}} \int_{\boldsymbol{\Theta}} \mathcal{R}(\theta, \widetilde{A}) d\pi(\theta).$$

Distribution π is called *prior distribution* on $\boldsymbol{\Theta}$. ■

The above examples give two possible choices of decision rules. The first one tries to minimize the worst case risk, whereas the second one uses additional knowledge in terms of a prior distribution π on $\boldsymbol{\Theta}$. This means that we impose stronger assumptions in the second case to get stronger conclusions. The difficult part in practice is to justify these stronger assumptions in order to validate the stronger conclusions. Below, we are going to introduce other criteria that should be satisfied by good decision rules, an important one in estimation will be unbiasedness.

3.2 Parameter Estimation

This section focuses on estimating the (unknown) parameter $\theta \in \boldsymbol{\Theta}$ from observation $\boldsymbol{Y} \sim P(\cdot; \theta)$. For this we consider decision rules $A : \mathbb{Y} \to \mathbb{A} = \boldsymbol{\Theta}$ with $A(\boldsymbol{Y})$ estimating θ. We assume there exist densities $p(\cdot; \theta)$ w.r.t. a fixed σ-finite measure ν on $\mathbb{Y} \subset \mathbb{R}^n$,

$$dP(\boldsymbol{y}; \theta) = p(\boldsymbol{y}; \theta) d\nu(\boldsymbol{y}),$$

for all distributions $P(\cdot; \theta) \in \mathcal{P}$, i.e., all $\theta \in \boldsymbol{\Theta}$.

Definition 3.4 (Maximum Likelihood Estimator, MLE) The maximum likelihood estimator (MLE) of θ for a given observation $\boldsymbol{Y} \in \mathbb{Y}$ is given by (subject to existence and uniqueness)

$$\widehat{\theta}^{\mathrm{MLE}} \;=\; \arg\max_{\widetilde{\theta} \in \boldsymbol{\Theta}} \; p(\boldsymbol{Y}; \widetilde{\theta}) \;=\; \arg\max_{\widetilde{\theta} \in \boldsymbol{\Theta}} \; \ell_{\boldsymbol{Y}}(\widetilde{\theta}),$$

where the log-likelihood function of $p(\boldsymbol{Y}; \theta)$ is defined by $\theta \mapsto \ell_{\boldsymbol{Y}}(\theta) = \log p(\boldsymbol{Y}; \theta)$.

The MLE $\boldsymbol{Y} \mapsto \widehat{\theta}^{\mathrm{MLE}} = \widehat{\theta}^{\mathrm{MLE}}(\boldsymbol{Y}) = A(\boldsymbol{Y})$ is nothing else than a specific decision rule with action space $\mathbb{A} = \boldsymbol{\Theta}$ for estimating θ. We can now start to explore the risk function $\mathcal{R}(\theta, \widehat{\theta}^{\mathrm{MLE}})$ of that decision rule for a given loss function L.

Example 3.5 (MLE within the EDF) We emphasize that this example is used throughout these notes. Assume that the (independent) components of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \sim P(\cdot; \theta)$ follow a given EDF distribution. That is, we assume that $Y_1, \ldots, Y_n$ are independent and have densities w.r.t. σ-finite measures on $\mathbb{R}$ given by, see (2.14),

$$Y_i \sim f(y_i; \theta, v_i/\varphi) = \exp\left\{\frac{y_i\theta - \kappa(\theta)}{\varphi/v_i} + a(y_i; v_i/\varphi)\right\},$$

for $1 \le i \le n$. Note that these random variables are not i.i.d. because they may differ in exposures $v_i > 0$. Throughout, we assume that Assumption 2.6 is fulfilled and that the cumulant function κ is steep, see Theorem 2.19. For the latter we also refer to Remark 2.20: the supports $\mathfrak{T}_{v_i/\varphi}$ of Y_i may differ; however, these supports share the same convex closure.

Independence between the Y_i's implies that the joint probability $P(\cdot; \theta)$ is the product distribution of the individual distributions $F(\cdot; \theta, v_i/\varphi)$, $1 \le i \le n$. Therefore, the MLE of θ in the EDF is found by solving

$$\widehat{\theta}^{\mathrm{MLE}} = \operatorname*{arg\,max}_{\widetilde{\theta} \in \boldsymbol{\Theta}} \ell_{\boldsymbol{Y}}(\widetilde{\theta}) = \operatorname*{arg\,max}_{\widetilde{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \frac{Y_i\widetilde{\theta} - \kappa(\widetilde{\theta})}{\varphi/v_i}.$$

Since the cumulant function κ is strictly convex we receive the MLE (subject to existence)

$$\widehat{\theta}^{\mathrm{MLE}} = \widehat{\theta}^{\mathrm{MLE}}(\boldsymbol{Y}) = (\kappa')^{-1}\left(\frac{\sum_{i=1}^n v_i Y_i}{\sum_{i=1}^n v_i}\right) = h\left(\frac{\sum_{i=1}^n v_i Y_i}{\sum_{i=1}^n v_i}\right).$$

Thus, the MLE is received by applying the canonical link $h = (\kappa')^{-1}$, see Definition 2.8, and strict convexity of κ implies that the MLE is unique. However, existence needs to be analyzed more carefully! It may happen that the MLE $\widehat{\theta}^{\mathrm{MLE}}$ is a boundary point of the effective domain $\boldsymbol{\Theta}$ which may not exist (if $\boldsymbol{\Theta}$ is open). We give an example. Assume we work in the Poisson model presented in Sect. 2.1.2. The canonical link in the Poisson model is the log-link $\mu \mapsto h(\mu) = \log(\mu)$, for $\mu > 0$. With positive probability we have in the Poisson case $\sum_{i=1}^n v_i Y_i = 0$.

Therefore, with positive probability the MLE $\widehat{\theta}^{\mathrm{MLE}}$ does not exist (we have a degenerate Poisson model in that case).

Since the canonical link is strictly increasing we can also perform MLE in the dual (mean) parametrization. The dual parameter space is given by $\mathcal{M} = \kappa'(\mathring{\boldsymbol{\Theta}})$, see Remarks 2.9, with mean parameters $\mu = \kappa'(\theta) \in \mathcal{M}$. This motivates

$$\widehat{\mu}^{\mathrm{MLE}} \;=\; \operatorname*{arg\,max}_{\widetilde{\mu}\in\mathcal{M}} \; \ell_{\boldsymbol{Y}}(h(\widetilde{\mu})) \;=\; \operatorname*{arg\,max}_{\widetilde{\mu}\in\mathcal{M}} \; \sum_{i=1}^{n} \frac{Y_i h(\widetilde{\mu}) - \kappa(h(\widetilde{\mu}))}{\varphi/v_i}. \tag{3.4}$$

Subject to existence, this provides the unique MLE

$$\widehat{\mu}^{\mathrm{MLE}} \;=\; \widehat{\mu}^{\mathrm{MLE}}(\boldsymbol{Y}) \;=\; \frac{\sum_{i=1}^{n} v_i Y_i}{\sum_{i=1}^{n} v_i}. \tag{3.5}$$

Also this dual MLE does not need to exist (in the dual parameter space $\mathcal{M}$). Under the assumption that the cumulant function κ is steep, we know that the closure of the dual parameter space $\overline{\mathcal{M}}$ contains the supports $\mathfrak{T}_{v_i/\varphi}$ of Y_i, see Theorem 2.19 and Remark 2.20. Thus, in that case we can close the dual parameter space and receive MLE $\widehat{\mu}^{\mathrm{MLE}} \in \overline{\mathcal{M}}$ (in a possibly degenerate model). In the aforementioned degenerate Poisson situation we receive $\widehat{\mu}^{\mathrm{MLE}} = 0$ which is in the boundary $\partial\mathcal{M}$ of the dual parameter space. ■

Definition 3.6 (Bayesian Estimator) The Bayesian estimator of θ for a given observation $\boldsymbol{Y} \in \mathbb{Y}$ and a given prior distribution π on $\boldsymbol{\Theta}$ is given by (subject to existence)

$$\widehat{\theta}^{\mathrm{Bayes}} \;=\; \widehat{\theta}^{\mathrm{Bayes}}(\boldsymbol{Y}) \;=\; \mathbb{E}_{\pi}[\theta|\boldsymbol{Y}],$$

where the conditional expectation on the right-hand side is calculated under the posterior distribution $\pi(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y};\theta)\pi(\theta)$ for a given observation $\boldsymbol{Y} = \boldsymbol{y}$.

