Chapter 3

Basic Theory of Point Estimation.

Suppose X is a random observable taking values in a measurable space $(\Xi, \mathcal{G})$ and let $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ denote the family of possible distributions of X. An *estimand* is a function $g(\theta)$ of the unknown parameter which we wish to estimate. We will typically assume that g takes values in $I\!R$. The goal of point estimation is to produce a single number (when g is one dimensional) which is (hopefully) "close" to $g(\theta)$ where θ is the unknown value of the parameter. An *estimator* of $g(\theta)$ is a measurable mapping $\delta : (\Xi, \mathcal{G}) \longrightarrow (I\!R, \mathcal{B})$.

3.1 General Principles of Estimation.

In this section, we consider some fairly general methods for deriving estimators, and some general properties of estimators useful for comparing and evaluating estimators.

3.1.1 Methods for Deriving Estimators.

Here we consider a variety of techniques or "principles" which are used for obtaining estimators in practice. In Sections 3 through 6 of this chapter, we will consider estimators which are optimal in some sense. However, in many practical problems, there either do not exist acceptable optimal procedures (such as UMVUE's or minimax estimators) or they are not readily computable (e.g. Bayesian estimators when one has an acceptable prior). It is safe to say that most often applied statisticians resort to such *ad hoc* procedures as we introduce in this section. We will see that some of these (in particular, maximum likelihood) have certain "asymptotic optimality" properties (Section 7), but this may not be a compelling justification for some statisticians.

Methods Based on the Empirical Distribution.

One fairly general method for constructing estimators we have already seen is based on the empirical distribution. This method is applicable when we can extend the definition of $g(\theta)$ to a family of distributions which will include the empirical distribution. For instance, suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. with a $N(\mu, \sigma^2)$ distribution and both μ and σ^2 are unknown. If we consider estimation of the two estimands:

$$g_1(\mu, \sigma^2) \;=\; \mu, \quad g_2(\mu, \sigma^2) \;=\; \sigma^2.$$

Since $g_1(\mu, \sigma^2)$ is just the mean of the distribution, and the mean is defined for a large class of distributions, namely

$$\mathcal{M}_1 \;=\; \{P : \; P \text{ is a Borel p.m. on } I\!R \text{ and } E_P[|X_1|] = \int_R |x| \, dP(x) < \infty\}.$$

We can extend g_1 to, say, $\bar{g}_1 : \mathcal{M}_1 \longrightarrow I\!R$ by $\bar{g}_1(P) = E_P[X_1] = \int_R x \, dP(x)$, for any $P \in \mathcal{M}_1$. Now, any realization of the empirical distribution will be in $\mathcal{M}_1$, so we can estimate $g_1(\mu, \sigma^2) = \bar{g}_1\left(N(\mu, \sigma^2)\right)$ by $\bar{g}_1\left(\hat{P}_n\right)$ where $\hat{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical distribution. Of course, as we have already seen, $\bar{g}_1\left(\hat{P}_n\right) = \int_R x \, d\hat{P}_n(x) = n^{-1} \sum_{i=1}^n X_i = \bar{X}$ is just the sample mean. In a similar manner, we can extend g_2 above from the family of i.i.d. $N(\mu, \sigma^2)$ distributions to $\bar{g}_2$ defined on the family of i.i.d. distributions with finite second moment, viz.

$$\mathcal{M}_2 \;=\; \{P : \; P \text{ is a Borel p.m. on } I\!R \text{ and } E_P[X_1^2] = \int_R x^2 \, dP(x) < \infty\}.$$

Then, the corresponding estimator $\bar{g}_2\left(\hat{P}_n\right)$ is the sample variance

$$S^2 \;=\; \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad .$$

Unfortunately, there are several difficulties with the simple notions we have presented here. First of all, in a given problem, there may be more than one way to extend the estimand to a class of distributions containing the empirical distribution. For instance, g_1 is not only the mean of the $N(\mu, \sigma^2)$ distribution, but it is also the median of the $N(\mu, \sigma^2)$ distribution (by symmetry), so we could use the median of the data to estimate $g_1(\mu, \sigma^2)$. Further, in many settings, there is no obvious extension of the estimand. For instance, suppose $X_1, X_2, \ldots, X_n$ are i.i.d. with a *Gamma*(α, β) distribution and α and β are unknown. If were interested in estimating α, there is no obvious extension of α to a family of distributions containing the empirical. In fact, the "shape" parameter α really only makes sense in the context of a *Gamma* family. But the scale parameter β also only makes sense in the context of the *Gamma* family. So there is no obvious way of applying this principle of estimation to the Gamma family.

Method of Moments.

Assume $X_1, X_2, \ldots, X_n$ are i.i.d. from the *Gamma*(α, β) distribution as in the last paragraph. Consider the two estimands

$$\begin{aligned} g_1(\alpha,\beta) &= \alpha\beta = E_{(\alpha,\beta)}[X_1] \\ g_2(\alpha,\beta) &= \alpha\beta^2 = Var_{(\alpha,\beta)}[X_1]. \end{aligned}$$

Now we can apply the technique based on the empirical distribution and estimate $g_1(\alpha,\beta)$ with $\bar{X}$ and $g_2(\alpha,\beta)$ with S^2. Observe that the parameters are related to these two estimands by

$$\beta = g_2(\alpha,\beta)/g_1(\alpha,\beta), \qquad \alpha = g_1^2(\alpha,\beta)/g_2(\alpha,\beta).$$

We can plug in the estimators of g_1 and g_2 into the algebraic relationships and obtain estimators of the parameters, viz.

$$\begin{aligned} \hat{\alpha}_{mm} &= \overline{X}^2/S^2 \quad , \\ \hat{\beta}_{mm} &= S^2/\overline{X} \quad . \end{aligned} \tag{3.1}$$

The subscript *mm* stands for *method of moments*, a terminology we shall now explain.

Suppose $\theta \in I\!R^p$, then usually the value of θ will uniquely determine the first p moments of the distribution, i.e. the map $\theta \mapsto (E_\theta[X], E_\theta[X^2], \ldots E_\theta[X^p])$ is a one to one correspondence (bijective map). Letting $M(\theta)$ denote this map, we obtain p equations in p unknowns if we set the first p sample moments equal to the corresponding theoretical moments, i.e. solve for the θ which gives

$$\frac{1}{n}\sum_{i=1}^{n}[X_i^p - M_j(\theta)] = 0 \quad , \quad 1 \le i \le p \quad . \tag{3.2}$$

Alternatively, for $i > 1$ we can replace the i'th noncentral moment by the i'th centeral moment, i.e. solve the equations

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\{X_i - E_\theta[X_1]\} &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_i - \bar{X}\right)^p - E_\theta[(X_1 - E_\theta[X_1])^p]\right\} &= 0, \quad 1 < i \le p. \end{aligned}$$

Quite clearly, the traditional moments are rather arbitrary here, and one can consider a *generalized method of moments estimator* as follows. Let $\beta : \mathcal{X} \times \Theta \longrightarrow I\!R^p$ be such that $\eta(\theta) := E_\theta[\beta(X,\theta)]$ is bijective. Then we can estimate θ by setting the theoretical expecatation equal to the sample expectation, i.e.

$$\frac{1}{n}\sum_{i=1}^{n}[\beta(X_i,\theta) - \eta(\theta)] = 0 \quad . \tag{3.3}$$

Still more generally, we can consider any function $\psi : \mathcal{X} \times \Theta \longrightarrow I\!R^p$ for which $E_{\theta_1}[\psi(X, \theta_2)] = 0$ if $\theta_1 = \theta_2$, and $E_{\theta_1}[\psi(X, \theta_2)] \neq 0$ if $\theta_1 \neq \theta_2$. Then we simply set $(1/n) \sum_i \psi(X_i, \theta) = 0$. Putting $\psi(x, \theta) = \beta(x, \theta) - \eta(\theta)$ as in (3.3) recovers the generalized method of moment estimators, which we see is a class of M–estimators, to be discussed next.

M–Estimators.

Recall that the mean μ is the value of b which minimizes

$$\lambda(b) \;=\; \int (x-b)^2 \, dP_X(x),$$

so if we wish to estimate the mean of a distribution on $I\!R$ from i.i.d. observations, a reasonable approach is to minimize

$$\hat{\lambda}(b) \;=\; \int (x-b)^2 \, d\hat{P}(x) \;=\; \frac{1}{n} \sum_{i=1}^{n} (X_i - b)^2$$

where we have simply plugged in the empirical $\hat{P}$ for P_X in the expression for λ. Of course, minimization of $\hat{\lambda}(b)$ leads to $\bar{X}$ as an estimator for the mean of the distribution. In a similar fashion, one can show that a median is a value of b which minimizes

$$\lambda(b) \;=\; \int |x-b| \, dP_X(x),$$

and plugging in the empirical distribution for the unknown P_X leads to the sample median as an estimator of the median of P_X.

Of course, we did not need to invent these optimization problems to obtain the sample mean and median as estimators of the mean and median of the unknown distribution. However, there are many estimands that can be expressed as solutions of optimization problems involving the unknown distribution, and then the technique above can be applied. Of course, when we solve the optimization problem, it will probably by setting a derivative equal to 0, i.e. solving an equation. An *M–estimator* (or *maximum likelihood–type estimator*) is one obtained either by a minimization or maximization problem, or as the root of a equation. For instance, if we model $X_1, \ldots, X_n$ as i.i.d. with Law$[X_i]$ in a dominated family with densities $f_\theta(x)$, then the M–estimator might be obtained from a minimization problem such as

$$\hat{\theta}_n \;=\; \arg\min_{\theta} \frac{1}{n} \sum \rho(X_i, \theta) \quad . \tag{3.4}$$

We will refer to ρ as a *criterion function* (some authors call it a "loss function" but it is certainly not a loss function in the decision theoretic sese of the term). If $\psi(x, \theta) = D\rho(x, \theta)$ (derivatives in this context are always w.r.t. the parameter, i.e. $\psi(x, \theta) = (\partial/\partial\theta)\rho(x, \theta)$ if θ is one dimensional) then we would usually seek stationary points of the objective function which are roots of

$$\frac{1}{n}\sum \psi(X_i, \theta) \; = \; 0 \quad . \tag{3.5}$$

Here, $\psi(x, \theta)$ is called the *score function.* Note that this is a version of the generalized method of moments discussed above. In general, we will be more interested in the root characterization (3.5) rather than the minimization characterization of (3.4), mostly because this permits an easier asymptotic analysis, and because we want to consider examples of estimators not obtained from optimization problems.

Maximum Likelihood Estimation.

Now we introduce the notion of maximum likelihood estimation. It is probably safe to say that this estimation methodology is used more often in practice than any other.

Suppose $\underline{X}$ is an observation vector with Law$[\underline{X}]$ assumed to be in a parametric family $\mathcal{P} \; = \; \{P_\theta : \theta \in \Theta\}$. Assume $\mathcal{P}$ is dominated by a σ-finite measure μ and let $f_\theta(\underline{x}) = [dP_\theta/d\mu](x)$ be the density under θ. We usually treat $f_\theta(\underline{x})$ as a function of $\underline{x}$ for fixed θ, but when we think of it as a function of θ for the fixed value of $\underline{X}$ which is actually observed, we call it the *likelihood function.* The (a) *maximum likelihood estimator* (abbreviated MLE) $\hat{\theta}_{mle}$ is defined the (a) value of θ which maximizes the likelihood function.

Remarks 3.1.1 (a) There are a number of problems with this definition. First, there is no guarantee that a maximizer of the likelihood exists. It does happen in practice that one sometimes attempts to compute an MLE and has problems because it doesn't exist. This issue will be taken up in the asymptotic theory below where we show (under regularity conditions) that the MLE exists with high probability in sufficiently large samples. There is also the question of uniqueness of the maximizer of the likelihood, of course, which is why we said "the (a) maximum likelihood estimator" above. Again, this can be a problem in practice, although it is safe to say less of a problem than nonexistence. Typically when the MLE is not unique, there is a relatively small set of values with nice properties (e.g. a small interval for unidimensional θ). In order to deal with these issues, we will almost always have to impose "regularity conditions" on the densities $f_\theta(\underline{x})$, e.g. continuity in θ for fixed $\underline{x}$. See the next remark for even more potentially troubling issues if such "regularity conditions" are not imposed. We will see below that for exponential families there is generally not a problem with uniqueness.

(b) Consider a single observation X from $N(\mu, 1)$. Suppose we use the following version of the density:

$$f_\mu(x) \; = \; \begin{cases} \phi(x-\mu) & \text{if } x \neq 2\mu; \\ \quad 100 & \text{if } x = 2\mu. \end{cases}$$

This is a perfectly good version of the density since we have changed it from the usual version only on a set of Lebesgue measure 0. The likelihood will be maximized at the value of μ where $x = 2\mu$, i.e. the MLE will be $x/2$, which is rather silly. If however we require that the likelihood be continuous for all x, then there is only one version of the density which works, namely $\phi(x-\mu)$, and the MLE under this version is x, which makes a lot more sense.

(c) Apparently the likelihood and hence also the MLE depend on which dominating measure we choose. Suppose $\mathcal{P} \ll \mu$ and also $\mathcal{P} \ll \nu$ and let $\tau = \mu + \nu$. By the chain rule for Radon-Nikodym derivatives,

$$\begin{aligned} \frac{dP_\theta}{d\mu}(x) &= \frac{dP_\theta}{d\tau}(x)\frac{d\mu}{d\tau}(x) \quad \tau\text{–a.e.} \\ \frac{dP_\theta}{d\nu}(x) &= \frac{dP_\theta}{d\tau}(x)\frac{d\nu}{d\tau}(x) \quad \tau\text{–a.e.} \end{aligned}$$

Forgetting about the τ–a.e. that follows each of the above equations (i.e. just assuming they hold everywhere), then we see that maximizing $dP_\theta/d\tau$ is the same as maximizing either $dP_\theta/d\mu$ or $dP_\theta/d\nu$ since the factors $d\mu/d\tau$ and $d\mu/d\tau$ don't depend on θ.

Now let's make a more rigorous argument. Take a fixed version of $dP_\theta/d\tau$ (e.g. one that is continuous in θ) and find the MLE using this, call it $\hat{\theta}_{\tau,mle}$. Since this version when multiplied by a version of $d\mu/d\tau$ gives a version of $dP_\theta/d\mu$, $\hat{\theta}_{\tau,mle}$ is also an MLE under this version of $dP_\theta/d\mu$. Now if we use the same version of $dP_\theta/d\tau$ to form a version of $dP_\theta/d\nu$, then we will get this same MLE $\hat{\theta}_{\tau,mle}$. Thus, the MLE doesn't depend on the dominating measure, in this sense of matching up the versions of the Radon-Nikodym derivatives. Of course, from (b) above we already knew there was a problem with using different versions of the density. In particular, if we start with a version of $dP_\theta/d\tau$ which is continuous in θ for fixed x, then all the versions of $dP_\theta/d\mu$ and $dP_\theta/d\nu$ that we constructed will be continuous.

(d) Why is maximum likelihood estimation a good idea? The answer to this is not at all simple and obvious. On a very crude intuitive level, we "expect" (in a loose sense of the word) that the observed value of $\underline{X}$ is more "likely" to come from a region where the true density f_{θ_0} is large, so finding a value of θ which maximizes the likelihood of the observed value seems reasonable. A better, albeit still heuristic justification might be based on Kullback-Leibler information, which is discussed below. In Section 5.5 we give some Bayesian justifications for maximum likelihood.

Ultimately, we have to admit that maximum likelihood is rather *ad hoc*, but it has been discovered that (under regularity conditions) it gives *asymptotically optimal estimators*, a subject to be taken up in Section 5.7. It turns out that maximum likelihood is not unique in this regard (e.g. Bayes estimators are usually asymptotically optimal, also), but it is also relatively easy to compute in practice in that one can generally compute the likelihood and numerical optimization to find the maximum is relatively well understood and good software exists for doing it.

□

If $g : \Theta \longrightarrow I\!R^k$ is an estimand, we define the *maximum likelihood estimator of* $g(\theta)$ to be $g(\hat{\theta}_{mle})$. Thus, within the framework of maximum likelihood estimation we solve all estimation problems at once when we find $\hat{\theta}_{mle}$. One can show that this definition is not entirely frivolous in the following sense. Suppose g is a "partial reparameterization" in that there exists an h defined on Θ such that $\theta \mapsto (g(\theta), h(\theta))$ is a 1-1 correspondence (so that $(g(\theta), h(\theta))$ is a genuine reparameterization). Then under the new parameterization, $(g(\hat{\theta}_{mle}), h(\hat{\theta}_{mle}))$ is the MLE. We will see also that $g(\hat{\theta}_{mle})$ enjoys the same "asymptotic optimality" property as $\hat{\theta}_{mle}$, which is the only general theoretical justification for maximum likelihood estimation.

It is commonly the case that the P_θ all have the same support, so restricting attention to a common support all the densities are (or can be taken to be) positive. Then it is usual to consider the log–likelihood

$$\ell(\theta) \ = \ \log f_\theta(\underline{X}) \quad .$$

We do not show the $\underline{X}$ in $\ell(\theta)$ as it is fixed once the observation is taken and plugged in, but it should be kept in mind that ℓ is a random function. Of course, if $\underline{X} = (X_1, X_2, \ldots, X_n)$ has i.i.d. components X_i with common marginal $f_\theta(x)$, then the log–likelihood is

$$\ell(\theta) \ = \ \sum_{i=1}^{n} \log f_\theta(X_i) \quad ,$$

and maximization of the likelihood is equivalent to maximization of the log–likelihood, which often is easier to do mathematically. Typically we will carry this out by taking derivatives and setting to 0, which gives rise to a (random) equation

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \theta) \ = \ 0$$

where the score function is

$$\psi(x, \theta) \ = \ - \bigtriangledown \log f_\theta(x)$$

and all derivatives are w.r.t. θ. Recall that a solution to the equation is called a "stationary point." One should be careful to check one of the conditions to ensure that the stationary point is indeed a maximizer (e.g. check that the matrix $-D^2\ell(\theta)$ is positive definite), otherwise one could end up with a "minimum likelihood estimator" which probably would be pretty bad. If multiple roots are found, one has to determine which is the global maximizer as well. We will see below that for natural parameter exponential families, it is not necessary to check the condition – there is at most one stationary point and if it exists, then it is the MLE.

