Hey! I’m getting married! -Romain.

1 Asymptotic Relative Efficiency

We can compare relative efficiencies in estimation by comparing asymptotic variance, and in testing by comparing slopes. Here we consider testing. We’ll start with some review.

In comparing the efficiency of two tests, we consider a series $K_n$ of critical regions for rejecting the null hypothesis and a series $T_n$ of statistics. We then have a series $\pi_n(\theta) = P_{\theta}(T_n \in K_n)$ of power functions, and we define an asymptotic power function $\pi(h) \equiv \lim_{n \to \infty} \pi_n(h\sqrt{n})$.

We assume that for some functions $\mu(\theta)$ and $\sigma(\theta)$, we have asymptotic normality when $\theta_n \to \theta_0$:

$$\frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \xrightarrow{\theta_n} N(0, 1).$$

If this is the case, then, assuming $\theta_0 = 0$ for simplicity,

$$\pi_n \left( \frac{h}{\sqrt{n}} \right) \to 1 - \Phi \left( z_\alpha - \frac{\mu(0)}{\sigma(0)} \right).$$

The quantity $\mu(0)/\sigma(0)$ is referred to as the slope of the test. It is a measure of how easy it is to distinguish alternatives that are of order $1/\sqrt{n}$ away from the null hypothesis. The asymptotic relative efficiency (ARE) of two tests is defined to be the ratio of the square of their slopes.

Example 1. Mann-Whitney vs. 2-sample $t$-test

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x)$$

$$Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F(x - \theta)$$

$H_0 : \theta = 0$

The $t$-test is as defined in the previous lecture. It has slope $1/\sigma$. The Mann-Whitney test rejects if $\frac{1}{nm} \sum_{i,j} I(X_i \leq Y_j)$ is large.

Note. The Mann-Whitney statistic has a relationship with the area under the ROC curve (AUC) for classification algorithms with a tunable parameter. The ROC plot has one axis for proportion of false positives and one axis for proportion of true positives; as we move to the right, the classifier becomes more sensitive to true positives, but misclassifies more negative points as positive. Nearly-flat ROC curves are bad and strongly humped ROC curves are good (see Figure 1). The empirical ROC curve shows the classifier’s performance on the training data and has the form of discrete stair-steps. If we consider the positive examples to be one population $Y_1, \ldots, Y_m$ and the negative examples to be another population $X_1, \ldots, X_n$, and consider
classifiers of the form \( f_\theta(z) = \text{sign}(z - \theta) \), then the AUC statistic for the empirical ROC curve is simply the Mann-Whitney statistic up to a constant factor. Because the Mann-Whitney statistic is a U-statistic, it is asymptotically normal (asymptotic normality for U-statistics can be established under a changing parameter \( \theta_n \)). Therefore, AUC is asymptotically normal as well.

To find the moments of the Mann-Whitney statistic, we can use the formulae for moments of U-statistics. One can compute the slope to be \( \int f dF / \sigma(0) \), from which we derive the ARE of Mann-Whitney versus 2-sample \( t \) to be \( 12 \cdot \text{Var}(X)(\int f(y)^2 dy)^2 \).

The following table lists the ARE of Mann-Whitney versus 2-sample \( t \) for several forms of the distribution \( F \).

<table>
<thead>
<tr>
<th>Distribution</th>
<th>ARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>( \pi^2/9 \approx 95% )</td>
</tr>
<tr>
<td>Normal</td>
<td>( 3/\pi )</td>
</tr>
<tr>
<td>Laplace</td>
<td>( 3/2 )</td>
</tr>
<tr>
<td>Uniform</td>
<td>1</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>1.24</td>
</tr>
<tr>
<td>( t_5 )</td>
<td>1.9</td>
</tr>
<tr>
<td>( c(1-x^2) \lor 0 )</td>
<td>108/125 \approx 86%</td>
</tr>
</tbody>
</table>

As is evident in the table, the Mann-Whitney statistic has good efficiency in a variety of settings.

### Interpreting ARE

Suppose the sequence of alternative hypotheses \( \theta_n \) is chosen ahead of time. Define \( n_{\nu} \) to be the minimum number of observations needed such that \( \pi_{n,\nu}(0) \leq \alpha \) and \( \pi_{n,\nu}(\theta_n) \geq \gamma \).

Do this for two tests, test 1 and test 2, and consider \( n_{\nu,1}/n_{\nu,2} \). In the statement of the theorem below, \( \|P_{n,\theta} - P_{n,0}\| \triangleq \sup_{A \in \mathcal{B}} |P_{n,\theta}(A) - P_{n,0}(A)| \), the variational distance between \( P_{n,\theta} \) and \( P_{n,0} \), where \( \mathcal{B} \) is the set of Borel sets.