Example 3.7 (Bayesian Estimator) Assume that $\mathbb{A} = \boldsymbol{\Theta} = \mathbb{R}$ and choose the square loss function $L(\theta, a) = (\theta - a)^2$. Assume that for ν-a.e. $\boldsymbol{y} \in \mathbb{Y}$ the following decision rule $A : \mathbb{Y} \to \mathbb{A}$ exists

$$A(\boldsymbol{y}) \;=\; \operatorname*{arg\,min}_{a\in\mathbb{A}} \; \mathbb{E}_{\pi}[(\theta - a)^2|\boldsymbol{Y} = \boldsymbol{y}], \tag{3.6}$$

where the expectation is calculated w.r.t. the posterior distribution $\pi(\theta|\boldsymbol{y})$. In this case, A is a Bayesian decision rule w.r.t. π and $L(\theta, a) = (\theta - a)^2$: by assumption (3.6) we have for any other decision rule $\widetilde{A} : \mathbb{Y} \to \mathbb{A}$, ν-a.s.,

$$\mathbb{E}_\pi[(\theta - A(\boldsymbol{Y}))^2|\boldsymbol{Y} = \boldsymbol{y}] \leq \mathbb{E}_\pi[(\theta - \widetilde{A}(\boldsymbol{Y}))^2|\boldsymbol{Y} = \boldsymbol{y}].$$

Applying the tower property we receive for any other decision rule $\widetilde{A}$

$$\int_{\boldsymbol{\Theta}} \mathcal{R}(\theta, A)d\pi(\theta) = \mathbb{E}[(\theta - A(\boldsymbol{Y}))^2] \leq \mathbb{E}[(\theta - \widetilde{A}(\boldsymbol{Y}))^2] = \int_{\boldsymbol{\Theta}} \mathcal{R}(\theta, \widetilde{A})d\pi(\theta),$$

where the expectation $\mathbb{E}$ is calculated over the joint distribution of $\boldsymbol{Y}$ and θ. This proves that A is a Bayesian decision rule w.r.t. π and $L(\theta, a) = (\theta - a)^2$, see Example 3.3. Finally, note that the conditional expectation given in Definition 3.6 is the minimizer of (3.6). This justifies the name Bayesian estimator in Definition 3.6 (for the square loss function). The case of the Bayesian estimator for a general loss function L is considered in Theorem 4.1.1 of Lehmann [244]. ■

Definition 3.8 (Method of Moments Estimator) Assume that $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ and that the components Y_i of $\boldsymbol{Y}$ are i.i.d. $F(\cdot; \boldsymbol{\theta})$ distributed with finite k-th moments for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The law of large numbers provides, a.s., for all $1 \leq l \leq k$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i^l = \mathbb{E}_{\boldsymbol{\theta}}[Y_1^l].$$

Assume that the following map is invertible (on suitable range definitions for (3.7)–(3.8))

$$\gamma : \boldsymbol{\Theta} \to \mathbb{R}^k, \qquad \boldsymbol{\theta} \mapsto \gamma(\boldsymbol{\theta}) = (\mathbb{E}_{\boldsymbol{\theta}}[Y_1], \ldots, \mathbb{E}_{\boldsymbol{\theta}}[Y_1^k])^\top. \tag{3.7}$$

The method of moments estimator of $\boldsymbol{\theta}$ is defined by

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MM}} = \widehat{\boldsymbol{\theta}}^{\mathrm{MM}}(\boldsymbol{Y}) = \gamma^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \ldots, \frac{1}{n}\sum_{i=1}^{n} Y_i^k\right)^\top. \tag{3.8}$$

The MLE, the Bayesian estimator and the method of moments estimator are the most commonly used parameter estimators. They may have additional properties (under certain assumptions) that we are going to explore below. In the remainder of this section we give an additional view on estimators which is based on the empirical distribution of the observation $\boldsymbol{Y}$.

Assume that the components Y_i of $\boldsymbol{Y}$ are real-valued and i.i.d. F distributed. The empirical distribution induced by the observation $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is given by

$$\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}} \qquad \text{for } y \in \mathbb{R}, \tag{3.9}$$

we also refer to Fig. 1.2 (lhs). The Glivenko–Cantelli theorem [64, 159] tells us that the empirical distribution $\widehat{F}_n$ converges uniformly to F, a.s., for $n \to \infty$.

Definition 3.9 (Fisher-Consistency) Denote by $\mathfrak{P}$ the set of all distribution functions on the given probability space. Let $Q : \mathfrak{P} \to \boldsymbol{\Theta}$ be a functional with the property

$$Q(F(\cdot; \theta)) = \theta \qquad \text{for all } F(\cdot; \theta) \in \mathcal{F} = \{F(\cdot; \theta); \theta \in \boldsymbol{\Theta}\} \subset \mathfrak{P}.$$

Such a functional is called *Fisher-consistent* for $\mathcal{F}$ and $\theta \in \boldsymbol{\Theta}$, respectively.

A given Fisher-consistent functional Q motivates the estimator $\widehat{\theta} = Q(\widehat{F}_n) \in \boldsymbol{\Theta}$. This is exactly what we have applied for the method of moments estimator (3.8) with Fisher-consistent functional induced by the inverse of (3.7). The next example shows that this also works for MLE.

Example 3.10 (MLE and Kullback–Leibler (KL) Divergence) The MLE can be received from a Fisher-consistent functional. Consider for $F \in \mathfrak{P}$ the functional

$$Q(F) = \arg\max_{\widetilde{\theta}} \int \log f(y; \widetilde{\theta}) dF(y),$$

assuming that $f(\cdot; \widetilde{\theta})$ are densities w.r.t. a σ-finite measure on $\mathbb{R}$. Assume that F has density f w.r.t. the σ-finite measure ν on $\mathbb{R}$. Then, we can rewrite the above as

$$Q(F) = \arg\min_{\widetilde{\theta}} \int \log \left(\frac{f(y)}{f(y; \widetilde{\theta})} \right) f(y) d\nu(y) = \arg\min_{\widetilde{\theta}} D_{\mathrm{KL}}(f || f(\cdot; \widetilde{\theta})).$$

The latter is the Kullback–Leibler (KL) divergence which we have met in Sect. 2.3. Lemma 2.21 states that the KL divergence is non-negative, and it is zero if and only if the two densities f and $f(\cdot; \widetilde{\theta})$ are identical, ν-a.s. This implies that $Q(F(\cdot; \theta)) = \theta$. Thus, Q is Fisher-consistent for $\theta \in \boldsymbol{\Theta}$, assuming identifiability, see Remarks 3.1.

Next, we use this Fisher-consistent functional (KL divergence) to receive the MLE. Replace the unknown distribution F by the empirical one to receive

$$\begin{aligned} Q(\widehat{F}_n) &= \arg\min_{\widetilde{\theta}} D_{\mathrm{KL}}(\widehat{f}_n || f(\cdot; \widetilde{\theta})) \\ &= \arg\max_{\widetilde{\theta}} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i; \widetilde{\theta}) \;=\; \widehat{\theta}^{\mathrm{MLE}}, \end{aligned}$$

where we have used that the empirical density $\widehat{f}_n$ allocates point masses of size $1/n$ to the i.i.d. observations $Y_1, \ldots, Y_n$. Thus, the MLE $\widehat{\theta}^{\text{MLE}}$ of θ can be obtained by choosing the model $f(\cdot; \widetilde{\theta})$, $\widetilde{\theta} \in \boldsymbol{\Theta}$, that is closest in KL divergence to the empirical distribution $\widehat{F}_n$ of i.i.d. observations $Y_i \sim F$. Note that in this construction we do not assume that the true distribution F is in $\mathcal{F}$, see Definition 3.9. ■

Remarks 3.11

- Many properties of estimators of θ are based on properties of Fisher-consistent functionals Q (in cases where they exist). For instance, asymptotic properties as $n \to \infty$ are obtained from smoothness properties of Fisher-consistent functionals Q, or using the influence function we can analyze the impact of individual observations Y_i on decision rules $\widehat{\theta} = \widehat{\theta}(\boldsymbol{Y}) = Q(\widehat{F}_n)$. The latter is the basis of robust statistics, see Huber [194] and Hampel et al. [180]. Since Fisher-consistent functionals do not require that the true distribution belongs to $\mathcal{F}$ it requires a careful consideration of the quantity to be estimated.
- The discussion on parameter estimation has implicitly assumed that the true data generating model belongs to the family $\mathcal{P} = \{P(\cdot; \theta); \theta \in \boldsymbol{\Theta}\}$, and the only problem was to find the true parameter in $\boldsymbol{\Theta}$. More generally, one should also consider model uncertainty w.r.t. the chosen family $\mathcal{P}$, i.e., the data generating model may not belong to this family. Of course, this problem is by far more difficult. We explore this in more detail in Sect. 11.1.4, below.

3.3 Unbiased Estimators

We introduce the property of uniformly minimum variance unbiased (UMVU) for decision rules in this section. This is a very attractive property in insurance pricing because it gives a quality statement to decision rules (and to the resulting prices). At the current stage it is not clear how unbiasedness is related, e.g., to the MLE of θ.

3.3.1 Cramér–Rao Information Bound

Above we have stated some quality criteria for decision rules like the minimax property. A crucial property in financial applications is the so-called *unbiasedness* (for mean estimates) because this guarantees that the overall (price) levels are correctly specified.

Definition 3.12 (Uniformly Minimum Variance Unbiased, UMVU) A decision rule $A : \mathbb{Y} \to \mathbb{A} = \mathbb{R}$ is unbiased for $\gamma : \boldsymbol{\Theta} \to \mathbb{R}$ if for all $\boldsymbol{Y} \sim P(\cdot;\theta)$, $\theta \in \boldsymbol{\Theta}$, we have

$$\mathbb{E}_\theta[A(\boldsymbol{Y})] = \gamma(\theta). \tag{3.10}$$

The decision rule A is called UMVU for γ if additionally to the unbiasedness (3.10) we have

$$\mathrm{Var}_\theta(A(\boldsymbol{Y})) \le \mathrm{Var}_\theta(\widetilde{A}(\boldsymbol{Y})),$$

for all $\theta \in \boldsymbol{\Theta}$ and for any other decision rule $\widetilde{A} : \mathbb{Y} \to \mathbb{R}$ that is unbiased for γ.

