STA 532: Theory of Statistical Inference

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

4 Parametric Inference II

In this section we will take $X = \{X_1, \cdots\}$ to be a sequence of independent random variables all with the same probability distribution (so they are "iid"), from some parametric family $\mathfrak{F} = \{F(x \mid \theta) : \ \theta \in \Theta \subset \mathbb{R}^k\}$ of distributions that either have pdfs or pmfs, either of which we will denote by $f(x \mid \theta)$. We will look at the behavior of sequences of estimators T_n of θ based on the first n observations.

Define the "Kullback-Leibler Divergence" from a distribution with pdf $f(x)$ to another $g(x)$ to be

$$\mathrm{KL}(f : g) := \int_{\mathcal{X}} f(x) \, \log \frac{f(x)}{g(x)} \, dx. \tag{1}$$

Although it isn't symmetric and doesn't obey the triangle inequality, and so can't be a true "distance" measure, the KL divergence has many of the properties of a metric. In particular, $\mathrm{KL}(f : f) = 0$ for any f and $0 < \mathrm{KL}(f : g) \le \infty$ for any distinct densities f, g, because

$$\mathrm{KL}(f : g) = \int_{\mathcal{X}} f(x) \, \left[-\log \frac{g(x)}{f(x)} \right] \, dx \ge \int_{\mathcal{X}} f(x) \, \left[1 - \frac{g(x)}{f(x)} \right] \, dx = 0$$

since $-\log y \ge (1 - y)$ for all $y > 0$ (because $e^x \ge x + 1$ for all $x \in \mathbb{R}$). For pdfs $f(x \mid \theta)$ and $f(x \mid \theta')$ from distributions in the parametric family $\mathfrak{F}$, we simplify the notation by writing $\mathrm{KL}(\theta : \theta') := \mathrm{KL}\big(f(x \mid \theta) : f(x \mid \theta')\big)$. Some denote this as $\mathrm{KL}(\theta \| \theta')$.

The family $\mathfrak{F}$ is called "identifiable" if $\mathrm{KL}(\theta : \theta') > 0$ whenever $\theta \ne \theta'$, *i.e.*, when different parameters lead to different distributions. This would *not* hold for the Normal $\mathsf{No}(\mu, \sigma^2)$ distribution if we took $\sigma \in \mathbb{R}$, for example, even though the normal pdf is well-defined for all $(\mu, \sigma) \in \mathbb{R}^2$. From now on we will assume all our parametric families are identifiable.

Asymptotic Properties of MLEs

Suppose the observations $\{X_i\} \overset{\text{iid}}{\sim} F(x \mid \theta_\star)$ for some $\theta_\star \in \Theta$, and let $\hat\theta_n$ be the MLE based on the first n observations, *i.e.*, the value of $\theta \in \Theta$ that maximizes $\mathcal{L}_n(\theta) := \prod_{j \le n} f(x_j \mid \theta)$ or, equivalently, that maximizes $\ell_n(\theta) := \sum_{j \le n} \log f(x_j \mid \theta)$.

Consistency

Since $\theta_\star$ and n are just constants, the MLE $\hat\theta_n$ also *minimizes*

$$n^{-1}[\ell_n(\theta_\star) - \ell_n(\theta)] = \frac{1}{n}\sum \log \frac{f(x_i \mid \theta_\star)}{f(x_i \mid \theta)}$$

which, by the law of large numbers, converges as $n \to \infty$ to

$$\int_{\mathcal{X}} \log \frac{f(x \mid \theta_\star)}{f(x \mid \theta)}\, f(x \mid \theta_\star)\, dx = \mathrm{KL}(\theta_\star : \theta).$$

But this limit is minimized at $\theta = \theta_\star$, where $\mathrm{KL}(\theta_\star : \theta)$ attains its minimum value of zero. The details of proving $\hat\theta_n \to \theta_\star$ for every model under very weak conditions are a bit more involved, ask me if you'd like more details.

Asymptotic Normality

Fisher Information

The k-dimensional random vector $Z(X;\theta)$ with components $Z_i(X;\theta) = (\partial/\partial\theta_i)\log f(X \mid \theta)$ for $1 \le i \le k$ is called the "score function". It isn't a *statistic*, because it depends explicitly on the parameter θ. The mean is $\mathsf{E}[Z(X;\theta)] = 0$ because

$$\begin{aligned}\mathsf{E}Z_i(X;\theta) &= \int_{\mathcal{X}} \{(\partial/\partial\theta_i)\log f(x \mid \theta)\}\ f(x \mid \theta)\, dx \\ &= \int_{\mathcal{X}} \frac{(\partial/\partial\theta_i) f(x \mid \theta)}{f(x \mid \theta)}\ f(x \mid \theta)\, dx \\ &= (\partial/\partial\theta_i)\int_{\mathcal{X}} f(x \mid \theta)\, dx = (\partial/\partial\theta_i)1 = 0,\end{aligned} \tag{2}$$

since any pdf integrates identically to one. The $(k\times k)$ *covariance* matrix

