Chapter 1

Heuristics

The official statistical dogma on estimation is: good estimators converge to the right thing and have limiting normal distributions. Moreover, the variance of the limiting distribution should not be smaller than a quantity defined by the Fisher information function. The estimators that achieve the asymptotic lower bound are called efficient. Maximum likelihood estimators are efficient.

The dogma is not quite correct, but much of it can be rescued in slightly altered form. Therein hangs a tale. This Chapter starts the story by describing one method for building estimators that typically have good properties, by explaining when the estimators should be efficient, and by showing what can go wrong.

1. Notation

There is a large body of statistical theory and literature regarding optimality and large sample approximation, some of it true, some of it almost true, and some of it a little bit wrong. As with most folklore, there are grains of truth buried amongst the chaff. Some ideas—such as efficiency and sufficiency—have survived mathematical indignations and counterexamples, by evolving to retain their secure place at the foundations of statistics. Some myths have died out.

Mostly we will be concerned with parts of the theory that are both useful and mathematically correct, but, to appreciate the virtues of rigor, you must first understand some of the folklore.

Many problems in mathematical statistics boil down to the following question. Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model—a family of probability measures all defined on the same sigma-field $\mathcal{F}$ on a set Ω. Let $T = T(\omega)$ be a random variable (or, more generally, a random vector, or even a random element of some wonderfully abstract space). What is the distribution of T under each $\mathbb{P}_\theta$ model?

Typically T is thought of as an estimator for some function $\tau(\theta)$ of the indexing parameter θ, or perhaps it represents a choice from a set of possible actions. From knowledge of the distribution of T under each $\mathbb{P}_\theta$, one calculates various expectations used to evaluate the performance of T.

Unfortunately, it is seldom possible to calculate all the distributions directly. Instead one must make simplifying approximations, or, more formally, find limiting forms of distributions for sequences of estimators $\{T_n\}$ under sequences of models

$\{\mathbb{P}_{n,\theta} : \theta \in \Theta_n\}$. The extra parameter n typically denotes a sample size, and the approximations are called ***large-sample*** (or ***asymptotic***) distributions.

Let me start with a more concrete example to establish some finer points of notation.

<1> **Example.** Let $\{f_\theta : \theta \in \Theta\}$ be a family of probability densities for probability measures P_θ on the (Borel sigma-field of the) real line. Let $X_1, \ldots, X_n$ be random variables on Ω, and let $\mathbb{P}_\theta$ be a probability measure on Ω under which the $\{X_i\}$ are independent, each with distribution P_θ. The method of maximum likelihood defines $\widehat{\theta}$ as the value of θ that maximizes $\prod_{i\le n} f_\theta(X_i(\omega))$. For the moment I will ignore all questions of existence, uniqueness, or measurability of a maximizing value.

Notice that $\widehat{\theta}$ depends on ω only through the vector of observations $\mathbf{X}(\omega) := (X_1(\omega), \ldots, X_n(\omega))$. For many purposes it is better to think of an estimator as a function on $\mathbb{R}^n$ (or $\mathcal{X}^n$, if the variables X_i take values in a set $\mathcal{X}$). That is, define the ***estimating function*** $\widehat{\theta}(\mathbf{x})$ to maximize

$$\prod_{i\le n} f(x_i, \theta) \qquad \text{for } \mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

then use the estimator $\widehat{\theta}(\mathbf{X}(\omega))$. This approach has the conceptual advantage of focussing attention on $\widehat{\theta}$ as a function of $\mathbf{x}$, before making any assumptions about how $\mathbf{x}$ is to be interpreted. It shows that the definition of the maximum likelihood estimator depends only on the model being fitted. The performance of the estimator under various probabilistic mechanisms for generation of the sample $\mathbf{X}(\omega)$—and not just for those mechanisms prescribed by the model—becomes a separate question. That is, the view of $\widehat{\theta}(\cdot)$ as a function of $\mathbf{x}$ disentangles the issue of definition via a model from the issue of behaviour of the estimator under those models.

The idea of an estimating function also helps to distinguish between the multiple roles played by θ. For the definition of the maximum likelihood estimator, θ is merely a dummy variable, a placeholder that indicates a function to be maximized. In its second role, θ identifies one particular model, under which performance is to be evaluated. It is traditional to use a separate symbol, such as θ_0, for this second use of the θ parameter. The θ_0 is usually held fixed throughout an asymptotic calculation. It is tempting to call θ_0 the "true value", or refer to $\mathbb{P}_{\theta_0}$ as the "true underlying mechanism", for the purposes of the calculation. Of course if we actually knew the truth we wouldn't need to estimate; the title "true" serves merely to distinguish one particular parameter value during the course of a calculation. A name like "test case" or "typical case" might be less misleading.