Note that unbiasedness is not invariant under transformations, i.e., if $A(\boldsymbol{Y})$ is unbiased for $\gamma(\theta)$, then, in general, $b(A(\boldsymbol{Y}))$ is not unbiased for $b(\gamma(\theta))$. For instance, if b is strictly convex then we get a counterexample by simply applying Jensen's inequality.

Our first step is to derive a general lower bound for $\mathrm{Var}_\theta(A(\boldsymbol{Y}))$. If this general lower bound is met for an unbiased decision rule A for γ, then we know that it is UMVU for γ. We start with the one-dimensional case given in Section 2.6 of Lehmann [244].

Theorem 3.13 (Cramér–Rao Information Bound) *Assume that the distributions $P(\cdot;\theta)$, $\theta \in \boldsymbol{\Theta}$, have densities $p(\cdot;\theta)$ for a given σ-finite measure ν on $\mathbb{Y}$, and that $\boldsymbol{\Theta} \subset \mathbb{R}$ is an open interval such that the set $\{\boldsymbol{y};\ p(\boldsymbol{y};\theta) > 0\}$ does not depend on $\theta \in \boldsymbol{\Theta}$. Let $A(\boldsymbol{Y})$ be unbiased for $\gamma : \boldsymbol{\Theta} \to \mathbb{R}$ having finite second moment. If the limit*

$$\frac{\partial}{\partial\theta}\log p(\boldsymbol{y};\theta) = \lim_{\Delta\to 0}\frac{1}{\Delta}\frac{p(\boldsymbol{y};\theta+\Delta) - p(\boldsymbol{y};\theta)}{p(\boldsymbol{y};\theta)}$$

exists in $L^2(P(\cdot;\theta))$ and if

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{Y};\theta)\right)^2\right] \in (0,\infty),$$

then the function $\theta \mapsto \gamma(\theta)$ is differentiable, $\mathbb{E}_\theta[\frac{\partial}{\partial\theta}\log p(\boldsymbol{Y};\theta)] = 0$ and we have information bound

$$\mathrm{Var}_\theta(A(\boldsymbol{Y})) \ge \frac{\gamma'(\theta)^2}{\mathcal{I}(\theta)}.$$

Proof We start from an arbitrary function $\psi : \boldsymbol{\Theta} \times \mathbb{Y} \to \mathbb{R}$ with finite variance $\mathrm{Var}_\theta(\psi(\theta, \boldsymbol{Y})) \in (0, \infty)$ for all $\theta \in \boldsymbol{\Theta}$. The Cauchy–Schwarz inequality implies

$$\mathrm{Var}_\theta(A(\boldsymbol{Y})) \geq \frac{\mathrm{Cov}_\theta(A(\boldsymbol{Y}), \psi(\theta, \boldsymbol{Y}))^2}{\mathrm{Var}_\theta(\psi(\theta, \boldsymbol{Y}))}. \tag{3.11}$$

If we manage to make the right-hand side of (3.11) independent of decision rule $A(\cdot)$ we have a general lower bound, we also refer to Theorem 2.6.1 in Lehmann [244].

The Cauchy–Schwarz inequality implies that for any $U \in L^2(P(\cdot;\theta))$ the following limit exists and is equal to

$$\lim_{\Delta \to 0} \mathbb{E}_\theta \left[\frac{1}{\Delta} \frac{p(\boldsymbol{Y}; \theta + \Delta) - p(\boldsymbol{Y}; \theta)}{p(\boldsymbol{Y}; \theta)} U \right] = \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log p(\boldsymbol{Y}; \theta) U \right]. \tag{3.12}$$

Setting $U \equiv 1$ gives average score $\mathbb{E}_\theta[\frac{\partial}{\partial \theta} \log p(\boldsymbol{Y}; \theta)] = 0$ because for sufficiently small Δ

$$\mathbb{E}_\theta \left[\frac{p(\boldsymbol{Y}; \theta + \Delta) - p(\boldsymbol{Y}; \theta)}{p(\boldsymbol{Y}; \theta)} \right] = \int_{\mathbb{Y}} \frac{p(\boldsymbol{y}; \theta + \Delta) - p(\boldsymbol{y}; \theta)}{p(\boldsymbol{y}; \theta)} p(\boldsymbol{y}; \theta) d\nu(\boldsymbol{y}) = 0,$$

where we have used that the support of the random variables does not depend on θ and that the domain $\boldsymbol{\Theta}$ of θ is open.

Secondly, we set $U = A(\boldsymbol{Y})$ in (3.12). We have similarly to above using unbiasedness w.r.t. γ

$$\begin{aligned} \mathrm{Cov}_\theta \left(A(\boldsymbol{Y}), \frac{p(\boldsymbol{Y}; \theta + \Delta) - p(\boldsymbol{Y}; \theta)}{p(\boldsymbol{Y}; \theta)} \right) &= \int_{\mathbb{Y}} A(\boldsymbol{y}) \frac{p(\boldsymbol{y}; \theta + \Delta) - p(\boldsymbol{y}; \theta)}{p(\boldsymbol{y}; \theta)} p(\boldsymbol{y}; \theta) d\nu(\boldsymbol{y}) \\ &= \gamma(\theta + \Delta) - \gamma(\theta). \end{aligned}$$

Existence of limit (3.12) provides the differentiability of γ. Finally, from (3.11) we have

$$\mathrm{Var}_\theta(A(\boldsymbol{Y})) \geq \lim_{\Delta \to 0} \frac{\mathrm{Cov}_\theta \left(A(\boldsymbol{Y}), \frac{p(\boldsymbol{Y}; \theta + \Delta) - p(\boldsymbol{Y}; \theta)}{p(\boldsymbol{Y}; \theta)} \right)^2}{\mathrm{Var}_\theta \left(\frac{p(\boldsymbol{Y}; \theta + \Delta) - p(\boldsymbol{Y}; \theta)}{p(\boldsymbol{Y}; \theta)} \right)} = \frac{\gamma'(\theta)^2}{\mathcal{I}(\theta)}. \tag{3.13}$$

This completes the proof. □

Remarks 3.14 (Fisher's Information and Score)

- $\mathcal{I}(\theta)$ is called *Fisher's information* or *Fisher metric*.
- $s(\theta, \boldsymbol{Y}) = \frac{\partial}{\partial \theta} \log p(\boldsymbol{Y}; \theta)$ is called *score*, and $\mathbb{E}_\theta[s(\boldsymbol{Y}; \theta)] = 0$ in Theorem 3.13 expresses that the average score is zero under the assumptions of that theorem.
- Under the regularity conditions of Lemma 6.1 in Section 2.6 of Lehmann [244]

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{Y};\theta)\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{Y};\theta)\right]. \quad (3.14)$$

Fisher's information $\mathcal{I}(\theta)$ expresses the variance of the score $s(\theta, \boldsymbol{Y})$. Identity (3.14) justifies the notion Fisher's information in Sect. 2.3 for the EF.

- In order to determine the Cramér–Rao information bound for unknown θ we need to estimate Fisher's information $\mathcal{I}(\theta)$ from the available data. There are two different ways to do so, either we choose

$$\mathcal{I}(\widehat{\theta}) = \mathbb{E}_{\widehat{\theta}}\left[\left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{Y};\theta)\right)^2\right],$$

or we choose the *observed Fisher's information*

$$\widehat{\mathcal{I}}(\widehat{\theta}) = \left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{Y};\theta)\right)^2\Bigg|_{\theta=\widehat{\theta}},$$

for given data $\boldsymbol{Y}$ and where $\widehat{\theta} = \widehat{\theta}(\boldsymbol{Y})$. Both estimated Fisher's information $\mathcal{I}(\widehat{\theta})$ and $\widehat{\mathcal{I}}(\widehat{\theta})$ play a central role in MLE of generalized linear models (GLMs). They are used in Fisher's scoring method, the iterated re-weighted least squares (IRLS) algorithm and the Newton–Raphson algorithm to determine the MLE.

- The Cramér–Rao information bound in Theorem 3.13 is stated in terms of the observation $\boldsymbol{Y} \sim p(\cdot;\theta)$. Assume that the components Y_i of $\boldsymbol{Y}$ are i.i.d. $f(\cdot;\theta)$ distributed. In this case, Fisher's information scales as

$$\mathcal{I}(\theta) = \mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta), \quad (3.15)$$

with single risk's Fisher's information (contribution)

$$\mathcal{I}_1(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(Y_1;\theta)\right)^2\right].$$

In general, Fisher's information is additive in independent random variables, because the product of densities is additive after applying the logarithm, and because the average score is zero.