There is one aspect of computing MLE's which is often useful. Suppose the parameter vector can be split into two components $\theta = (\theta_1, \theta_2)$ such that for a fixed value of θ_2, the likelihood can be easily maximized analytically over θ_1. In symbols we have

$$\hat{\theta}_1(\theta_2) \;=\; \arg\max_{\theta_1} \ell(\theta_1, \theta_2) \quad . \tag{3.6}$$

Note that the maximizer over θ_2 depends on the fixed value of θ_1. Then we can plug back in the maximizer over θ_2 and obtain a *concentrated log–likelihood*

$$\ell_2(\theta_2) \;=\; \ell(\hat{\theta}_1(\theta_2), \theta_2) \quad . \tag{3.7}$$

Now this can be maximized over θ_2 numerically to find $\hat{\theta}_2$, which is then plugged into the formula for $\hat{\theta}_1(\theta_2)$ to obtain $\hat{\theta}_1$. This situation arises for instance in the *Gamma* family (see Example 3.1.1 below). From the practical point of computing the MLE this is important to recognize and take advantage of it since it reduces the dimensionality of the optimization problem and thereby improves numerical efficiency.

Kullback-Leibler Information. Now we turn to the "justification" or motivation of maximum likelihood based on Kullback-Leibler information. This concept will also be useful in hypothesis testing.

Definition 3.1.1 *Let P and Q be probability measures on a measurable space $(\Omega, \mathcal{F})$. The* Kullback-Leibler information *or* divergence *between Q and P is*

$$K(Q,P) \;:=\; \begin{cases} \int \log \frac{dP}{dQ}\, dP & \text{if both } P \ll Q \text{ and } Q \ll P; \\ \\ \infty & \text{otherwise.} \end{cases}$$

□

Proposition 3.1.1 *We have $K(Q,P) \geq 0$ with $K(Q,P) = 0$ if and only if $P = Q$.*

Proof. Note that log is strictly concave, so by Jensen's inequality,

$$-\int \log \frac{dQ}{dP}\, dP \;\geq\; -\log \int \frac{dQ}{dP}\, dP \;=\; -\log 1 \;=\; 0 \quad ,$$

with equality if and only if $dQ/dP = 1$, P–a.s., i.e. if and only if $Q = P$.

□

Suppose μ is a σ–finite measure which dominates both P and Q, and let p and q denote their respective densities, so of course $dP/dQ = p/q$, μ–a.e. Then under the first line of the definition, we can write

$$K(Q,P) \ = \ \int (\log p) p \, d\mu \ - \ \int (\log q) p \, d\mu \quad . \tag{3.8}$$

Now consider the setup of maximum likelihood estimation with a parametric family $\{P_\theta : \theta \in \Theta\} \ll \mu\}$ of mutually absolutely continuous probabilities with densities $\{f_\theta\}$. Denoting the unknown true value by θ_0, we have for any other θ,

$$K(P_\theta, P_{\theta_0}) \ = \ \int (\log f_{\theta_0}) f_{\theta_0} \, d\mu \ - \ \int (\log f_\theta) f_{\theta_0} \, d\mu \quad . \tag{3.9}$$

Now minimization over θ of $K(P_\theta, P_{\theta_0})$ gives θ_0, by Proposition 3.1.1. But the first term on the r.h.s. of (3.9) doesn't depend on θ, so we can write

$$\theta_0 \ = \ \arg\max_\theta \int [\log f_\theta(x)] f_{\theta_0}(x) \, d\mu(x) \ = \ E_{\theta_0}[\log f_\theta(X)] \quad .$$

Of course, we don't know θ_0, so we can't compute this latter expectation, but given a sample X_1, X_2, …, X_n of i.i.d. observations from P_{θ_0}, we can estimate this expectation with the sample average

$$\frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i) \ = \ n^{-1} \ell(\theta) \quad .$$

Thus, maximizing the log likelihood $\ell(\theta)$ should provide us approximately with the value of θ which minimizes Kullback-Leiber information between P_θ and the true probability, at least for large n where we expect the sample average to be close to the true expectation. In fact, this interpretation is very useful in that it also tells us what to expect if our data are generated from distribution not in our parametric model. It has also been used as the basis for a very general proof of the consistency of MLE's; see Wald (???).

Maximum Likelihood in Exponential Families. As mentioned above, for exponential families, we can obtain some nice results about the existence and uniqueness of the MLE, which we now detail.

Theorem 3.1.2 *Suppose $\{f_\eta(x) : \eta \in \Lambda\}$ is a natural parameter exponential family of full rank with*

$$f_\eta(x) \ = \ \exp\left[\eta^t T(x) - A(\eta)\right] h(x) \quad .$$

If a solution $\hat{\eta}$ of the equation

$$\bigtriangledown A(\eta) \ = \ T(X) \tag{3.10}$$

exists in the interior of the natural parameter space and is in Λ, then it is the unique MLE. Conversely, if Λ is an open set and if an MLE exists, then it is unique and given by the solution of (3.10).

Proof. We know that $A(\eta)$ is strictly convex (in fact, D^2A is strictly positive definite). This implies that the negative log–likelihood

$$\ell(\eta) \;=\; -\eta^t T(x) \;+\; A(\eta)$$

is also strictly convex in η. (We have dropped the unimportant $\log h(x)$ which doesn't depend on η.) To see this, note that for $t \in (0,1)$,

$$\ell((1-t)\eta_0 + t\eta_1) \;=\; -(1-t)\eta_0^t T(x) \;+\; t\eta_1^t T(x) \;+\; A((1-t)\eta_0 + t\eta_1)$$

$$<\; -(1-t)\eta_0^t T(x) - t\eta_1^t T(x) + (1-t)A(\eta_0) + tA(\eta_1) \;=\; (1-t)\ell(\eta_0) + t\ell(\eta_1) \quad ,$$

the inequality following from strict convexity of A. (In general, the sum of a convex function and a strictly convex function gives a strictly convex function, and a linear function like $\eta \mapsto -\eta' T$ is convex.)

Letting Λ_0 denote the natural parameter space (which is convex) we show that $\ell(\eta)$ can have at most one minimizer in Λ_0, because of strict convexity. Suppose η_0 and η_1 both minimize ℓ in Λ_0, then for any $t \in (0,1)$, of course $\eta_t \in \Lambda_0$ and

$$\ell(\eta_t) \;<\; (1-t)\ell(\eta_0) \;+\; t\ell(\eta_1) \;=\; \ell(\eta_0) \qquad (3.11)$$

since $\ell(\eta_0) = \ell(\eta_1)$ the minimum value of ℓ on Λ_0 by assumption, but this last display contradicts this assumption, so any minimizer of ℓ on Λ_0 must be unique. If the minimizer on Λ_0 exists and is an interior point, then the derivative must vanish (recall A is infinitely differentiable, so clearly also is ℓ since it is just a linear function added to A) and we obtain

$$\bigtriangledown \left[-\eta^t T(X) \;+\; A(\eta) \right] \;=\; -T(X) \;+\; \bigtriangledown A(\eta) \;=\; 0$$

which gives equation (3.10).

Conversely, we claim that a solution to (3.10) which is an interior point must be a minimizer of ℓ. To see this, we have by Taylor expansion that if $\bigtriangledown \ell(\eta_0) = 0$ then

$$\ell(\eta) \;=\; \ell(\eta_0) \;+\; [\eta - \eta_0]^t D^2\ell(\eta_0)[\eta - \eta_0] \;+\; o(\|\eta - \eta_0\|^2)$$

and since $D^2\ell(\eta_0) = D^2A(\eta_0)$ is strictly positive definite, it follows that for η in some sufficiently small neighborhood of η_0, we must have $\ell(\eta) > \ell(\eta_0)$ unless $\eta = \eta_0$. So at least a stationary point of ℓ is a local minimizer, but this is impossible unless it is also a global minimizer since if η_0 is a local minimizer but $\ell(\eta_1) < \ell(\eta_0)$ then we have $\ell(\eta_t) < \ell(\eta_0)$ for all η_t on the line segment joining η_0 and η_1 (see equation (3.11)), which contradicts the assumption that η_0 is a local minimizer.

Now all of the above argument is for Λ_0, the natural parameter space of which Λ is a subset with nonempty interior (from the full rank assumption). A solution of equation (3.10) in the interior of Λ_0 would be the MLE over Λ_0, so if it is in Λ then it would also be the MLE over Λ, and also unique by our argument above. Conversely, if Λ is open, then an MLE over Λ is of course an interior point of Λ_0, hence a stationary point, and so a solution of equation (3.10).

□

Note that equation (3.10) may be interpreted as setting the observed value of the complete and sufficient statistic $T(X)$ equal to its expected value. Thus, for exponential families, the maximum likelihood estimator is a kind of method of moments estimator. Also, if we have i.i.d. observations X_1, X_2, ..., X_n from the exponential family, then (3.10) becomes

$$\bigtriangledown A(\eta) \; = \; \overline{T}_n \; := \; \frac{1}{n}\sum_{i=1}^{n} T(X_i) \quad . \tag{3.12}$$

One final remark which applies generally to MLE's: when we have a parameter space Θ or Λ, to say the MLE exists means that the maximizer of the likelihood is an element of the parameter space. Thus for instance, if we believe the true variance $\sigma^2 < 1$ for i.i.d. observations from $N(\mu, \sigma^2)$, then the MLE must satisfy $\hat{\sigma}^2_{mle} < 1$. If the sample variance (which is the MLE when σ^2 is unrestricted) happens to be bigger than 1 (which can happen with positive probability even if our assumption $\sigma^2 < 1$ is true), then the MLE simply doesn't exist, although we would probably at least entertain the idea of enlarging the parameter space to $\sigma^2 \leq 1$ so that the MLE would be $\hat{\sigma}^2_{mle} = 1$ if the sample variance is bigger than 1 (see Exercise 3.1.1). The student should perhaps reread the statement and last paragraph of the proof above with these remarks in mind and remembering that the natural parameter space is the largest possible parameter space for a natural parameter exponential family, but one may wish to consider only a subset of the natural parameter space.

Examples of MLE's. Here we consider a number of special cases which are easy to treat, and we leave the verifications to the student (Exercises 3.1.2, 3.1.3, 3.1.4, and x8.2.7).

Example 3.1.1 In each of the following, assume we have i.i.d. observations X_1, X_2, ..., X_n from the specified family of distributions. We denote the MLE of a parameter by putting the hat over the notation for that parameter.

$N(\mu, \sigma^2)$: $\hat{\mu} = \overline{X}$, the sample mean, and

$$\hat{\sigma^2} \; = \; \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

This is a biased estimator of σ^2 which however has smaller MSE than the UMVUE.

$N_d(\underline{\mu}, \Sigma)$: The MLE of $\underline{\mu}$ is the sample mean $\underline{\overline{X}}$ and the MLE of the covariance matrix Σ is the sample covariance

$$\hat{\Sigma} \; = \; \frac{1}{n}\sum_{i=1}^{n}(\underline{X}_i - \underline{\overline{X}})(\underline{X}_i - \underline{\overline{X}})^t \quad .$$

$B(n,p)$: With a single observation X, $\hat{p} = X/n$, the UMVUE.

$Expo(\mu)$: $\hat{\mu} = \overline{X}$.

$Laplace(\mu, s)$: Recall that the density for a single observation is given by

$$f_{\mu,s}(x) \;=\; \frac{1}{2s}\exp\left[-\frac{|x-\mu|}{s}\right] \quad .$$

Here, $\mu \in I\!\!R$ is a location parameter and $s > 0$ is a scale parameter. The MLE $\hat{\mu}$ of the location is any sample median M_n (the MLE is not unique with probability 1 when n is even for this example). The MLE of the scale parameter is

$$\hat{s} \;=\; \frac{1}{n}\sum_{i=1}^{n}|X_i - M_n|$$

which is the average absolute deviation from any median.

$U(0,\theta)$: $\hat{\theta} = X_{(n)} = \max\{X_i : 1 \le i \le n\}$, the maximal order statistic. To derive this result, first note that with probability 1, all X_i are positive, so we assume they are all positive. We will use the version of the density given by $f_\theta(x) = I_{[0,\theta]}(x)$, i.e. include the endpoints of the interval. Then the likelihood is

$$\begin{aligned} f_\theta(\underline{X}) \;&=\; \prod_{i=1}^{n}\theta^{-1}I_{[0,\theta]}(X_i) \\ &=\; \begin{cases} 0 & \text{if } X_{(n)} > \theta; \\ \theta^{-n} & \text{if } X_{(n)} \le \theta. \end{cases} \end{aligned}$$

This function is maximized over θ by making θ as small as possible subject to the constraint $X_{(n)} \le \theta$, i.e. at $\theta = X_{(n)}$. Notice in this case that the likelihood is not continuous and there is no version of $f_\theta(x)$ which will make it continuous. Also, if we had chosen the version of the density $f_\theta(x)$ $= I_{(0,\theta)}(x)$, i.e. left out the endpoints, then we would have maximized the likelihood by making θ as small as possible subject to the constraint $X_{(n)} < \theta$, and the MLE would technically not have existed, although it is clear what value one should use. This is an example that *will not* be covered by the asymptotic theory to be discussed in the Section 5.7.

$U(\theta_1, \theta_2)$: Adapting the argument from the $U(0,\theta)$ example above, one can show that (with appropriate choice of the density again) the MLE's are $\hat{\theta}_1 = X_{(1)}$ $= \min\{X_i : 1 \le i \le n\}$, and $\hat{\theta}_2 = X_{(n)}$, the minimal and maximal order statistics.

□

Example 3.1.2 Assume we have i.i.d. observations X_1, X_2, ..., X_n from the *Gamma*(α, β) family. It is convenient to replace the scale parameter β by $1/\beta = \eta$ which gives the natural parameterization. Then by Theorem 3.1.2, the MLE $(\hat{\alpha}, \hat{\eta})$ is the unique solution (if it exists) of

$$\psi_0(\alpha) \; - \; \log \eta \; = \; \overline{L}_n \quad ,$$

$$\alpha / \eta \; = \; \overline{X}_n \quad ,$$

where

$$\overline{L}_n \; = \; \frac{1}{n} \sum_{i=1}^{n} \log X_i \quad ,$$

and $\psi_0(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the so–called *digamma function*. One can solve for $\hat{\eta}$ from the second equation,

$$\hat{\eta} \; = \; \frac{\hat{\alpha}}{\overline{X}}$$

or what is the same

$$\hat{\beta} \; = \; \frac{\overline{X}}{\hat{\alpha}}$$

and substitute it out of the first equation to obtain a single equation for $\hat{\alpha}$, viz.

$$\psi_0(\hat{\alpha}) \; - \; \log \hat{\alpha} \; = \; \overline{L}_n \; - \; \log \overline{X} \quad . \tag{3.13}$$

Note that we have to solve only one equation in one unknown – i.e. we have reduced the dimensionality of our optimization problem from 2 to 1 by "concentrating out" one of the unknown parameters. Now it turns out that $\psi_0(\alpha) - \log \alpha$ is strictly increasing and takes on all values between $-\infty$ and 0 (see Abramowitz and Stegun ??? for properties of the digamma function). Further, by Jensen's inequality, since log is a strictly concave function, if there are at least 2 distinct X_i's (which happens with probability 1 under the model as soon as $n \geq 2$), then the r.h.s. of (3.13) is strictly negative, so we are guaranteed a unique root $\hat{\alpha}$ of this equation. One must solve for this root numerically in practice. Software for computing the digamma function and its derivative (known as the trigamma function) is generally available, e.g. in the cmlib special functions package of fortran subprograms. It is also available as a supplied function in the Mathematica system.

□

3.1.2 Properties of Estimators.

Here we consider various properties of an estimator that may be desirable.

The Principle of Unbiasedness.

Assume $g : \Theta \longrightarrow I\!R$. We say an estimator $\delta(X)$ is *unbiased* for $g(\theta)$ iff

$$E_\theta[\delta(X)] \ = \ g(\theta) \quad , \qquad \text{for all } \theta \in \Theta.$$

(Note: in the above we are implicitly assuming that $E_\theta[|\delta(X)|] < \infty$ for all θ.) Recall from Section 4.2 that an unbiased estimator is a *uniform minimum variance unbiased estimator* (abbreviated UMVUE) iff it has smallest variance (and hence mean squared error) among all unbiased estimators of $g(\theta)$. Classically, unbiasedness has been a fundamental requirement, so there is a well developed theory for unbiased estimators, which we explore in more detail in Section 3.

Let $g(\theta)$ be an estimand in the setup described above. The class of unbiased estimators may be empty. We say $g(\theta)$ is *U–estimable* iff there exists an unbiased estimator for $g(\theta)$. U–estimability is equivalent to the existence of a solution $\delta(x)$ of an integral equation of the form

$$g(\theta) \ = \ \int_\Xi \delta(x)\, dP_\theta(x) \ = \ \int_\Xi \delta(x)\, f_\theta(x)\, d\mu(x) \quad , \text{ for all } \theta \in \Theta \quad .$$

In the last expression we have assumed that the family $\mathbf{P}$ is dominated by μ and $f_\theta = dP_\theta/d\mu$ are the densities.