One must be careful not to confuse the two roles for θ. For example, it would usually be a fatal error to replace a dummy θ by a fixed θ_0 before optimizing over the dummy value. One way to avoid confusion between dummies and truth is to consider behaviour of the estimator under a fixed $\mathbb{P}$, or under a sequence of fixed distributions $\{\mathbb{P}_n\}$, which might—or might not—correspond to a particular θ_0 model. The θ_0 can then be thought of as some value of θ that just happens to be picked out by some procedure related to $\mathbb{P}$; it is a value defined by $\mathbb{P}$, and not necessarily the index value that selects $\mathbb{P}$ from a parametric class of possible distributions. □

2. Limit theory heuristics

With the preliminaries about truth out of the way, let me turn to a general problem that illustrates a number of important asymptotic ideas. Suppose the data are given by random quantities $\mathbf{X} := (X_1, \ldots, X_n)$ taking values in a set $\mathcal{X}^n$ (such as $\mathbb{R}^n$). Suppose Θ is a set, perhaps with some interpretation as an index for a model, or perhaps not. Suppose $\{g(\cdot, \theta) : \theta \in \Theta\}$ is a collection of real-valued functions on $\mathcal{X}$. Define an estimating function $\widehat{\theta}_n(\mathbf{x})$ as the value of θ that minimizes

$$G_n(\mathbf{x}, \theta) = n^{-1} \sum\nolimits_{i \le n} g(x_i, \theta) \qquad \text{for } \mathbf{x} := (x_1, \ldots, x_n) \in \mathcal{X}^n.$$

That is, $\widehat{\theta}_n(\mathbf{x}) := \text{argmin}_{\theta \in \Theta} G_n(\mathbf{x}, \theta)$. In the language of Huber (1964), the correrspondin g $\widehat{\theta}_n(\mathbf{X})$ is an ***M-estimator***.

> TYPICAL QUESTION: *What can we say about the behaviour of the estimator $\widehat{\theta}_n$ when the X_i are independent, each with marginal distribution P?*

For the purposes of an asymptotic answer to the QUESTION, we might regard the data as the initial segment of an infinite sequence of independent $\mathcal{X}$-valued random variables $X_1, X_2, \ldots$, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with each X_i having distribution P. Alternatively, we might treat the data as one row in a triangular array of random variables, defined on a probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ that can change with n. The X_1 for the nth row of the array might be completely unrelated to the X_1 element in other rows. It might even be better to make this possibility explicit, by writing the data as $\mathbf{X}_n := (X_{n,1}, \ldots, X_{n,n})$. The distribution P could also be replaced by a P_n that changes with n, a generalization that will be needed when we consider behavior of estimators under sequences of alternatives.

For the moment I will work with the the simpler setting of a fixed underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a fixed distribution P. The traditional answer to the QUESTION then comes in three stages. The first two steps require a distance function on Θ. The third step has meaning only when Θ is a subset of a vector space.

(i) Consistency

Show that $\widehat{\theta}_n$ converges to some fixed θ_0 as $n \to \infty$. If the underlying probability space does not change with n, it makes sense to ask about convergence at $\mathbb{P}$-almost all ω; that is, it makes sense to enquire about ***strong consistency*** of the estimators. If Ω or $\mathbb{P}$ could change with n, then strong consistency is ill defined. In that case, it is better to enquire about possible ***weak consistency***, that is, convergence in probability of $\widehat{\theta}_n$ to θ_0, or even to a value θ_n that changes with n.

The weaker form usually suffices when consistency is just the prelude to a more detailed analysis of asymptotic behaviour. Strong consistency is sometimes of interest only because it implies weak consistency.

(ii) Root-n consistency

If we know that $\widehat{\theta}_n$ converges to some θ_0, then it makes sense to ask how rapidly it converges. Again we have a choice between asking about rates at almost all ω or about rates of convergence in probability. Again the in-probability assertion is often the more useful, in part because of its role as a necessary preliminary to the next stage in the analysis.