$$I(\theta) = \mathsf{E}_\theta[Z(X;\theta)Z(X;\theta)'] \tag{3a}$$

(here $Z(X;\theta)'$ denotes the $(1\times k)$ transpose of the $(k\times 1)$ vector $Z(X;\theta)$) is called the "Fisher Information". Differentiating (2) w.r.t θ_j gives

$$\begin{aligned}0 &= (\partial/\partial\theta_j)\int_{\mathcal{X}} \{(\partial/\partial\theta_i \log f(x \mid \theta)\}\ f(x \mid \theta)\, dx \\ &= \int_{\mathcal{X}} \left\{(\partial^2/\partial\theta_i\partial_j)\log f(x \mid \theta)\right\}\ f(x \mid \theta)\, dx + \int_{\mathcal{X}} \{(\partial/\partial\theta_i \log f(x \mid \theta)\}\ (\partial/\partial\theta_j) f(x \mid \theta)\, dx \\ &= \int_{\mathcal{X}} \left\{(\partial^2/\partial\theta_i\partial_j)\log f(x \mid \theta)\right\}\ f(x \mid \theta)\, dx + \mathsf{E}_\theta[Z_i(X;\theta)Z_j(X;\theta)],\end{aligned}$$

showing that the $(k\times k)$ Fisher Information can also be written as

$$I_{ij}(\theta) = \mathsf{E}_\theta\left[\frac{-\partial^2}{\partial\theta_i\partial\theta_j}\log f(X \mid \theta)\right]; \tag{3b}$$

sometimes one is easier to compute, sometimes the other. For a multivariate normal $\mathsf{No}(\mu,\Sigma)$ distribution with uncertain mean $\mu\in\mathbb{R}^k$ and known $(k\times k)$ covariance matrix Σ, for example, the log likelihood is

$$-\tfrac{1}{2}\log\det(2\pi\Sigma)-\tfrac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)$$

whose negative second partial derivative wrt μ_i and μ_j is Σ^{-1}_{ij}, so $I(\mu):=\Sigma^{-1}$ for the MVN distribution. This is worth remembering: in some sense $I(\theta)$ acts like a *precision matrix* (inverse of a covariance) for $\hat\theta_n$ in almost all problems, as we'll see below.

The score function for the one-parameter Poisson $\mathsf{Po}(\lambda)$ distribution is

$$Z(X;\lambda)=(\partial/\partial\lambda)[X\log\lambda-\log X!-\lambda]=X/\lambda-1,$$

so $\mathsf{E}_\lambda[Z(X;\lambda)]=0$ (as always) and the Fisher information is $I(\lambda)=\mathsf{E}_\lambda Z(X;\lambda)^2=1/\lambda$, again one over the variance of X.

The Fisher Information for a sample of size n is just n times that for a sample of size one, $I_n(\theta)=nI(\theta)$. Fisher Information depends on the particular choice of parameterization: if $\eta=H(\theta)$ is another parameterization, then the Fisher informations for the two parameterization are related by

$$I^\theta(\theta)=J(\theta)^2I^\eta\big(H(\theta)\big)$$

where $J(\theta)=\partial H(\theta)/\partial\theta$ is the Jacobian of the transformation $\theta\mapsto\eta=H(\theta)$. In $k>1$ a matrix version of the same formula applies: $I^\theta(\theta)=J(\theta)'I^\eta\big(H(\theta)\big)J(\theta)$.

MLE and Information

Suppose that the likelihood function $\mathcal{L}$ or equivalently the joint pdf $f(X\mid\theta)$ has at least two continuous derivatives in θ, and that for increasing samples of size $n\in\mathbb{N}$ from $X\sim f(x\mid\theta_\star)$ with $\theta_\star$ in the interior of Θ it takes its maximum $\hat\theta_n(X)$ for each n at a sequence of points $\hat\theta_n\to\theta_\star$ (mild regularity conditions on the statistical model will ensure this). By Taylor's theorem the log likelihood function in a neighborhood of $\theta_\star$ will be approximately

$$0=\ell_n'(\hat\theta)\approx\ell_n'(\theta_\star)+[\hat\theta_n-\theta_\star]\ell_n''(\theta_\star),\quad\text{so}$$

$$\sqrt{n}[\hat\theta_n-\theta_\star]\approx\frac{n^{-1/2}\ell_n'(\hat\theta)}{-n^{-1}\ell_n''(\theta_\star)}\approx\frac{\sqrt{n}\;\overline{Z(X;\theta)}}{I(\theta)}$$

by (3b), since the gradient of ℓ_n will vanish at $\theta_\star$. By the CLT, $\overline{Z(X;\theta)}$ is approximately normally distributed with mean zero and variance $I(\theta)/n$, so

$$\sqrt{n}[\hat\theta_n-\theta_\star]\approx\mathsf{No}\big(0,1/I(\theta_\star)\big).\tag{4}$$

A similar result holds for multivariate $\theta\in\Theta\subset\mathbb{R}^k$, where $\sqrt{n}\;\big[\hat\theta_n-\theta_\star\big]$ has approximately a k-variate $\mathsf{No}\big(0,I(\theta_\star)^{-1}\big)$ distribution.

