Lecture 7
"By Shape of likelihood the news was told" Henry IV. part 1
Introduction to the MLE via Sufficiency.

We will begin with a brief introduction to statistical decision theory

$\theta \in \Theta$ states of the world, e.g. rain or no rain today.
$a \in \mathcal{A}$ actions, e.g. umbrella, or no umbrella today.
$L(a, \theta)$ loss function
$Z \in \mathcal{Z}$ data (information relevant to Θ, like the weather forecast.)
$a = d(z)$ decision function (rule), how the forecast influences the action.

Assuming Z has a density, given $\theta, f(z|\theta)$, we will consider the risk of the decision rule $d(\cdot)$ as,

$$R(\theta, d(Z)) = E_Z L(\theta, d(Z)) = \int L(\theta, d(z)) f(z|\theta) dz$$

Obviously, some rules will be good for some θ's, e.g., $d(z) = \text{argmin}_d L(\theta_0, d)$ is especially good if $\theta = \theta_0$, but we would like the rule to be good in some broader sense.

Def: We say d^* *dominates* d, or $d^* \succeq d$ iff $R(\theta, d^*) \leq R(\theta, d)$ for all $\theta \in \Theta$. And d^* strictly dominates d, $d^* \succ d$ if $d^* \succeq d$ and $R(\theta, d^*) < R(\theta, d)$ for some $\theta \in \Theta$.

Definition d is *admissible* iff no d' strictly dominates d.

Finding admissible decision rules was one of the major party games of the statistics profession in the 1950's and 1960's inspired by the work of Wald, Stein, LeCam and others. It slowly lost popularity as participants began to realize that it was extremely difficult to get beyond what had already been done by their forefathers.

Example The most celebrated admissibility result is Stein's (1956) discovery that the multivariate sample mean is inadmissible in dimension 3 or more. An explicit improvement is achieved by the James-Stein (1961) estimator. Suppose we have (possibly correlated) jointly normal random variables, $Y_1, \cdots, Y_n$ with $Y_i \sim \mathcal{N}(\mu_i, 1), i = 1, \cdots, n$ and would like to estimate the n-vector μ under squared error loss:

$$\mathcal{L}(\hat{\mu}, \mu) = \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 = \|\hat{\mu} - \mu\|^2$$

The MLE for each μ is just the (unbiased) vector Y itself,[1] but James and Stein demonstrated that the estimator

$$\hat{\mu}_i = (1 - \frac{n-2}{S}) Y_i$$

with $S = \sum Y_i^2$ has better $\mathcal{L}$ than the MLE whenever $n \geq 3$.

[1] Consider the regression of Y on the n-dimensional identity matrix, regardless of the form of the error covariance matrix, the MLE is $\hat{\mu} = Y$.

To see this we review the Efron's version of Stein's (1981) argument. Given the identity,

$$(\hat{\mu}_i - \mu_i)^2 = (\hat{\mu}_i - Y_i)^2 - (Y_i - \mu_i)^2 + 2(\hat{\mu}_i - \mu_i)(Y_i - \mu_i),$$

summing and taking expectations yields,

$$\mathbb{E}_\mu \|\hat{\mu} - \mu\|^2 = \mathbb{E}_\mu \|Y - \hat{\mu}\|^2 - n + 2 \sum \mathrm{Cov}(\hat{\mu}_i, Y_i),$$

where the covariance can be computed [Why?] as if $Y \sim \mathcal{N}(\mu, I_n)$. By Stein's lemma[2]

$$\mathrm{Cov}(\hat{\mu}_i, Y_i) = \mathbb{E}_\mu \frac{\partial \hat{\mu}_i}{\partial Y_i}$$

implying that

$$\begin{aligned} \sum \mathrm{Cov}(\hat{\mu}_i, Y_i) &= \mathbb{E}_\mu \sum [(1 - \frac{n-2}{S}) + 2\frac{n-2}{S^2} Y_i^2] \\ &= n - \mathbb{E}_\mu \frac{(n-2)^2}{S}, \end{aligned}$$

while,

$$\mathbb{E}_\mu \|Y_i - \hat{\mu}\|^2 = \mathbb{E}_\mu \sum (\frac{(n-2)}{S} Y_i)^2 = \mathbb{E}_\mu \frac{(n-2)^2}{S}.$$

Thus,

$$\mathbb{E}_\mu \|\hat{\mu} - \mu\|^2 = n - \mathbb{E}_\mu \frac{(n-2)^2}{S},$$

and the last term is positive for $n \geq 3$.