Example 3.1.3 Suppose $X \sim Poisson(\mu)$ where $\mu > 0$. Then $g(\mu)$ is U–estimable if and only if there is a sequence $\{\delta(k) :, \, k \in I\!N\}$ such that

$$g(\mu) \ = \ \sum_{k=0}^{\infty} e^{-\mu} \frac{\delta(k)}{k!} \mu^k \quad , \quad \text{all } \mu > 0\,.$$

This is true just in case

$$e^\mu g(\mu) \ = \ \sum_{k=0}^{\infty} \frac{\delta(k)}{k!} \mu^k \quad , \quad \text{all } \mu > 0\,.$$

Evidently, $e^\mu g(\mu)$ must have a Taylor series expansion which converges for all $\mu > 0$. Recall that if a power series of the form $\sum_k a_k \mu^k$ converges form some μ_0, then it converges for all μ such that $|\mu| < |\mu_0|$. Thus, the Taylor series above for $e^\mu g(\mu)$ converges for all $\mu \in I\!R$. (Of course, for the statistical problem at hand, it doesn't make sense to speak of $\mu < 0$, but the point is that mathematically we can extend the function g from $(0, \infty)$ to all of $I\!R$.) Furthermore, if an unbiased estimator exists, then $e^\mu g(\mu)$ is infinitely differentiable at all μ, so by an easy chain rule argument (multiply by $e^{-\mu}$), $g(\mu)$ must be infinitely differentiable at all μ. (In fact, the student who knows a little complex analysis can see that $g(\mu)$ is U–estimable in this setup if and only if $g(\mu)$ is analytic in the entire complex plane, i.e. g is an entire function. Again, we don't consider complex values of the Poisson parameter, but mathematically the function can be extended to all of the complex plane.) If $\delta(X)$ is an unbiased estimator of $g(\mu)$ then by uniqueness of Taylor series coefficients,

$$\delta(k) \ = \ \frac{d^k}{d\mu^k} \left[\, e^{\mu} g(\mu) \, \right]_{\mu=0} \ .$$

In this setting, δ is the only unbiased estimator of g and hence is UMVUE. In particular, the standard deviation $\sqrt{\mu}$ is not U–estimable since the square root function it is not differentiable at $\mu = 0$. One estimand which is U–estimable is

$$g(\mu) \ = \ (\, P_\mu[X = 0] \,)^2 \ = \ e^{-2\mu} \ .$$

For this estimand, the unique unbiased estimator is

$$\frac{d^k}{d\mu^k} \left[e^{\mu} e^{-2\mu} \right]_{\mu=0} \ = \ \left[(-1)^k e^{-\mu} \right]_{\mu=0} \ = \ (-1)^k \ .$$

So the UMVUE of $e^{-2\mu}$ is $(-1)^X$, which is a rather silly estimator (especially if X is odd). This example illustrates that restricting oneself to unbiased estimation is not always a wise choice. In the last two estimands we have seen two problems that can exist with the class of unbiased estimators: it may be empty, and it may contain only ridiculous estimators. In general one should always examine a decision rule to see if it "makes sense," no matter what optimality properties it satisfies.

□

The class of U–estimable estimands depends on sample size, and for a given U–estimable estimand the class of unbiased estimators depends on sample size. In the previous example, if we had a random sample X_1, X_2, ..., X_n which were i.i.d. $Poisson(\mu)$ with sample size $n > 1$, then there would generally be more than one unbiased estimator for any U–estimable estimand.

3.1.3 Equivariance and Invariance.

We consider here other "reasonable restrictions" on the class of estimators similar to unbiasedness. Consider a location-scale family X_1, X_2, $\ldots$, X_n i.i.d. with Lebesgue density

$$f_{ab}(x) \ = \ \frac{1}{b} f\left(\frac{x - a}{b} \right)$$

where f is a given Lebesgue probability density function, the unknown location parameter $a \in I\!R$, and the unknown scale parameter $b \in (0, \infty)$. Now suppose someone adds a fixed constant $\alpha \in I\!R$ to the data, say,

$$\tilde{X}_i \ = \ X_i \, + \, \alpha.$$

We will denote the new data vector

$$\tilde{\underline{X}} \;=\; \underline{X} + \alpha\underline{1} \;=\; \underline{X} + \alpha$$

where $\underline{1} = (1, 1, \ldots, 1)$ is a vector of all 1's of dimension n. Note that "$\underline{X} + \alpha$" is an abuse of notation since one can't add a vector and scalar, but we will use it to mean $\underline{X} + \alpha\underline{1}$. Now the distribution of the shifted data $\tilde{\underline{X}}$ still comes from the location-scale family but the location parameter has been shifted from a to $a + \alpha$ (and the scale parameter is unchanged). If $\hat{a}$ is an estimator of a, it makes sense to require that it be *location equivariant*, which means

$$\hat{a}(\underline{x} + \alpha) \;=\; \hat{a}(\underline{x}) \;+\; \alpha, \quad \forall \alpha \in \mathbb{R}, \quad \forall \underline{x} \in \mathbb{R}^n.$$

Furthermore, if someone multiplies the data by a positive scalar β, then the new data vector $\tilde{\underline{X}} = \beta\underline{X}$ still comes from the location-scale family but the location parameter is βa (and the scale parameter has changed to βb). Thus, it also would be reasonable to require that the location estimator $\hat{a}$ be *scale equivariant*, which means

$$\hat{a}(\beta\underline{x}) \;=\; \beta\hat{a}(\underline{x}), \quad \forall \beta \in (0, \infty), \quad \forall \underline{x} \in \mathbb{R}^n.$$

In a similar manner, it is reasonable that an estimator $\hat{b}$ of the scale parameter should be *location invariant*, meaning

$$\hat{b}(\underline{x} + \alpha) \;=\; \hat{b}(\underline{x}), \quad \forall \alpha \in \mathbb{R}, \quad \forall \underline{x} \in \mathbb{R}^n.$$

and *scale equivariant*, i.e.

$$\hat{b}((\beta\underline{x}) \;=\; \beta\hat{b}\hat{a}(\underline{x}) \;+\; \alpha, \quad \forall \beta \in (0, \infty), \quad \forall \underline{x} \in \mathbb{R}^n.$$

These are "reasonable" requirements since the corresponding scale parameter is transformed in this way when the data is shifted or rescaled.

Example 3.1.4 Clearly the sample mean is location-scale equivariant. In fact, given scalars c_i, consider a linear location estimator of the form

$$\hat{a}(\underline{x}) \;=\; \sum_{i=1}^{n} c_i x_i.$$

Let us determine conditions under which this is location-scale equivariant.

$$\hat{a}(\underline{x} + \alpha) \;=\; \hat{a}(\underline{x}) \;+\; \alpha \sum_{i=1}^{n} c_i$$

which equals $\hat{a}(\underline{x}) + \alpha$ for all $\alpha \in \mathbb{R}$ if and only if $\sum_i c_i = 1$. Now $\hat{a}(\beta\underline{x}) = \beta\hat{a}(\underline{x})$ by the distributive law no matter what the c_i are. Since the order statistics are sufficient, it makes sense to require the c_i to be all equal so that $\hat{a}$ does not depend on the order of the data. Hence, we would expect that the choice $c_i = 1/n$ is optimal in the sense of minimizing MSE.

Also, the sample median is location-scale equivariant, and in fact any sample quantile is. Further, one could consider a linear combination of order statistics

$$\hat{a}(\underline{x}) \;=\; \sum_{i=1}^{n} c_i x_{(i)},$$

where again $\sum_i c_i = 1$ is necessary and sufficient to guarantee location-scale equivariance. Such an estimator is called an *L-estimator*. For this type of estimator, there is not an obvious choice for the c_i. In fact, the optimal c_i for minimizing the MSE will depend on the shape of the generating density f.

Now for scale estimation, the sample standard deviation is location invariant and scale equivariant. We can construct a large class of location invariant and scale equivariant estimators of scale by

$$\hat{b}(\underline{x}) \;=\; \left(\sum_{i=1}^{n} |x_i - \hat{a}(\underline{x})|^p \right)^{1/p}$$

where $p > 0$ and $\hat{a}$ is location-scale equivariant. Note that for any $\alpha \in I\!\!R$ and $\beta > 0$,

$$\begin{aligned}
\hat{b}(\beta\underline{x} + \alpha) \;&=\; \left(\sum_{i=1}^{n} |\beta x_i + \alpha - \hat{a}(\beta\underline{x} + \alpha)|^p \right)^{1/p} \\
&=\; \left(\sum_{i=1}^{n} |\beta x_i + \alpha - [\beta\hat{a}(\underline{x}) + \alpha]|^p \right)^{1/p} \\
&=\; \left(\sum_{i=1}^{n} |\beta x_i - \beta\hat{a}(\underline{x})|^p \right)^{1/p} \\
&=\; \beta \left(\sum_{i=1}^{n} |x_i - \hat{a}(\underline{x})|^p \right)^{1/p} \\
&=\; \beta\hat{b}(\underline{x})
\end{aligned}$$

Other location invariant and scale equivariant estimators are given in the exercises.

□

Exercises for Section 5.1.

3.1.1 Suppose we have i.i.d. observations from $N(\mu, \sigma^2)$ where $0 < \sigma^2 \leq 1$. Show that the MLE of σ^2 is $\min\{S^2, 1\}$ where $S^2 = n^{-1} \sum_i (X_i - \overline{X})^2$.

3.1.2 Verify the claims about the MLE's for the $N(\mu, \sigma^2)$, $B(n, p)$, and $Expo(\mu)$ distributions in Example 3.1.1.

3.1.3 Verify the claims about the MLE's for the $N_d(\underline{\mu}, \Sigma)$ in Example 3.1.1.

3.1.4 Verify the claims about the MLE's for the $Laplace(\mu, s)$ distribution in Example 3.1.1.

3.1.5 Verify the claims about the MLE's for the $U(\theta_1, \theta_2)$ distribution in Example 3.1.1.

3.1.6 *(Logistic Regression)* The following is a common example where existence of the MLE can be a problem. Suppose we have data (x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n) where the x_i's are fixed real numbers and the Y_i's are independent Bernoulli random variables with

$$P[Y_i = 1] \;=\; \frac{\exp[a + bx_i]}{1 + \exp[a + bx_i]} \quad .$$

Note that

$$a + bx_i \;=\; \log\left(\frac{P[Y_i = 1]}{1 - P[Y_i = 1]}\right) .$$

Here a and b are unknown real numbers. Assume for convenience that $x_1 \leq x_2 \leq \ldots \leq x_n$ and that there are at least 3 distinct values in the x_i's. Show that the MLE of (a, b) exists if and only if there exists $i_1 < i_2 < i_3$ such that either

$$Y_{i_1} = 1, \quad Y_{i_2} = 0, \quad Y_{i_3} = 1,$$

or

$$Y_{i_1} = 0, \quad Y_{i_2} = 1, \quad Y_{i_3} = 0.$$

Loosely speaking, this says there are at least 2 “sign changes” in the Y_i's. (Hints: Use exponential family theory. If there are less than 2 “sign changes” in the Y_i's, then the log likelihood can be increased by going off to ∞ in a certain direction.)

3.1.7 Let X_1, X_2, ..., X_n be i.i.d. random variables with finite second moment. Consider affine estimators of $\mu = E[X_i]$ which have the form

$$\hat{\mu}(\underline{x}) \;=\; b \;+\; \sum_{i=1}^{n} c_i x_i$$

where b and c_1, c_2, ..., c_n are constants. (Such estimators are sometimes called "linear" although strictly speaking one must set $b = 0$ to get a linear estimator). Determine the values of these which minimize the variance of the estimator subject to the unbiasedness constraint

$$E[\hat{\mu}(\underline{X})] \; = \; \mu$$

for all possible distributions.

3.1.8 Show that each of the following estimators is location invariant and scale equivariant:

$$\begin{aligned}
\hat{b}_1(\underline{x}) &= |x_n - x_1| \\
\hat{b}_2(\underline{x}) &= x_{(n)} - x_{(1)} \\
\hat{b}_3(\underline{x}) &= \frac{1}{n(n-1)} \sum_{i \neq j} |x_i - x_j| \\
\hat{b}_4(\underline{x}) &= \sum_{i=1}^{n} c_i x_{(i)}
\end{aligned}$$

where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ denote the order statistics and for $\hat{b}_4$ the constants c_i satisfy $\sum_i c_i = 0$.

3.2 Rao-Cramer Inequality.

In this section we derive a lower bound on mean squared error of an estimator under certain “regularity” conditions. The basic techniques involved are (i) some tricky calculus (which is where the “regularity” conditions come in), and (ii) the Cauchy-Schwartz inequality. The resulting lower bound will prove very useful for many purposes. Of course, one can always use the lower bound of 0 on means squared error, but the lower bound derived here is much sharper. We will see in particular that when we cannot find an optimal procedure (such as an UMVUE), then the lower bound provides some standard for assessing the performance of an estimator. For instance, if we have an estimator whose MSE is “close” to the lower bound, then we know that it is “close” to optimal.

3.2.1 The Main Result.

Suppose a random observable X has a distribution in a family $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ and we wish to estimate $g(\theta)$. Let $\delta(X)$ be any estimator of $g(\theta)$ and

$$\text{MSE}(\theta, \delta) \;=\; E_\theta[(\delta(X) - g(\theta))^2]$$

its mean squared error. Let us assume $\Theta \subset I\!\!R^p$ and a dominated family with densities

$$f_\theta(x) \;=\; \frac{dP_\theta}{d\mu}(x) \;.$$

The lower bound we shall derive will require “regularity” conditions, mostly to allow differentiation and interchangeability of differentiation and integration. Rather than state the regularity conditions at the outset, we will derive the result and determine what is needed as we go along, which is the natural way to proceed.

Start with the equation

$$\int_\Xi \delta(x) f_\theta(x)\, d\mu(x) \;=\; E_\theta[\delta(X)] \;=\; m(\theta) \quad , \tag{3.14}$$

where $m(\theta)$ just denotes the mean of the statistic $\delta(X)$ as a function of θ. Recalling that $\delta(X)$ is meant as an estimator of $g(\theta)$, we have

$$m(\theta) \;=\; g(\theta) + b(\theta) \quad , \tag{3.15}$$

where

$$b(\theta) \;=\; E_\theta[\delta(X)] - g(\theta)$$

is the bias. Assuming we can differentiate both sides of (3.14) w.r.t. θ and interchange $\int$ and $\partial/\partial\theta_j$, for $1 \le j \le p$, we obtain

$$\int \delta(x) \nabla f_\theta(x)\, d\mu(x) \;=\; \nabla m(\theta) \quad , \tag{3.16}$$

where ∇f_θ denotes the derivative (gradient) of $f_\theta(x)$ w.r.t. θ when x is held fixed. (Note: Many students get confused on which variable to differentiate w.r.t. Just remember, X may be a discrete r.v. so that the density may only be defined on the integers, as a function of X, and then it makes no sense to differentiate w.r.t. x. *All derivatives are w.r.t.* θ.) The l.h.s. of (3.16) can be rewritten as

$$\int \delta(x)\frac{\nabla f_\theta(x)}{f_\theta(x)} f_\theta(x)\, d\mu(x) \;=\; \int \delta(x)\, \underline{\Psi}(x,\theta)\, f_\theta(x)\, d\mu(x) \tag{3.17}$$

$$= \; E_\theta\left[\, \delta(X)\, \underline{\Psi}(X,\theta)\,\right] \;,$$

where

$$\underline{\Psi}(x,\theta) \;=\; \nabla \log\, f_\theta(x) \;=\; \frac{\nabla f_\theta(x)}{f_\theta(x)}\;. \tag{3.18}$$

Furthermore, by interchanging differentiation and integration in the equation

$$\int f_\theta(x)\, d\mu(x) \;=\; 1$$

we obtain

$$\underline{0} \;=\; \int \nabla f_\theta(x)\, d\mu(x) \;=\; E_\theta[\underline{\Psi}(X,\theta)]\;. \tag{3.19}$$

In particular, since $m(\theta)$ is nonrandom,

$$E_\theta[\, m(\theta)\, \underline{\Psi}(X,\theta)\,] \;=\; 0\;. \tag{3.20}$$

If we substitute (3.17) into (3.16) and subtract (3.20) from the result, we obtain

$$E_\theta\left\{\, [\, \delta(X) - m(\theta)\,]\; \underline{\Psi}(X,\theta)\,\right\} \;=\; \nabla m(\theta) \quad . \tag{3.21}$$

For simplicity, first assume $p = 1$, i.e. that we have a one dimensional parameter θ, referred to simply as the one parameter case. Then the ∇'s in (3.21) are simply $d/d\theta$, and in particular are scalar. Now apply the Cauchy-Schwartz inequality (Theorem 2.1.5) to the l.h.s. of (3.21) to obtain

$$m'(\theta)^2 \;\le\; E_\theta\left\{\, [\, \delta(X) - m(\theta)\,]^2\,\right\}\, E_\theta\left\{\, \left[\, \frac{d}{d\theta}\log\, f_\theta(X)\,\right]^2\,\right\}$$

$$= \; \mathrm{Var}_\theta[\delta(X)]\, I_X(\theta)\;, \tag{3.22}$$

where $I_X(\theta)$ is called the *Fisher information* (about θ contained in X) and is given in the one parameter case by

$$I(\theta) \;=\; E_\theta\left\{\, \left[\, \frac{d}{d\theta}\log\, f_\theta(X)\,\right]^2\,\right\} \;=\; E_\theta[\Psi(X,\theta)^2]\;, \tag{3.23}$$

where we usually drop the subscript X when it is clear from context. If we assume that $I(\theta) > 0$ then from (3.22) we get

$$\mathrm{Var}\theta[\delta(X)] \ \geq \ \frac{m'(\theta)^2}{I(\theta)} \ . \tag{3.24}$$

Here, m' denotes $dm/d\theta$. Recalling that mean squared error (MSE) is bias squared plus variance (see (4.6) we obtain

$$\mathrm{MSE}(\theta, \delta) \ \geq \ b(\theta)^2 \ + \ \frac{m'(\theta)^2}{I(\theta)} \ . \tag{3.25}$$

This is the *Rao–Cramer lower bound* in the one parameter case. If δ is unbiased or if the estimand is the parameter itself, then this can be simplified (see Corollary 3.2.2 below).

