1 Asymptotic Comparisons of Estimators

Consider the following generic version of an estimation problem. One observes data \( X_i, i = 1, \ldots, n \) i.i.d. with distribution \( P \in \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \). Suppose we wish to estimate \( \psi(\theta) \) using the data and that we have an estimator \( T_n = T_n(X_1, \ldots, X_n) \) such that for each \( \theta \in \Theta \)

\[
\sqrt{n}(T_n - \psi(\theta)) \xrightarrow{d} L_\theta
\]

under \( \theta \). What is the “best” possible limit distribution for such an estimator?

It is natural to measure “best” in terms of concentration, and we can measure concentration with a loss function. A loss function \( \ell(x) \) is simply any function that takes values in \([0, \infty)\). A loss function is said to be “bowl-shaped” if the sublevel sets \( \{ x : \ell(x) \leq c \} \) are convex and symmetric about the origin. A common bowl-shaped loss function on \( \mathbb{R} \) is mean-squared error loss, that is, \( \ell(x) = x^2 \). For a given loss function \( \ell(x) \), a limit distribution will be considered “good” if

\[
\int \ell(x) dL_\theta
\]

is small.

If the estimator \( T_n \) is asymptotically normal in the sense that

\[
L_\theta = N(\mu(\theta), \sigma^2(\theta))
\]

then in order to minimize the mean-squared error loss it is optimal to have \( \mu(\theta) = 0 \) and \( \sigma^2(\theta) \) as small as possible. Of course, for estimators that are not asymptotically normal, this may not be true, and we do not wish to restrict attention \textit{a priori} to asymptotically normal estimators.

2 Hodge’s Estimator and Superefficiency

Suppose \( \mathcal{P} = \{ N(\theta, 1) : \theta \in \mathbb{R} \} \) and \( \psi(\theta) = \theta \). A natural estimator of \( \theta \) is the sample mean, that is, \( T_n = \bar{X}_n \). As you already know, this estimator
has many finite-sample optimality properties (it’s minimax for every bowl-shaped loss function, it’s minimum variance unbiased, etc.), so we might reasonably expect it to be optimal asymptotically as well.

A second estimator of \( \theta \), \( S_n \), can be defined as follows:

\[
S_n = \begin{cases} 
T_n & \text{if } |T_n| \geq n^{-1/4} \\
0 & \text{if } |T_n| < n^{-1/4}
\end{cases}
\]

In words, \( S_n = T_n \) when \( T_n \) is “far” from zero and \( S_n = 0 \) when \( T_n \) is “close” to zero.

It is easy to see that

\[
\sqrt{n}(T_n - \theta) \sim N(0, 1).
\]

But how does \( S_n \) behave asymptotically? To answer this question, first consider \( \theta \neq 0 \). For any such \( \theta \),

\[
P_\theta\{|T_n| \geq n^{-1/4}\} \to 1.
\]

To see this, let \( Z_n = \sqrt{n}(T_n - \theta) \) and note that

\[
Pr_\theta\{|T_n| < n^{-1/4}\} = Pr_\theta\{-n^{-1/4} < T_n < n^{-1/4}\} = Pr_\theta\{\sqrt{n}(-n^{-1/4} - \theta) < Z_n < \sqrt{n}(n^{-1/4} - \theta)\}.
\]

For \( \theta > 0 \), \( n^{-1/4} - \theta < 0 \) for \( n \) sufficiently large, so the probability tends to 0. For \( \theta < 0 \), \( -n^{-1/4} - \theta > 0 \) for \( n \) sufficiently large, so the probability tends to 0. The desired result thus follows. From the definition of \( S_n \), we have that \( S_n = T_n \) with probability approaching 1 for \( \theta \neq 0 \).

Now consider \( \theta = 0 \). In this case,

\[
P_\theta\{|T_n| \geq n^{-1/4}\} \to 0.
\]

To see this, note that

\[
Pr_\theta\{|T_n| \geq n^{-1/4}\} = Pr_\theta\{T_n \geq n^{-1/4} \cup T_n \leq -n^{-1/4}\} = Pr_\theta\{Z_n \geq n^{1/4} \cup Z_n \leq -n^{1/4}\} \leq Pr_\theta\{Z_n \geq n^{1/4}\} + Pr_\theta\{Z_n \leq -n^{1/4}\}.
\]
Both of the probabilities in the last expression tend to 0, so the result follows. From the definition of $S_n$, we have that $S_n = 0$ with probability approaching 1 for $\theta = 0$.

