2.2 Estimation and convexity. In an estimation problem, we are given a family $\mathcal{P}$ of laws (probability measures) on a sample space $(X, \mathcal{B})$, we observe a point x of X, where often X is an n-fold product, $x = (X_1, \ldots, X_n)$, and we want to estimate a real-valued function g on $\mathcal{P}$. Usually $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ and g is written as a function of the parameter(s), $g(P_\theta) = g(\theta)$.

Estimation problems are decision problems as treated in Secs. 1.2 and 1.3 where in estimation, if one is trying to estimate θ itself, the action space is equal to the parameter space Θ or, possibly, a larger space including the parameter space, such as its closure if Θ is not closed. In estimation of a function $g(\theta)$, the action space will be a subset of some measurable space S which includes the range of g, where g is assumed measurable. The loss function in a problem of estimating $g(\theta)$ will be assumed to satisfy $L(\theta, T) \geq 0$ and $L(\theta, T) = 0$ if and only if $T = g(\theta)$. Often, for a metric (distance) d defined on S, the loss function $L(\theta, T)$ will be an increasing function of $d(T, g(\theta))$, being small when T is close to $g(\theta)$ and large when T is far from $g(\theta)$. In any case, a (non-randomized) decision rule for an estimation problem is a measurable function from X^n into S, called an *estimator*. A particular value of an estimator obtained for some given data is called an *estimate*. For an estimator $T(X_1, \ldots, X_n)$ we then have the *risk*

$$r(\theta, T(\cdot)) \ = \ E_\theta L(\theta, T(X_1, \ldots, X_n)).$$

Here E_θ is the expectation when $X_1, \ldots, X_n$ are i.i.d. (P_θ), so that

$$E_\theta \ := \ \int \cdots \int \cdot dP_\theta(X_1) \cdots dP_\theta(X_n).$$

The parameter spaces and spaces S statisticians have considered up to now have often been subsets of Euclidean spaces $\mathbb{R}^k$ with their usual norms $\|x\| \ := \ (x_1^2 + \cdots + x_k^2)^{1/2}$. The loss function most treated in classical statistics has been $L(\theta, T) = \|T - g(\theta)\|^2$. In one dimension, this is just the squared difference between the estimate and the quantity to be estimated, called "squared-error loss." This is not to say that in applications of statistics, individuals in fact suffer losses, even approximately, proportional to the squared errors. Rather, the theory based on squared-error losses has been easier to work out and is relatively traditional in statistics. It may be hoped that results for squared-error losses will shed light on the situation for other, possibly more realistic loss functions.

For example, let $X = \Theta = \mathbb{R}$ and $P_\theta = N(\theta, 1),\ \theta \in \mathbb{R}$. Then the classical estimator for θ is $T(X_1, \ldots, X_n) = \overline{X} \ := \ (X_1 + \cdots X_n)/n$. It does minimize the mean squared-error loss under some conditions, as will be seen later.

Testing between two simple hypotheses P and Q, as in Secs. 1.1 and 1.5-1.7, is a special case of estimation in which the parameter space has only two points. But note that in that case, if the losses L_{PQ} and L_{QP} are different, the loss is not any function of a metric d on $\mathcal{P} = \{P, Q\}$, which would satisfy $d(P, Q) = d(Q, P)$.

Admissibility for estimators (for a given loss function) is defined from the general definition for decision rules. Likewise, if a prior is given on the parameter space, an estimator which minimizes the overall risk is called a *Bayes* estimator.

For a fixed $\phi \in \Theta$, consider the constant estimator $T \equiv g(\phi)$. This makes $r(T, \phi) = 0$, a minimum. If this T is not admissible, there is some other statistic U such that $U(X_1, \ldots, X_n) = g(\phi)$ almost surely for P_ϕ^n. This implies that for any $\theta \in \Theta$ such that P_θ is absolutely continuous with respect to P_ϕ, we also have $U = g(\phi)$ almost surely for P_θ^n, so that U cannot distinguish θ from ϕ. If all the P_θ are equivalent (mutually absolutely continuous), for example if each $P_\theta = N(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$, the constant T will always be admissible.

This trivial estimator T is admissible because it does well when $g(\theta) = g(\phi)$. For other θ, T will tend to do badly. Unlike reasonable estimators, the estimator T does not provide a better approximation to the true $g(\theta)$ as n increases. So it will not be enough for an estimator to be admissible; some other good properties must be sought. One way to rule out such bad behavior as that of constant estimators is the following property:

Definition. If g is a function from Θ into some $\mathbb{R}^k$, a statistic $T(X_1, \ldots, X_n)$ with values in $\mathbb{R}^k$ is called an *unbiased* estimator of g iff for all $\theta \in \Theta$, $E_\theta T = g(\theta)$.