Now we return to the multiparameter case, i.e. the dimension of θ is $p \geq 1$. We wish to apply Cauchy-Schwartz to (3.21) but it is a vector equation. Let $\underline{v}$ be any nonrandom vector, and take inner products of both sides of (3.21) with $\underline{v}$ to obtain

$$E_\theta \left\{ \left[\delta(X) - m(\theta) \right] \left[\underline{\Psi}(X, \theta)' \underline{v} \right] \right\} \ = \ \nabla m(\theta)' \underline{v} \ . \tag{3.26}$$

Here, A' denotes the transpose of the matrix A. Now by Cauchy-Schwartz we have

$$\left\{ \nabla m(\theta)' \underline{v} \right\}^2 \ \leq \ \mathrm{Var}_\theta[\delta(X)] E_\theta \left\{ \left[\underline{\Psi}(X, \theta)' \underline{v} \right]^2 \right\} \tag{3.27}$$

Note that

$$\begin{aligned} E_\theta \left\{ \left[\underline{\Psi}(X, \theta)' \underline{v} \right]^2 \right\} \ &= \ E_\theta \left\{ \left(\underline{v}' \underline{\Psi}(X, \theta) \right) \left(\underline{\Psi}(X, \theta)' \underline{v} \right) \right\} \\ &= \underline{v}' \, E_\theta \left[\underline{\Psi}(X, \theta) \, \underline{\Psi}(X, \theta)' \right] \underline{v} \ = \ \underline{v}' \, I_X(\theta) \, \underline{v} \ , \end{aligned} \tag{3.28}$$

where $I_X(\theta)$ is a $p \times p$ matrix given by

$$\begin{aligned} I(\theta) \ &= \ E_\theta \left[\underline{\Psi}(X, \theta) \, \underline{\Psi}(X, \theta)' \right] \\ &= \ E_\theta \left\{ \left[\nabla \log f_\theta(X) \right] \left[\nabla \log f_\theta(X) \right]' \right\} \ . \end{aligned} \tag{3.29}$$

$I(\theta)$ is called the *Fisher information matrix.* Assuming $I(\theta)$ is strictly positive definite and $\underline{v} \neq \underline{0}$, then (3.27) yields

$$\mathrm{Var}_\theta[\delta(X)] \ \geq \ \frac{\left\{ \nabla m(\theta)' \underline{v} \right\}^2}{\underline{v}' I(\theta) \underline{v}} \ . \tag{3.30}$$

We are free to choose $\underline{v}$ in (3.30) (in a nonrandom way subject to $\underline{v} \neq \underline{0}$) so as to obtain the largest possible r.h.s. Set $\underline{u} = I(\theta)^{1/2} \underline{v}$ so that $\underline{v} = I(\theta)^{-1/2} \underline{u}$ (see Exercise 2.1.18 for definition and construction of $I(\theta)^{1/2}$). Note that as $\underline{v}$ varies over $I\!R^p - \{\underline{0}\}$ then so also does $\underline{u}$. Then (3.30) becomes

$$\mathrm{Var}_\theta[\delta(X)] \ \geq \ \frac{\left\{ \nabla m(\theta)' I(\theta)^{-1/2} \underline{u} \right\}^2}{\|\underline{u}\|^2} \ . \tag{3.31}$$

Note that if we multiply $\underline{u}$ in (3.31) by any positive scalar, it does not change the r.h.s. Thus, we may constrain

$$\|\underline{u}\|^2 \;=\; 1 \tag{3.32}$$

Then, subject to (3.32) we wish to maximize the following expression over $\underline{u} \in I\!\!R^p$:

$$\left\{ \nabla m(\theta)' I(\theta)^{-1/2} \underline{u} \right\}^2 \;=\; \left\{ \left[I(\theta)^{-1/2} \nabla m(\theta) \right]' \underline{u} \right\}^2 . \tag{3.33}$$

An inner product of a unit vector with a given vector is maximized by choosing the unit vector in the direction of the given vector, i.e. the last expression in (3.33) is maximized if

$$\underline{u} \;=\; \frac{I(\theta)^{-1/2} \nabla m(\theta)}{\|I(\theta)^{-1/2} \nabla m(\theta)\|} . \tag{3.34}$$

Putting this back into the r.h.s. of (3.33) gives

$$\left\{ \frac{\nabla m(\theta)' I(\theta)^{-1} \nabla m(\theta)}{\|I(\theta)^{-1/2} \nabla m(\theta)\|} \right\}^2 \;=\; \nabla m(\theta)' I(\theta)^{-1} \nabla m(\theta) . \tag{3.35}$$

Finally, using (3.35) in (3.31) gives

$$\mathrm{Var}_\theta[\delta(X)] \;\geq\; \nabla m(\theta)' I(\theta)^{-1} \nabla m(\theta) . \tag{3.36}$$

We state this formally in the following where the regularity conditions are spelled out.

Theorem 3.2.1 (Rao-Cramer Inequality.) *Suppose* $Law[X] \in \mathbf{P} = \{P_\theta : \theta \in \Theta\} \ll \mu$ *σ–finite. Let* $dP_\theta/d\mu = f_\theta$ *denote the densities w.r.t.* μ*, let* $g : \Theta \to I\!\!R$ *be an estimand, and let* $\delta(X)$ *be an estimator of* $g(\theta)$*. Assume the following regularity conditions:*

(i) $\Theta \subset I\!\!R^p$ *is an open set.*

(ii) $E_\theta[\delta(X)^2] < \infty$ *for all* $\theta \in \Theta$*.*

(iii) *For* μ*-almost all* x*,* $\underline{\Psi}(x,\theta) = \nabla \log f_\theta(x)$ *exists for all* $\theta \in \Theta$*.*

(iv) $E_\theta[\|\underline{\Psi}(X,\theta)\|^2] < \infty$ *for all* $\theta \in \Theta$*.*

(v) *For* $\phi(x) \equiv 1$ *and* $\phi(x) = \delta(x)$*,* $\gamma(\theta) = E_\theta[\phi(x)]$ *is differentiable for all* θ*, and the derivative can be computed by interchanging differentiation and integration.*

(vi) *The Fisher information matrix in (3.29) is strictly positive definite for all* $\theta \in \Theta$*.*

Then

$$MSE(\theta, \delta) \;\geq\; b(\theta)^2 \;+\; \nabla m(\theta)' I(\theta)^{-1} \nabla m(\theta)\ . \tag{3.37}$$

gives a lower bound on the MSE of any estimator of $g(\theta)$ with bias $b(\theta)$ and mean function $m(\theta) = g(\theta) + b(\theta)$.

Proof. First, we discuss the "self–consistency" of the regularity conditions. Condition (i) is merely to guarantee that we can discuss differentiability of functions of θ. It is needed, for instance, in order that (iii) and (iv) make sense. If for a given θ there is not some neighborhood of θ contained in the parameter space, then we have potential difficulties with differentiation at θ. Condition (i) can be weakened considerably, but it is not particularly useful to do so.

If one attempts to compute derivatives in (v) by interchanging $\int \ldots d\mu$ with $\partial/\partial\theta_j$, it would be necessary to differentiate $f_\theta(x)$, so it is implicit in (v) that $f_\theta(x)$ is differentiable in θ. This also follows from (iii) for values of θ and x for which $f_\theta(x) > 0$ so that $\log f_\theta(x)$ is finite. However, it seems a little clearer to explicitly state (iii) which also guarantees that (iv) makes sense because $\underline{\Psi}$ exists. Note that the $\phi = \delta$ case of (v) makes sense by (ii) (i.e. $E_\theta[|\delta(X)|] < \infty$ for all θ) and this case of (v) also states that $m(\theta)$ is differentiable, which is of course necessary for the lower bound to make sense as ∇m appears there. Finally, (iv) is needed for (vi), i.e. we need to know that the entries of I are well defined real numbers.

Now we will indicate where the regularity conditions are needed in the calculations leading up to the theorem. The first result requiring justification is (3.16), which follows from condition (v) with $\phi = \delta$. Equation (3.17) requires condition (iii) to justify the existence of $\underline{\Psi}$ and also,

$$\begin{aligned} &\int |\delta(x)|\, \|\underline{\Psi}(x,\theta)\|\, f_\theta(x)\, d\mu(x) \\ &\leq \;\; \left\{\int \delta(x)^2 f_\theta(x)\, d\mu(x)\right\}^{1/2} \left\{\int \|\underline{\Psi}(x,\theta)\|^2 f_\theta(x)\, d\mu(x)\right\}^{1/2} \\ &< \;\; \infty \;\; , \end{aligned}$$

where the last line follows from conditions (ii) and (iv). This shows integrability, i.e. that the integrals in (3.17) are defined (and in fact, finite). Of course, condition (v) with $\phi \equiv 1$ is needed for (3.19), and we know from (iv) that $E_\theta[\|\underline{\Psi}(X,\theta)\|] < \infty$, so the integral in (3.19) is defined (and finite), and also (3.21) is valid (i.e. all integrals are finite so there is no $\infty - \infty$). Of course, condition (vi) is needed at (3.30) (to know the denominator is not 0), and in the definition of $\underline{v}$ which appears in (3.31) (so we know that $I(\theta)^{-1/2}$ exists). This completes the verification.

□

Remarks 3.2.1 (a) The result is sometimes called the *Information inequality*, *Rao-Cramer lower bound*, or *Information lower bound*.

(b) If (i) doesn't hold, then we simply replace Θ by its interior.

(c) Assumption (iii) that $\log f_\theta(x)$ be differentiable in θ for μ-almost all x implicitly requires that $\mu(\{\, x : f_\theta(x) = 0 \text{ for any } \theta \,\}) = 0$, for otherwise $\log f_\theta(x)$ is not even finite μ-a.e. If $f_\theta(x) = 0$ on a set if positive μ measure, then we may be able to change dominating measures so as to avoid the problem.

For example, if we use Lebesgue measure to dominate the $Exp[\beta, 0]$ family with densities $\beta^{-1} \exp[-x/\beta] I_{(0,\infty)}(x)$, then we do not have condition (iii). However, we can simply change the dominating measure to Lebesgue measure restricted to $(0, \infty)$, i.e. $\mu(B) = m(B \cap (0, \infty))$.

On the other hand, for the $Unif[a, b]$ family, there is no way to change dominating measures so as to achieve (iii). The same holds for the shifted exponential family $Exp[\beta, b]$. In general, in order for condition (iii) to hold it is necessary that the support of the P_θ's not depend on θ.

(d) Condition (v) is applied with $\phi(x) = \delta(x)$ in (3.16) and $\phi(x) = 1$ in (3.19). In general, one will use Theorem 1.2.10 to check condition (v), as in the proof of Theorem 2.2.1 (b). A sufficient condition for (v) to hold is the following: for all $\theta_0 \in \Theta$ there is a neighborhood $B(\theta_0, \epsilon) \subset \Theta$ and a constant $K < \infty$ such that $\|\nabla f_\theta(x)\| \le K f_{\theta_0}(x)$ for all $\theta \in B(\theta_0, \epsilon)$ and all $x \in \Xi$. If this condition holds, then we may apply Theorem 1.2.10 to the integral $\int g(x, \theta_i) d\mu(x)$ where $g(x, \theta_i) = \phi(x) f_\theta(x)$ (where we hold all components of θ constant except θ_i), and the dominating function for $\partial g/\partial\theta_i$ is $G(x) = K|\phi(x)| f_{\theta_0}(x)$. Since $E_\theta|\phi(X)| < \infty$ is assumed, we have $\int G(x) d\mu(x) = K E_{\theta_0}|\phi(X)| < \infty$.

□

Corollary 3.2.2 *Under the same Assumptions as Theorem 3.2.1, the following hold:*

(a) If $\delta(X)$ is unbiased for $g(\theta)$, then

$$Var_\theta[\delta(X)] \;\ge\; \nabla g(\theta)' I(\theta)^{-1} \nabla g(\theta) \,.$$

(b) If θ is 1 dimensional and $g(\theta) = \theta$, then

$$MSE(\theta, \delta) \;\ge\; b(\theta)^2 \;+\; \frac{[1 + b'(\theta)]^2}{I(\theta)} \,.$$

(c) If θ is 1 dimensional and $\delta(X)$ is any unbiased estimator of θ, then

$$Var_\theta[\delta(X)] \;\ge\; \frac{1}{I(\theta)} \,.$$

□

It is worthwhile to consider exponential families since some of the regularity condtions are obtained “for free” in that case.

Proposition 3.2.3 *Suppose X has distribution in an exponential family with densities of the form*

$$dP_\theta(x) \;=\; f_\theta(x)\,d\mu(x) \;=\; \exp\left[\,\eta(\theta)'T(x) \;-\; B(\theta)\,\right]\,h(x)\,d\mu(x) \quad , \quad \theta \in \Theta \,.$$

Suppose $\Theta \subset I\!\!R^p$ is open and $\eta : \Theta \longrightarrow I\!\!R^k$ is differentiable with derivative (Jacobian matrix) $D\eta$ which is of full rank $p \leq k$. Assume further that the natural parameter exponential family with natural parameter space Λ_0 is of full rank, and $\eta(\Theta)$ is contained in the interior of Λ_0. Let $g : \Theta \longrightarrow I\!\!R$ be continuously differentiable. Suppose $\delta(X)$ is an estimator of $g(\theta)$ for which

$$\int \delta(X)^2 \exp[\eta' T(x)]\,d\mu(x) \;<\; \infty \quad , \textit{ for all } \eta \in \Lambda_0 \,. \tag{3.38}$$

Then

$$E_\theta\{[(\delta(X) - g(\theta)]^2\} \;\geq$$

$$\{E_\theta[(\delta(X) - g(\theta)]\}^2 \;+\; \nabla E_\theta[\delta(X)]'\;\left[\,D\eta(\theta)'Cov_\theta[T(X)]D\eta(\theta)\,\right]^{-1}\,\nabla E_\theta[\delta(X)]\,.$$

Proof. We verify that the conditions of the previous theorem apply. Conditions (i) and (ii) are already assumed in the Proposition (condition (ii) in (3.38)). For (iii), it may be necessary to change dominating measures μ so that $h(x) > 0$, but this is easily done as in Remark 2.3.1 (b). Then differentiability of log f_θ w.r.t. θ is easy from the chain rule since log f_θ is clearly differentiable in η and η is assumed differentiable in θ. In fact, we have

$$\Psi(x,\theta)' \;=\; \frac{d}{d\theta}\log\, f_\theta(x) \;=\; \left[\frac{d}{d\eta}\log\, f_\theta(x)\right]\frac{d\eta}{d\theta} \tag{3.39}$$

$$=\; \left[T(x)' \;-\; \frac{dA(\eta)}{d\eta}\right]\frac{d\eta}{d\theta} \;=\; (\,T(x) - E_\theta[T(X)]\,)'\,\frac{d\eta}{d\theta}\,.$$

In the above, we use $A(\eta) = B(\theta)$ when $\eta = \eta(\theta)$, and Proposition 2.3.1 (b).

Condition (v) holds for any $\phi(x)$ such that $E_\theta[|\phi(X)|] < \infty$ for all θ by Proposition 2.3.1 (c).

Also,

$$I(\theta) \;=\; E_\theta[\Psi(X,\theta)\Psi(X,\theta)'] \tag{3.40}$$

$$=\; E_\theta\left\{\left(\frac{d\eta}{d\theta}\right)'(\,T(X) - E_\theta[T(X)]\,)\,(\,T(X) - E_\theta[T(X)]\,)'\,\frac{d\eta}{d\theta}\right\}$$

$$=\; \left(\frac{d\eta}{d\theta}\right)'\,\mathrm{Cov}_\theta[T(X)]\,\left(\frac{d\eta}{d\theta}\right)\,.$$

Note in particular that $E_\theta[\|\Psi(X,\theta)\|^2] = \text{Trace}[I(\theta)] < \infty$ by Proposition 2.3.1 so (iv) holds. Since $d\eta/d\theta$ is full rank and the natural family is full rank, it follows that $I(\theta)$ is strictly positive definite. To see this, note that if $v \neq 0$ is a p-vector and $u = (d\eta/d\theta)v$ (which is nonzero a k-vector since $(d\eta/d\theta)$ is full rank of dimension $k \times p$ with $p \leq k$), then

$$v'I(\theta)v \; = \; u'\text{Cov}_\theta[T(X)]u \; > \; 0$$

since $\text{Cov}_\theta[T(X)]$ is strictly positive definite (see the proof of Proposition 2.3.1 (b)), so condition (vi) holds. Plugging (3.40) into the Rao-Cramer inequality, we obtain the result.

□

3.2.2 Facts about Fisher Information.

Here we develop some useful results about Fisher information. The first one states that the Fisher information from independent observations is the sum of the individual (or marginal) Fisher informations, which is intuitively appealing. The second result concerns transformation of Fisher information under reparameterization. The final result gives an alternative formula for computing Fisher information which was already evident in the last result above (see Remark 3.2.2 below).

Proposition 3.2.4 *Suppose X and Y are independent observables from $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ and $\mathbf{Q} = \{Q_\theta : \theta \in \Theta\}$, respectively. (Note that the parameter space is the same for both families.) Assume both families satisfy the conditions of Theorem 3.2.1, and denote the Fisher informations, $I_X(\theta)$ and $I_Y(\theta)$, respectively. Then the Fisher information for the joint observation (X,Y), denoted $I_{XY}(\theta)$, is defined and $I_{XY}(\theta) = I_X(\theta) + I_Y(\theta)$.*

In particular, if X_1, X_2, ..., X_n are i.i.d. with common distribution from $\mathbf{P}$ as above, then

$$I_{(X_1,X_2,\ldots,X_n)}(\theta) \; = \; nI_{X_1}(\theta) \; .$$

Proof. We are implicitly assuming both families are dominated, and let μ and ν denote the dominating measures for $\mathbf{P}$ and $\mathbf{Q}$, respectively. Also, let $f_\theta(x)$ and $g_\theta(y)$ denote the respective densities. Then on the product of the respective observation spaces, the joint distribution is dominated by $\mu \times \nu$ (Proposition 1.4.3), and the logarithm of the joint density is the sum of the logarithms of the marginals. Hence,

$$E_\theta\Big[\,(\nabla \log\, f_\theta(X) \,+\, \nabla \log\, g_\theta(Y)\,)\,(\nabla \log\, f_\theta(X) \,+\, \nabla \log\, g_\theta(Y)\,)'\,\Big]$$

$$= \; \text{Cov}_\theta\,[\,\nabla \log\, f_\theta(X) \,+\, \nabla \log\, g_\theta(Y)\,]$$

$$= \mathrm{Cov}_\theta[\nabla \log f_\theta(X)] + \mathrm{Cov}_\theta[\nabla \log g_\theta(Y)] = I_X(\theta) + I_Y(\theta) .$$

Independence of X and Y implies the second equality above. Note that the general equation

$$I(\theta) = \mathrm{Cov}_\theta[\Psi(X, \theta)]$$

follows from the definition of covariance and (3.19).