Thus, for $\theta \neq 0$

$$\sqrt{n}(S_n - \theta) \overset{d}{\to} N(0, 1)$$

under $\theta$ and for $\theta = 0$

$$r_n(S_n - \theta) \overset{d}{\to} 0$$

under $\theta$ for any sequence $r_n$, including $r_n = \sqrt{n}$. The estimator $S_n$ is said to be “superefficient” at $\theta = 0$.

Let $L_\theta$ denote the limit distribution of $T_n$ and $L'_\theta$ denote the limit distribution of $S_n$. It follows from the above discussion that for $\theta \neq 0$

$$\int x^2 dL'_\theta = \int x^2 dL_\theta$$

and for $\theta = 0$

$$\int x^2 L'_\theta = 0 < 1 = \int x^2 L_\theta .$$

Thus, $S_n$ appears, at least in terms of its limiting distribution, to be a better estimator of $\theta$ than $T_n$. But appearances can be deceiving. This reasoning again reflects the poor use of asymptotics. Our hope is that

$$\int x^2 L'_\theta$$

is a reasonable approximation to the finite-sample expected loss

$$E_\theta[|\sqrt{n}(S_n - \theta)|^2] .$$

In finite-samples, for $\theta$ “far” from zero, we might expect $S_n = T_n$, and so we might expect $L'_\theta$ to be a reasonable approximation to the distribution of $\sqrt{n}(S_n - \theta)$; for $\theta$ “close” to zero, on the other hand, $S_n$ will frequently differ from $T_n$, so the distribution of $\sqrt{n}(S_n - \theta)$ may be quite different from $L'_\theta$. As before, the definition of “close” and ”far” will differ with the sample size $n$. We must therefore consider the behavior of $S_n$ under sequences $\theta_n \to 0$. 

To illustrate this point, consider $\theta_n = \frac{h}{n^{1/4}}$ where $0 < h < 1$. (Implicitly, we are redefining $T_n = \bar{X}_{n,n}$, where $X_{i,n}, i = 1, \ldots, n$ are i.i.d with distribution $P_{\theta_n} = N(\theta_n, 1)$.) As before,

$$\sqrt{n}(T_n - \theta_n) \sim N(0, 1)$$

but how does $S_n$ behave under $\theta_n$? To answer this, note that

$$\Pr_{\theta_n}\{|T_n| < n^{-1/4}\} = \Pr_{\theta_n}\{-n^{-1/4} < T_n < n^{-1/4}\} = \Pr_{\theta_n}\{\sqrt{n}(-n^{-1/4} - \theta_n) < Z_n < \sqrt{n}(n^{-1/4} - \theta_n)\} = \Pr_{\theta_n}\{-n^{1/4}(1 + h) < Z_n < n^{1/4}(1 - h)\}$$

We saw earlier that this probability tended to 0 under $\theta \neq 0$, but under $\theta_n = \frac{h}{n^{1/4}}$, this probability tends to 1. Thus, under $\theta_n$, we have that $S_n = 0$ with probability approaching 1. Hence, under $\theta_n$,

$$\sqrt{n}(S_n - \theta_n) = -n^{1/4}h$$

with probability approaching 1, and $-n^{1/4}h \to -\infty$. Denote by $L$ the limiting distribution of $T_n$ under $\theta_n$ and by $L'$ the limiting distribution of $S_n$ under $\theta_n$ (in this case $L'$ is degenerate at $-\infty$). It follows that

$$\int x^2 dL' = +\infty > 1 = \int x^2 dL.$$

Thus, $S_n$ “buys” its better asymptotic performance at 0 at the expense of worse behavior for points “close” to zero. The definition of “close” changes with $n$, so this feature is not borne out by a pointwise asymptotic comparison for every $\theta \in \Theta$, but we can see it if we consider a sequence $\theta_n$. We can also see it graphically by plotting the finite-sample expected losses, $E_\theta[\ell(\sqrt{n}(S_n - \theta))]$ versus $E_\theta[\ell(\sqrt{n}(T_n - \theta))] = 1$, for different samples sizes $n$.