The sample mean $\overline{X}$ is evidently unbiased as an estimator of the true mean for any distribution having a finite mean. Constant estimators of a non-constant function $g(\theta)$ will not be unbiased. A requirement that estimators be unbiased can, however, lead to bad estimators as in the following:

Example. A function $g(\theta)$ may have a unique unbiased estimator which is inadmissible. For $\lambda > 0$, let C_λ be the Poisson distribution with parameter λ, conditioned on $k \geq 1$, in other words

$$C_\lambda(k) \ := \ e^{-\lambda}\lambda^k/(k!(1 - e^{-\lambda})) \ \text{ for } \ k = 1, 2, \ldots .$$

Let $g(\lambda) \ := \ e^{-\lambda}$. For $V(\cdot)$ to be an unbiased estimator of g we have $\sum_{k \geq 1} \lambda^k V(k)/k! = 1 - e^{-\lambda}$ for all $\lambda > 0$. Comparing power series coefficients gives $V(k) = (-1)^{k+1}$. This is the unique unbiased estimator of $e^{-\lambda}$. For k even, $V(k) = -1$, a bad estimate for $e^{-\lambda}$ which is positive. For any loss function which increases as the estimate moves away from the true value, it will reduce the risk to replace -1 by 0. Even then, the choice between the maximum value 1 and the minimum value 0 for the estimator based on the parity of k seems unreasonable since larger values of k indicate larger values of λ and so smaller values of $e^{-\lambda}$. So, at least for large values of k, e^{-k} would be a much more reasonable estimate of $e^{-\lambda}$.

A sequence $T_n = T_n(X_1, \ldots, X_n)$ of estimators for some $g(\theta)$, where T_n and g take values in a space S with a metric d, is called *consistent* if $T_n \to g(\theta)$ in probability as $n \to \infty$ for each P_θ, that is, for each $\varepsilon > 0$ and each θ,

$$\lim_{n \to \infty} P_\theta^n \{d(T_n(X_1, \ldots, X_n), g(\theta)) > \varepsilon\} \ = \ 0.$$

$\{T_n\}$ are called *strongly* consistent iff for each θ, $T_n(X_1, \ldots, X_n)$ converge to $g(\theta)$ as $n \to \infty$ almost surely for P_θ^∞.

For example, if $S = X = \mathbb{R}$ and $g(\theta) = \int x dP_\theta$, where $\int |x| dP_\theta < \infty$ for all θ, and $T_n = \overline{X} = (X_1 + \cdots X_n)/n$, then T_n are strongly consistent for $g(\theta)$ by the strong law of large numbers (RAP, Theorem 8.3.5).

As mentioned, one much-used loss function for real-valued $g(\theta)$ has been squared-error loss $(T - g(\theta))^2$. Perhaps the next most often considered is the absolute deviation $|T - g(\theta)|$. Both these functions are convex functions f of $T - g(\theta)$, defined as follows. A set $C \subset \mathbb{R}^k$ is *convex* if for any $u, v \in C$ and $0 \leq t \leq 1$, we have $tu + (1-t)v \in C$. Then a real-valued function f on C is *convex* if for any $u, v \in C$ and $0 \leq t \leq 1$,

$$f(tu + (1-t)v) \ \leq \ tf(u) + (1-t)f(v).$$

Then f is continuous on the interior of C (RAP, Theorem 6.3.4) but not necessarily on the boundary of C. A basic fact about convex functions and probability is Jensen's inequality (RAP, 10.2.6), which says that if X is a random variable having expectation EX and f is a convex function then $f(EX) \leq Ef(X)$ (under suitable measurability conditions). When X has just two values u and v, with $P(X = u) = t = 1 - P(X = v)$, Jensen's inequality reduces to the definition of convexity.

When loss functions are convex, decision rules can be improved or simplified in some ways. For one, randomization can be dispensed with, in rather general decision problems, as follows:

2.2.1 Theorem. If in a decision problem the action space A is a Borel measurable convex subset of some Euclidean space $\mathbb{R}^k$, and the loss function $L(\theta, \cdot)$ for any fixed $\theta \in \Theta$ is convex and Borel measurable on A, then if d is any randomized decision rule such that $\int \|u\| d(x)(du) < \infty$ for P_θ-almost all x for all θ, d can be replaced by a non-randomized rule $b(\cdot)$ without increasing the risk for any θ.