□

Proposition 3.2.5 *Suppose $\{P_\theta : \theta \in \Theta\}$ satisfies the conditions of Theorem 3.2.1 with Fisher information $I(\theta)$. Let $\zeta = h(\theta)$ where $h : \mathbb{R}^p \to \mathbb{R}^p$ is one to one and continuously differentiable with continuously differentiable inverse. Denote the Fisher information for the new parameter ζ by $J(\zeta)$ and let $G(\zeta) = (dh/d\theta)(h^{-1}(\zeta))$. Then*

$$J(\zeta) = [G(\zeta)^{-1}]' \, I(h^{-1}(\zeta)) \, [G(\zeta)^{-1}] \quad .$$

Proof. Let $g = h^{-1}$. Then the densities are $f_\theta(x) = f_{g(\zeta)}(x)$. Thus, by the chain rule for multivariate differentiation and the formula for differentiation of an inverse function,

$$\frac{d}{d\zeta} \log f_{g(\zeta)}(x) = \left[\frac{d}{d\theta} \log f_\theta(x) \right]_{\theta = g(\zeta)} \frac{dg}{d\zeta}(\zeta)$$

$$= \left[\frac{d}{d\theta} \log f_\theta(x) \right]_{\theta = g(\zeta)} [G(\zeta)]^{-1} .$$

Using $\theta = g(\zeta)$ and plugging this into the formula for Fisher information gives

$$J(\zeta) = E_\theta \left\{ \left[\left(\frac{d}{d\theta} \log f_\theta(x) \right) [G(\zeta)]^{-1} \right] \left[\left(\frac{d}{d\theta} \log f_\theta(x) \right) [G(\zeta)]^{-1} \right]' \right\}$$

$$= ([G(\zeta)]^{-1})' E_\theta \left[\left(\frac{d}{d\theta} \log f_\theta(x) \right) \left(\frac{d}{d\theta} \log f_\theta(x) \right)' \right] [G(\zeta)]^{-1} \quad ,$$

which is the desired formula. Note that we can factor $G(\zeta)^{-1}$ out of the expectation since it is nonrandom.

□

Proposition 3.2.6 *Suppose that in addition to the assumptions of Theorem 3.2.1, the family* $\{P_\theta : \theta \in \Theta\}$ *has densities* $f_\theta(x)$ *satisfying*

(vi) $\log f_\theta(x)$ *is twice continuously differentiable for* μ*-almost all* x*, and if* $H(X,\theta)$ *denotes the Hessian of* $\log f_\theta(X)$*, then*

$$E_\theta[|H_{ij}(X,\theta)|] \; = \; E_\theta\left[\left| \; \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j} \log f_\theta(X) \right| \; \right] \; < \; \infty \quad , \quad \textit{for } 1 \le i,j \le p \quad .$$

(vii) We can interchange differentiation and integration as in the following:

$$\frac{\partial}{\partial\theta_i} \int \Psi_j(x,\theta) f_\theta(x)\, d\mu(x) \; = \; \int H_{ij}(X,\theta) f_\theta(x)\, d\mu(x) \; .$$

Then the Fisher information is given by

$$I(\theta) \; = \; -E_\theta[H(X,\theta)] \; .$$

Proof. Note that $H(X,\theta) = \frac{d}{d\theta}\Psi(X,\theta)$. Applying condition (vii) to (3.19) and using the product rule for differentiation yields,

$$0 \; = \; \frac{d}{d\theta} E_\theta[\Psi(X,\theta)] \tag{3.41}$$

$$= \frac{d}{d\theta} \int \Psi(x,\theta) f_\theta(x)\, d\mu(x) \; = \; \int \frac{d}{d\theta} \left[\Psi(x,\theta) f_\theta(x) \right] d\mu(x)$$

$$= \int \left[\left(\frac{d}{d\theta}\Psi(x,\theta) \right) f_\theta(x) \; + \; \Psi(x,\theta) \left(\frac{d}{d\theta} f_\theta(x) \right) \right] d\mu(x) \; .$$

Now recall that

$$\frac{d}{d\theta} f_\theta(x) \; = \; \Psi(x,\theta)' f_\theta(x) \; .$$

Using this and the definition of H, the last expression in (3.41) equals

$$\int\int \left[H(x,\theta) \; + \; \Psi(x,\theta)\Psi(x,\theta)' \right] f_\theta(x)\, d\mu(x) \; = \; E_\theta[H(X,\theta)] \; + \; I(\theta) \; = \; 0 \; .$$

Subtracting $E_\theta[H(X,\theta)]$ from both sides gives the desired result.

□

Remarks 3.2.2 We can use the previous two Propositions to derive part of Proposition 3.2.3, namely the formula for the Fisher information in the case where the natural parameter mapping $\eta(\theta)$ is one to one, continuously differentiable and has continuously differentiable inverse. Assume first the family is in canonical form, and let $J(\eta)$ denote the information for the natural parameter. Then it follows from Proposition 3.2.6 (see Exercise 3.2.2) that

$$J(\eta) \; = \; -E_\eta \left[\frac{d^2}{d\eta^2} \left(\eta' T(X) \; - \; A(\eta) \right) \right]$$

$$= \; -E_\eta[-A(\eta)] \; = \; \frac{d^2}{d\eta^2} A(\eta) \; = \; \mathrm{Cov}_\eta[T(X)] \; ,$$

where the last equality follows from Proposition 2.3.1 (b). Now, let $\theta = h(\eta)$ where h is the inverse of the natural parameter mapping $\eta(\theta)$. Note that θ takes on the role of ζ in Proposition 3.2.5 and $\eta(\cdot)$ takes on the role of h^{-1} in that Proposition. Also, the roles of I and J are reversed. Then $G(\theta)$ in Proposition 3.2.5 is $(d\eta/d\theta)(\theta)^{-1}$. Hence, the Fisher information for θ is by Proposition 3.2.5

$$I(\theta) \; = \; \left(\frac{d\eta}{d\theta}(\theta) \right)' J(\eta) \left(\frac{d\eta}{d\theta}(\theta) \right) \; .$$

Combining the previous two displays, one sees the formula for Fisher information which appears in the lower bound in Proposition 3.2.3 (see (3.40)).

□

3.2.3 Applications.

Example 3.2.1 Let f be a positive probability density function w.r.t. Lebesgue measure on $\mathbb{R}$, and assume f is continuously differentiable. Put

$$\psi(x) \; = \; -f'(x)/f(x) \; = \; -\frac{d}{dx} \log f(x) \; . \qquad (3.42)$$

(Note: This is one of the few times we will differentiate w.r.t. the variable x.) Then the location family of densities generated by f is

$$f_b(x) \; = \; f(x-b) \; ,$$

where $-\infty < b < \infty$ is the location parameter. Assume that there is an $\epsilon_0 > 0$ and a function $G_0(x)$ such that

$$|\beta| < \epsilon_0 \; \Rightarrow \; |f'(x-\beta)| \; < \; G_0(x) \quad , \text{ for all } x \quad , \qquad (3.43)$$

and

$$\int G_0(z)\, dz \; < \; \infty \; . \qquad (3.44)$$

This will imply condition (v) of Theorem 3.2.1 with $\phi \equiv 1$. To see this, note that for any b_0, $|-f'(x-b)| \le G_{b_0}(x)$ for $b \in (b_0 - \epsilon_0, b_0 + \epsilon_0)$ where $G_{b_0}(x)$

$= G_0(x - b_0)$. Then, Theorem 1.2.10 can be used to interchange d/db and $\int$ in $\int f(x-b)\,dx$ for $b \in (b_0 - \epsilon_0, b_0 + \epsilon_0)$, and since b_0 is arbitrary, for all b.

Assume also that

$$\iota \equiv \int \psi(x)^2 f(x)\,dx \ < \ \infty \ . \tag{3.45}$$

Then the Fisher information for location estimation is

$$I(b) \ = \ E_b\left[\left(\frac{d}{db}\log f(X-b)\right)^2\right] \tag{3.46}$$

$$= \int_{-\infty}^{\infty} \psi(x-b)^2 f(x-b)\,dx = \int_{-\infty}^{\infty} \psi(z)^2 f(z)\,dz \ = \ \iota \ ,$$

where the second to last equation follows from the simple change of variables, $z = x - b$. Note that in this case, the Fisher information is independent of the parameter value. Of course the Fisher information for a random sample of size n from this model is n times this. Thus, if $\hat{b} = \delta(\underline{X})$ is an unbiased estimator based on a random sample of size n and satisfying condition (ii) of Theorem 3.2.1, then

$$\mathrm{Var}[\hat{b}] \ \geq \ \frac{1}{n\iota} \ .$$

Now we consider three special cases.

(a) Suppose f is the $N(0,1)$ density, i.e.

$$f(x) \ = \ e^{-x^2/2}$$

where we have included the constant $1/\sqrt{2\pi}$ in the dominating measure for convenience. Then

$$f'(x-\beta) \ = \ -(x-\beta)e^{-(x-\beta)^2/2} \ .$$

Note that

$$|-(x-\beta)| \ \leq \ |x| + \epsilon_0 \quad , \quad \text{for } |\beta| < \epsilon_0 \ .$$

Also,

$$-(x-\beta)^2 \ = \ -(x^2 - 2\beta x + \beta^2)$$
$$\leq \ -(x^2 - 2\epsilon_0|x| + \epsilon_0^2 - \epsilon_0^2) \ = \ (|x| - \epsilon_0)^2 - \epsilon_0^2 \ ,$$

for $|\beta| \leq \epsilon_0$, so we may take

$$G_0(x) \ = \ (|x| + \epsilon_0)e^{-\epsilon_0^2/2}\left[e^{-(x-\epsilon_0)^2/2}I_{[0,\infty)}(x) \ + \ e^{-(x+\epsilon_0)^2/2}I_{(-\infty,0)}(x)\right]$$

which is clearly integrable w.r.t. Lebesgue. Then,

$$\psi(x) \ = \ -\frac{d}{dx}(-x^2/2) \ = \ x \ ,$$

and condition (3.45) clearly holds. Thus,

$$\iota \ = \ 1 \ , \tag{3.47}$$

for the $N(0,1)$ problem. By Corollary 3.2.2 (c), a lower bound on the variance of any unbiased estimator $\delta(X)$ is $1/n$, which the variance achieved by $\bar{X}$. This shows $\bar{X}$ to be the UMVUE, a fact we already knew.

(b) Suppose f is the logistic density, given by

$$f(x) \;=\; \frac{e^{-x}}{[1+e^{-x}]^2}\;. \tag{3.48}$$

Note that this is the density for the c.d.f.

$$F(x) \;=\; \frac{1}{1+e^{-x}}\;.$$

Using the quotient rule for differentiation,

$$f'(x) \;=\; \frac{-e^{-x}[1-e^{-x}]}{[1+e^{-x}]^3}\;.$$

Since $|1-\exp[-x+\beta]| \le 1+\exp[-x+\beta]$, we have for $|\beta| \le \epsilon_0$,

$$|f'(x-\beta)| \;\le\; \frac{e^{-x+\beta}}{[1+e^{-x+\beta}]^2}$$

$$\le\; e^{2\epsilon_0}\,\frac{e^{-x-\epsilon_0}}{[1+e^{-x-\epsilon_0}]^2} \;=\; G_0(x)$$

and

$$\int G_0(x)\,dx \;=\; e^{2\epsilon_0}\int f(x-\epsilon_0)\,dx \;=\; e^{2\epsilon_0} \;<\; \infty\;.$$

Note that

$$\psi(x) \;=\; \frac{1-e^{-x}}{1+e^{-x}} \;=\; \frac{e^{x/2}-e^{-x/2}}{e^{x/2}+e^{-x/2}} \;=\; \tanh(x/2)\;,$$

is bounded between ±1, so $\psi^2(x)f(x)$ is Lebesgue integrable. Thus,

$$\iota \;=\; \int \psi^2(x)\,f(x)\,dx \;=\; \int_{-\infty}^{\infty} \frac{e^{-x}[1-e^{-x}]^2}{[1+e^{-x}]^4}\,dx$$

$$=\; \int_1^{\infty} \frac{[2-u]^2}{u^4}\,du \;=\; \frac{1}{3}\;,$$

where the second to last equation follows from the change of variables $u = 1+\exp[-x]$, and the evaluation can be done with elementary calculus. Thus, the lower bound on variance of an unbiased estimator of location based on n independent observations from the logistic density in (3.48) is $3/n$.

(c) In order to make a fair comparison of the logistic density in (b) with the $N(0,1)$ density in (a), it is useful to rescale so that the variance of the density in (3.48) is 1. One can show that (e.g. using residue calculus, or more simply looking in a table of integrals)

$$\int_{-\infty}^{\infty} \frac{x^2 e^{-x}}{[1+e^{-x}]^2}\, dx \;=\; \frac{\pi^2}{3} \; .$$

So the rescaled logistic density with variance 1 is $(\pi/\sqrt{3})f(x\pi/\sqrt{3})$, and the Fisher information for location for this family is.

$$\tilde{\iota} \;=\; \frac{1}{3}\left(\frac{\pi^2}{3}\right) \;=\; \left(\frac{\pi}{3}\right)^2 \;\doteq\; 1.0966 \; .$$

Thus, the lower bound on variance for an unbiased estimator (for this rescaled logistic family with variance 1) of location based on a sample of size n is $(3/\pi)^2/n$ $\doteq .9119/n$. Of course, the variance of $\bar{X}$, the sample mean is $1/n$, which is not much larger. So if we are unbiasedly estimating location in the logistic location family, then not much is lost by simply using $\bar{X}$, since no unbiased estimator could have a variance less than about 91% of the sample mean, anyway.

(d) Next, we consider the Cauchy density,

$$f(x) \;=\; \frac{1}{\pi(1+x^2)} \; , \tag{3.49}$$

which is the Lebesgue density of the p.m. with c.d.f.

$$F(x) \;=\; \frac{1}{\pi}\,\mathrm{Tan}^{-1}x \;+\; \frac{1}{2} \; .$$

Now, if $|\beta| < \epsilon_0 < 1$, then

$$|f'(x-\beta)| \;=\; \frac{2}{\pi}\,\frac{|x-\beta|}{[1+(x-\beta)^2]^2}$$

$$\leq\; \frac{2}{\pi}\,\frac{|x|+\epsilon_0}{[1-\epsilon_0^2+x^2]^2} \;:=\; G_0(x)$$

and

$$\int_{-\infty}^{\infty} G_0(x)\, dx \;\leq\; \int_{-1}^{1} G_0(x)\, dx \;+\; \frac{4}{\pi}\int_{1}^{\infty} \frac{x+\epsilon_0}{x^4}\, dx \;<\; \infty \; .$$

This verifies (3.44). Also,

$$\psi(x) \;=\; \frac{2x}{1+x^2}$$

is bounded between ±2, so (3.45) holds and

$$\iota \;=\; \frac{4}{\pi}\int_{-\infty}^{\infty} \frac{x^2}{[1+x^2]^3}\, dx \;=\; \frac{1}{2} \; . \tag{3.50}$$

The integral can be evaluated by residue calculus, elementary (but tedious methods) or partial fractions, or with a table of integrals. Thus, the variance of an unbiased estimator of location in the Cauchy location family (generated by (3.49)) is always no less than $1/(2n)$, where n is sample size.

(e) Unlike with the logistic, there are difficulties comparing the Cauchy with $N(0, 1)$ since (i) the Cauchy density doesn't have a mean, or a variance, and (ii) consequently $\bar{X}$ is not an unbiased estimator of location (its bias is undefined). In particular, we cannot rescale the Cauchy to variance 1. However, we can "match" another scale functional, namely the interquartile range

$$\mathrm{IQR}(F) \;=\; F^{-1}(3/4) \;-\; F^{-1}(1/4)\;. \tag{3.51}$$

(One must use a more general definition of the quantile function to define IQR, such as given in Section 2.4, if F is not strictly increasing and continuous.) For the standard normal,

$$\Phi^{-1}(3/4) \;-\; \Phi^{-1}(1/4) \;\doteq\; 0.6741892 \;-\; (-0.6741892) \;\doteq\; 1.348378\;. \tag{3.52}$$

while for the "standard" Cauchy with density as given in (3.49), the IQR is

$$\tan\left[\pi\left(\frac{3}{4}-\frac{1}{2}\right)\right] \;-\; \tan\left[\pi\left(\frac{1}{4}-\frac{1}{2}\right)\right] \;=\; 2\;. \tag{3.53}$$

Thus, if we rescale the Cauchy to have the same IQR as the $N(0, 1)$, then the information for location

$$\tilde{\iota} \;\doteq\; (2/1.348378)^2/2 \;\doteq\; 1.100035\;.$$

Thus, the lower bound on variance of an unbiased estimate of this rescaled Cauchy location family is (approximately) $1/(1.100035n) \doteq 0.909062/n$. Thus, we see that it may be possible to estimate location in a Cauchy more accurately than in a normal (by about 9% in terms of variance), although we have not yet seen a way to realize this since this is only a lower bound and we have not even shown that an unbiased estimate of location for this family exists. See however Exercise 3.2.12.

As a final point, we note that the ψ function of (3.42) reveals some rather interesting features about the underlying density. This function is called the *location score function.* In Figure #4.1, we have plotted the corresponding location score functions for the densities of parts (a), (b), and (d). Note especially that the $N(0, 1)$, which has "light" tails, has a linear asymptotic behavior for the ψ function, the logisitic (with "exponential" tails) has a nonzero constant asymptotic behavior, and the Cauchy (with "heavy" tails behaving "algebraically" or "like a power") has a ψ function which tends to 0 as the variable → $\pm\infty$.

□

Example 3.2.2 Now we consider simultaneous estimation of location and scale. Suppose f is a given Lebesgue density and consider the location-scale family generated by f, which has Lebesgue density

$$f_{ab}(x) \;=\; \frac{1}{a} f\left(\frac{x-b}{a}\right) \;, \tag{3.54}$$

where $a > 0$ is the scale parameter and b is the location parameter. Assume the following "regularity conditions":

(i) f is positive and continuously differentiable on $I\!\!R$.