This leads to an asymptotic $100\gamma\%$ confidence intervals of the form

$$\left[\;\hat\theta_n-\frac{z_{\alpha/2}}{\sqrt{nI(\hat\theta_n)}},\quad\hat\theta_n+\frac{z_{\alpha/2}}{\sqrt{nI(\hat\theta_n)}}\;\right]$$

for θ, where $\alpha=1-\gamma$ and $\Phi(-z_{\alpha/2})=\alpha/2$.

Optimality

One form of the "Cramér-Rao" or "Information" Inequality asserts that *any* statistic $T : \mathcal{X} \to \mathbb{R}$ with mean $\psi(\theta) = \mathsf{E}_\theta T(X)$ satisfies

$$\mathsf{V}_\theta T \geq \frac{|\psi'(\theta)|^2}{I(\theta)}$$

for one-dimensional parametric families. Any unbiased estimator $T(X)$ has mean $\psi(\theta) = \theta$ with $|\psi'(\theta)| \equiv 1$, so every unbiased estimator based on a sample of size n has $\mathsf{mse}[T_n] \geq 1/nI(\theta)$. By (4) the MLE achieves this lower bound asymptotically, so it is optimal in the sense that no other asymptotically unbiased estimator has strictly lower mse. Also we see that, under mild regularity, $\mathsf{mse}(T_n)$ decreases no faster than rate $\propto 1/n$ for *any* sequence T_n of estimators.

Cautionary Note

Sometimes "mild regularity" doesn't hold. For example, consider estimating the parameter θ for $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Un}(0,\theta)$ on the basis of the first n observations $X = (X_1, \cdots, X_n)$. The likelihood function

$$\mathcal{L}_n(\theta \mid X) = \prod_{i=1}^n \left\{\theta^{-1}\mathbf{1}_{[0,\theta]}(X_i)\right\} = \theta^{-n}\mathbf{1}_{[0,\theta]}(X_n^*)$$

vanishes for $\theta < X_n^*$ and decreases monotonically for $\theta > X_n^*$, the sufficient statistic $X_n^* := \max(X_i : \ 1 \leq i \leq n)$, so the MLE is in fact $\hat\theta_n = X_n^*$. The CDF for X_n^* is

$$F_n(x) = \mathsf{P}_\theta[X_n^* \leq x] = \{\mathsf{P}_\theta[X_1 \leq x]\}^n = (x/\theta)^n$$

for $0 \leq x \leq \theta$, and zero for $x < 0$, one for $x > \theta$. Thus the MLE has mean square error

$$\begin{aligned}\mathsf{mse}(\theta) &= \mathsf{E}_\theta\big|X_n^* - \theta\big|^2 \\ &= \int_0^\theta (x-\theta)^2 n x^{n-1}\theta^{-n}\,dx \\ &= 2\theta^2/(n+1)(n+2),\end{aligned}$$

shrinking at a rate faster than $\propto 1/n$. The error $[\hat\theta_n - \theta]$ does not have a Normal asymptotic distribution. Rather, it is always negative (since $\mathsf{P}_\theta[\hat\theta_n < \theta] = 1$) and, for $z > 0$, satisfies

$$\begin{aligned}\mathsf{P}_\theta\left\{n[\hat\theta_n - \theta] \leq -z\right\} &= \mathsf{P}_\theta\left\{\hat\theta_n \leq \theta - z/n\right\} \\ &= (1 - z/n\theta)^n \to \exp\big(-z/\theta\big).\end{aligned}$$

Thus $n[\theta - \hat\theta_n]$ has approximately the Weibull $\mathsf{We}(1, 1/\theta)$ distribution for large n, with mean θ and variance θ^2. For more on the Information Inequality, see (or click on)

https://stat.duke.edu/courses/Spring16/sta532/lec/fish.pdf

The Delta Method

Let $\tau = g(\theta)$ for some smooth function $g : \Theta \to \mathbb{R}$, and let $\{T_n\}$ be a consistent sequence of estimators of θ (like the MLEs $\hat\theta_n$ based on the first n observations) that are asymptotically normal in the sense that

$$\sqrt{n}[T_n - \theta] \approx \mathsf{No}(0, \sigma^2)$$

for large n, for some constant $\sigma^2 > 0$. By Taylor's theorem with remainder, there exist numbers $\tilde\theta_n$ between θ and T_n such that

$$\begin{aligned} g(T_n) &= g(\theta) + g'(\tilde\theta_n)[T_n - \theta], \quad \text{so} \\ \sqrt{n}[g(T_n) - g(\theta)] &= g'(\tilde\theta_n)\sqrt{n}[T_n - \theta] \\ &\approx \mathsf{No}(0, |g'(\theta)|^2\sigma^2), \end{aligned}$$