Proof. Let $b(x) := \int_A u d(x)(du)$, meaning that the (vector-valued) identity function $u \mapsto u$ on A is integrated with respect to the law $d(x)(\cdot)$. The integral is well-defined for P_θ-almost all x for all θ, and is a measurable non-randomized decision rule by Proposition 1.2.7. Then by Jensen's inequality (RAP, 10.2.6), for any such x,

$$L(\theta, b(x)) \ \leq \ \int L(\theta, u) d(x)(du).$$

Integrating with respect to P_θ gives

$$r(\theta, b) \ = \ \int L(\theta, b(x)) dP_\theta(x) \ \leq \ \int \int L(\theta, u) d(x)(du) dP_\theta(x) \ = \ r(\theta, d),$$

finishing the proof. □

Next, estimators can be taken measurable for a sufficient σ-algebra without increasing any risk:

2.2.2 Theorem (Rao-Blackwell). Let $\mathcal{A}$ be a sufficient σ-algebra for a family $\mathcal{P}$ of laws in a decision problem satisfying the conditions in Theorem 2.2.1. Suppose U is any non-randomized decision rule with $\int \|U\| dP < \infty$ for every $P \in \mathcal{P}$. Then $T := E(U|\mathcal{A})$ is a non-randomized decision rule with $r(P, T) \leq r(P, U)$ for each $P \in \mathcal{P}$.

Proof. By Theorem 2.1.8, under the assumptions, $E(U|\mathcal{A})$ exists and doesn't depend on $P \in \mathcal{P}$ (conditional expectations of vector-valued functions can be taken coordinatewise).

Thus $r(P,T)$ is well-defined. If $r(P,U) = +\infty$ there is nothing to prove, so we can restrict to the class of $P \in \mathcal{P}$ for which $r(P,U) < \infty$. Thus $L(P,U) \in \mathcal{L}^1(P)$ and has a conditional expectation $E_P(L(P,U)|\mathcal{A})$. By the conditional Jensen inequality (RAP, 10.2.7), $E_P(L(P,U)|\mathcal{A}) \geq L(P,T)$ a.s. for each $P \in \mathcal{P}$. Integrating both sides with respect to P gives the result. □

The Rao-Blackwell theorem applies to estimation problems where the loss function is a convex function W of $Y - g(P)$ for an estimator Y, such as $W(v) = |v|$ or $W(v) = v^2$.

Essentially complete and complete classes of decision rules were defined in Sec. 1.2. These notions will also be defined here relative to any given classes of decision rules. Let $\mathcal{U}$ be a class of decision rules and $\mathcal{P}$ a class of laws on a sample space. A class $\mathcal{D} \subset \mathcal{U}$ will be called essentially complete for $\mathcal{P}$ *relative to* $\mathcal{U}$ iff for each $U \in \mathcal{U}$ there is some $d \in \mathcal{D}$ such that for each $P \in \mathcal{P}$, $r(P,d) \leq r(P,U)$.

2.2.3 Corollary. If $\mathcal{A}$ is a sufficient σ-algebra for $\mathcal{P}$, then for any action space A which is a convex, Borel measurable subset of some $\mathbb{R}^k$ and any loss function L which is convex and Borel measurable on A for each $P \in \mathcal{P}$, the $\mathcal{A}$-measurable statistics form an essentially complete class for $\mathcal{P}$ relative to the class of all decision rules U such that $\int \|U\| dP < \infty$ for all $P \in \mathcal{P}$. Also, the family of unbiased $\mathcal{A}$-measurable estimators of a function $g(P)$ is essentially complete relative to the class of all unbiased estimators.

Proof. We need only apply the definitions and Theorem 2.2.2, and note that if U is an unbiased estimator, so is $E(U|\mathcal{A})$ for any sub-σ-algebra $\mathcal{A}$. □

If an estimator U for $g(P)$ has $\int \|U\| dP = +\infty$ for some P, then $E\|U - g(P)\| = +\infty$ and $E(\|U - g(P)\|^2) \geq [E\|U - g(P)\|]^2 = +\infty$, so there would be infinite risk for P for both of the two most-studied loss functions. Such a U seems quite undesirable, so the hypothesis on U in the first half of Corollary 2.2.3 seems not too restrictive.