(ii) There is an $\epsilon_0 \in (0,1)$ and a function $G_0(x)$ which is Lebesgue integrable such that if $|\beta| < \epsilon_0$ and $|\alpha - 1| < \epsilon_0$, then for all x

$$\left|\; f'\left(\frac{x-\beta}{\alpha}\right)\right| \;\leq\; G_0(x)\;.$$

(iii) The following holds:

$$\int_{-\infty}^{\infty} x^2\psi(x)^2 f(x)\,dx \;<\; \infty\;.$$

Then the Rao-Cramer lower bound holds and the Fisher information matrix $I(a,b)$ is given by

$$I_{11}(a,b) \;=\; \frac{1}{a^2}\int [x\psi(x)-1]^2 f(x)\,dx \quad, \tag{3.55}$$

$$I_{12}(a,b) \;=\; \frac{1}{a^2}\int x\psi(x)^2 f(x)\,dx \quad, \tag{3.56}$$

$$I_{22}(a,b) \;=\; \frac{1}{a^2}\int \psi(x)^2 f(x)\,dx \quad. \tag{3.57}$$

The verification of this is left as Exercise 3.2.5.

One interesting feature of this is that we can assess the increase in the Rao-Cramer lower bound for location estimation from the scale is known case to the scale is unknown case. Assuming a random sample of size n and a scale $a = 1$, if a is known then the lower bound on variance of unbiased estimates of location is

$$\lambda_1 \;=\; \frac{1}{n\int\,\psi(x)^2 f(x)\,dx}\;. \tag{3.58}$$

This follows of course from the previous example. However, when a is unknown then the lower bound changes in general. Our estimand is $g(\theta) = g(a,b) = b$, so $\nabla g = (0,1)$, and so the quadratic form in Corollary 3.2.2 (a) is the $(2,2)$ entry of $I(\theta)^{-1}$. Now the inverse of a 2×2 matrix is easily given by a formula:

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} \;=\; \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad. \tag{3.59}$$

Thus, we obtain for the lower bound on variance of an unbiased estimator of the location parameter based on a sample of size n with unknown scale $a = 1$,

$$\lambda_2 \;=\; \frac{1 \,-\, I_{12}(1,b)^2/[I_{11}(1,b)I_{22}(1,b)]}{nI_{22}(1,b)}$$

$$= \lambda_1 \left[1 \,-\, \frac{\left(\int x\psi(x)^2 f(x)\,dx \right)^2}{\int [x\psi(x)-1]^2 f(x)\,dx \, \int \psi(x)^2 f(x)\,dx} \right] . \tag{3.60}$$

One can show that the r.h.s. $\leq \lambda_1$. So in general, there is (Fisher) information lost about location if we do not know scale. See Exercises 3.2.13, 3.2.15, and 3.2.16 for more on this example.

□

Example 3.2.3 Consider the *Gamma*(α,β) family, as in Example 2.3.2. This is an exponential family, and it will be convenient to use the natural parameters α and $\eta = 1/\beta$. Then the density can be written

$$\exp[\, \alpha \log x \;+\; \eta\,(-x) \;-\; (\,-\alpha \log \eta \;+\; \log\Gamma(\alpha)\,)\,]$$

so the logarithmic normalizing constant for this parameterization is

$$A(\alpha,\eta) \;=\; -\alpha \log \eta \;+\; \log\Gamma(\alpha) \;.$$

By Remark 3.2.2, the Fisher Information matrix is the Hessian of A. It will be convenient to consider some special functions. The *digamma* function is

$$\psi(t) \;=\; \frac{d}{dt} \log\Gamma(t) \;=\; \frac{\Gamma'(t)}{\Gamma(t)} \,, \tag{3.61}$$

and the *trigamma* function is

$$\psi^{(1)}(t) \;=\; \frac{d}{dt}\,\psi(t) \;. \tag{3.62}$$

With this notation, one can write the Hessian as

$$I(\alpha,\eta) \;=\; \left[\begin{array}{cc} \psi^{(1)}(\alpha) & -1/\eta \\ & \\ -1/\eta & \alpha/\eta^2 \end{array} \right] . \tag{3.63}$$

Using equation (3.59) we obtain for the inverse Fisher information matrix,

$$I(\alpha,\eta)^{-1} \;=\; \frac{1}{\alpha\psi^{(1)}(\alpha) - 1} \left[\begin{array}{cc} \alpha & \eta \\ & \\ \eta & \eta^2\psi^{(1)}(\alpha) \end{array} \right] . \tag{3.64}$$

Now suppose we wish to estimate the scale parameter $\beta = 1/\eta = g(\alpha, \eta)$, based on a random sample of size n. Of course

$$\nabla g(\alpha, \eta) \ = \ \begin{bmatrix} 0 \\ \\ -1/\eta^2 \end{bmatrix} \, , \tag{3.65}$$

and hence the lower bound on variance for an unbiased estimator is

$$\begin{aligned} \mathrm{Var}[\hat{\beta}] \ &\geq \ \frac{1}{\eta^4} \frac{\eta^2 \psi^{(1)}(\alpha)}{n[\alpha \psi^{(1)}(\alpha) - 1]} \ = \ \frac{\psi^{(1)}(\alpha)}{n\eta^2 [\alpha \psi^{(1)}(\alpha) - 1]} \\ &= \ \frac{\beta^2 \psi^{(1)}(\alpha)}{n[\alpha \psi^{(1)}(\alpha) - 1]} \, . \end{aligned} \tag{3.66}$$

It is instructive to note that the lower bound on variance of an unbiased estimator of β when the *shape* parameter α is known is

$$\mathrm{Var}[\hat{\beta}] \ \geq \ \frac{1}{\eta^4} \frac{\eta^2}{n\alpha} \ = \ \frac{\beta^2}{n\alpha} \, . \tag{3.67}$$

One can verify that the lower bound in (3.66) is larger than the lower bound in (3.67) (Exercise 3.2.17).

There of course remains the challenge of obtaining the lower bounds in (3.66) and (3.67). Consider the latter one first. When α is known, $\sum X_i$ is complete and sufficient for β, and

$$E_\beta \left[\sum_{i=1}^n X_i \right] \ = \ n\alpha\beta \tag{3.68}$$

so

$$\hat{\beta} \ = \ \frac{1}{n\alpha} \sum_{i=1}^n X_i \tag{3.69}$$

is the UMVUE of β. Also, its variance is

$$\mathrm{Var}[\hat{\beta}] \ = \ \frac{1}{n\alpha^2} \alpha\beta^2 \ = \ \frac{\beta^2}{n\alpha} \, . \tag{3.70}$$

Thus, we see that the lower bound in (3.67) is obtained. This provides an alternative proof that the $\hat{\beta}$ of (3.69) is UMVUE, since it is clearly unbiased and no unbiased estimator can have variance smaller than the Rao-Cramer lower bound, which is the variance of $\hat{\beta}$. Unfortunately, there is not an obvious UMVUE for β when α is unknown.

□

Exercises for Section 4.4.

3.2.1 Suppose the family $\{P_\theta : \theta \in \Theta\} \ll \mu$ σ–finite and satisfies the conditions of Theorem 3.2.1. Let $f_\theta(x)$ denote the densities w.r.t. μ. Suppose ν is another σ–finite measure which is equivalent to μ, and let $g_\theta(x)$ denote the densities w.r.t. ν. Show that if the Fisher information is computed using the densities g_θ the result is the same as using f_θ.

3.2.2 In Remark 3.2.2, verify that condition (vii) from Proposition 3.2.6 holds when using the natural parameterization.

3.2.3 Assume X is a r.v. with density from a location family $f(x-b)$ as in Example 3.2.1. Let $Y = aX$ where $a > 0$ is given, so Y has Lebesgue density $a^{-1}f(a^{-1}y - \tilde{b})$ where $\tilde{b} = a^{-1}b$ is the location parameter for Y. Find the information $\tilde{\iota}$ in Y for $\tilde{b}$ in terms of the information ι in X for b. Use this to verify the results in Examples 3.2.1 (c) and (e).

3.2.4 In Example 3.2.1 (c), we compared the location information for a logistic with variance 1 with a $N(0,1)$, and in Example 3.2.1 (e) we compared the Cauchy with IQR 1.348378 to a $N(0,1)$. Rescale the logistic to have IQR 1.348378, and compare its location information with both $N(0,1)$ and the rescaled Cauchy of Example 3.2.1 (e).

3.2.5 (a) Show that the regularity conditions (i), (ii), and (iii) assumed in Example 3.2.2 imply the regularity conditions (iv), (v) with $\phi(x) \equiv 1$, and (vi) of Theorem 3.2.1.

(b) Verify equations (3.55),(3.56), and (3.57).

3.2.6 Verify equation (3.59).

3.2.7 Verify equations (3.63) through (3.67)

3.2.8 Verify the claims made in the last paragraph of Example 3.2.3.

3.2.9 Find the Fisher informations for each of the following cases. Verify that the regularity conditions hold.

(a) $Poisson(\theta)$, $\theta > 0$.

(b) $B(n, \theta)$, $0 < \theta < 1$.

(c) $NB(m, \theta)$, $0 < \theta < 1$. See Exercise 2.3.9 (d).

3.2.10 For each of the examples of Exercise 3.2.9, give a lower bound on variance of an unbiased estimator for each of the following estimands. Wherever possible, find the UMVUE and compare its variance with the lower bound, or else show that no unbiased estimator of the given estimand exists.

(a) $g(\theta) = \theta$.

(b) $g(\theta) = \theta^2$.

(c) $g(\theta) = e^\theta$.

3.2.11 Show that under futher regularity conditions the Fisher information of a single obseration for location estimation in a pure location family as in Example 3.2.1 is given by

$$\iota \;=\; \int \psi'(X) f(x)\, dx \;.$$

State what extra regularity conditions are required. Apply your result to verify the Fisher informations for location given in Examples 3.2.1 (a), (b), and (d), checking that your regularity conditions hold in each case.

3.2.12 Consider a random sample of size n from a Cauchy distribution, and let $\hat{b}$ be the sample median. Assume for convenience that sample size $n = 2k+1$ is odd. Show that for integers p, $E[|\hat{b}|^p] < \infty$ if and only if $p \leq k$. Conclude that the sample median is unbiased for location in odd sample sizes if $n \geq 3$, and has finite variance if $n \geq 5$.

Hints: The Lebesgue density for $X_{(k)}$ is

$$f_{X_{(k)}}(x) \;=\; n \begin{pmatrix} n-1 \\ k \end{pmatrix} F(x)^k [1 - F(x)]^{n-k-1} f(x) \;.$$

For the c.d.f. of Example 3.2.1 (d), note that

$$1 - F(x) \;=\; \int_x^\infty \frac{1}{\pi(1+u^2)}\, du \;\leq\; \int_x^\infty \frac{1}{\pi u^2}\, du \;=\; \frac{1}{\pi x} \;,$$

for all $x > 0$, and for $u \geq x \geq 1$, we have $1 + u^2 \leq 2u^2$ so

$$1 - F(x) \;\geq\; \int_x^\infty \frac{1}{\pi 2 u^2}\, du \;=\; \frac{1}{2\pi x} \;.$$

3.2.13 (a) Show that $\lambda_2 \leq \lambda_1$ (see (3.58) and (3.60)) by using the fact that the information matrix is positive definite.

(b) Show that same result as in (a) by using the fact that a correlation coefficient is ≤ 1 in magnitude. Hint: Show that $\int x\psi(x) f(x)\, dx = 1$.

3.2.14 Assume X_1, X_2, ..., X_n are i.i.d. with Lebesgue density from the pure scale family

$$f_a(x) \;=\; a^{-1} f(x/a) \;.$$

Assume f is positive on $I\!R$, and let $\psi = -f'/f$ as in Example 3.2.1.

(a) Find a formula for the Fisher information for estimation of a, the scale parameter. Give a lower bound on MSE for any "regular" estimator and on variance for unbiased estimators.

(b) Calculate as explicitly as possible the Fisher information for the scale family in the following cases:

(i) logistic

$$f(x) \;=\; \frac{e^{-x}}{(1+e^{-x})^2} \;.$$

(ii) Normal

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}}e^{-x/2} \;.$$

(c) Similarly to Exercise 3.2.4, after renormalizing so each generating distribution has the same IQR, compare the informations for scale estimation.

3.2.15 In the setup of Exercise 3.2.14 (a), assume a full location–scale family (as in Example 3.2.2) and derive a lower bound on the variance of an unbiased scale estimate. Show that the lower bound is larger than the one in 3.2.14 (a) when location was known.

3.2.16 (a) Assume that the generating density f in Example 3.2.2 is symmetric, i.e. $f(-x) = f(x)$. Show that the Rao-Cramer lower bound for unbiased location estimation is the same irrespective of whether or not scale is known.

(b) Do the analogue of (a) for scale estimation.

(c) What does this say about the three generating densities of Example 3.2.1: normal, logistic, and Cauchy?

3.2.17 (a) Suppose $\theta = (\theta_1, \theta_2)$ and $I(\theta_1, \theta_2)$ is the information matrix for both parameters. Let $J(\theta_1, \theta_2) = I(\theta_1, \theta_2)^{-1}$. Show that

$$\frac{1}{I_{11}(\theta_1, \theta_2)} \;\leq\; J_{11}(\theta_1, \theta_2) \;.$$

(b) State the meaning of this concerning lower bounds on the variance of unbiased estimators of θ_1 when θ_2 is unknown vs. when θ_2 is known.

3.3 Unbiased Estimation.

When a complete and sufficient statistic exists, it is easy to choose which among the unbiased estimators is "best" in a very strong sense, as we shall see next.

3.3.1 Lehmann-Scheffe Theorem.

The next result provides a way for finding the UMVUE in the class of unbiased estimators. It also tells us that in many settings, the UMVUE is (essentially) unique and simultaneously uniformly minimizes the risk for a wide class of loss functions.

Theorem 3.3.1 (Lehmann-Scheffe Theorem.) *Suppose T is a complete and sufficient statistic for θ and $\delta_0(X)$ is an unbiased estimator of $g(\theta)$. Put*

$$\delta(T) \;=\; E[\delta_0(X)|T] \quad .$$

The following hold:

(i) *$\delta(T)$ is the essentially unique function of T which is unbiased for $g(\theta)$.*

(ii) *$\delta(T)$ uniformly minimizes the risk under any convex loss function. In particular, it is a UMVUE.*

(iii) *For a strictly convex loss function, if there is an unbiased estimator $\delta^\star(X)$ with finite risk at any value of θ, say $R(\theta_0, \delta^\star) < \infty$, then $R(\theta_0, \delta(T) < R(\theta_0, \delta^\star)$ unless $\delta^\star(X) = \delta(T)$ P_{θ_0}–a.s.*

(iv) *In particular, if $Var_\theta[\delta_0(X)] < \infty$ for all $\theta \in \Theta$, then $\delta(T)$ is the essentially unique UMVUE.*

Proof. First of all, $\delta(T)$ is unbiased since

$$\begin{aligned}
E_\theta[\delta(T)] &= E_\theta[E_\theta[\delta_0(X)|T]] && \text{(by definition of } \delta(T))\\
&= E_\theta[E[\delta_0(X)|T]] && \text{(by sufficiency of } T)\\
&= E_\theta[\delta_0(X)]) && \text{(by the Law of Total Expectation (Theorem 1.5.5(d))}\\
&= g(\theta) && \text{(by unbiasedness off } \delta_0(X).)
\end{aligned}$$

If $\delta_1(T)$ is any other unbiased estimator which is a function of T, then

$$E_\theta[\delta(T) - \delta_1(T)] \;=\; 0 \quad , \quad \text{for all } \theta \; .$$

Hence, by completeness of T,

$$\delta(T) \;=\; \delta_1(T) \quad , \quad \mathbf{P} - a.s. \tag{3.71}$$

This establishes essential uniqueness of $\delta(T)$ among unbiased estimators of $g(\theta)$ which are a function of T.

Now assume the loss is convex and then by the Rao–Blackwell theorem (or just Jensen's inequality), $\delta(T)$ satisfies

$$R(\theta, \delta(T)) \leq R(\theta, \delta_0) \quad , \quad \text{for all } \theta \ . \tag{3.72}$$

(Application of Jensen's inequality requires finite risk, but if $R(\theta, \delta_0) = \infty$, then (3.72) holds trivially.) If $\delta^\star(X)$ is any other unbiased estimator of $g(\theta)$, then the "Rao–Blackwellized" version of $\delta_0^\star(T) = E[\delta^\star(X)|T]$, is as good as $\delta^\star$ as in (3.72), and $\delta_0^\star(T)$ must be essentially equal to $\delta(T)$ by (3.71). (Thus, Rao–Blackwellization of an unbiased estimator with a complete and sufficient statistic always leads to essentially the same result.) This shows that $\delta(T)$ uniformly minimizes risk.

Assuming L is strictly convex as in part (ii) and that some unbiased estimator $\delta_0^\star(X)$ has finite risk at $\theta = \theta_0$, then (3.71) establishes that $\delta(T)$ has finite risk at θ_0 and since strict inequality holds in (3.71) unless $\delta_0(X)$ is already a function of T by the Rao–Blackwell theorem (or Jensen's inequality) P_{θ_0}–a.s., it follows that $\delta(T)$ is the unique uniform minimum risk unbiased estimator. Taking the loss to be squared error loss (which is strictly convex) leads to the final claim in the statement of the theorem.

□

This theorem allows us to assert that any function of a complete and sufficient statistic is automatically the UMVUE of its expectation. If is often easy to come up with such a function by inspection. For instance, if X_1, X_2, ..., X_n are i.i.d. with $N(\mu, \sigma^2)$ density, then the sample mean $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and sample variance $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are jointly complete and sufficient (i.e. $\underline{T} = (\bar{X}, S^2)$ is complete and sufficient for (μ, σ^2)), and since $E_{(\mu,\sigma^2)}[\bar{X}] = \mu$ and $E_{(\mu,\sigma^2)}[S^2] = \sigma^2$, it follows that $\bar{X}$ is the UMVUE for μ (note that $\text{Var}_{(\mu,\sigma^2)}[\bar{X}] < \infty$) and S^2 is the UMVUE for σ^2 (note that its variance, i.e. risk, is also finite). A more nontrivial example is the following.