i.e., $g(T_n)$ is approximately normally-distributed with mean $g(\theta)$ and variance $|g'(\theta)|^2\sigma^2/n$.
Example: Let $S_n \sim \mathsf{Po}(n\lambda)$ be the sum of n iid RVs $X_i \sim \mathsf{Po}(\lambda)$ and let $\bar X_n := S_n/n$ be their sample mean. Then $\bar X_n \to \lambda$ by the LLN and $\sqrt{n}(\bar X_n - \lambda) \approx \mathsf{No}(0, \lambda)$ by the CLT. Set $g(x) := \log(1+x)$. By the Delta Method, the log of one plus the sample mean $g(X_n/n) = \log(1+\bar X_n)$ is approximately normally-distributed with mean and variance

$$\mathsf{E}[g(\bar X_n)] \approx \log(1+\lambda) \qquad\qquad \mathsf{V}[g(\bar X_n)] \approx \frac{\lambda}{n\,(1+\lambda)^2}.$$

A similar result holds for multivariate distributions with $\Theta \subset \mathbb{R}^k$.
Warning: What goes wrong in a similar argument with the functions $g(x) = \log(x)$ or $g(x) = 1/x$? Note that success of the Delta Method hinges on the assumption that, for large n, all the "action" is close to $T_n \approx \theta$, and that nothing too bad happens away from there.

The Parametric Bootstrap

Earlier we computed approximate features of the sampling distributions of statistics $T : \mathcal{X} \to \mathbb{R}$ by replicating draws $X_j^{(b)}$ from the empirical distribution $\hat F_n$ (*i.e.*, by draws with replacement from the random sample). In parametric models an alternative is to estimate $\theta \in \Theta$ by some estimator $\tilde\theta_n(X)$ (for example, the MLE or a Bayes estimator), then use simulation to approximate the sampling distribution of $T(X)$ by that of the "parametric bootstrap sample" $\left\{T\big(X_i^{(b)}\big) : \ X_i^{(b)} \overset{\text{iid}}{\sim} \hat F_n\big(x \mid \tilde\theta_n\big)\right\}$. This approach is typically more efficient than the usual nonparametric bootstrap when the parametric model is correct, but can be grossly misleading if the model is misspecified. It is prudent to check model assumptions, at least informally.

Checking Model Assumptions

Famously George Box quipped "All models are wrong, but some are useful" and, less famously,

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

One way to explore "how wrong" a model is is to make an overlay plot of a histogram of the data along with the pdf $f(x \mid \hat\theta_n)$ at the estimated value of $\theta \in \Theta$. If these differ grossly, the model may not be useful. Another is to make a "Q-Q plot" of data quantiles (on the vertical axis) *vs.* theoretical quantiles (on the horizontal axis) for the best-fitting distribution from the model family. For example, Figure (1) shows a Q-Q plot for a sample of $N = 100$ replicates for a normal $\mathsf{No}(\mu, \sigma^2)$ model, along with a histogram (same sample) with best $\mathsf{No}(\mu, \sigma^2)$ pdf fit overlaid. In fact these

data were generated from the t_2 distribution, with substantially heavier tails; the sigmoid shape of the curve is a tip-off about this problem, as are the outliers above the $x = y$ line on the right and below it on the left.

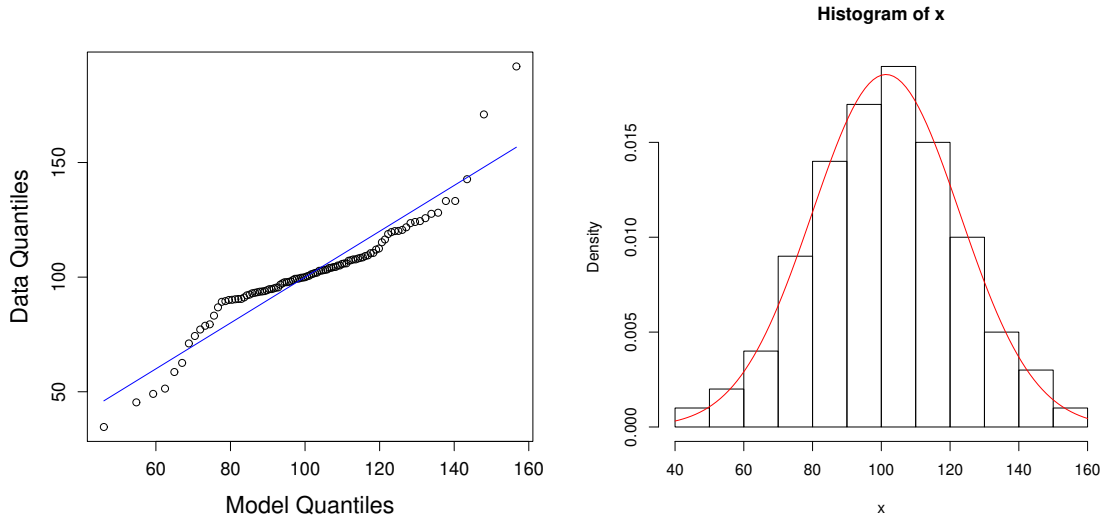

Figure 1: Q-Q plot and histogram to explore how well a $\mathsf{No}(\mu, \sigma^2)$ model fits data $\{X_i\}$. In fact, data were generated with $t_2(100, 10)$ distribution with heavier tails than the normal, but it's hard to see that from the histogram.