Shrinking the MLE toward the origin dominates the MLE itself irrespective of what the original μ_i might look like. This still seems quite astonishing after 50 years, and still hasn't penetrated most textbook treatments of econometrics. As data becomes more plentiful, this result and related ideas arising from empirical Bayes methods are becoming increasingly important.

Two approaches to choosing $d(z)$:

(1) *Minimaxity* Find the "least favorable" case for each d, i.e.,

$$M(d) = \max_{\theta \in \Theta} R(\theta, d)$$

and then choose d to minimize $M(d)$. (May be unduly pessimistic.)

(2) *Bayes Rules* Assign (subjective) probabilities to various states and compute expected risk

$$\begin{aligned} B(d) &= E_\theta R(\theta, d) = \int_\Theta R(\theta, d) g(\theta) d\theta \\ &= \int_\Theta \int L(\theta, d(z)) f(z|\theta) g(\theta) dz d\theta \\ &= \int \{ \int_\Theta L(\theta, d(z)) h(\theta|z) d\theta \} p(z) dz \end{aligned}$$

[2]Stein's lemma, apparently attributable to Chernoff, asserts that if $X \sim \mathcal{N}(0, I_n)$ and $g(\cdot)$ is differentiable and well behaved (enough) as $\|x\| \to \infty$, then $\mathrm{Cov}(g(X), X_i) = \mathbb{E}\partial g(X)/\partial x_i$. This follows from the fact that, $\phi'(x) = -\phi(x)x$ so integrating by parts and invoking the well behavedness of g, we have, $\int g(x)x\phi(x)dx = -\int g(x)\phi'(x)dx = \int g'(x)\phi(x)dx$.

where we have written

$$\begin{array}{ccccc} f(z|\theta) & g(\theta) & = & h(\theta|z) & p(z) \\ \downarrow & \downarrow & & \downarrow & \\ \text{likelihood} & \text{prior on } \theta & & \text{Posterior on } \theta & \end{array}$$

A *Bayes Rule* is a $d(\cdot)$ which minimizes this Bayes Risk, which is accomplished by minimizing the term in $\{\quad\}$'s. There is a general principle and a whole genre of theorems that assert that admissible decision rules are limits of Bayes rules. These so-called "complete class theorems" become very technical, but the intuitive appeal is obvious: there should a prior that would justify any admissible rule. If there isn't such a prior, it is hard to imagine how the decision rule could be admissible.

Often as a first shot at a problem we may wish to make a point estimate of θ as our decision rule, i.e., $\hat{\theta} = d(z)$. This is an obvious "academic" setting: we are interested in publishing a paper based on some empirical analysis, we have no power to implement policies, or engage in real actions. Instead we are able to make some assertion about the parameters in a model that we have estimated. In this case

$$(*) \qquad\qquad L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|^p$$

is a commonly used family of loss functions. We will remark in passing that

$$\begin{array}{lcll} p = 2 & \Rightarrow & \hat{\theta} = \int \theta h(\theta, y) d\theta & \text{mean of posterior} \\ p = 1 & \Rightarrow & \hat{\theta} = \inf\{t | \int_{-\theta}^{t} h(\theta, y) d\theta \geq 1/2\} & \text{median} \\ p = \infty & \Rightarrow & \hat{\theta} = \text{argmax}\{h(\theta, y)\} & \text{mode} \end{array}$$

The former case is easily proven by a familiar argument. Let Z be a random variable with density h we seek $\hat{\theta}$ to minimize

$$E(Z - \hat{\theta})^2 = E(Z - \mu)^2 + (\mu - \hat{\theta})^2$$

which decomposes MSE into variance plus squared bias as before and is obviously minimal when $\hat{\theta} = \mu$.

The median example is a special case of the following more general asymmetric linear loss function

$$L(\theta, \hat{\theta}) = \tau(\theta - \hat{\theta})^+ + (1 - \tau)(\theta - \hat{\theta})^-$$

$$\min \int L(\theta, \hat{\theta}) h(\theta, z) d\theta$$

$$\begin{array}{ll} f.o.c. & -\tau \int_{\hat{\theta}}^{\infty} h d\theta + (1 - \tau) \int_{-\infty}^{\hat{\theta}} h d\theta = 0 \\ & \Rightarrow -\tau(1 - H(\hat{\theta})) + (1 - \tau) H(\hat{\theta}) = 0 \\ & \Rightarrow \tau = H(\hat{\theta}) \Rightarrow \hat{\theta} = H^{-1}(\tau). \end{array}$$

The third commonly considered loss function is $0 - 1$ loss. For a continuous parameter we can approximate this form of loss by

$$L(\theta, \hat{\theta}) = I(|\theta - \hat{\theta}| > \varepsilon)$$

which for sufficiently small ϵ leads to choosing the mode of the posterior.