> REMARK. The name root-n consistency is slightly misleading: it is a $1/\sqrt{n}$ rate that is typical. That is, we seek to prove that $\widehat{\theta}_n = \theta_0 + O_p(1/\sqrt{n})$.

(iii) Limiting distribution

Convergence at a $1/\sqrt{n}$-rate in probability need not imply existence of a limiting distribution for the standardized estimator $\sqrt{n}(\widehat{\theta} - \theta_0)$. And even if there is a limiting distribution, it might be concentrated at zero, which would mean that the estimator actually converges at a rate faster than $1/\sqrt{n}$. To settle the matter, it would suffice if we could demonstrate existence of a nontrivial limiting distribution for the standardized estimator. Sometimes the existence of that limit is made to follow from an explicit asymptotic representation, $\sqrt{n}(\widehat{\theta}_n - \theta_0) = W_n + o_p(1)$, where W_n has known limiting behaviour. You will learn in Chapter 3 how such a representation can be more useful than mere existence of a limit distribution.

Typically M-estimators are well behaved, under mild regularity assumptions. Consider the simplest case where Θ is a subset of the real line. To understand $\widehat{\theta}_n$ we need to know what G_n is doing. The key idea is approximation of G_n by another process, whose minimizing value is more easily analyzed. For a rigorous analysis we would have to determine the effect of the errors in approximation to G_n, to ensure that the minimizing values are close. The rigorous treatment will begin with Chapter 2, where error will be expressed as remainder terms from Taylor expansions. For the moment, I will approximate with abandon.

Consistency for M-estimators

For each fixed θ, a law of large numbers (strong or weak?) implies that $G_n(\theta)$ should be close to its expected value $G(\theta) := P^x g(x, \theta)$. That is, as a first approximation, we should have $G_n(\theta) \approx G(\theta)$ for every θ. We might then hope that $\text{argmin}_\theta\, G_n(\theta) \approx \text{argmin}_\theta\, G(\theta)$. That is, we might hope that $\widehat{\theta}_n$ lies close to the value $\theta_0 := \text{argmin}_\theta\, G(\theta)$, the value that minimizes the approximating G. Notice that θ_0 depends on P.

> REMARK. You will learn in Chapter 2 one way, essentially due to Wald (1949), to make the approximation idea more precise and establish consistency. Later Chapters will generalize the method. The crucial idea will always be that the approximations should hold uniformly in θ, at least in regions of Θ that matter.

Asymptotic normality for M-estimators

If we know that $\widehat{\theta}_n$ has large probability of lying close to θ_0, then the behaviour of G_n near θ_0 becomes our main concern. A Taylor expansion of $g(x, \cdot)$ about θ_0, with dots denoting partial derivatives with respect to θ,

$$g(x, \theta) \approx g(x, \theta_0) + (\theta - \theta_0)\dot{g}(x, \theta_0) + \tfrac{1}{2}(\theta - \theta_0)^2 \ddot{g}(x, \theta_0)$$

leads to a quadratic approximation for G_n near θ_0:

$$\begin{aligned} G_n(\theta) &= n^{-1}\sum\nolimits_{i\le n} g(X_i,\theta) \\ &\approx n^{-1}\sum\nolimits_{i\le n}\Big(g(X_i,\theta_0)+(\theta-\theta_0)\dot g(X_i,\theta_0)+\tfrac12(\theta-\theta_0)^2\ddot g(X_i,\theta_0)\Big). \end{aligned}$$

The random variables $\dot g(X_i,\theta_0)$ should have zero expected value,

<2> $$\mathbb{P}\dot g(X_i,\theta_0)=P^x\frac{\partial}{\partial\theta}g(x,\theta)\Big|_{\theta=\theta_0}\overset{?}{=}\left(\frac{\partial}{\partial\theta}P^x g(x,\theta)\right)\Big|_{\theta=\theta_0}=\dot G(\theta_0)=0,$$

because G is minimized at θ_0. (Of course some regularity conditions would be needed to justify the interchange in the order of differentiation and integration.) The random variables have variance $\sigma^2 := P^x\dot g(x,\theta_0)^2$. The standardized average

$$Z_n := \sum\nolimits_{i\le n}\dot g(X_i,\theta_0)/\sqrt{n}$$

should be approximately $N(0,\sigma^2)$ distributed.