A more formal approach is to construct a "goodness of fit" test to check the hypothesis that the data come from the parametric family; we'll discuss that more when we study hypothesis testing in a few weeks.

Sufficiency

In some statistical models one can find a low-dimensional statistic $T(X)$ that captures all the information that a sample $X = \{X_1, \cdots, X_n\}$ has about the uncertain parameter $\theta \in \Theta$. For example, if $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Po}(\lambda)$, then the likelihood function

$$\begin{aligned}\mathcal{L}(\lambda) &\propto \prod_{i=1}^{n} \left[\lambda^{X_i} e^{-\lambda}\right] \\ &= \lambda^{n\bar{X}_n} e^{-n\lambda}\end{aligned}$$

depends on the data *only* through $\bar{X}_n$ (or, equivalently, through $\sum X_i$). A statistic $T(X)$ is said to be *sufficient* (for $\theta \in \Theta$) if the conditional distribution of the data vector X, given the value $t = T(X)$ of the statistic, does not depend on θ. In that case there is no more to be learned about θ from the entire vector than there is from $T(X)$ alone.

More examples:

If X is a vector of n iid Bernoulli $\mathsf{Bi}(1,p)$ random variables, then its joint pmf is

$$f(x \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{S_n}(1-p)^{n-S_n},$$

a function that depends on X only through $S_n = \sum_{i=1}^n X_i$ or (equivalently) of $\bar{X}_n = S_n/n$.

If X is a vector of n iid Beta $\mathsf{Be}(\alpha,\beta)$ random variables, then you can show that the two-dimensional statistic $T(X) := [\sum \log X_i, \sum \log(1-X_i)]$ is sufficient (you should do that).

If X is a vector of n iid Gamma $\mathsf{Ga}(\alpha,\lambda)$ random variables, then you can show that the two-dimensional statistic $T(X) := [\sum \log X_i, \sum X_i]$ is sufficient (do that too). If α is known then $\bar{X}_n$ is sufficient for λ; if λ is known then $\overline{\log X}_n$ is sufficient for α.

For iid Normal random variables $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu,\sigma^2)$, the sample mean $\bar{X}_n$ and sample variance S_n^2 are easily shown to be sufficient.

Two caveats:

If X is a vector of n iid Student $t_\nu(m,s)$ random variables with ν degrees of freedom, with location m and scale s, then there is no sufficient statistic for m (whether or not ν and s are known) with dimension less than n. The ordered sample $\left[X_{(1)} < X_{(2)} < \cdots < X_{(n)}\right]$ is sufficient, for example, but not any low-dimensional collection of moments as there were for the Poisson, Bernoulli, Beta, Gamma, and Normal examples. Thus, (1) sufficient statistics don't always exist, and (2) they are highly model-dependent. If one investigator models $\{X_i\}$ as Normal and reports only the (sufficient) sample mean and variance, another investigator concerned that the data may have heavier tails will be unable to estimate the parameters of a t_ν distribution from these statistics.

Finally— sometimes reporting only sufficient statistics will lose the opportunity to check the model. In the Bernoulli example, the likelihood function for a result of ten successes and four failures will be $\mathcal{L}(p) \propto p^{10}(1-p)^4$, for example, but if the original data were $[0,0,0,0,1,1,1,1,1,1,1,1,1,1]$ then there is reason to doubt the iid model, and ask questions like "did something happen between the fourth and fifth observations?"

Is T Sufficient?

A statistic T is sufficient for θ in a family $\{f(x \mid \theta) : \theta \in \Theta\}$ if and only if the pdf or pmf factors in the form of a product

$$f(x \mid \theta) = h(x)\ g\big(T(x),\theta\big)$$

of a function of x (that doesn't depend on θ) and a function of $T(x)$ and θ (that doesn't depend on x except through $T(x)$). This happens if and only if the *likelihood* function can be written as a function of only $T(X)$ and θ. Here's the most common place this happens:

Examples: Exponential Families

The pdf $f(x \mid \theta)$ for many parametric families of distributions can be written in the form

$$f(x \mid \theta) = h(x) \exp\big(\eta(\theta)T(x) - B(\theta)\big) \tag{5}$$

for some functions $h(x)$, $T(x)$ of $x \in \mathcal{X}$ and $\eta(\theta)$, $B(\theta)$ of $\theta \in \Theta$. For example:

$\mathsf{Po}(\lambda)$	$h(x) = \frac{1}{x!}$	$T(x) = x$	$\eta(\lambda) = \log\lambda$	$B(\lambda) = \lambda$
$\mathsf{Ex}(\lambda)$	$h(x) = 1$	$T(x) = x$	$\eta(\lambda) = -\lambda$	$B(\lambda) = -\log\lambda$
$\mathsf{No}(\theta, 1^2)$	$h(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$T(x) = x$	$\eta(\theta) = \theta$	$B(\theta) = \theta^2/2$
$\mathsf{No}(0, \theta^2)$	$h(x) = \frac{1}{\sqrt{2\pi}}$	$T(x) = x^2$	$\eta(\theta) = -1/2\theta^2$	$B(\theta) = \log\theta$
$\mathsf{Bi}(m, p)$	$h(x) = \binom{m}{x}$	$T(x) = x$	$\eta(p) = \log\frac{p}{1-p}$	$B(p) = -m\log(1-p)$

Note that if iid $\{X_i\}$ come from an exponential family with functions h, T, η, B then the vector $X = (X_1, \cdots, X_n)$ has an exponential family distribution too, with

$$h_n(x) = \prod h(x_i) \qquad T_n(x) = \sum T(x_i) \qquad \eta_n(\theta) = \eta(\theta) \qquad B_n(\theta) = nB(\theta)$$

Exponential families are especially pleasant to work with, for several reasons:

- They have a natural sufficient statistic, T (or $\sum T(x_i)$ for vectors), whose mean $B'(\theta)/\eta'(\theta)$ is easy to compute from the score $Z_\theta = \eta'(\theta)T(X) - B'(\theta)$;
- Their Fisher information $I(\theta) = B''(\theta) - \eta''(\theta)B'(\theta)/\eta'(\theta)$ is easy to compute, especially for "natural" exponential families where $\eta(\theta) = \theta$ and so $\eta'' \equiv 0$ and $I(\theta) = B''(\theta)$;
- The MLE $\hat\theta_n$ can be found by solving $\overline{T}_n = \mathsf{E}_{\hat\theta} T = B'(\hat\theta_n)/\eta'(\hat\theta_n)$;
- They have conjugate families of prior distributions (see below).

Note the name "exponential family" has nothing to do with the "exponential distribution" (except that $\mathsf{Ex}(\lambda)$ *is* one). Exponential families with multivariate parameter $\theta \in \Theta \subset \mathbb{R}^k$ arise too (usually with $k = 2$ or so); examples include $\mathsf{No}(\mu, \sigma^2)$, $\mathsf{Ga}(\alpha, \lambda)$, and $\mathsf{Be}(\alpha, \beta)$. The only change in the definition (5) is that now $T(x)$ and $\eta(\theta)$ are k-dimensional vector functions, and that the scalar product "$\eta(\theta)T(x)$" must be replaced with a dot product "$\eta(\theta) \cdot T(x)$". The sufficient statistic T has the same dimension (k) as the parameter space Θ.

Many distributions commonly used in modeling do have an exponential family representation, for fixed values of all but one (or sometimes two) of their parameters. Those include the beta, binomial, exponential, gamma, geometric, negative binomial, normal, Pareto, Poisson, and Weibull. Sufficient statistics and conjugate prior distributions are available for each of these. Many other useful distributions do *not* have exponential family form, and typically these will not have sufficient statistics or conjugate priors— including the logistic, Student t, α-stable, and others. One way to check is to write out the log likelihood function and see if its dependence on θ and x includes only terms that are functions of either θ or x but not both, plus a product (or dot-product) of a function of θ (that will be η) and a function of x (that will be $T(x)$).

Efficiency and Sufficiency

Let $T(\mathbf{x})$ be an estimator of a parameter θ, with MSE

$$m(\theta) = \mathsf{E}_\theta\big[\,|T(\mathbf{x})-\theta|^2\,\big].$$

The celebrated theorem of Raô and Blackwell asserts that if T does *not* depend on the data $\mathbf{x}$ through a sufficient statistic, then it can be improved by replacing it with one that does, in the following sense.

Theorem 1 *Let $\mathcal{F} = \{f_\theta(\mathbf{x}) : \ \theta\in\Theta\subset\mathbb{R},\ \mathbf{x}\in\mathcal{X}\}$ be a parametric statistical model. Let $T:\mathcal{X}\to\mathbb{R}$ be any statistic and let $S:\mathcal{X}\to\mathcal{S}$ be a sufficient statistic. Set*

$$T^*(\mathbf{x}) := \mathsf{E}_\theta\big[T(\mathbf{x})\mid S(\mathbf{x})\big],$$

*the conditional expectation of $T(\mathbf{x})$ given the value $s=S(\mathbf{x})$ of the sufficient statistic. Then T^* is a statistic (*i.e.*, does not depend on θ) with mse*

$$\mathsf{E}_\theta\big[|T^*(\mathbf{x})-\theta|^2\big] \le \mathsf{E}_\theta\big[|T(\mathbf{x})-\theta|^2\big],$$

with strict inequality unless $T(\mathbf{x})$ already depends only on $S(\mathbf{x})$, almost-surely.