**Theorem 1** (14.19 in Van der Vaart). Consider a sequence of models \( \{P_{n,\theta} : \theta \geq 0\} \) such that \( \|P_{n,\theta} - P_{n,0}\| \to 0 \) as \( \theta \to 0 \). Let \( T_{n,1} \) and \( T_{n,2} \) satisfy

\[
\sqrt{n} \left( \frac{T_{n,i} - \mu_i(\theta_n)}{\sigma_i(\theta_n)} \right) \xrightarrow{d} N(0, 1)
\]
for every \( \theta_n \searrow 0 \), with \( \mu'_i(0) > 0 \) and \( \sigma_i \) continuous at 0. Then

\[
\frac{\left( \frac{\mu'_1(0)}{\sigma_1(0)} \right)^2}{\frac{\mu'_2(0)}{\sigma_2(0)}} = \lim_{\nu \to \infty} \frac{n_{\nu,2}}{n_{\nu,1}}
\]

for every \( \theta_\nu \searrow 0 \).

**Proof.** This is just a sketch. First show that as \( n_{\nu,i} \to \infty \),

\[
\pi_{n_{\nu,i}}(\theta_\nu) = 1 - \Phi \left( z_\gamma + o(1) - \sqrt{n_{\nu,1} \theta_\nu} \frac{\mu'_i(0)}{\sigma_i(0)} (1 + o(1)) \right) + o(1).
\]

This tends to \( \gamma < 1 \) if the argument of \( \Phi \) tends to \( z_\gamma \).

\[\square\]

## 2 Rates of Convergence

Rates are a stronger measure of the goodness of a test than consistency (since they assume consistency), and are weaker than convergence in distribution. They are an active topic in statistics today; many statistics don’t converge to a Gaussian distribution, but we’d still like to say things about how fast they converge. In fact, if I have a procedure of some kind, I don’t particularly care about its limiting distribution—I just want to know how well it does, in terms of risk or other performance measures.

For instance, in what sense is a kernel density estimate better than a histogram estimate? There’s no Gaussian limit, so we can’t use ARE.

**Example 2** (Kernel density estimation). For densities, plugin estimation is useless, since the plugin density estimate is \( f = \phi(F) = dF/dx \), and \( \phi(F_n) \) just returns spikes at the data.

The histogram estimate breaks up the domain of the data \( X_i \) into bins, and estimates the height of the density at each bin by the number of data points falling in that bin.

Kernel density estimation estimates the density by

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left( \frac{x - X_i}{h} \right)
\]

where \( X_i \) are the observed data points and \( K \) is a kernel function. The scale factor \( h \) is referred to as the kernel bandwidth. If \( \int K(x)dx = 1 \) then \( \int \hat{f}(x)dx = 1 \). Note that \( K \) is not a kernel in the sense of an RKHS. Some typical kernels are shown in Figure 2. The Epanechnikov kernel is optimal in a sense to be defined later.
Performance measures

We could use Hellinger or KL to measure the difference between estimated and true density, but for the present discussion we’ll use the mean integrated squared error,

\[ \text{MISE}_f(\hat{f}) \triangleq \int E_f(\hat{f}(x) - f(x))^2 dx \]

\[ = \int \text{Var}_f \hat{f}(x) dx + \int \text{Bias}^2_f \hat{f}(x) dx \]

We’ll see later that the integrated variance term goes as \( 1/\n h \), and the integrated bias term goes as \( h^4 \). We need both to go to zero, so we let \( h \to 0 \) but \( nh \to \infty \). What’s the optimal choice of \( h \) to minimize \( \text{MISE}_f(\hat{f}) \)? This is a simple calculation: we set \( 0 = \frac{d}{dh}((nh)^{-1} + h^4) = -(nh)^{-1} + 4h^3 \), and solving for \( h \) yields \( h = O(n^{-1/5}) \). With this value of \( h \), both the bias and variance terms are \( O(n^{-4/5}) \), so \( \text{MISE}^* \sim n^{-4/5} \).

In the parametric case, when we are only considering densities indexed by a parameter \( \theta \), we typically have \( \hat{\theta}_n = \theta + O_p(n^{-1/2}) \), where \( \hat{\theta}_n \) is our estimate of \( \theta \). Then \( \text{MISE}_{\theta}(\hat{f}_\theta) = \int E_{\theta}(\hat{f}_\theta(x) - f_\theta(x))^2 dx \approx E_{\theta}(\hat{\theta} - \theta)^2 \) by a Taylor series expansion. This latter term goes as \( n^{-1} \). Thus, we do slightly worse with nonparametric density estimation, but we are able to handle any density (subject to some regularity conditions), not just those in a parametric family.