Note. If an unbiased estimator performs badly, as in the example of estimating $e^{-\lambda}$ for a Poisson variable X observed only for $X \geq 1$ earlier in this section, or in problem 5(b), then conditioning on a sufficient σ-algebra will not necessarily make the estimator a good one.

PROBLEMS

1. Let $X_1, \ldots, X_n$ be i.i.d. from a distribution with a finite variance σ^2. Let

$$s^2 := (n-1)^{-1} \sum_{j=1}^{n} (X_j - \overline{X})^2 \text{ and } s'^2 := (n-1)s^2/n \text{ for } n \geq 2.$$

Show that the sequences of estimators for σ^2 defined by s^2 and s'^2 are (a) both consistent, while (b) only s^2 is unbiased ($n \geq 2$).

(c) Show that for $n = 1$, there is an unbiased estimator of σ^2 for Poisson distributions.

2. In problem 1, show that for $n = 1$ there is no unbiased estimator of σ^2 for general distributions, specifically for normal distributions. *Hints:* Suppose there is a measurable function f on $\mathbb{R}$ such that $Ef(X) = \sigma^2$ whenever X has a $N(\mu, \sigma^2)$ distribution. Let

Y, Z be i.i.d. $N(0,1)$. Find $Ef(Y+Z)$ and the conditional expectation $E(f(Y+Z)|Y)$ for each Y, whose expectation with respect to Y should equal $Ef(Y+Z)$.

3. Give an example to show that when T is an unbiased estimator of some $g(\theta)$, T^2 may not be an unbiased estimator for $g(\theta)^2$. Find under what conditions, if any, T^2 will also be unbiased for $g(\theta)^2$.
4. Give an example where a constant c is in the range of the function g on the parameter space but $T \equiv c$ is not an admissible estimator. *Hint*: consider families of laws which are not equivalent, e.g. laws $U[\theta, \theta+1]$ on $\mathbb{R}$ for $-\infty < \theta < \infty$, with squared-error loss and $g(\theta) \equiv \theta$.
5. For a given n consider the family of binomial distributions $b(k,n,p) := \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$. Here the sample space consists of the integers $0, 1, \ldots, n$ and the parameter is p.
   (a) Show that a function $g(p)$ has an unbiased estimator if and only if g is a polynomial of degree $\leq n$, and then the unbiased estimator is unique.
   (b) Show that the unbiased estimator for $g(p) = p^2$ is 0 not only when $k = 0$ but also when $k = 1$.
6. Let $\mathcal{P}_1$ be the class of all laws on $\mathbb{R}$ with finite mean. Let $X_1, ..., X_n$ be i.i.d. with a distribution in $\mathcal{P}_1$. For any $c_1, ..., c_n$ such that $\sum_{j=1}^n c_j = 1$, $T := \sum_{j=1}^n c_j x_j$ is an unbiased estimator of $\mu = EX_1$. For the σ-algebra $\mathcal{S}_n$ of permutation-invariant (symmetric) measurable events in $\mathbb{R}^n$, as in Problem 5 of Section 2.1, show that $E(T|\mathcal{S}_n) = \overline{X}$.
7. Let $X_1, \ldots, X_5$ be observed, i.i.d. $N(\mu, 1)$ for μ unknown. Let $S_n := X_1 + \cdots + X_n$ for $n = 1, \ldots, 5$. Show that S_5 is a sufficient statistic for μ. Let $U := S_3/3$. As an example of the Rao-Blackwell theorem 2.2.2, find $T = E(U|S_5)$, where the conditional expectation given a function is the same as the conditional expectation given the smallest σ-algebra for which it is measurable. For squared-error loss, evaluate the risks $r(P,T)$ and $r(P,U)$ for any law $P = N(\mu, 1)$.

NOTES

The example of Poisson distributions conditioned on $k \geq 1$ in showing what can go wrong with unbiased estimation was found in Kendall and Stuart, 1967, p. 34. At this writing the latest edition, the 6th, is Kendall, Stuart and Ord (1994-1999), see vol. 2A.

REFERENCE

Kendall, Maurice G., and Alan Stuart (1967), *The Advanced Theory of Statistics*, vol. 2, 2d ed. Griffin, London.

Kendall, Maurice G. [posth.], Alan Stuart, and J. Keith Ord (1994-99), *[Kendall's] Advanced Theory of Statistics*, vols. 1 (Distribution Theory), 2A (Classical Inference and the Linear Model, by A. Stuart, J. K. Ord, and S. Arnold), 2B (Bayesian Inference, by A. O'Hagan, 1st ed., 1994).