The analogous approximation for G near θ_0,

$$\begin{aligned} G(\theta) &\approx G(\theta_0)+(\theta-\theta_0)P^x\dot g(x,\theta_0)+\tfrac12(\theta-\theta_0)^2P^x\ddot g(x,\theta_0) \\ &= G(\theta_0)+\tfrac12(\theta-\theta_0)^2 J \qquad \text{where } J := P^x\ddot g(x,\theta_0), \end{aligned}$$

tells us that J should be nonnegative if G is to have a minimum at θ_0, at least when θ_0 is an interior point of Θ. It would be awkward if J were zero, for then we would need to consider the contributions from the higher-order derivatives.

The average $\sum_{i\le n}\ddot g(X_i,\theta_0)/n$ should be close to J. The random criterion function is approximately a quadratic in $\theta-\theta_0$,

$$G_n(\theta)\approx G_n(\theta_0)+(\theta-\theta_0)Z_n/\sqrt{n}+\tfrac12(\theta-\theta_0)^2 J \qquad \text{near } \theta_0.$$

The minimizing $\widehat\theta_n$ for G_n should be close to the value $\theta_0-Z_n/(J\sqrt{n})$ that minimizes the quadratic. The standardized estimator $\sqrt{n}(\widehat\theta_n-\theta_0)$ should be close to $-Z_n/J$, which has an approximate $N(0,\sigma^2/J^2)$ distribution.

3. Efficiency heuristics for M-estimators

If the asymptotic heuristics from the previous Section are to be believed, there is a wide class of estimators that have approximate normal distributions, with means and variance that decrease like $1/n$. It is natural to look for a g that gives the smallest possible multiple of $1/n$ for the approximate variance.

Actually, the task is slightly more complicated than choosing g to minimize the variance at a fixed P. After all, we would not bother to estimate if we already knew the underlying distribution exactly. The real challenge is to minimize the asymptotic variance *under a whole class of possible P's.* Specifically, suppose $\{\mathbb{P}_\theta : \theta\in\Theta\}$ is a statistical model, with $X_1, X_2,\ldots$ independently distributed as P_θ under $\mathbb{P}_\theta$.

REMARK. Specifically, we could take $\mathbb{P}_\theta$ to be a countable product of P_θ measures, on the product sigma-field of $\mathcal{X}^{\mathbb{N}}$. For calculations involving only $X_1,\ldots,X_n$, we could also work with P_θ^n on the product sigma-field of $\mathcal{X}^n$. In that case it would be better to write $\mathbb{P}_{n,\theta}$, with the dependence on n made explicit.

For each θ in Θ, we first need to ensure that the estimator $\widehat{\theta}_n$ converges in $\mathbb{P}_\theta$ probability to θ. Then we need to consider the asymptotic variance as a function of θ, and look for a g to minimize that function at every θ.

Consider the question of consistency. For independent observations from a fixed P the heuristics suggested that $\widehat{\theta}_n$ converges in probability to $\operatorname{argmin}_\theta P^x g(x,\theta)$. Clearly there is going to be some confusion if we use θ both to identify the underlying P and as the dummy variable for the minimization. Let me, therefore, temporarily replace P_θ by P_t, where t also ranges over Θ. For samples from P_t, the estimator converges in probability to the value $\theta_0(t)$ minimizing the function $\theta \mapsto P_t^x g(x,\theta)$, that is, $\theta_0(t) := \operatorname{argmin}_\theta P_t^x(x,\theta)$. For consistency we need $\theta_0(t) = t$ for all t in Θ. By <2>, we then have

<3> $$P_t^x \dot{g}(x,t) = 0 \qquad \text{for all } t \text{ in } \Theta.$$

Define $\sigma_g^2(\theta) := P_\theta^x \dot{g}(x,\theta)^2 = \operatorname{var}_\theta \dot{g}(x,\theta)$ and $J_g(\theta) := P_\theta^x \ddot{g}(x,\theta)$. We hope to find a g to minimize the asymptotic variance $\sigma_g^2(\theta)/J_g(\theta)^2$ for every θ, subject to the constraint <3>.

Now suppose P_θ is given by a density f_θ (with respect to Lebesgue measure on the real line, for simplicity, although the argument works for every dominating measure). Write $\ell_\theta(x)$ for $\log f_\theta(x)$. Classical theory asserts that the minimum is achieved by $g(x,\theta) := -\ell_\theta(x)$, that is, by the maximum likelihood estimator. Jensen's inequality ensures that $-P_t^x \log f_\theta(x)$ is minimized at t,

$$P_t^x \log f_\theta(x) - P_t^x \log f_t(x) \le \log \int f_t(x) \left(f_\theta(x)/f_t(x) \right) dx \le \log 1 = 0.$$

According to the heuristics, the maximum likelihood estimator is therefore a consistent M-estimator.