Proof. Because S is sufficient, the conditional distribution of $\mathbf{x}$ given $S(\mathbf{x})$ doesn't depend on θ, so for any $t\in\mathbb{R}$

$$\mathsf{P}_\theta\big[T^*(\mathbf{x})\le t\mid S(\mathbf{x})\big]$$

is a function $g(s)$ of the value $s=S(\mathbf{x})$ that doesn't depend on θ either— *i.e.*, T^* is a statistic. The MSE of T is

$$\begin{aligned}\mathsf{E}_\theta\big[|T(\mathbf{x})-\theta|^2\big] &= \mathsf{E}_\theta\big[|T(\mathbf{x})-T^*(\mathbf{x})+T^*(\mathbf{x})-\theta|^2\big]\\ &= \mathsf{E}_\theta\big[|T(\mathbf{x})-T^*(\mathbf{x})|^2\big] + \mathsf{E}_\theta\big[|T^*(\mathbf{x})-\theta|^2\big]\end{aligned}$$

because the cross-product $\mathsf{E}_\theta\big(T(\mathbf{x})-T^*(\mathbf{x})\big)\big(T^*(\mathbf{x})-\theta\big)$ vanishes

$$\ge \mathsf{E}_\theta\big[|T^*(\mathbf{x})-\theta|^2\big],$$

with equality only if $\mathsf{E}_\theta\big[|T(\mathbf{x})-T^*(\mathbf{x})|^2\big]=0$— *i.e.*, only if $\mathsf{P}_\theta[T=T^*]=1$, a function of $S(\mathbf{x})$. □

Thus it's always best to use estimators that depend on the data only through sufficient statistics, when possible.

Examples

Let $\mathbf{x}=\{X_i\}_{1\le i\le 9}\overset{\text{iid}}{\sim}\mathsf{No}(\theta,1)$ be a sample of size $n=9$ from the normal distribution with unit variance. Then $\bar X_9$ and X_1 are both unbiased estimators of θ, but $\bar X_9$ is a function of the sufficient statistic $S_9=\sum X_i$ and X_1 is not. This is reflected in their MSEs,

$$\mathsf{E}|\bar X_9-\theta|^2 = 1/9 < 1 = \mathsf{E}|X_1-\theta|^2.$$

The inferior estimator X_1 can be improved, however, by the Raô-Blackwell procedure, to

$$X_1^* := \mathsf{E}_\theta\big[X_1 \mid S_9\big] = S_9/9 = \bar{X}_9$$

where we evaluated $\mathsf{E}_\theta\big[X_1 \mid S_9\big] = S_9/9$ by invoking symmetry. Similarly, the sample median $\mathring{X}_9$ has conditional expectation $\mathring{X}_9^* := \mathsf{E}_\theta\big[\mathring{X}_9 \mid S_9\big] = \bar{X}_9$ with lower MSE. The sample mean $\bar{X}_n$ cannot be improved, however; a naïve attempt at doing so yields

$$\bar{X}_9^* := \mathsf{E}_\theta\big[\bar{X}_9 \mid S_9\big] = S_9/9 = \bar{X}_9,$$

unchanged.

Conjugate Prior Distributions

A family $\{\pi_\xi(\theta) :\ \xi \in \Xi\}$ of prior density (or pmf) functions on Θ is called "conjugate" for a statistical model $\{f(x \mid \theta) :\ \theta \in \Theta\}$ if the posterior distribution upon observing $X \in \mathcal{X}$ is again in the family $\{\pi_\xi(\theta) :\ \xi \in \Xi\}$, *i.e.*, if for each $\xi \in \Xi$ and $x \in \mathcal{X}$ there is some $\xi^{(x)} \in \Xi$ such that

$$\pi_\xi(\theta) f(x \mid \theta) \propto \pi_{\xi^{(x)}}(\theta).$$

For example, the family $\mathsf{Be}(\alpha, \beta)$ of prior distributions for $\theta \in \Theta = [0,1]$ is conjugate for the Binomial statistical model $X \sim \mathsf{Bi}(n, \theta)$ for any fixed $n \in \mathbb{N}$: if $\theta \sim \mathsf{Be}(\alpha, \beta)$ and $X \sim \mathsf{Bi}(n, \theta)$, then the posterior distribution is $\theta \mid X \sim \mathsf{Be}(\alpha + X, \beta + n - X)$, *i.e.*,

$$\begin{aligned}\Big\{\pi_\xi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}\Big\}\ \Big\{f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\Big\}& \\ \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \propto \Big\{\pi_{\xi^x}(\theta) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\ \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\Big\}&\end{aligned}$$

Similarly if $\lambda \sim \mathsf{Ga}(\alpha, \beta)$ and $X \sim \mathsf{Po}(\lambda)$, then $\lambda \mid X \sim \Gamma(\alpha+X, \beta+1)$. Note that both binomial and Poisson are exponential family distributions— *every* exponential family (see (5)) has a conjugate family of prior distributions of the form

$$\pi_\xi(\theta) = c_\xi^{-1}\pi_\star(\theta)\exp\big(\eta(\theta)\xi_1 - \xi_2 B(\theta)\big)$$

for any function $\pi_\star(\theta) \ge 0$ for which the set Ξ of $\xi = (\xi_1, \xi_2)$ for which

$$0 < c_\xi := \int_\Theta \pi_\star(\theta)\exp\big(\eta(\theta)\xi_1 - \xi_2 B(\theta)\big)\,d\theta < \infty$$

is non-empty and closed under the addition of $(t, 1)$ for any possible value t of $T(X)$. Evidently $\xi^{(x)} = \big(\xi_1 + T(X), \xi_2 + 1\big)$ or, for a sample of size n, $\xi^{(\mathbf{x})} = \big(\xi_1 + \sum T(X_i), \xi_2 + n\big)$.