Proposition 3.15 *The unbiased decision rule A for γ attains the Cramér–Rao information bound if and only if the density is of the form $p(\boldsymbol{y};\theta) = \exp\{\delta(\theta)T(\boldsymbol{y}) - \beta(\theta) + a(\boldsymbol{y})\}$ with $T = A$. In that case we have $\gamma(\theta) = \beta'(\theta)/\delta'(\theta)$.*

Proof of Proposition 3.15 The Cauchy–Schwarz inequality provides equality in (3.13) if and only if $\frac{\partial}{\partial\theta}\log p(\boldsymbol{y};\theta) = \delta'(\theta)A(\boldsymbol{y}) - \beta'(\theta)$, ν-a.s, for some functions $\delta'(\theta)$ and $\beta'(\theta)$ on $\boldsymbol{\Theta}$. Integration and the fact that $p(\cdot;\theta)$ is a density whose support does not depend on the explicit choice of $\theta \in \boldsymbol{\Theta}$ provide the implication "$\Rightarrow$". For the implication "$\Leftarrow$" we study for $A = T$

$$0 = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p(\boldsymbol{Y};\theta)\right] = \int_{\mathbb{Y}} (\delta'(\theta)A(\boldsymbol{y}) - \beta'(\theta))p(\boldsymbol{y};\theta)d\nu(\boldsymbol{y}) = \delta'(\theta)\mathbb{E}_\theta[A(\boldsymbol{Y})] - \beta'(\theta).$$

In that case we have $\gamma(\theta) = \mathbb{E}_\theta[A(\boldsymbol{Y})] = \beta'(\theta)/\delta'(\theta)$. Moreover, we have equality in the Cauchy–Schwarz inequality. This finishes the proof. □

The single-parameter EF fulfills the properties of Proposition 3.15 with $\delta(\theta) = \theta$ and $\beta(\theta) = \kappa(\theta)$, and decision rule $A(y) = T(y)$ attains the Cramér–Rao information bound for $\gamma(\theta) = \kappa'(\theta)$.

We give a multi-dimensional version of the Cramér–Rao information bound.

Theorem 3.16 (Multi-Dimensional Version of the Cramér–Rao Information Bound, Without Proof) *Assume that the distributions $P(\cdot;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, have densities $p(\cdot;\boldsymbol{\theta})$ for a given σ-finite measure ν on $\mathbb{Y}$, and that $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ is an open convex set such that the set $\{\boldsymbol{y};\ p(\boldsymbol{y};\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $A(\boldsymbol{Y})$ be unbiased for $\gamma : \boldsymbol{\Theta} \to \mathbb{R}$ having finite second moment. Under additional regularity conditions, see Theorem 7.3 in Section 2.7 of Lehmann [244], we have*

$$\mathrm{Var}_{\boldsymbol{\theta}}(A(\boldsymbol{Y})) \geq (\nabla_{\boldsymbol{\theta}}\gamma(\boldsymbol{\theta}))^\top \mathcal{I}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\gamma(\boldsymbol{\theta}),$$

with (positive definite) Fisher's information matrix $\mathcal{I}(\boldsymbol{\theta}) = (\mathcal{I}_{l,j}(\boldsymbol{\theta}))_{1\leq l,j\leq k}$ given by

$$\mathcal{I}_{l,j}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_l}\log p(\boldsymbol{Y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log p(\boldsymbol{Y};\boldsymbol{\theta})\right],$$

for $1 \leq l, j \leq k$.

Remarks 3.17

- Whenever an unbiased decision rule $A(\boldsymbol{Y})$ for $\gamma(\boldsymbol{\theta})$ meets the Cramér–Rao information bound it is UMVU. Thus, it minimizes the risk function $\mathcal{R}(\boldsymbol{\theta}, A)$ being based on the square loss $L(\boldsymbol{\theta}, a) = (\gamma(\boldsymbol{\theta}) - a)^2$ among all unbiased decision rules, because unbiasedness for $\gamma(\boldsymbol{\theta})$ gives $\mathcal{R}(\boldsymbol{\theta}, A) = \mathrm{Var}_{\boldsymbol{\theta}}(A(\boldsymbol{Y}))$.
- The regularity conditions in Theorem 3.16 include that Fisher's information matrix $\mathcal{I}(\boldsymbol{\theta})$ is positive definite.
- Under additional regularity conditions we have the following identity for Fisher's information matrix

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[(\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}; \boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}; \boldsymbol{\theta}))^{\top}\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}; \boldsymbol{\theta})\right] \in \mathbb{R}^{k \times k}.$$

  Thus, Fisher's information matrix can either be calculated from a quadratic form of the score $s(\boldsymbol{\theta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}; \boldsymbol{\theta})$ or from the Hessian $\nabla^2_{\boldsymbol{\theta}}$ of the log-likelihood $\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \log p(\boldsymbol{Y}; \boldsymbol{\theta})$. Since the score has mean zero, Fisher's information matrix is equal to the covariance matrix of the score $s(\boldsymbol{\theta}, \boldsymbol{Y})$.

In many situations we do not work under the canonical parametrization $\boldsymbol{\theta}$. Considerations then require a change of variable. Assume that

$$\boldsymbol{\zeta} \in \mathbb{R}^r \;\mapsto\; \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\zeta}) \in \mathbb{R}^k,$$

such that all derivatives $\partial\theta_l(\boldsymbol{\zeta})/\partial\zeta_j$ exist for $1 \le l \le k$ and $1 \le j \le r$. The Jacobian matrix is given by

$$J(\boldsymbol{\zeta}) = \left(\frac{\partial}{\partial \zeta_j}\theta_l(\boldsymbol{\zeta})\right)_{1 \le l \le k, 1 \le j \le r} \in \mathbb{R}^{k \times r}.$$

Fisher's information matrix w.r.t. $\boldsymbol{\zeta}$ is given by

$$\mathcal{I}^*(\boldsymbol{\zeta}) = \left(\mathbb{E}_{\boldsymbol{\theta}(\zeta)}\left[\frac{\partial}{\partial \zeta_l} \log p(\boldsymbol{Y}; \boldsymbol{\theta}(\boldsymbol{\zeta}))\frac{\partial}{\partial \zeta_j} \log p(\boldsymbol{Y}; \boldsymbol{\theta}(\boldsymbol{\zeta}))\right]\right)_{1 \le l, j \le r} \in \mathbb{R}^{r \times r},$$

and we have the identity

$$\mathcal{I}^*(\boldsymbol{\zeta}) = J(\boldsymbol{\zeta})^{\top}\, \mathcal{I}(\boldsymbol{\theta}(\boldsymbol{\zeta}))\, J(\boldsymbol{\zeta}). \tag{3.16}$$

This formula is used quite frequently, e.g., in generalized linear models when changing the parametrization of the models.

3.3.2 Information Bound in the Exponential Family Case

The purpose of this section is to summarize the Cramér–Rao information bound results for the EF and the EDF, since these families play a distinguished role in statistical and actuarial modeling.

Cramér–Rao Information Bound in the EF Case

We start with the EF case. Assume we have i.i.d. observations $Y_1, \ldots, Y_n$ having densities w.r.t. a σ-finite measure ν on $\mathbb{R}$ given by the EF, see (2.2),

$$dF(y;\boldsymbol{\theta}) \;=\; f(y;\boldsymbol{\theta})d\nu(y) \;=\; \exp\left\{\boldsymbol{\theta}^\top T(y) - \kappa(\boldsymbol{\theta}) + a(y)\right\} d\nu(y),$$

for canonical parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$. We assume to work under a minimal representation implying that the cumulant function κ is strictly convex on the interior $\mathring{\boldsymbol{\Theta}}$, see Assumption 2.6. Moreover, we assume that the cumulant function κ is steep in the sense of Theorem 2.19. Consider the (aggregated) *statistics* of the joint EF $\mathcal{P} = \{P(\cdot;\theta); \theta \in \boldsymbol{\Theta}\}$

$$\boldsymbol{y} \mapsto S(\boldsymbol{y}) \stackrel{\text{def.}}{=} \left(\sum_{i=1}^n T_1(y_i), \ldots, \sum_{i=1}^n T_k(y_i)\right)^\top \in \mathbb{R}^k. \tag{3.17}$$

We calculate the score of this EF

$$s(\boldsymbol{\theta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{Y};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}^\top \sum_{i=1}^n T(Y_i) - n\kappa(\boldsymbol{\theta})\right) = S(\boldsymbol{Y}) - n\nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}).$$

An immediate consequence of Corollary 2.5 is that the expected value of the score is zero for any $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$. This then reads as

$$\mu = \mathbb{E}_\theta\left[T(Y_1)\right] = \mathbb{E}_\theta\left[S(\boldsymbol{Y})/n\right] = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \in \mathbb{R}^k. \tag{3.18}$$

Thus, the statistics $S(\boldsymbol{Y})/n$ is an unbiased decision rule for the mean $\mu = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})$, and we can study its Cramér–Rao information bound. Fisher's information matrix is given by the positive definite matrix

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[s(\boldsymbol{\theta}, \boldsymbol{Y})s(\boldsymbol{\theta}, \boldsymbol{Y})^\top\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y};\boldsymbol{\theta})\right] = n\nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \in \mathbb{R}^{k\times k}.$$

Note that the multi-dimensionally extended Cramér–Rao information bound in Theorem 3.16 applies to the individual components of vector $\mu = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \in \mathbb{R}^k$. Assume we would like to estimate its j-th component, set $\gamma_j(\boldsymbol{\theta}) = \mu_j =$

$(\nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}))_j = \partial\kappa(\boldsymbol{\theta})/\partial\theta_j$, for $1 \le j \le k$. This corresponds to the j-th component $S_j(\boldsymbol{Y})$ of the statistics $S(\boldsymbol{Y})$. We have unbiasedness of $S_j(\boldsymbol{Y})/n$ for $\gamma_j(\boldsymbol{\theta}) = \mu_j = (\nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}))_j$, and this unbiased statistics attains the Cramér–Rao information bound

$$\mathrm{Var}_{\boldsymbol{\theta}}(S_j(\boldsymbol{Y})/n) = \frac{1}{n}\left(\nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\right)_{j,j} = (\nabla_{\boldsymbol{\theta}}\gamma_j(\boldsymbol{\theta}))^\top \mathcal{I}(\boldsymbol{\theta})^{-1}(\nabla_{\boldsymbol{\theta}}\gamma_j(\boldsymbol{\theta})). \tag{3.19}$$

Recall that $\mathcal{I}(\boldsymbol{\theta})^{-1}$ scales as n^{-1}, see (3.15). This provides us with the following corollary.