The derivative $\dot{\ell}_\theta(x) = \dot{f}_\theta(x)/f_\theta(x)$ is usually called the ***score function*** for the model. The $J_g(\theta)$ corresponding to $-\ell$ equals

$$\mathbb{I}(\theta) := -P_\theta^x \left(\frac{\partial^2}{\partial\theta^2} \log f_\theta(x) \right),$$

the ***(Fisher) information function*** for the model.

To prove that $-\ell$ achieves the constrained minimum for the asymptotic variance, write <3> as $0 \equiv \int \dot{g}(x,\theta)\, f_\theta(x)\, dx$, then differentiate to derive another constraint,

$$0 \equiv \int \ddot{g}(x,\theta)\, f_\theta(x)\, dx + \int \dot{g}(x,\theta)\, \dot{f}_\theta(x)\, dx = P_\theta^x \ddot{g}(x,\theta) + P_\theta \left(\dot{g}(x,\theta) \dot{\ell}_\theta(x) \right).$$

That is, $J_g(\theta) = -P_\theta \left(\dot{g}(x,\theta) \dot{\ell}_\theta(x) \right)$ for all θ. Thus

$$\begin{aligned} \left(J_g(\theta) \right)^2 &= \left(P_\theta \left(\dot{g}(x,\theta) \dot{\ell}_\theta(x) \right) \right)^2 \\ &\le \left(P_\theta \dot{g}(x,\theta)^2 \right) \left(P_\theta \dot{\ell}_\theta(x)^2 \right) \qquad \text{by Cauchy-Schwarz} \\ &= \sigma_g^2(\theta) \sigma_{-\ell}^2(\theta). \end{aligned}$$

When g equals $-\ell$ we have equality in the second line, thereby implying that

<4> $$\mathbb{I}(\theta) = J_{-\ell}(\theta) = \sigma_{-\ell}^2(\theta) = \operatorname{var}_\theta \left(\dot{\ell}_\theta \right).$$

The Cauchy-Schwartz bound then gives

$$\frac{\sigma_g^2(\theta)}{J_g(\theta)^2} \ge \frac{1}{\sigma_{-\ell}^2(\theta)} = \mathbb{I}(\theta)^{-1} = \frac{\sigma_{-\ell}^2(\theta)}{J_{-\ell}(\theta)^2} \qquad \text{where } \ell(x,\theta) := \log f_\theta(x).$$

The asymptotic normal distribution for the maximum likelihood estimator has variance equal to the lower bound. At least that is what the heuristics suggest.

REMARK. Readers familiar with the usual proof of the information inequality should recognize the technique used in the preceding paragraphs. By restricting myself to M-estimators, I have avoided the usual handwaving arguments by which one tries to downplay the assumption that the estimators be unbiased.

The dual representation for the information function, as in <4>, will turn out (Chapter 3) to be a requirement for a basic property known as contiguity.

I have not been rigorous about the conditions required for the arguments leading to "asymptotic optimality" of the maximum likelihood estimator amongst the class of M-estimators. For example, the argument surely fails when f_θ denotes the Uniform$(0,\theta)$ density, which is not everywhere differentiable.

As the next Section explains, optimality is a slippery concept even for models that seem unlikely candidates for making trouble. A completely rigorous treatment can seem quite difficult—if one does not have the right tools. The development of the rigorous theory has been a major theme in modern theoretical statistics.

4. Fisher's concept of efficiency

If the heuristics are to be believed, in typical cases M-estimators cannot do better than mimic the limiting behaviour of the maximum likelihood estimator, which asymptotically achieves the information bound. In fact, it was long accepted in the statistics literature that the maximum likelihood estimator has optimality properties amongst an even wider class of estimators. As Fisher (1922, page 277) put it, "The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation." Unfortunately, the unqualified assertion about the limit distributions is not quite valid, although it can be rescued. There exist estimators that beat the efficiency bound, as shown by a famous construction due to Hodges.