Often we are loath to admit that we *have* prior information, but prefer to let the current data analysis "speak for itself." In this case it may be reasonable to take the prior density $g(\theta)$ as uniform over some relevant range (move quickly here, or we will be lost in the Bayesian swamp) in which case the posterior is proportional to the likelihood. From this Bayesian point of view it is natural that all the information about the parameter available from the experimental realization, z, is contained in the likelihood.

A somewhat more classical point of view may be illustrated by considering a problem with a discrete sample space, $\mathcal{Z} \in \{z_1, \dots\}$ and only two possible states $\Theta = \{\theta_0, \theta_1\}$. Let $A = \{z | f(z,\theta_1)/f(z,\theta_0) = c\}$, then if $z_i \in A$, we have

$$\frac{f(z_i,\theta_1)}{\sum_{z\in A} f(z_i,\theta_1)} = \frac{cf(z_i,\theta_0)}{\sum_{z\in A} cf(z_i,\theta_0)} = \frac{f(z_i,\theta_0)}{\sum_{z\in A} f(z_i,\theta_0)}$$

I.e., the θ_1-conditional probabilities given $z \in A$ are the same as the θ_0-conditional probabilities given $z \in A$. So the conditional distribution of any statistic $T(z)$ will be the same for both values of the parameter, i.e., once we know $\lambda = f(z,\theta_1)/f(z,\theta_0), T(z)$ will yield no new information about θ. We will say λ is a sufficient statistic in this context. Generalizing somewhat, when θ takes countably many values we may regard the function

$$\lambda(\theta) = f(z,\theta)/f(z,\theta_0)$$

as sufficient. Once we know the values taken by this function for the various values of θ, nothing more of use can be extracted from the sample about which θ is more likely. Note however, that when $\mathcal{Z}, \Theta$ are uncountable then the conditioning is dicey, but we should emphasize that in some problems a dramatic reduction of the data is possible via sufficiency. This is a concept which we will now introduce somewhat more formally.

Sufficiency

Given a r.v. Z, we may define a statistic $T(Z)$, e.g., $Z = (X_1, \dots, X_n)$ and $T(Z) = \bar{X}_n$. We say T is *sufficient* for the family $\mathcal{F}$ if

$$f_{Z|T}(z|t) = f_{Z,T}(z,t)/f_T(t)$$

is the same for all distributions in $\mathcal{F}$. Typically, $\mathcal{F}$ is indexed by a parameter θ so T sufficient for $\mathcal{F}$, really means T sufficient for θ and that $f_{Z|T}(z|t,\theta)$ doesn't depend upon θ. In other words, once we know $T = t$, knowing the conditional density of Z given t isn't informative about θ. Kiefer(1987) gives a nice characterization of sufficiency in terms of a probabilistic ability to reconstruct the sample.

Thm: (Factorization) Given a parametric family $\mathcal{F}_\theta$ $T(Z)$ is sufficient for θ iff there exist functions g and h such that for all $\theta \in \Theta$,

$$f(z|\theta) = g(T(z),\theta)h(z)$$

Proof: See Bickel and Doksum. The fundamental paper is Halmos and Savage (1949).

Example 1: Gaussian Linear Model

$$z = y \qquad f(y,\theta) = (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{\mathcal{S}}{2\sigma^2} - \frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{2\sigma^2}\right\}$$

$$\theta = (\beta,\sigma^2) \qquad \hat{\beta} = (X'X)^{-1}X'y; \quad \mathcal{S} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

hence $T(z) = (\hat{\beta}, \mathcal{S})$ is sufficient for $\theta \equiv (\beta, \sigma^2)$. In this model we can throw away y once we have computed $(\hat{\beta}, \mathcal{S})$.

Example 2: Uniform $Y_i \sim U[0, \theta] \quad i = 1, \ldots, n$

$$\begin{aligned} f_Y(y|\theta) &= \frac{1}{\theta^n} \prod_{i=1}^{n} u(0, y_i) u(y_i, \theta) \\ &\qquad \text{where } u(a,b) = \begin{cases} 1 & \text{if } a < b \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\theta^n} u(T(y), \theta) \prod_{i=1}^{n} u(0, y_i) \end{aligned}$$

where $T(y) = y_{(n)}$, showing $T(y)$ is sufficient for θ.