This example is quite famous and is due to Hodges. The estimator $S_n$ is often referred to as Hodges’ estimator.

### 3 Efficiency of Maximum Likelihood

The above example shows that it is impossible to give a nontrivial definition of “best” to the limit distributions $L_\theta$. In fact, it is not even enough to
consider $L_\theta$ under every $\theta \in \Theta$. For some fixed $\theta' \in \Theta$, we could always construct an estimator whose limit distribution was equal to $L_\theta$ for $\theta \neq \theta'$, but “better” at $\theta = \theta'$ by using the trick due to Hodges.

Under certain conditions, it turns out that the “best” limit distributions are in fact those the limit distributions of maximum likelihood estimators, but to make this idea precise is a bit tricky.

One of the conditions we will require in the statement of the result is that $\mathcal{P}$ is a reasonably nice family of distributions. The precise condition is that $\mathcal{P}$ is “differentiable in quadratic mean”. Many commonly encountered families of distributions are differentiable in quadratic mean, including, e.g., exponential families (which include the family of normal distributions) and location models with smooth underlying densities. See Chapter 12 of Lehmann and Romano (2005) for a precise definition of differentiability in quadratic mean.

The notation $I_\theta$ will be used to denote the Fisher Information matrix. If $p_\theta$ is the density of $P_\theta$ w.r.t. some measure $\mu$ (e.g., Lebesgue measure, counting measure, etc.) and $l_\theta = \log p_\theta$ is differentiable, then

$$I_\theta = E_\theta[\dot{l}_\theta \dot{l}_\theta'] .$$

The Fisher Information can be defined more generally for families of distributions that are differentiable in quadratic mean, but we won’t go into that right now.

We can now state the following result:

**Theorem 3.1** Suppose that $\mathcal{P}$ is differentiable in quadratic mean, that $I_\theta$ is nonsingular for every $\theta$, and that $\psi$ is differentiable at every $\theta$. Let $T_n$ be *any* estimator such that for every $\theta$

$$\sqrt{n}(T_n - \psi(\theta)) \overset{d}{\to} L_\theta$$

under $\theta$. Then, there exist distributions $M_\theta$ such that for almost every $\theta$ w.r.t. Lebesgue measure

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta') \ast M_\theta .$$
The notation \( \star \) denotes the "convolution" operation between two distributions and should be interpreted as follows: If \( X \sim F \) and \( Y \sim G \) and \( X \perp Y \), then \( X + Y \sim F \star G \). Theorem 3.1 is often referred to as the (almost-everywhere) convolution theorem.

This theorem does not contradict the results of the previous section. In that case, \( P = \{ N(\theta, 1) : \theta \in \mathbb{R} \} \), \( \psi(\theta) = \theta \), and \( N(0, \psi \theta^{-1} \psi') = N(0, 1) \). For every \( \theta \neq 0 \), \( \sqrt{n}(S_n - \theta) \overset{d}{\to} N(0, 1) \) under \( \theta \), so the theorem is satisfied for \( M_\theta \) the distribution with unit mass at 0.

Note that \( N(0, \psi \theta^{-1} \psi') \) is the limit distribution of the maximum likelihood estimator of \( \psi(\theta) \). In order to assert that this is in fact the "best" limit distribution, we need the following lemma:

**Lemma 3.1** For any bowl-shaped loss function \( \ell \) on \( \mathbb{R}^k \), every probability distribution \( M \) on \( \mathbb{R}^k \), and every covariance matrix \( \Sigma \),

\[
\int \ell(x) dN(0, \Sigma) \leq \int \ell(x) d(N(0, \Sigma) \star M).
\]

Thus, if "best" is measured by any bowl-shaped loss function (including mean-squared error loss), then, under the assumptions of Theorem 3.1, maximum likelihood estimators are "best" for almost every \( \theta \) w.r.t. Lebesgue measure.

For a proof of these two results, see van der Vaart (1998).