<5> **Example.** Let $\{\widehat{\theta}_n\}$ be a sequence of estimators that is efficient in Fisher's sense, for the framework described in Section 3. Let $\{\alpha_n\}$ be a sequence of positive real numbers converging to zero more slowly than $1/\sqrt{n}$, such as $\alpha_n := n^{-1/4}$. For a fixed θ_0 in Θ define $U_n := \{\theta \in \Theta : |\theta - \theta_0| \le \alpha_n\}$. Modify $\widehat{\theta}_n$ so that it performs superefficiently if θ_0 happens to be the true value, without disturbing its performance elsewhere, by defining $\theta_n^* := \widehat{\theta}_n\{\widehat{\theta}_n \notin U_n\} + \theta_0\{\widehat{\theta}_n \in U_n\}$.

REMARK. Notice that θ_n^* is not an M-estimator.

Under $\mathbb{P}_{\theta_0}$ the modification takes effect with probability tending to one, that is, $\mathbb{P}_{\theta_0}\{\theta_n^* = \theta_0\} \ge \mathbb{P}_{\theta_0}\{\widehat{\theta}_n \in U_n\} \to 1$, which results in an estimator with obvious merits,

$$\sqrt{n}(\theta_n^* - \theta_0) \to 0 \qquad \text{in } \mathbb{P}_{\theta_0} \text{ probability.}$$

In particular, the efficiency bound is well beaten at θ_0. Effectively θ_n^* behaves like the constant estimator, θ_0, when the true value is θ_0. But unlike the constant estimator, θ_n^* can adapt when the true value is not θ_0,

$$\mathbb{P}_\theta\{\theta_n^* = \widehat{\theta}_n\} \geq \mathbb{P}_{\theta_0}\{\widehat{\theta}_n \notin U_n\} \to 1 \qquad \text{if } \theta \neq \theta_0,$$

Under $\mathbb{P}_\theta$ for $\theta \neq \theta_0$, the estimator θ_n^* has the same asymptotic behaviour as $\widehat{\theta}_n$. The estimator θ_n^* achieves the efficiency bound at all points of Θ, except at θ_0, where it does much better than the Fisherian concept of efficiency would allow. □

The Hodges phenomenon has nothing to do with the smoothness or regularity of the parametrization of the model. It occurs even with the estimation of the mean of a $N(\theta, 1)$ distribution, where the maximum likelihood estimator is none other than the sample mean. Clearly the efficiency heuristics don't tell the whole story. The concept of efficiency as a desirable property of estimators—the property that they asymptotically achieve the information lower bound for variance—will be rescued in Chapter 4, where a requirement of good behaviour of the estimator along sequences of alternative models will be used to exclude the Hodges estimator and its ilk from the optimality competition, in a sense that I will soon explain.

Fisher (1924) asserted another property for efficient estimators. He regarded maximum likelihood as the basic method for constructing an efficient estimator. He described the effect of inefficient estimation as equivalent, asymptotically (a qualification that was seldom made explicit during the period when Fisher first contributed to the subject), to the addition of an independent source of error beyond what one should expect of an efficient estimator.

> Let A be the efficient statistic with variance σ^2/n, and B the inefficient statistic with variance σ^2/En; ... the correlation of A with $(B - A)$ is zero, so that the deviations of B from the population value may be regarded as made up of two parts: one, an error of random sampling, properly so called, is the deviation of A from the population value; the other, distributed independently of the first, is the error of estimation by which the inferior estimate, B, differs from the superior estimate, A.
>
> [Fisher 1924, page 446]

Fisher's assertion corresponds to an asymptotic assertion for an estimator T_n,

$$\sqrt{n}(T_n - \theta_0) = \sqrt{n}(T_n - \widehat{\theta}_n) + \sqrt{n}(\widehat{\theta}_n - \theta_0) \qquad \text{with } \widehat{\theta}_n \text{ efficient},$$

where, in some sense, the two terms on the right-hand side should be asymptotically independent. If limiting distributions existed, we could interpret asymptotic independence to mean $\left(\sqrt{n}(T_n - \widehat{\theta}), \sqrt{n}(\widehat{\theta} - \theta_0)\right) \rightsquigarrow (M, Z)$, with Z distributed $N(0, \mathbb{I}(\theta_0)^{-1})$ independently of the "noise" M. Consequently, we would have $\sqrt{n}(T_n - \theta_0) \rightsquigarrow M + Z$. The limit distribution would be least dispersed when M were degenerate. For example, when variances were finite, as would be the case when M had a normal distribution, the equality $\mathbb{P}_{\theta_0}|M + Z|^2 = \mathbb{P}_{\theta_0}|M|^2 + \mathbb{P}_{\theta_0}|Z|^2$ would show that the mean-squared error were a minimum if $M \equiv 0$. (An assumption of asymptotic normality was implicit in Fisher's concept of efficiency.) More generally, if $\rho(\cdot)$ were nonnegative, symmetric, and convex, the symmetry of the distribution of Z would give