Example 3: Cauchy Location $Y_i \sim \text{Cauchy}(\theta, 1)$

$$f_Y(y|\theta) = \pi^{-n} \prod_{i=1}^{n} (1 + (y_i - \theta)^2)^{-1}$$

here there is very little reduction possible. Trivially, one can say that the original sample $(y_1, \ldots, y_n)$ is sufficient, and we can go a little bit further. Since the observations are independent, the ordering of the observations is immaterial, and we can take the order statistics, $Y_{(1)}, \ldots Y_{(n)}$ as sufficient.

Sufficiency needs to be clarified somewhat since we would like to convey the notion that there is not redundant information in T. Obviously we can always say trivially "the sample is sufficient for the sample", as we did in the Cauchy example, but this isn't too interesting.

Def: A sufficient statistic T is said to be *minimal* if of all sufficient statistics it provides the greatest possible reduction in the data, i.e., if for any sufficient statistic U there exists a function H such that $T = H(U)$.

Remark: If T has redundant components, then they won't be computable based only on U. Obvious example: you can't reconstruct the whole sample from $(\hat{\beta}, \mathcal{S})$ in Example 1 above.

This leads to the question what should we call statistics which are redundant once we have computed the sufficient statistics?

Def: A statistic $V(X)$ is said to be *ancillary* (with respect to θ) if its distribution does not depend upon θ.

An awkward aspect of sufficiency is that a minimal sufficient statistic may contain much ancillary information.

Def: A sufficient statistic T is said to be *complete* if

$$E_\theta f(T) = 0 \quad \text{for} \quad \text{all} \quad \theta \in \Theta \Rightarrow f(t) = 0 \quad a.e.P.$$

where $P = \{\mathcal{F}_\theta, \theta \in \Theta\}$.

Examples:

: 1. A nice example of this is $X_i \sim U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. Here $T = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic, but T isn't complete since

$$E\left(X_{(n)} - X_{(1)} - \frac{n-1}{n+1}\right) = 0$$

To see this, take $\theta = 1/2$, so $X_i \sim U[0,1]$ then

$$\begin{aligned} P(X_{(n)} < x) &= P(X_i < x, \quad i = 1, \ldots, n) \\ &= P(X_i < x)^n \\ &= x^n \quad \text{for } x \in [0,1] \end{aligned}$$

Thus,

$$F_{X_{(n)}}(x) = x^n I_{[0,1]}(x) + I_{[1,\infty)}(x)$$

and

$$f_{X_{(n)}}(x) = n x^{n-1} I_{[0,1]}(x)$$

so

$$EX_{(n)} = n \int_0^1 x^n dx = \frac{n}{n+1}$$

and similarly

$$EX_{(1)} = \frac{1}{n+1}.$$

Since $X_i \sim U[\theta - 1/2, \theta + 1/2]$ is just a $U[0,1]$ r.v. with the constant $\theta - 1/2$ subtracted, the expectations are just shifted by $\theta - 1/2$ and the difference in expectations is the same as it is in the $U[0,1]$ case.

: 2. In the $X_i \sim U[0,\theta]$ it *is* possible to show that $X_{(n)}$ is a complete sufficient statistic.

: 3. In the exponential family models introduced in the next lecture it is possible to show that the sufficient statistics are also complete provided the parameter space includes a p-dimensional rectangle. However, the proof of this requires a bit a complex analysis which seems beyond the purview of this course, see Lehmann (1959 p 132-33).

: 4. Not surprisingly, one can show that if T is a complete sufficient statistic, then it has to be minimal. See Lehmann and Scheffé (1950).

Thm: (Basu) If T is a complete sufficient statistic for $\mathcal{F}_\theta$, then any ancillary statistic V is independent of T.

Proof: If V is ancillary, $p_A = P(V \in A)$ is independent of θ for all A. Let

$$\pi_A(t) = P(V \in A | T = t)$$

then

$$E\pi_A(T) = p_A$$

hence by completeness

$$\pi_A(t) = p_A.$$

Remark: The subtlety here is that completeness means that $E_\theta f(T) = c \Rightarrow f(t) = c$ since we can always redefine f to subtract (add) c. This result, although it may seem obscure at this point, underlies the Hausman test variance computation as we shall see later in the course.