Example 3.3.1 Let X_1, X_2, ..., X_n be i.i.d. r.v.'s with $Unif[\theta_1, \theta_2]$ where $\theta_1 < \theta_2$. We know $T = (X_{(1)}, X_{(n)})$ is complete and sufficient for $\theta = (\theta_1, \theta_2)$ (Example 4.2.2 and Exercise 4.3.7 (a)). It is reasonable to guess that $X_{(n)}$ might be a good estimator of θ_2, but if we compute its distribution

$$P_\theta[X_{(n)} \leq x_{(n)}] \ = \ \frac{(x_{(n)} - \theta_1)^n}{(\theta_2 - \theta_1)^n}$$

(see Example 4.3.2) and then its expectation we obtain

$$E_\theta[X_{(n)}] \ = \ \int_{\theta_1}^{\theta_2} x_{(n)} \, \frac{n(x_{(n)} - \theta_1)^{n-1}}{(\theta_2 - \theta_1)^n} \, dx_{(n)}$$

$$= \frac{n}{n+1}(\theta_2 - \theta_1) \int_{\theta_1}^{\theta_2} (n+1)(\theta_2 - \theta_1)^{-(n+1)}(x_{(n)} - \theta_1)^n \, dx_{(n)} \;+\; \theta_1$$

$$= \frac{n}{n+1}\theta_2 \;+\; \frac{1}{n+1}\theta_1 \; .$$

Thus, we cannot use $X_{(n)}$ itself to obtain the desired UMVUE. However, by a symmetry argument (look at $-X_1$, ..., $-X_n$), or direct calculation,

$$E_\theta[X_{(1)}] \;=\; \frac{1}{n+1}\theta_2 \;+\; \frac{n}{n+1}\theta_1 \; .$$

Now it is clear that we can find a and b so that the linear combination $aX_{(1)} + bX_{(n)}$ is unbiased for θ_2. To this end we want

$$a\left(\frac{n}{n+1}\theta_1 + \frac{1}{n+1}\theta_2\right) \;+\; b\left(\frac{1}{n+1}\theta_1 + \frac{n}{n+1}\theta_2\right) \;=\; \theta_2$$

for all $\theta_1 < \theta_2$. For this, it suffices that a and b solve the 2×2 linear system

$$\frac{n}{n+1}a \;+\; \frac{1}{n+1}b \;=\; 0$$

$$\frac{1}{n+1}a \;+\; \frac{n}{n+1}b \;=\; 1$$

which gives

$$a \;=\; -\frac{1}{n-1} \quad , \quad b \;=\; \frac{n}{n-1} \; .$$

Thus, the UMVUE for θ_2 is

$$\hat{\theta}_2 \;=\; \frac{nX_{(n)} - X_{(1)}}{n-1} \; .$$

By symmetry, the UMVUE for θ_1 is

$$\hat{\theta}_1 \;=\; \frac{nX_{(1)} - X_{(n)}}{n-1} \; .$$

□

Another method for obtaining a UMVUE is to compute the distribution for the complete and sufficient statistic T and solve the first kind integral equation

$$g(\theta) \;=\; \int \delta(t) d\,\mathrm{Law}_\theta[T](t) \; . \tag{3.73}$$

This method has already been applied in Example 3.1.3.

Example 3.3.2 Suppose X_1, X_2, ..., X_n are i.i.d. with $Unif(0, \theta)$ where $\theta > 0$. Then $T = X_{(n)}$ is complete and sufficient and has Lebesgue density

$$f_\theta(t) \ = \ \frac{nt^{n-1}}{\theta^n} \quad , \quad 0 < t < \theta \ .$$

See Example 4.3.2. Hence, plugging a general estimand $g(\theta)$ into (3.73) we need to find the solution of the integral equation

$$n^{-1}\theta^n g(\theta) \ = \ \int_0^\theta \delta(t) t^{n-1} \, dt \ . \tag{3.74}$$

This is a so-called Volterra integral equation of the first kind, and the solution is easily obtained by the fundemental theorem of calculus

$$\frac{d}{d\theta}\left(n^{-1}\theta^n g(\theta) \right) \theta^{-(n-1)} \ = \ \delta(\theta) \tag{3.75}$$

provided $g(\theta)$ is continuously differentiable, which is a sufficient condition for g to be U–estimable. (Note that if g does not satisfy some smoothness condition, then there will be no solution to (3.74) since the r.h.s. of that equation is obviously a smooth function of θ, and hence g will not be U–estimable since if an unbiased estimator exists there is an unbiased estimator which is a function of the sufficient statistic T.) Hence, the UMVUE for any such g is

$$\begin{aligned} \delta(X_{(n)}) \ &= \ n^{-1} X_{(n)}^{-(n-1)} \left[\frac{d}{d\theta}\left(n^{-1}\theta^n g(\theta) \right) \right]_{\theta = X_{(n)}} \\ &= n^{-1} X_{(n)}^{-(n-1)} \left[n X_{(n)}^{(n-1)} g(X_{(n)}) + X_{(n)}^n g'(X_{(n)}) \right] \\ &= \ g(X_{(n)}) \ + \ n^{-1} X_{(n)} g'(X_{(n)}) \ . \end{aligned}$$

For example, the UMVUE of $g(\theta) = \theta$ is

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

and the UMVUE of $\mathrm{Var}_\theta[X_1] = \theta^2/12$ is

$$\hat{\sigma}^2 \ = \ \frac{1}{12}[X_{(n)}^2 + 2n^{-1}X_{(n)}^2] \ = \ \left(\frac{n+2}{12n} \right) X_{(n)}^2 \ .$$

□

Both of the above examples exemplify what Lehmann in his book *Point Estimation* calls "Method 1" for obtaining UMVUE's which involves finding the function of the complete and sufficient statistic which is unbiased for the given estimand. Lehmann's "Method 2" relies on the proof of Theorem 3.3.1: if $\delta_0(X)$ is any unbiased estimator of $g(\theta)$ and T is complete and sufficient, then $\delta(T) = E[\delta_0(X)|T]$ is the UMVUE. Ostensibly, one must compute the conditional distribution Law$[X|T = t]$ to use this method, but one can frequently avoid this by using methods based on ancillarity.

Example 3.3.3 Let X_1, X_2, ..., X_n be i.i.d. $N(\mu_X, 1)$ and Y_1, Y_2, ..., Y_m be i.i.d. $N(\mu_Y, 1)$, and assume the $\underline{X}$ and $\underline{Y}$ samples are independent. The unknown parameter is $\theta = (\mu_X, \mu_Y)$. The joint density (w.r.t. Lebesgue on $I\!R^{(n+m)}$) is

$$f_\theta(\underline{x}, \underline{y}) \;=\; C(\theta) \exp\left[\mu_X \sum_{i=1}^{n} x_i \;+\; \mu_Y \sum_{i=1}^{m} y_i\right] h(\underline{x}, \underline{y})$$

so $T = (\bar{X}, \bar{Y}) = (n^{-1}\sum X_i,\, m^{-1}\sum Y_i)$ is complete and sufficient (the student should check that the family is full rank). Suppose we wish to estimate

$$g(\theta) \;=\; P_\theta[X_1 < Y_1]\;.$$

A trivial unbiased estimator is

$$\delta_0 \;=\; I_{(0,\infty)}(Y_1 - X_1)$$

so the UMVUE is

$$\delta(T) \;=\; E[I_{(0,\infty)}(Y_1 - X_1)|\bar{X}, \bar{Y}] \;=\; P[X_1 < Y_1|\bar{X}, \bar{Y}]\;.$$

Now we claim that $V = V(\underline{X}, \underline{Y}) = (X_1 - \bar{X}, Y_1 - \bar{Y})$ is ancillary. To this end, note that V is invariant of location shifts in the $\underline{X}$ and $\underline{Y}$ separately, i.e.

$$V(\underline{X} - \mu_X\underline{1}, \underline{Y} - \mu_Y\underline{1}) \;=\; V(\underline{X}, \underline{Y})$$

and so

$$\begin{aligned} \text{Law}_{(\mu_X,\mu_Y)}[V(\underline{X}, \underline{Y})] \;&=\; \text{Law}_{(\mu_X,\mu_Y)}[V(\underline{X} - \mu_X\underline{1}, \underline{Y} - \mu_Y\underline{1})] \\ &=\; \text{Law}_{(0,0)}[V(\underline{X}, \underline{Y})] \end{aligned}$$

since

$$\text{Law}_{(\mu_X,\mu_Y)}[\,\underline{X} - \mu_X\underline{1}\,,\, \underline{Y} - \mu_Y\underline{1}\,] \;=\; \text{Law}_{(0,0)}[\underline{X}, \underline{Y}]\;.$$

Thus, V is ancillary, and so by Basu's theorem, V is independent of $T = (\bar{X}, \bar{Y})$ for all $\theta = (\mu_X, \mu_Y)$. Hence,

$$\begin{aligned} \delta(t_1, t_2) \;&=\; P[Y_1 - X_1) > 0|\bar{X} = t_1, \bar{Y} = t_2] \\ &=\; P[(Y_1 - \bar{Y}) - (X_1 - \bar{X}) > -(\bar{Y} - \bar{X})|\bar{X} = t_1, \bar{Y} = t_2] \\ &=\; P[V_2 - V_1 > -(t_2 - t_1)|\bar{X} = t_1, \bar{Y} = t_2] \\ &=\; P[V_2 - V_1 > -(t_2 - t_1)] \end{aligned}$$

where the latter follows by independence of V from T (see Exercise 4.2.2). Thus, we need only derive the distribution of $U = V_2 - V_1 = Y_1 - \bar{Y} - X_1 + \bar{X}$. Since this is a linear combination of jointly normal random variables, it is normal with $E[U] = 0$ and

$$\text{Var}[U] \;=\; \text{Var}\left[\frac{m-1}{m}Y_1 \;-\; \frac{1}{m}\sum_{i=2}^{m} Y_i \;-\; \frac{n-1}{n}X_1 \;+\; \frac{1}{n}\sum_{i=2}^{n} X_i\right]$$

$$= \left(\frac{m-1}{m}\right)^2 + \left(\frac{1}{m}\right)^2 (m-1) + \left(\frac{n-1}{n}\right)^2 + \left(\frac{1}{n}\right)^2 (n-1)$$

$$= 2 - \left(\frac{1}{m} + \frac{1}{n}\right) \quad .$$

Hence, if $n > 1$ and $m > 1$,

$$\begin{aligned} & P[U > -(t_2 - t_1)] \\ &= P\left[\frac{U}{\sqrt{2 - (1/m + 1/n)}} > \frac{-(t_2 - t_1)}{\sqrt{2 - (1/m + 1/n)}}\right] \\ &= \Phi\left(\frac{(t_1 - t_2)}{\sqrt{2 - (1/m + 1/n)}}\right) \end{aligned}$$

where Φ is the $N(0,1)$ c.d.f. Recalling that $T_2 = \bar{Y}$ and $T_1 = \bar{X}$, we see that the UMVUE of $P_{(\mu_X, \mu_Y)}[X_1 < Y_1]$ is given by

$$\delta(\bar{X}, \bar{Y}) = \Phi\left(\frac{\bar{X} - \bar{Y}}{\sqrt{2 - (1/m + 1/n)}}\right) \quad .$$

Compare this with the formula

$$P_{(\mu_X, \mu_Y)}[X_1 < Y_1] = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{2}}\right) \quad ,$$

which follows from the fact that $X_1 - Y_1 \sim N(\mu_X - \mu_Y, 2)$.

□

3.3.2 Nonparametric Models.

In this section, we consider estimation for some "nonparametric" models. Suppose X_1, X_2, ..., X_n are i.i.d. r.v.'s with unknown Lebesgue density f, and we wish to estimate

$$g(f) = \int h(x) f(x)\, dx = E_f[h(X)] \tag{3.76}$$

where X denotes a generic r.v. with density f. We will suppose f is in the family of densities which satisfies the additional restriction

$$E_f[h(X)^2] = \int h(x)^2 f(x)\, dx < \infty \quad , \tag{3.77}$$

so that the mean squared error for estimating g will be finite for some estimator, namely $h(X_1)$, which also happens to be an unbiased estimator. We will consider unbiased estimation of g. Now we know from Example 4.2.1 that the vector of order statistics $T = \mathbf{Sort}(\underline{X})$ is sufficient, and as long as h is bounded on finite intervals, we can apply the argument of Example 4.3.3 to show that T is complete since the densities used in that argument will satisfy (3.77) (see Exercise 4.3.1). Let $\gamma : I\!\!R^n \longrightarrow I\!\!R$ denote the projection map onto the first coordinate, i.e. $\gamma(\underline{x}) = x_1$. Now clearly $h(X_1)$ is an unbiased estimator of $g(f)$. Thus, by the Lehmann-Scheffe and Theorem 2.5.1 the UMVUE of $g(f)$ is given by

$$E[h(X_1)\,|\,T\,] \;=\; \frac{1}{n!}\sum_{\pi\in\mathbf{Perm}} h(\gamma(\tilde{\pi}T)) \;=\; \frac{1}{n}\sum_{i=1}^{n} h(X_i)\;. \tag{3.78}$$

The last equation follows easily as in Remark (c) after Theorem 4.2.1. Note in particular that if $h = I_A$ is an indicator function of a set A, then the empirical probability measure of A, $\hat{P}(A) = n^{-1}\#\{i : X_i \in A\}$ is the UMVUE for $P[X_1 \in A]$.

Suppose for instance we wish to estimate the "population" mean $E_f[X]$ (i.e. $h(x) = x$), then (assuming finite second moments), the sample mean $\bar{X}$ is UMVUE for this nonparametric family. We consider estimation of $\mu_k(f) = \int x^k f(x)\,dx$, the k'th moment. Assume that f has finite $2k$'th moment so that X_1^k is an unbiased estimator with finite variance. As in (3.78), the UMVUE is given by

$$\hat{\mu}_k \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^k\;, \tag{3.79}$$

which is the sample k'th moment.

Now one might guess that the UMVUE of the "population" variance σ^2 (assuming fourth moments) is

$$\tilde{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\;, \tag{3.80}$$

but this estimator is not unbiased for σ^2. It easy to see that

$$E[\tilde{\sigma}^2] \;=\; \frac{n}{n-1}\sigma^2\;, \tag{3.81}$$

provided that $n \geq 2$, (See Exercise 3.3.14), and also that

$$\tilde{\sigma}^2 \;=\; \hat{\mu}_2 \,-\, \hat{\mu}_1^2\;, \tag{3.82}$$

so $\tilde{\sigma}^2$ is a function of $\mathbf{Sort}(\underline{X})$. Hence,

$$\hat{\sigma}^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.83}$$

is the UMVUE of σ^2, provided that $n \geq 2$.

In general, when estimating central moments $E[(X - E[X])^k]$ or cumulants (see Section 2.2), one can expand using elementary algebra into a linear combination of powers of moments. For instance in (3.82), we see a linear combination of the second moment and the square of the first moment. Thus, one needs to estimate powers of moments such as $\mu_k^p = \mu_{k,p}$. Assuming $n \geq p$, then an easy unbiased estimator of $\mu_{k,p}$ is obtained from

$$E_f[\prod_{i=1}^{p} X_i^k] \;=\; \prod_{i=1}^{p} E[X_i^k] \;=\; \mu_{k,p}\ .$$

Now let $\gamma : I\!\!R^n \longrightarrow I\!\!R^p$ denote the projection onto the first p coordinates, i.e. $\gamma(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_p)$, and let $h : I\!\!R^p \longrightarrow I\!\!R$ be given by $h(x_1, x_2, \ldots, x_p) = x_1^k x_2^k \ldots x_p^k$. As in (3.78), the UMVUE is given by

$$\hat{\mu}_{k,p} \;=\; E[X_1^k X_2^k \ldots X_p^k \,|\, \mathbf{Sort}(\underline{X})] \;=\; \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} h(\gamma(\tilde{\pi} T)) \qquad (3.84)$$

$$= \frac{1}{n(n-1)\ldots(n-p+1)} \quad \begin{array}{l} \sum_{i_1} \sum_{i_2} \ldots \sum_{i_p} \\ \text{all } i_j\text{'s distinct} \end{array} \quad X_{i_1}^k X_{i_2}^k \ldots X_{i_p}^k$$

Note that the constant in front of the summations is 1 over the number of summands, since there are n ways of choosing i_1, and for each of these there are $n-1$ ways of choosing i_2 distinct from i_1, and so forth until given i_1, i_2, …, i_{p-1}, there are $n-p+1$ ways of choosing i_p distinct from i_1, i_2, …, i_{p-1}. Applying (3.84) to find the UMVUE of μ_1^2, we see that it is given by

$$\begin{aligned} \hat{\mu}_{1,2} \;&=\; \frac{1}{n(n-1)} \sum_{i_1=1}^{n} \sum_{i_2=1,\, i_2 \neq i_1}^{n} X_{i_1} X_{i_2} \qquad (3.85) \\ &=\; \frac{1}{n(n-1)} \left[\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} X_{i_1} X_{i_2} \;-\; \sum_{i_1=1}^{n} X_{i_1}^2 \right] \\ &=\; \frac{1}{n(n-1)} \left[\sum_{i_1=1}^{n} X_{i_1} \right] \left[\sum_{i_2=1}^{n} X_{i_2} \right] - \frac{1}{n(n-1)} \sum_{i_1=1}^{n} X_{i_1}^2 \\ &=\; \frac{n}{n-1} \hat{\mu}_1^2 - \frac{1}{n-1} \hat{\mu}_2 \end{aligned}$$

If one uses this with the UMVUE of μ_2 in (3.82), then (3.83) can be obtained with a little algebra (Exercise 3.3.15).