$$\mathbb{P}_{\theta_0}\rho(M + Z) = \tfrac{1}{2}\mathbb{P}_{\theta_0}\rho(M + Z) + \tfrac{1}{2}\mathbb{P}_{\theta_0}\rho(-M + Z) \geq \mathbb{P}_{\theta_0}\rho(Z),$$

with strict inequality if $\rho(\cdot)$ were strictly convex and if M were not degenerate at zero. Efficient estimators (in the sense of asymptotic mean squared error) would be those for which $M \equiv 0$. The distribution of Z would provide an asymptotic lower bound for the accuracy of estimation; only efficient estimators could achieve that bound. For an efficient T_n the difference $\sqrt{n}(T_n - \widehat{\theta}_n)$ would converge in probability to zero; T_n and $\widehat{\theta}_n$ would be asymptotically equivalent.

Unfortunately, this second view of efficiency is also not quite valid, although it too can be rescued.

The superefficient estimator from Example <5> does well under $\mathbb{P}_\theta$ if θ does not change with n, but the modification has unfortunate consequences at alternatives $\mathbb{P}_{\theta_n}$ for $\{\theta_n\}$ that approaches θ_0 at an $O(1/\sqrt{n})$ rate through Θ. For simple cases, such as $P_\theta := N(\theta, 1)$, it is easy to prove directly that $\sqrt{n}(\widehat{\theta}_n - \theta_n) \rightsquigarrow N(0, \mathbb{I}(\theta_0)^{-1})$ under $\mathbb{P}_{\theta_n}$. (In fact, $\sqrt{n}(\widehat{\theta}_n - \theta_n)$ has exactly a $N(0, 1)$ distribution, for observations from the $N(\theta_n, 1)$. See Chapter 3 for a way to handle more general models.)

The neighborhood U_n captures $\widehat{\theta}_n$ with high $\mathbb{P}_{\theta_0}$ probability, because α_n decreases more slowly than the $O_p(1/\sqrt{n})$ rate at which $\widehat{\theta}_n$ converges to θ_0. Unfortunately, U_n has the same effect under $\mathbb{P}_{\theta_n}$, because $|\widehat{\theta}_n - \theta_0| \le |\widehat{\theta}_n - \theta_n| + |\theta_n - \theta_0| = O_p(1/\sqrt{n})$, implying $\mathbb{P}_{\theta_n}\{\theta_n^* = \theta_0\} \to 1$. In particular, if $\theta_n := \theta_0 + \delta_n/\sqrt{n}$ with $\delta_n \to \delta$, then $\sqrt{n}(\theta_n^* - \theta_n) \to -\delta$ in $\mathbb{P}_{\theta_n}$ probability, which is not good if $|\delta|$ is large. If we allow δ_n to wander off to infinity more slowly than $\sqrt{n}\alpha_n$, we can even arrange $|\sqrt{n}(\theta_n^* - \theta_n)| \to \infty$ in $\mathbb{P}_{\theta_n}$ probability. The estimator θ_n^* has achieved its superefficient status at the expense of poor behaviour under certain types of local alternative.

Acceptable behaviour under alternatives close to $\mathbb{P}_{\theta_0}$ will rule out superlative behaviour at θ_0. With some added local uniformity requirements, Fisher's concepts of efficiency will be rescued in Chapter 4, in the forms of the Convolution and Local Asyymptotic Minimax Theorems.

References

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London, A* **222**, 309–368.

Fisher, R. A. (1924), 'The conditions under which χ^2 measures the discrepency between observation and hypothesis', *Journal of the Royal Statistical Society* **87**, 442–450.

Huber, P. J. (1964), 'Robust estimation of a location parameter', *Annals of Mathematical Statistics* pp. 73–101.

Wald, A. (1949), 'Note on the consistency of the maximum likelihood estimate', *Annals of Mathematical Statistics* **20**, 595–601.