Corollary 3.18 *Assume $Y_1, \ldots, Y_n$ are i.i.d. and follow an EF (under a minimal representation). The components of the statistics $S(\boldsymbol{Y})/n$ are UMVU for $\gamma_j(\boldsymbol{\theta}) = \partial\kappa(\boldsymbol{\theta})/\partial\theta_j$, $1 \le j \le k$ and $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$, with*

$$\mathrm{Var}_{\boldsymbol{\theta}}\left(\frac{1}{n}S_j(\boldsymbol{Y})\right) = \frac{1}{n}\,\frac{\partial^2}{\partial\theta_j^2}\kappa(\boldsymbol{\theta}).$$

The corresponding covariance terms are for $1 \le j, l \le k$ given by

$$\mathrm{Cov}_{\boldsymbol{\theta}}\left(\frac{1}{n}S_j(\boldsymbol{Y}), \frac{1}{n}S_l(\boldsymbol{Y})\right) = \frac{1}{n}\,\frac{\partial^2}{\partial\theta_j\partial\theta_l}\kappa(\boldsymbol{\theta}).$$

The UMVU property stated in Corollary 3.18 is, in general, not related to MLE, but within the EF there is the following link. We have (subject to existence)

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}} = \operatorname*{arg\,max}_{\widetilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta}} p(\boldsymbol{Y};\widetilde{\boldsymbol{\theta}}) = \operatorname*{arg\,max}_{\widetilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta}} \left(\widetilde{\boldsymbol{\theta}}^\top S(\boldsymbol{Y}) - n\kappa(\widetilde{\boldsymbol{\theta}})\right) = h\left(\frac{1}{n}S(\boldsymbol{Y})\right), \tag{3.20}$$

where $h = (\nabla_{\boldsymbol{\theta}}\kappa)^{-1}$ is the canonical link of this EF, see Definition 2.8; and where we need to ensure that a solution to (3.20) exists; e.g., the solution to (3.20) might be at the boundary of $\boldsymbol{\Theta}$ which may cause problems, see Example 3.5.[1] Because the cumulant function κ is strictly convex (in a minimal representation), we receive the

[1] Another example where there does not exist a proper solution to the MLE problem (3.20) is, for instance, obtained within the 2-dimensional Gaussian EF if we have only one single observation Y_1. Intuitively this is clear because we cannot estimate two parameters from one observation $T(Y_1) = (Y_1, Y_1^2)$.

MLE for the mean parameter $\mu = \mathbb{E}_\theta [T(Y_1)]$

$$\widehat{\mu}^{\text{MLE}} = \underset{\widetilde{\mu} \in \overline{\mathcal{M}}}{\arg\max} \left(h(\widetilde{\mu})^\top S(\boldsymbol{Y}) - n\kappa(h(\widetilde{\mu})) \right) = \frac{1}{n} S(\boldsymbol{Y}),$$

the dual parameter space $\mathcal{M} = \nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{\Theta}) \subseteq \mathbb{R}^k$ has been introduced in Remarks 2.9. If $S(\boldsymbol{Y})/n$ is contained in $\mathcal{M}$, then this MLE is a proper solution; otherwise, because we have assumed that the cumulant function κ is steep, the MLE exists in the closure $\overline{\mathcal{M}}$, see Theorem 2.19, and it is UMVU for μ, see Corollary 3.18.

Corollary 3.19 (Balance Property) *Assume $Y_1, \ldots, Y_n$ are i.i.d. and follow an EF with $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$ and $T(Y_i) \in \overline{\mathcal{M}}$, a.s. The MLE $\widehat{\mu}^{\text{MLE}} \in \overline{\mathcal{M}}$ is UMVU for μ, and it fulfills the balance property on portfolio level, i.e.,*

$$\sum_{i=1}^n \mathbb{E}_{\widehat{\mu}^{\text{MLE}}} [T(Y_i)] = n\widehat{\mu}^{\text{MLE}} = S(\boldsymbol{Y}).$$

Remarks 3.20

- The balance property is a very important property in insurance pricing because it implies that the portfolio is priced on the right level: we have unbiasedness

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^n \mathbb{E}_{\widehat{\mu}^{\text{MLE}}} [T(Y_i)] \right] = \mathbb{E}_{\boldsymbol{\theta}} [S(\boldsymbol{Y})] = n\mu. \tag{3.21}$$

- We emphasize that the balance property is much stronger than unbiasedness (3.21), note that the balance property provides unbiasedness even if $\boldsymbol{Y}$ follows a completely different model, i.e., even if the chosen EF $\mathcal{P}$ is completely misspecified.
- In general, the MLE $\widehat{\boldsymbol{\theta}}^{\text{MLE}}$ is not unbiased for $\boldsymbol{\theta}$. E.g., if the canonical link $h = (\nabla_{\boldsymbol{\theta}} \kappa)^{-1}$ is strictly concave, we have from Jensen's inequality, subject to existence at the boundary of $\boldsymbol{\Theta}$,

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\widehat{\boldsymbol{\theta}}^{\text{MLE}} \right] = \mathbb{E}_{\boldsymbol{\theta}} \left[h \left(\frac{1}{n} S(\boldsymbol{Y}_n) \right) \right] < h \left(\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{1}{n} S(\boldsymbol{Y}_n) \right] \right) = h(\mu) = \boldsymbol{\theta}. \tag{3.22}$$

- The statistics $S(\boldsymbol{Y})$ is a sufficient statistics of $\boldsymbol{Y}$, this follows from the factorization criterion; see Theorem 1.5.2 of Lehmann [244].

Cramér–Rao Information Bound in the EDF Case

The single-parameter linear EDF case is very similar to the above vector-valued parameter EF case. We briefly summarize the main results in the EDF case.

Recall Example 3.5: assume that $Y_1, \ldots, Y_n$ are independent having densities w.r.t. a σ-finite measures on $\mathbb{R}$ (not being concentrated in a single point) given by, see (2.14),

$$Y_i \sim f(y_i; \theta, v_i/\varphi) = \exp\left\{\frac{y_i\theta - \kappa(\theta)}{\varphi/v_i} + a(y_i; v_i/\varphi)\right\}, \tag{3.23}$$

for $1 \le i \le n$. Note that these random variables are not i.i.d. because they may differ in the exposures $v_i > 0$. The MLE of $\mu = \kappa'(\theta)$, $\theta \in \mathring{\boldsymbol{\Theta}}$, is found by, see (3.5),

$$\widehat{\mu}^{\mathrm{MLE}} = \underset{\widetilde{\mu}\in\overline{\mathcal{M}}}{\arg\max} \sum_{i=1}^{n} \frac{Y_i h(\widetilde{\mu}) - \kappa(h(\widetilde{\mu}))}{\varphi/v_i} = \frac{\sum_{i=1}^n v_i Y_i}{\sum_{i=1}^n v_i}, \tag{3.24}$$

we assume that κ is steep to ensure $\widehat{\mu}^{\mathrm{MLE}} \in \overline{\mathcal{M}}$. The convolution formula of Corollary 2.15 says that the MLE $\widehat{\mu}^{\mathrm{MLE}} = Y_+$ belongs to the same EDF with the same canonical parameter θ and the same dispersion φ, only the weight changes to $v_+ = \sum_{i=1}^n v_i$.