**Theorem 2** (24.1 in Van Der Vaart). Assume \( f \) is twice continuously differentiable, and that the following conditions hold: \( \int K(y)dy = 1, \int f''(x)^2 dx < \infty, \int yK(y)dy = 0, \int y^2K(y)dy < \infty, \) and \( \int K(y)^2 dy < \infty \). Then there exists a constant \( C_f \) such that for the kernel density estimate \( \hat{f} \), \( \text{MISE}_f(\hat{f}) \leq C_f((nh)^{-1} + h^4) \).

For the histogram estimate, the bias term is \( h^2 \) rather than \( h^4 \) (see Scott, 1992).

**Proof.** We begin by showing that the variance term is \( O((nh)^{-1}) \).

\[
\text{Var}_f \hat{f}(x) = \frac{1}{n} \text{Var} \frac{1}{h} K \left( \frac{x-X_i}{h} \right) \leq \frac{1}{nh^2} E K \left( \frac{x-X_i}{h} \right)^2 \text{ (throwing out } (EK)^2 \text{)} = \frac{1}{nh} \int K(y)^2 f(x-hy)dy.
\]

But then

\[
\int \text{Var}_f \hat{f}(x) dx \leq \frac{1}{nh} \int \int K(y)^2 f(x-hy)dy dx = \frac{1}{nh} \int K(y)^2 dy \quad \text{by Fubini—the density goes away.}
\]

This integral is a finite constant by assumption.

Now we address the bias term. We have

\[
f(x+h) - f(x) = hf'(x) + h^2 \int_0^1 f''(x+sh)(1-s)ds
\]
by a slightly unusual form of the Taylor expansion with a Laplacian representation of the remainder, so
\[
\mathbb{E}_f \hat{f}(x) - f(x) = \int \frac{1}{h} K \left( \frac{x-t}{h} \right) f(t) dt - f(x)
= \int K(y)(f(x-hy) - f(x)) dy
= \int \int_0^1 K(y) (-hyf'(x) + (hy)^2 f''(x - shy)(1-s)) ds dy.
\]

But \( \int K(y) y dy = 0 \), so the \( f'(x) \) term vanishes and we are left with
\[
\mathbb{E}_f \hat{f}(x) - f(x) = \int \int_0^1 K(y) ((hy)^2 f''(x - shy)(1-s)) ds dy. \quad (1)
\]

Define random variables \( Y \sim K(y), \ S \sim \text{Unif}[0, 1], \) and \( Z = Y f''(x - ShY)(1 - S) \). Then (1) can simply be written \( h^2 \mathbb{E}[Y Z] \), and
\[
\text{Bias}^2(x) = h^4 (\mathbb{E}[Y Z])^2 \leq h^4 (\mathbb{E}Y^2)(\mathbb{E}Z^2) \quad \text{by Cauchy-Schwartz}
= h^4 \left( \int y^2 K(y) dy \right) \left( \int \int y^2 f''(x - shy)^2 (1-s)^2 ds K(y) dy \right).
\]

By integrating both sides with respect to \( x \), we can conclude that
\[
\int \text{Bias}^2(x) dx \leq h^4 \left( \int y^2 K(y) dy \right) \left( \int f''(x)^2 dx \cdot \frac{1}{3} \right)
\]
(because \( \int f''(x - shy)^2 dx = \int f''(x)^2 dx \) regardless of the values of \( s, h, \) and \( y \)). By assumption, these integrals are finite constants depending on \( f \). In all, then, \( \text{MISE} \leq C_f((nh)^{-1} + h^4) \) for some constant \( C_f \). \( \square \)

If \( f \) has \( m > 2 \) derivatives, then we get a rate of \( n^{-\frac{2m}{2m+1}} \) for kernel density estimation. This turns out to be minimax—a lower bound of order \( n^{-\frac{2m}{2m+1}} \) on the MISE for density estimation can be derived using Assouad’s lemma. We’ll see this in the next lecture. Kernel density estimation is therefore minimax optimal (up to a constant).

There is some leeway to affect the constant term \( C_f \) by manipulating the kernel \( K(y) \); this is a variational optimization problem, subject to the constraints \( \int y K(y) dy = 0 \) and \( \int K(y) dy = 1 \). The optimal kernel in this sense is the Epanechnikov kernel \( K(y) = (3/4)(1 - y^2) \) for \(-1 < y < 1\) with \( K(y) = 0 \) elsewhere.

Kernel bandwidth selection is a large topic. One method is to use a large \( h \) to estimate \( f'' \), use that estimate to reoptimize \( h \), and iterate.

References