In fact, among certain regular families of densities {smooth, with support independent of parameters} only a relatively small class of "happy families" permit dimension reduction through sufficiency. We shall look at this class next time. In the non-regular case $U[0,\theta]$ offers an example

of a density which admits a 1-dimensional sufficient statistic: the sample maximum. This means that all of the foregoing theory is really of quite restricted applicability.

The existence of a complete sufficient statistic in a particular practical problem leads to the existence of an optimal estimator (in the sense of convex loss) through the following result.

Thm: (Rao-Blackwell) Let T be a complete sufficient statistic for the parameter θ, and suppose that there exists an unbiased estimator of θ, then there exists a uniformly best unbiased estimator, of θ, for any convex loss function, and it is the function of T which is an unbiased estimator of θ.

Proof: By sufficiency $P(X = x|T(x) = t)$ doesn't depend on θ, thus if S is any unbiased estimator, we can write it as

$$\tilde{S}(t) = \int S(x)P(X = x|T(x) = t)dx$$

(We may interpret $\tilde{S}$ is just the estimation rule $\tilde{S}(T(X))$ so it really is a mapping from the sample space.) It is easy to see that

$$E_\theta \tilde{S}(T(X)) = E_\theta S(X) = \theta$$

Now let $L(\theta, D)$ be a loss function that is, for any θ, convex in D. By Jensen's inequality,

$$L(\theta, \tilde{S}(t)) \leq \int L(\theta, S(x))P(X = x|T(x) = t)dx$$

for every possible value t of $T(X)$. Multiplying both sides by $P_\theta(T(X) = t)$ and summing we have,

$$\begin{aligned} EL(\theta, \tilde{S}(T(X)) &\leq \int\int L(\theta, S(x))P(X = x|T(x) = t)P_\theta(T(X) = t)dxdt \\ &= EL(\theta, S(X)) \end{aligned}$$

so $\tilde{S}(T(X))$ is at least as good as any other unbiased estimator for any convex loss function.

Finally, suppose $\tilde{S}_1$ and $\tilde{S}_2$ are two functions such that

$$E_\theta \tilde{S}_1(T(X)) = E_\theta \tilde{S}_2(T(X))$$

for all θ. By completeness, the only function h of $T(X)$ for which

$$E_\theta h(T(X)) = 0 \qquad \text{for all } \theta \in \Theta$$

is the function h which is itself zero with probability one for all θ. So

$$\tilde{S}_1(t) - \tilde{S}_2(t) = 0$$

and thus

$$\tilde{S}_1(t) = \tilde{S}_2(t)$$

with probability one. Therefore if S_1 and S_2 are unbiased estimators of θ and if we obtain $\tilde{S}_1$ and $\tilde{S}_2$ as above, it follows that $\tilde{S}_1(T(X)) = \tilde{S}_2(T(X))$ $wp1$ for all θ. So for any unbiased estimator S the same $\tilde{S}$ is at least as good as S, and this $\tilde{S}$ is at least as good as any such S which is what was asserted.

Remark: This result yields the celebrated method of Rao-Blackwell: Given any unbiased estimator of θ, say S, we can then construct $\tilde{S}$, and it will be better than S in terms of convex loss.

Example: Suppose that $X_1, ..., X_n$ are iid from $U[0, \theta]$. We saw earlier that $X_{(n)}$ is a sufficient statistic with density,

$$f(t, \theta) = n\theta^{-n} t^{n-1} I_{[0,\theta]}(t),$$

and

$$E_\theta g(X_{(n)}) = n\theta^{-n} \int_0^\theta g(t) t^{n-1} dt.$$

Completeness requires that $E_\theta g(X_{(n)}) = 0$ implies that $g(x) \equiv 0$. This follows by differentiating with respect to θ. Since $E_\theta X_{(n)} = \frac{n\theta}{n+1}$, we have

$$\hat{\theta} = X_{(n)} \frac{n+1}{n}$$

is UMVU.

References:

A.W.F. Edwards, *Likelihood*, Johns Hopkins Press.

C. Gourieroux and A. Monfort, *Statistics and Econometric Models*, Cambridge U. Press.

P. Halmos, and L.J. Savage, Applications of the Radon-Nikodyn Theorem to the theory of sufficient statistics, *Annals of Math. Statistics*, 20, 225-41.

J. Kiefer, (1987) *Introduction to Statistical Inference*, Springer.

James, W., and C. Stein (1961): "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, p. 361. Univ of California Press.