Corollary 3.21 (Balance Property) *Assume $Y_1, \ldots, Y_n$ are independent with EDF distribution (3.23) for $\theta \in \mathring{\boldsymbol{\Theta}}$ and $Y_i \in \overline{\mathcal{M}}$, a.s. The MLE $\widehat{\mu}^{\mathrm{MLE}} \in \overline{\mathcal{M}}$ is UMVU for $\mu = \kappa'(\theta)$, and it fulfills the balance property on portfolio level, i.e.,*

$$\sum_{i=1}^{n} \mathbb{E}_{\widehat{\mu}^{\mathrm{MLE}}}[v_i Y_i] = \sum_{i=1}^{n} v_i \widehat{\mu}^{\mathrm{MLE}} = \sum_{i=1}^{n} v_i Y_i.$$

The score in this EDF is given by

$$s(\theta, \boldsymbol{Y}) = \frac{\partial}{\partial\theta} \log p(\boldsymbol{Y}; \theta) = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \frac{v_i}{\varphi} (\theta Y_i - \kappa(\theta)) = \sum_{i=1}^{n} \frac{v_i}{\varphi} \left(Y_i - \kappa'(\theta)\right).$$

Of course, we have $\mathbb{E}_\theta[s(\theta, \boldsymbol{Y})] = 0$ and we receive Fisher's information for $\theta \in \mathring{\boldsymbol{\Theta}}$

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log p(\boldsymbol{Y}; \theta)\right] = \sum_{i=1}^{n} \frac{v_i}{\varphi} \kappa''(\theta) > 0. \tag{3.25}$$

Corollary 2.15 gives for the variance of the MLE

$$\mathrm{Var}_\theta\left(\widehat{\mu}^{\mathrm{MLE}}\right) = \frac{\varphi}{\sum_{i=1}^n v_i}\kappa''(\theta) = \frac{(\kappa''(\theta))^2}{\mathcal{I}(\theta)} = \frac{(\partial\mu(\theta)/\partial\theta)^2}{\mathcal{I}(\theta)}.$$

This verifies that $\widehat{\mu}^{\mathrm{MLE}}$ meets the Cramér–Rao information bound and is UMVU for the mean $\mu = \kappa'(\theta)$.

Example 3.22 (Poisson Case) For this example, we consider independent Poisson random variables $N_i \sim \mathrm{Poi}(v_i\lambda)$. In Sect. 2.2.2 we have seen that $Y_i = N_i/v_i$ can be modeled within the single-parameter linear EDF framework using as cumulant function the exponential function $\kappa(\theta) = e^\theta$, and setting $\omega_i = v_i$ and $\varphi = 1$. Thus, the probability weights of a single observation Y_i are given by, see (2.15),

$$f(y_i;\theta,v_i) = \exp\left\{v_i\left(\theta y_i - e^\theta\right) + a(y_i;v_i)\right\},$$

with canonical parameter $\theta = \log(\lambda) \in \boldsymbol{\Theta} = \mathbb{R}$. The MLE in the mean parametrization is given by, see (3.24),

$$\widehat{\lambda}^{\mathrm{MLE}} = \frac{\sum_{i=1}^n v_i Y_i}{\sum_{i=1}^n v_i} = \frac{\sum_{i=1}^n N_i}{\sum_{i=1}^n v_i} \in \overline{\mathcal{M}} = [0,\infty).$$

This estimator is unbiased for λ. Having independent Poisson random variables we can calculate the variance of this estimator as

$$\mathrm{Var}\left(\widehat{\lambda}^{\mathrm{MLE}}\right) = \frac{\lambda}{\sum_{i=1}^n v_i}.$$

Moreover, from Corollary 3.21 we know that this estimator is UMVU for λ, which can easily be seen, and uses Fisher's information (3.25) with dispersion parameter $\varphi = 1$

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{Y};\theta)\right] = \sum_{i=1}^n v_i\kappa''(\theta) = \lambda\sum_{i=1}^n v_i.$$

■

One could study many other properties of decision rules (and corresponding estimators), for instance, admissibility or uniformly minimum risk equivariance (UMRE), and we could also study other families of distribution functions such as group families. We refrain from doing so because we will not need this for our purposes.

3.4 Asymptotic Behavior of Estimators

All results above have been based on a finite sample $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top$, we add a lower index n to $\boldsymbol{Y}_n$ to indicate the finite sample size $n \in \mathbb{N}$. The aim of this section is to analyze properties of decision rules when the sample size n tends to infinity.

3.4.1 Consistency

Assume we have an infinite sequence of observations Y_i, $i \geq 1$, which allows us to construct an infinite sequence of decision rules $A_n = A_n(\boldsymbol{Y}_n)$, $n \geq 1$, where A_n always considers the first n observations $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top \sim P_n(\cdot; \theta)$, for $\theta \in \boldsymbol{\Theta}$ not depending on n. To fix ideas, one may think of i.i.d. random variables Y_i.

Definition 3.23 (Consistency) The sequence $A_n = A_n(\boldsymbol{Y}_n) \in \mathbb{R}^r$, $n \geq 1$, is consistent for $\gamma : \boldsymbol{\Theta} \to \mathbb{R}^r$ if for all $\theta \in \boldsymbol{\Theta}$ and for all $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}_\theta \left[\|A_n(\boldsymbol{Y}_n) - \gamma(\theta)\|_2 > \varepsilon \right] = 0.$$

Definition 3.23 says that $A_n(\boldsymbol{Y}_n)$ converges in probability to $\gamma(\theta)$ as $n \to \infty$. If we (even) have a.s. convergence, we call $A_n, n \geq 1$, *strongly consistent* for $\gamma : \boldsymbol{\Theta} \to \mathbb{R}^r$. Consistency is a minimal property that decision rules should fulfill. Typically, in applications, this is not enough, and we are interested in (fast) rates of convergence, i.e., we would like to know the error rates between $A_n(\boldsymbol{Y}_n)$ and $\gamma(\theta)$ for $n \to \infty$.

Example 3.24 (Consistency of the MLE in the EF) We revisit Corollary 3.19 and consider an i.i.d. sequence of random variables Y_i, $i \geq 1$, belonging to an EF, and we assume to work under a minimal representation and to have a steep cumulant function κ. The MLE for μ is given by the statistics

$$\widehat{\mu}_n^{\mathrm{MLE}} = \frac{1}{n} S(\boldsymbol{Y}_n) = \frac{1}{n} \sum_{i=1}^n (T_1(Y_i), \ldots, T_k(Y_i))^\top \in \overline{\mathcal{M}}.$$

We add a lower index n to the MLE to indicate the sample size. The i.i.d. property of Y_i, $i \geq 1$, implies that we can apply the strong law of large numbers which tells us that we have $\lim_{n \to \infty} \widehat{\mu}_n^{\mathrm{MLE}} = \mathbb{E}_{\boldsymbol{\theta}} [T(Y_1)] = \nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta}) = \mu$, a.s., for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This implies strong consistency of the sequence of MLEs $\widehat{\mu}_n^{\mathrm{MLE}}$, $n \geq 1$, for μ.

We have seen that these MLEs are also UMVU for μ, but if we transform them to the canonical scale $\widehat{\boldsymbol{\theta}}_n^{\mathrm{MLE}}$ they are, in general, biased for $\boldsymbol{\theta}$, see (3.22). However, since the cumulant function κ is strictly convex (under a minimal representation) we receive $\lim_{n \to \infty} \widehat{\boldsymbol{\theta}}_n^{\mathrm{MLE}} = \boldsymbol{\theta}$, a.s., which provides strong consistency also on the canonical scale. ■

Proposition 3.25 *Assume the real-valued random variables Y_i, $i \geq 1$, are i.i.d. $F(\cdot;\theta)$ distributed with fixed $\theta \in \boldsymbol{\Theta}$. The resulting empirical distributions $\widehat{F}_n$, $n \geq 1$, are given by (3.9). Assume Q is a Fisher-consistent functional for $\gamma(\theta)$, i.e., $Q(F(\cdot;\theta)) = \gamma(\theta)$ for all $\theta \in \boldsymbol{\Theta}$. Moreover, assume that Q is continuous in $F(\cdot;\theta)$, for all $\theta \in \boldsymbol{\Theta}$, w.r.t. the supremum norm. The functionals $Q(\widehat{F}_n)$, $n \geq 1$, are consistent for $\gamma(\theta)$.*

Sketch of Proof The Glivenko–Cantelli theorem [64, 159] says that the empirical distribution $\widehat{F}_n$ converges uniformly to $F(\cdot;\theta)$, a.s., for $n \to \infty$. Using the assumptions made, we are allowed to exchange the corresponding limits, which provides consistency. □

In view of Proposition 3.25, we discuss the case of the MLE of $\theta \in \boldsymbol{\Theta}$. In Example 3.10 we have seen that the MLE of $\theta \in \boldsymbol{\Theta}$ is obtained from a Fisher-consistent functional Q for θ on the set of probability distributions $\mathfrak{P}$ given by

$$Q(F) = \arg\max_{\widetilde{\theta}} \int \log f(y;\widetilde{\theta}) dF(y) = \arg\min_{\widetilde{\theta}} D_{\mathrm{KL}}(f || f(\cdot;\widetilde{\theta})),$$

in the second step we assumed that F has a density f w.r.t. a σ-finite measure ν on $\mathbb{R}$.