Conjugate prior distributions simplify analyses by making it unnecessary to do any numerical integration to perform Bayesian statistical inference, *i.e.*, to evaluate features of posterior probability distributions (means, credible sets, *etc.*) For the first two centuries after the emergence of Bayesian methods, this simplification was critical— and led investigators to limit themselves to conjugate prior distributions or low-dimensional models.

With the emergence in the 1990s of hardware and algorithms capable of evaluating essentially any posterior distribution it's no longer necessary to limit attention to these families. This permits modelers to use more realistic models with both more realistic sampling distributions and more realistic prior distributions.

5 Summary: Desirable Properties for Estimators

We have seen a number of properties that an estimator $T(X)$, or a sequence of estimators $T_n(\mathbf{x})$, might have in a parametric statistical model $\mathfrak{F} = \{f(x \mid \theta) : \ \theta \in \Theta\}$. Some of these are:

Consistent: A sequence $T_n(\mathbf{x})$ of estimators of a feature $\tau = g(\theta)$ of $\theta \in \Theta$ is *consistent* if

$$\lim_{n\to\infty} \mathsf{P}_\theta[\ |T_n(\mathbf{x}) - \tau| > \epsilon\ ] = 0$$

for every $\epsilon > 0$.

MS Consistent: A sequence $T_n(\mathbf{x})$ of estimators of a feature $\tau = g(\theta)$ of $\theta \in \Theta$ is *mean square consistent* (or MS consistent) if

$$\lim_{n\to\infty} \mathsf{E}_\theta\ |T_n(\mathbf{x}) - \tau|^2 = 0.$$

This is a stronger condition than consistency, *i.e.*, if $T_n(\mathbf{x})$ is MS consistent then it is also consistent, since $\mathsf{P}_\theta[\ |T_n(\mathbf{x}) - \tau| > \epsilon\ ] \le \mathsf{E}_\theta\ |T_n(\mathbf{x}) - \tau|^2/\epsilon^2$ for any $\epsilon > 0$.

Efficient: An estimator $T(X)$ with mean $m(\theta) = \mathsf{E}_\theta T(X)$ is *efficient* if it satisfies the Information Inequality lower bound,

$$\mathsf{E}_\theta|T(X) - \tau|^2 = \frac{[m'(\theta)]^2}{I(\theta)} + \big[m(\theta) - \tau\big]^2.$$

In particular, an unbiased estimator $T(X)$ of θ is efficient if $\mathsf{E}_\theta|T(X) - \theta|^2 = 1/I(\theta)$. Some authors limit the definition of "efficient" to this case.

Asymptotically Efficient: A sequence $T_n(\mathbf{x})$ of estimators of θ is *asymptotically efficient* if it satisfies

$$\lim_{n\to\infty} n\mathsf{E}_\theta|T_n(\mathbf{x}) - \theta|^2 = \frac{1}{I(\theta)}$$

Unbiased: An estimator $T(X)$ of a feature $\tau = g(\theta)$ is *unbiased* if $\mathsf{E}_\theta T(X) = \tau$ for every $\theta \in \Theta$.

Asymptotically Unbiased: A sequence $T_n(\mathbf{x})$ of estimators of $\tau = g(\theta)$ is *asymptotically unbiased* if $\lim_{n\to\infty} \mathsf{E}_\theta T_n(\mathbf{x}) = \tau$ for every $\theta \in \Theta$.

Sufficient: A statistic S is *sufficient* if the conditional distribution of X, conditional on $S(X) = s$, does not depend on θ. Equivalently, the pdf can be written as a product

$$f(x \mid \theta) = h(x)\ g\big(S(x), \theta\big)$$

of a function $h(x)$ of the data only, and a function $g\big(S(x), \theta\big)$ that depends on θ and also on the data, but only through the value of the statistic $S(X)$.

Admissible: An estimator $T(X)$ of a feature $\tau = g(\theta)$ is *admissible* if there does *not* exist an estimator $S(X)$ that is "better" than $T(X)$ in the sense that

$$(\forall \theta \in \Theta)\ \mathsf{E}_\theta\big|S(X) - \tau\big|^2 \le \mathsf{E}_\theta\big|T(X) - \tau\big|^2,$$

with strict inequality for at least one $\theta \in \Theta$.

Last edited: January 9, 2019