Assume we have i.i.d. data $Y_i \sim f(\cdot;\theta)$, $i \geq 1$. Thus, the true data generating distribution is described by the parameter $\theta \in \boldsymbol{\Theta}$. MLE requires the study of the log-likelihood function (we scale with the sample size n)

$$\widetilde{\theta} \mapsto \frac{1}{n} \ell_{\boldsymbol{Y}_n}(\widetilde{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i;\widetilde{\theta}).$$

The law of large numbers gives us, a.s.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i;\widetilde{\theta}) = \mathbb{E}_\theta \left[\log f(Y;\widetilde{\theta}) \right]. \tag{3.26}$$

Thus, *if* we are allowed to exchange the arg max operation and the limit in $n \to \infty$ we receive, a.s.,

$$\begin{aligned} \lim_{n\to\infty} \widehat{\theta}_n^{\mathrm{MLE}} &= \lim_{n\to\infty} \left(\arg\max_{\widetilde{\theta}} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i;\widetilde{\theta}) \right) \\ &\stackrel{?}{=} \arg\max_{\widetilde{\theta}} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i;\widetilde{\theta}) \right) \\ &= \arg\max_{\widetilde{\theta}} \mathbb{E}_\theta \left[\log f(Y;\widetilde{\theta}) \right] = Q(F(\cdot;\theta)) = \theta. \end{aligned} \tag{3.27}$$

That is, we receive consistency of the MLE for θ if we are allowed to exchange the arg max operation and the limit in $n \to \infty$. This requires regularity conditions on the considered family of distributions $\mathcal{F} = \{F(\cdot;\theta); \theta \in \mathbf{\Theta}\}$. The case of a finite parameter space $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_J\}$ is easy, this is a simplified version of Wald's [374] consistency proof,

$$\mathbb{P}_{\theta_j}\left[\theta_j \notin \arg\max_{\theta_k} \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i;\theta_k)\right] \le \sum_{k\neq j} \mathbb{P}_{\theta_j}\left[\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i;\theta_k) > \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i;\theta_j)\right].$$

The right-hand side converges to 0 as $n \to \infty$ for all $\theta_k \neq \theta_j$, which gives consistency. For regularity conditions on more general parameter spaces we refer to Section 5.2 in Van der Vaart [363]. Basically, one needs that the arg max of the limiting function given on the right-hand side of (3.26) is well-separated from other large values of that function, see Theorem 5.7 in Van der Vaart [363].

Remarks 3.26

- The estimator from the arg max operation in (3.27) is also called M-estimator, and $(y, a) \mapsto \log(f(y;a))$ plays the role of a scoring function (similar to a loss function). The the last line of (3.27) says that this scoring function is strictly consistent for the functional $Q : \mathcal{F} \to \mathbf{\Theta}$, and Fisher-consistency of this functional Q implies

$$\mathbb{E}_\theta\left[\log f(Y;\widetilde{\theta})\right] \le \mathbb{E}_\theta\left[\log f(Y;\, Q(F(\cdot;\theta)))\right] = \mathbb{E}_\theta\left[\log f(Y;\theta)\right],$$

  for all $\widetilde{\theta} \in \mathbf{\Theta}$. Strict consistency of loss and scoring functions is going to be defined formally in Sect. 4.1.3, below, and we have just seen that this plays an important role for the consistency of M-estimators in the sense of Definition 3.23.
- Consistency (3.27) assumes that the data generating model $Y \sim F$ belongs to the specified family $\mathcal{F} = \{F(\cdot;\theta); \theta \in \mathbf{\Theta}\}$. Model uncertainty may imply that the data generating model does not belong to $\mathcal{F}$. In this situation, and if we are allowed to exchange the arg max operation and the limit in n in (3.27), the MLE will provide the model in $\mathcal{F}$ that is closest in KL divergence to the true model F. We come back to this in Sect. 11.1.4, below.

3.4.2 Asymptotic Normality

As mentioned above, typically, we would like to have stronger results than just consistency. We give an introductory example based on the EF.

Example 3.27 (Asymptotic Normality of the MLE in the EF) We work under the same EF as in Example 3.24. This example has provided consistency of the sequence of MLEs $\widehat{\mu}_n^{\text{MLE}}$, $n \ge 1$, for μ. Note that the i.i.d. property together with the finite

variance property immediately implies the following convergence in distribution

$$\sqrt{n}\left(\widehat{\mu}_n^{\mathrm{MLE}} - \mu\right) \;\Rightarrow\; \mathcal{N}(0, \nabla_{\boldsymbol{\theta}}^2 \kappa(\boldsymbol{\theta})) \overset{(\mathrm{d})}{=} \mathcal{N}(0, \mathcal{I}_1(\boldsymbol{\theta})) \qquad \text{as } n \to \infty,$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}(\mu) = (\nabla_{\boldsymbol{\theta}} \kappa)^{-1}(\mu) \in \boldsymbol{\Theta}$ for $\mu \in \mathcal{M}$, and $\mathcal{N}$ denotes the Gaussian distribution. This is the multivariate version of the central limit theorem (CLT), and it tells us that the rate of convergence is $1/\sqrt{n}$. This asymptotic result is stated in terms of Fisher's information matrix under parametrization $\boldsymbol{\theta}$. We transform this to the dual mean parametrization and call Fisher's information matrix under the dual mean parametrization $\mathcal{I}_1^*(\mu)$. This involves the change of variable $\mu \mapsto \boldsymbol{\theta} = \boldsymbol{\theta}(\mu) = (\nabla_{\boldsymbol{\theta}} \kappa)^{-1}(\mu)$. The Jacobian matrix of this change of variable is given by $J(\mu) = \mathcal{I}_1(\boldsymbol{\theta}(\mu))^{-1}$ and, thus, the transformation of Fisher's information matrix gives, see also (3.16),

$$\mu \;\mapsto\; \mathcal{I}_1^*(\mu) = J(\mu)^\top \mathcal{I}_1(\boldsymbol{\theta}(\mu))\, J(\mu) = \mathcal{I}_1(\boldsymbol{\theta}(\mu))^{-1}.$$

This allows us to express the above CLT w.r.t. Fisher's information matrix corresponding to μ and it gives us

$$\sqrt{n}\left(\widehat{\mu}_n^{\mathrm{MLE}} - \mu\right) \;\Rightarrow\; \mathcal{N}\left(0, \mathcal{I}_1^*(\mu)^{-1}\right) \qquad \text{as } n \to \infty. \tag{3.28}$$

We conclude that the appropriately normalized MLE $\widehat{\mu}_n^{\mathrm{MLE}}$ converges in distribution to the centered Gaussian distribution having as covariance matrix the *inverse of Fisher's information matrix* $\mathcal{I}_1^*(\mu)$, and the *rate of convergence* is $1/\sqrt{n}$.

Assume that the effective domain $\boldsymbol{\Theta}$ is open, and that $\boldsymbol{\theta} = \boldsymbol{\theta}(\mu) \in \boldsymbol{\Theta}$. This allows us to transform asymptotic normality (3.28) to the canonical scale. Consider again the change of variable $\mu \mapsto \boldsymbol{\theta} = \boldsymbol{\theta}(\mu) = (\nabla_{\boldsymbol{\theta}} \kappa)^{-1}(\mu)$ with Jacobian matrix $J(\mu) = \mathcal{I}_1(\boldsymbol{\theta}(\mu))^{-1} = \mathcal{I}_1^*(\mu)$. Theorem 1.9 in Section 5.2 of Lehmann [244] tells us how the CLT transforms under such a change of variable, namely,

$$\begin{aligned} \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{MLE}} - \boldsymbol{\theta}\right) &= \sqrt{n}\left((\nabla_{\boldsymbol{\theta}} \kappa)^{-1}\left(\widehat{\mu}_n^{\mathrm{MLE}}\right) - (\nabla_{\boldsymbol{\theta}} \kappa)^{-1}(\mu)\right) \\ &\Rightarrow \mathcal{N}\left(0, J(\mu)\mathcal{I}_1^*(\mu)^{-1} J(\mu)\right) \overset{(\mathrm{d})}{=} \mathcal{N}\left(0, \mathcal{I}_1(\boldsymbol{\theta})^{-1}\right) \qquad \text{as } n \to \infty. \end{aligned} \tag{3.29}$$

We have exactly the same structural form in the two asymptotic results (3.28) and (3.29). There is a main difference, $\widehat{\mu}_n^{\mathrm{MLE}}$ is unbiased for μ whereas, in general, $\widehat{\boldsymbol{\theta}}_n^{\mathrm{MLE}}$ is not unbiased for $\boldsymbol{\theta}$, but we receive the same asymptotic behavior. ■

There are many different versions of asymptotic normality results similar to (3.28) and (3.29), and the main difficulty often is to verify the assumptions made. For instance, one can prove asymptotic normality based on a Fisher-consistent functional Q. The assumptions made are, among others, that Q needs to be Fréchet differentiable in $P(\cdot;\theta)$ which, unfortunately, is rather difficult to verify. We make a list of assumptions here that are easier to check and then we give a version of the asymptotic normality result which is stated in the book of Lehmann [244]. This list of assumptions in the one-dimensional case $\boldsymbol{\Theta} \subseteq \mathbb{R}$ reads as follows:

(i) $\boldsymbol{\Theta} \subseteq \mathbb{R}$ is an open interval (possibly infinite).
(ii) The real-valued random variables $Y_i \sim F(\cdot;\theta)$, $i \geq 1$, have common support $\mathfrak{T} = \{y \in \mathbb{R};\ f(y;\theta) > 0\}$ which is independent of $\theta \in \boldsymbol{\Theta}$.
(iii) For every $y \in \mathfrak{T}$, the density $f(y;\theta)$ is three times continuously differentiable in θ.
(iv) The integral $\int f(y;\theta)d\nu(y)$ is twice differentiable under the integral sign.
(v) Fisher's information satisfies $\mathcal{I}_1(\theta) = \mathbb{E}_\theta[(\partial \log f(Y_1;\theta)/\partial\theta)^2] \in (0,\infty)$.
(vi) For every $\theta_0 \in \boldsymbol{\Theta}$ there exist a positive constant c and a function $M(y)$ (both may depend on θ_0) such that $\mathbb{E}_{\theta_0}[M(Y_1)] < \infty$ and

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(y;\theta)\right| \leq M(y) \qquad \text{for all } y \in \mathfrak{T} \text{ and } \theta \in (\theta_0 - c, \theta_0 + c).$$

Theorem 3.28 (Theorem 2.3 in Section 6.2 of Lehmann [244]) *Assume Y_i, $i \geq 1$, are i.i.d. $F(\cdot;\theta)$ distributed satisfying* (i)–(vi) *from above. Assume that $\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{Y}_n)$, $n \geq 1$, is a sequence of roots that solves the score equations*

$$\frac{\partial}{\partial\widetilde{\theta}}\sum_{i=1}^{n}\log f(Y_i;\widetilde{\theta}) = \frac{\partial}{\partial\widetilde{\theta}}\ell_{\boldsymbol{Y}_n}(\widetilde{\theta}) = 0,$$

and which is consistent for θ, i.e. this sequence of roots $\widehat{\theta}_n(\boldsymbol{Y}_n)$ converges in probability to the true parameter θ. Then we have asymptotic normality

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \ \Rightarrow\ \mathcal{N}\left(0, \mathcal{I}_1(\theta)^{-1}\right) \qquad \text{as } n \to \infty. \tag{3.30}$$

Sketch of Proof Fix $\theta \in \boldsymbol{\Theta}$ and consider a Taylor expansion of the score $\ell'_{\boldsymbol{Y}_n}(\cdot)$ in θ for $\widehat{\theta}_n$. It is given by

$$\ell'_{\boldsymbol{Y}_n}(\widehat{\theta}_n) = \ell'_{\boldsymbol{Y}_n}(\theta) + \ell''_{\boldsymbol{Y}_n}(\theta)\left(\widehat{\theta}_n - \theta\right) + \frac{1}{2}\ell'''_{\boldsymbol{Y}_n}(\theta_n)\left(\widehat{\theta}_n - \theta\right)^2,$$

for $\theta_n \in [\theta, \widehat{\theta}_n]$. Since $\widehat{\theta}_n$ is a root of the score, the left-hand side is equal to zero. This allows us to re-arrange the above Taylor expansion as follows

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) = \frac{\frac{1}{\sqrt{n}}\ell'_{\boldsymbol{Y}_n}(\theta)}{-\frac{1}{n}\ell''_{\boldsymbol{Y}_n}(\theta) - \frac{1}{2n}\ell'''_{\boldsymbol{Y}_n}(\theta_n)\left(\widehat{\theta}_n - \theta\right)}.$$

The enumerator on the right-hand side converges in distribution to $\mathcal{N}(0, \mathcal{I}_1(\theta))$, see (18) in Section 6.2 of [244], the first term in the denominator converges in probability to $\mathcal{I}_1(\theta)$, see (19) in Section 6.2 of [244], and in the second term of the denominator we have $\frac{1}{2n}\ell'''_{\boldsymbol{Y}_n}(\theta_n)$ which is bounded in probability, see (20) in Section 6.2 of [244]. The claim then follows from Slutsky's theorem.

□

Remarks 3.29

- A sequence $(\widehat{\theta}_n)_{n\geq 1}$ satisfying Theorem 3.28 is called *efficient likelihood estimator* (ELE) of θ. Typically, the sequence of MLEs $\widehat{\theta}_n^{\text{MLE}}$ gives such an ELE sequence, but there are counterexamples where this is not the case, see Example 3.1 in Section 6.2 of Lehmann [244]. In that example $\widehat{\theta}_n^{\text{MLE}}$ exists for all $n \geq 1$, but it converges in probability to ∞, regardless of the value of the true parameter θ.
- Any sequence of estimators that fulfills (3.30) is called *asymptotically efficient*, because, similarly to the Cramér–Rao information bound of Theorem 3.13, it attains $\mathcal{I}_1(\theta)^{-1}$ (which under certain assumptions is a lower variance bound except on Lebesgue measure zero, see Theorem 1.1 in Section 6.1 of Lehmann [244]). However, there are two important differences here: (1) the Cramér–Rao information bound statement needs unbiasedness of the decision rule, whereas (3.30) only requires consistency (but not unbiasedness nor asymptotically vanishing bias); and (2) the lower bound in the Cramér–Rao statement is an effective variance (on a finite sample), whereas the quantity in (3.30) is only an asymptotic variance. Moreover, any other sequence that differs in probability from an asymptotically efficient one less than $o(1/\sqrt{n})$ is asymptotically efficient, too.
- If we consider a differentiable function $\theta \mapsto \gamma(\theta)$, then Theorem 3.28 implies

$$\sqrt{n}\left(\gamma\left(\widehat{\theta}_n\right) - \gamma(\theta)\right) \;\Rightarrow\; \mathcal{N}\left(0, \frac{(\gamma'(\theta))^2}{\mathcal{I}_1(\theta)}\right) \qquad \text{as } n \to \infty. \tag{3.31}$$

  This follows from asymptotic normality, consistency and considering a Taylor expansion around θ.
- We were starting from the MLE problem

$$\widehat{\theta}_n^{\text{MLE}} = \arg\max_{\widetilde{\theta}} \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i; \widetilde{\theta}). \tag{3.32}$$

In statistical theory a parameter estimator that is obtained through a maximization operation is called M-estimator (for maximizing or minimizing), see also Remarks 3.26. If the log-likelihood is differentiable in $\widetilde{\theta}$ we can turn the above problem into a root search problem for $\widetilde{\theta}$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\widetilde{\theta}}\log f(Y_i;\widetilde{\theta})=0. \tag{3.33}$$

If a parameter estimator is obtained through a root search problem it is called Z-estimator (for equating to zero). The Z-estimator (3.33) does not require a maximum of the original function, but only a critical point; this is exactly what we have been exploring in Theorem 3.28. More generally, for a sufficiently nice function $\psi(\cdot;\theta)$ a Z-estimator $\widehat{\theta}_n^{\mathrm{Z}}$ for θ is obtained by solving the following equation for $\widetilde{\theta}$

$$\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i;\widetilde{\theta})=0, \tag{3.34}$$

for i.i.d. data $Y_i\sim F(\cdot;\theta)$. Suppose that the first moment of $\psi(Y_i;\widetilde{\theta})$ exists. The law of large numbers gives us, a.s., see also (3.26),

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i;\widetilde{\theta})=\mathbb{E}_\theta\left[\psi(Y;\widetilde{\theta})\right]. \tag{3.35}$$

Consistency of the Z-estimator $\widehat{\theta}_n^{\mathrm{Z}}$, $n\geq 1$, for θ is related to the right-hand side of (3.35) being zero for $\widetilde{\theta}=\theta$. Under additional regularity conditions (and consistency) it then holds asymptotic normality

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{Z}}-\theta\right)\ \Rightarrow\ \mathcal{N}\left(0,\frac{\mathbb{E}_\theta\left[\psi(Y;\theta)^2\right]}{\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\psi(Y;\theta)\right]^2}\right)\qquad\text{as } n\to\infty. \tag{3.36}$$

For rigorous statements we refer to Theorems 5.21 and 5.41 in Van der Vaart [363]. A modification to the regression case is given in Theorem 11.6 below.

Example 3.30 We consider the single-parameter linear EF for given strictly convex and steep cumulant function κ and w.r.t. a σ-finite measure ν on $\mathbb{R}$. The score equation gives requirement

$$\frac{1}{n}S(\boldsymbol{Y}_n)\ \stackrel{!}{=}\ \kappa'(\theta)=\mathbb{E}_\theta[Y_1]. \tag{3.37}$$

Strict convexity implies that the right-hand side strictly increases in θ. Therefore, we have at most one solution of the score equation here. We assume that the

effective domain $\boldsymbol{\Theta} \subseteq \mathbb{R}$ is open. It is easily verified that assumptions (ii)–(vi) hold, in particular, (vi) saying that the third derivative should have a uniformly bounded integrable bound holds because the third derivative is independent of y and continuous in θ. With probability converging to 1, (3.37) has a solution $\widehat{\theta}_n$ which is unique, consistent and Theorem 3.28 holds. Note that in Example 3.5 we have mentioned the Poisson case which can be degenerate. For the asymptotic normality result we use here that this degeneracy asymptotically vanishes with probability converging to one. ■

Remark 3.31 (Multi-Dimensional Extension) For an extension of Theorem 3.28 to the multi-dimensional case $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ we refer to Section 6.4 in Lehmann [244]. The assumptions made in the multi-dimensional case do not essentially differ from the ones in the 1-dimensional case.