We will say a function $\gamma : I\!\!R^p \longrightarrow I\!\!R$ is *symmetric* iff $\gamma(\tilde{\pi}\underline{z}) = \gamma(\underline{z})$ for all permutations $\tilde{\pi} \in \mathbf{Perm}_p$, i.e. if γ is invariant under reordering the arguments. A *one sample U-statistic* of order p with kernel γ which is a symmetric function of p variables is given by

$$\begin{aligned} U \;&=\; \frac{1}{n(n-1)\ldots(n-p+1)} \quad \begin{array}{l} \sum_{i_1} \sum_{i_2} \ldots \sum_{i_p} \\ \text{all } i_j\text{'s distinct} \end{array} \quad \gamma(X_{i_1}, X_{i_2}, \ldots, X_{i_p}) \\ &=\; \frac{1}{\binom{n}{p}} \sum_{i_1 <} \sum_{i_2 <} \cdots \sum_{< i_p} \gamma(X_{i_1}, X_{i_2}, \ldots, X_{i_p}) \quad . \qquad (3.86) \end{aligned}$$

The last equation follows since we may rearrange the X_{i_j}'s in the first sum so the indices are in increasing order (which doesn't change its values since γ is symmetric), and there are $p!$ summands which all have the same value. From our discussion above, one can see that U is the UMVUE of its expectation, which is

$$E[U] \;=\; E[\gamma(X_1, X_2, \ldots, X_p)]\;. \tag{3.87}$$

We briefly consider a two sample problem. Let X_1, X_2, …, X_n be i.i.d. r.v.'s with unknown Lebesgue density f and let Y_1, Y_2, …, Y_m be i.i.d. r.v.'s with unknown Lebesgue density g. Then $T = (\mathbf{Sort}(\underline{X}),\ \mathbf{Sort}(\underline{Y}))$ is complete and sufficient (Exercise 4.3.4). One can show that (Exercise 3.3.19)

$$\begin{aligned} &E[h(\underline{X},\underline{Y})\,|\,(\mathbf{Sort}(\underline{X}),\mathbf{Sort}(\underline{Y}))\,] \;= \qquad\qquad (3.88)\\ &\frac{1}{n!}\frac{1}{m!}\sum_{\pi\in\mathbf{Perm}_n}\;\sum_{\zeta\in\mathbf{Perm}_m} h(\tilde{\pi}\mathbf{Sort}(\underline{X}),\tilde{\zeta}\mathbf{Sort}(\underline{Y}))\end{aligned}$$

Hence, for instance, the UMVUE of $P[X < Y]$ is

$$\begin{aligned} &P[\,X_1 < Y_1\,|\,(\mathbf{Sort}(\underline{X}),\mathbf{Sort}(\underline{Y}))] \;=\\ &\frac{1}{n!}\frac{1}{m!}\sum_{\pi\in\mathbf{Perm}_n}\;\sum_{\zeta\in\mathbf{Perm}_m} I_{(0,\infty)}(Y_{\zeta^{-1}(1)} - X_{\pi^{-1}(1)}) \qquad\qquad (3.89)\\ &=\;\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{(0,\infty)}(Y_j - X_i) \;=\; \frac{1}{nm}\#\{(i,j) : X_i < Y_j\}\;.\end{aligned}$$

This latter expression is known as the Mann–Whitney statistic, although this is not the usual formula for it. To derive the usual formula, let $\underline{Z} = (\underline{Y},\underline{X})$ be the combined sample (a random $(m+n)$-vector), and let $\underline{R} = \mathrm{Rank}(\underline{Z})$ be the ranks of the combined sample. Let $\underline{Q} = \mathrm{Rank}(\underline{Y})$ be the ranks of the Y's among themselves. Then for $1 \le j \le m$,

$$\#\{\,i\;:\;X_i < Y_j\,\} \;=\; R_j \;-\; Q_j \quad ,$$

since R_j is the number of observations in the combined sample less than or equal to Y_j, and Q_j is the number of Y's less than or equal to Y_j. Since Q_1, Q_2, …, Q_m is just a permutation of $\{1, 2, \ldots, m\}$,

$$\sum_{j=1}^{m} Q_j \;=\; \sum_{j=1}^{m} j \;=\; m(m+1)/2$$

and so the UMVUE of $P[X < Y]$ can be written as

$$\frac{1}{nm}\#\{(i,j) : X_i < Y_j\} \;=\; \frac{1}{nm}\left[\sum_{j=1}^{m} R_j \;-\; \sum_{j=1}^{m} Q_j\right] \;=\; \frac{1}{nm}\sum_{j=1}^{m} R_j \;-\; \frac{(m+1)}{2n}\;. \tag{3.90}$$

This is a more usual form of the Mann–Whitney statistic.

Exercises for Section 4.1

In each exercise below where you derive a UMVUE, examine it to see if it is "sensible". Your answer may depend on sample size n.

3.3.1 (a) Suppose $g_1(\theta)$ and $g_2(\theta)$ are U-estimable. Let a_1 and a_2 be constants. Show that $g(\theta) = a_1 g_1(\theta) + a_2 g_2(\theta)$ is U-estimable.

(b) Assume a complete and sufficient statistic T exists and that $\delta_1(T)$ and $\delta_2(T)$ are the UMVUE's of $g_1(\theta)$ and $g_2(\theta)$, respectively. What is the UMVUE of $g(\theta)$ in part (a)? Prove your answer.

3.3.2 Suppose X_1, X_2, ..., X_n are i.i.d. $Poisson(\mu)$, $\mu \geq 0$.

(a) Show that $T = \sum_{i=1}^n X_i$ is complete and sufficient. What is $\mathrm{Law}_\mu[T]$?

(b) Give a general necessary and sufficient condition for U-estimability of an estimand $g(\mu)$, and give a formula for the UMVUE when it exists.

(c) For each of the following estimands, find UMVUE's, or show that they don't exist.

(i) $g(\mu) = \mu^k$, where k is a positive integer.

(ii) $g(\mu) = P_\mu[X_1 = k]$ where k is a given nonnegative integer.

(iii) $g(\mu) = log(\mu)$.

(iv) $g(\mu) = e^{\mu t}$ where t is a given real number.

(v) $g(\mu) = e^{\mu^2}$.

(vi) $g(\mu) = e^{\mu^{-2}}$.

3.3.3 In the setup of Example 3.3.1, find the distribution of the complete and sufficient statistic and give an integral equation to be solved to find the UMVUE of a U-estimable estimand, similarly to Example 3.3.2.

3.3.4 In the setup of Example 3.3.2, determine whether each of the following estimands are U-estimable, and find UMVUE's for those that are.

(a) $g(\theta) = \theta^p$ where p is a real number, positive or negative. (Warning: This is not U-estimable for some values of p.)

(b) $g(\theta) = e^{a\theta}$, where a is a real number.

(c) $g(\theta) = F_\theta(x)$ where x is a given positive number and F_θ denotes the distribution function.

(d) $g(\theta) = \psi_\theta(u)$ where u is a given real number and ψ_θ denotes the m.g.f.

3.3.5 Let X_1, X_2, ..., X_n be i.i.d. $N(\mu, \sigma^2)$. Determine for which real numbers p, positive or negative, σ^p is U-estimable and find the corresponding UMVUE.

3.3.6 Let X_1, X_2, ..., X_n be i.i.d. $N(\mu, \sigma^2)$. Consider estimators of σ^2 of the form

$$\hat{\sigma}^2(a) \; = \; a \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad ,$$

where $a > 0$. Note that $a = (n-1)^{-1}$ gives the UMVUE. Find all values of a such that the MSE of $\hat{\sigma}^2(a)$ is smaller than the variance of the UMVUE.

3.3.7 Let X_1, X_2, ..., X_n be i.i.d. $N(\mu_X, \sigma_X^2)$ and Y_1, Y_2,, $\ldots$, Y_m be i.i.d. $N(\mu_Y, \sigma_Y^2)$, and assume the $\underline{X}$ and $\underline{Y}$ samples are independent. Also, assume that all of μ_X, μ_Y, σ_X^2, and σ_Y^2 are unknown.

(a) Find the UMVUE of $\mu_X - \mu_Y$.

(b) Give conditions under which σ_X^2 / σ_Y^2 is U-estimable, and find the UMVUE under those conditions.

3.3.8 (Simple Linear Regression.) Suppose Y_i, $1 \leq i \leq n$, are given by

$$Y_i \; = \; a x_i \, + \, b \, + \, \epsilon_i \;\; ,$$

where ϵ_1, ..., ϵ_n are i.i.d. $N(0, \sigma^2)$. Here, the unknown parameters are a, b, and $\sigma^2 > 0$, and x_1, ..., x_n are known constants.

(a) Show that $T = (\sum x_i Y_i, \sum Y_i, \sum Y_i^2)$ is complete and sufficient.

(b) Let $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{Y}_i = Y_i - \bar{Y}$. Show that

$$\hat{a} \; = \; \frac{\sum_{i=1}^{n} \tilde{x}_i \tilde{Y}_i}{\sum_{i=1}^{n} \tilde{x}_i^2}$$

and

$$\hat{b} \; = \; \bar{Y} - \hat{a}\bar{x}$$

are the UMVUE's for a and b, respectively.

(c) Show the the UMVUE of σ^2 is

$$\hat{\sigma}^2 \; = \; \frac{1}{n-2} \sum_{i=1}^{n} \left(Yi - \hat{a} x_i - \hat{b} \right)^2$$

3.3.9 Let X be $B(n, p)$ with $0 < p < 1$ unknown.

(a) Show that an estimand $g(p)$ is U-estimable if and only if if is a polynomial of degree $\leq n$, and find the $UMVUE$. (Hint: To show that such polynomials are U-estimable, it suffices to show p^k is U-estimable for $1 \leq k \leq n$, by Exercise 3.3.1.)

(b) Find the UMVUE's of $E_p[X]$ and $\mathrm{Var}_p[X]$.

3.3.10 Let $\underline{X}$ be $Mult(n, \underline{p})$ where $\underline{p} = (p_1, p_2, ..., p_k)$ (see Example 2.3.3). Find the UMVUE's for $E_{\underline{p}}[X_i]$, $\mathrm{Var}_{\underline{p}}[X_i]$, and $\mathrm{Cov}_{\underline{p}}[X_i, X_j]$.

3.3.11 Suppose X_1 and X_2 are independent Poisson r.v.'s with $E[X_i] = \mu_i \geq 0$ where the μ_i's satisfy the constraint

$$\mu_1 + \mu_2 = 1 \quad .$$

(a) Is $T = (X_1,\ X_2)$ sufficient for the family? … complete? … minimal sufficient?

(b) Consider "linear" estimators (really, affine estimators) of μ_1, i.e. estimators of the form

$$\delta(X_1, X_2) = aX_1 + bX_2 + c$$

for some constants a, b, and c. Find a necessary and sufficient condition for such a linear estimator to be unbiased.

(c) Find a formula for the variance of any unbiased linear estimator from (b). Are any of these unbiased linear estimators UMVUE?

3.3.12 Suppose $\underline{X}_1$, $\underline{X}_2$, ..., $\underline{X}_n$ are i.i.d. random 2-vectors with uniform distribution on the disk of radius $\theta > 0$, i.e. their common density w.r.t. m^2 is

$$f_\theta(\underline{x}) = \frac{1}{2\pi\theta^2} I_{[0,\theta]}(\|\underline{x}\|) \ .$$

(a) Find a complete and sufficient statistic.

(b) Find the UMVUE for θ.

(c) Similarly to Example 3.3.2, find a formula for the UMVUE of a U-estimable estimand, and give a simple sufficient condition for U-estimability.

3.3.13 Let X be a discrete random variable with

$$P[X = n] = \begin{cases} \theta & \text{if } n = -1, \\ (1-\theta)^2\theta^n & \text{if } n \in I\!N \\ 0 & \text{otherwise.} \end{cases}$$

Here, $0 < \theta < 1$.

(a) Show that X is minimal sufficient for θ.

(b) Show that a function of X is unbiased for 0 if and only if it is of the form cX for some constant c. (We say $h(X)$ is unbiased for 0 if and only if $E_\theta[h(X)] = 0$ for all $\theta \in \Theta$.)

(c) Show that X is not complete for θ.

(d) Show that the UMVUE of $(1-\theta)^2$ is

$$\delta(X) = \begin{cases} 1 & \text{if } X = 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

(e) Show that there exists an unbiased estimator of θ, but there is no UMVUE of θ.

In the exercises below, X, X_1, X_2, $\ldots$, X_n are i.i.d. r.v.'s with unknown Lebesgue density f, and Y, Y_1, Y_2, $\ldots$, Y_m are i.i.d. r.v.'s with unknown Lebesgue density g. Assume any moment conditions or minimum sample size requirements you need.

3.3.14 Verify equations (3.81) and (3.82).

3.3.15 Verify that the UMVUE of σ^2 can be given by $\hat{\mu}_2 - \hat{\mu}_{1,2}$ by using (3.82) and Exercise 3.3.1 (b). Then use this to show that (3.83) gives the UMVUE of σ^2 from (3.85).

3.3.16 Find the UMVUE for $E[(X - E[X])^p]$ when $p = 3$ and 4. What minimal sample sizes do you need? Using this, give "natural" estimates of skewness and kurtosis defined as the "normalized" third and fourth cumulants:

$$\text{skewness} \;=\; \frac{\kappa_3}{\sigma^3} \quad , \quad \text{kurtosis} \;=\; \frac{\kappa_4}{\sigma^4} + 3 \; .$$

For your estimator, just use the UMVUE of the numerator and the corresponding power of the UMVUE of σ^2 in the denominator.

3.3.17 Find the UMVUE of σ^4.

3.3.18 (a) Find the UMVUE of $\psi_X(u)$ for fixed u where ψ_X is the m.g.f. What moment conditions are required?

(b) Find the UMVUE of $\psi_{X_1+X_2}(u)$ for fixed u where $\psi_{X_1+X_2}$ is the m.g.f. for $X_1 + X_2$. Hint: it will be a U-statistic of order 2.

(c) Find the UMVUE of $\psi_{(X_1,X_2)}(\underline{u})$ for fixed $\underline{u}$, where $\psi_{(X_1,X_2)}$ is the joint m.g.f. of (X_1, X_2).

3.3.19 Verify (3.88).

3.3.20 In the two sample problem, find UMVUE's for the following estimands. Simplify your answers as much as possible.

(a) $E[X]E[Y]$.

(b) $P[X + Y < r]$ where r is given.

3.4 Equivariant Estimation: General Theory

3.4.1 The Principle of Equivariance.

In this chapter, we consider another reasonable restriction on the class of estimators similar to unbiasedness, and then we seek an optimal estimator in the restricted class. Unlike unbiasedness, the principle of equivariance requires a more elaborate structure involving the model, the estimand, and the loss function. These will be set forth in the Assumptions below. When these assumptions hold, we will be able to find minimum risk equivariant (MRE) estimators fairly easily.

In order to facilitate understanding, we will constantly refer to the following "running example" as each new concept is introduced.

Example 3.4.1 We suppose that $\underline{X}$ is a random n-vector with Lebesgue density

$$f_b(\underline{X}) \;=\; q(\underline{X} - b\underline{1}) \;, \tag{3.91}$$

where q is a fixed, known Lebesgue density and $\underline{1}$ is the n-vector of all components 1. Here, $b \in I\!\!R$ is the unknown location parameter. We will refer to this as the *one sample location model.* Of course, if the observations are i.i.d., then q will be a product of one dimensional densities.

□

Assumption 3.4.1 *We assume* $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ *is a group family generated by a single probability measure and a group of transformations* $\mathbf{T}$*, i.e. there is a p.m.* Q *on the observation space* $(\Xi, \mathcal{G})$*,* $\mathbf{T}$ *is a group of transformations on* $(\Xi, \mathcal{G})$*, and*

$$\mathbf{P} \;=\; Q \circ \mathbf{T}^{-1} \,. \tag{3.92}$$

Thus, each $P_\theta = Q \circ g^{-1}$ *for some* $g \in \mathbf{T}$*. We further assume that we have a parameter* θ *is identifiable.*

□

One may use g as the parameter and $\mathbf{T}$ as the parameter space, but it has already been noted in Example 2.3.7, this parameterization may not be identifiable, i.e. different values of g may lead to the same probability measure. However, there is a "parameterization" mapping $p : \mathbf{T} \to \Theta$ defined by

$$P_p(g) \;=\; Q \circ g^{-1} \,. \tag{3.93}$$

Note that $\mathbf{T}$ gives an identifiable parameterization if and only if p is injective, i.e. one to one. From Assumption 3.4.1, it follows that p is surjective, i.e. p maps $\mathbf{T}$ onto Θ. We will need further measurability properties of p, which makes the following necessary.

Assumption 3.4.2 *(a) We assume that* $(\mathbf{T}, \mathbf{E})$ *is a measurable space which is* compatible *with the group structure on* $\mathbf{T}$*, i.e. for any* $B \in \mathbf{E}$ *and any* $g \in \mathbf{T}$*,* $gB := \{gh : h \in B\}$ *and* $Bg := \{hg : h \in B\}$ *are both in* $\mathbf{E}$*.*

(b) We also assume that $(\Theta, \mathcal{H})$ *is a measurable space.*

(c) Assume that the map p *is measurable* $(\mathbf{T}, \mathbf{E}) \to (\Theta, \mathcal{H})$*, and that for any* $B \in \mathbf{E}$*,* $p(B) \in \mathcal{H}$*, i.e. that forward images of measurable sets are measurable.*

□

Example 3.4.2 (This is a continuation of Example 3.4.1.) For the location model of (3.91), Q is the p.m. with Lebesgue density q, and $\mathbf{T}$ is the set of all transformations of the form

$$g_b(\underline{x}) \;=\; \underline{x} + b\underline{1}\;,$$

where $b \in I\!\!R$ determines the transformation. Such transformations are called location shifts or translations by a multiple of the $\underline{1}$ vector. For this example, we could identify $\mathbf{T}$ with the parameter space $\Theta = I\!\!R$, and then p is the identity. Of course, the "natural" σ-field to use is $\mathcal{B}$, and clearly this is compatible with the group structure by measurability of translations (see Exercise 3.4.1). Since Θ and $\mathbf{T}$ are both $I\!\!R$, and p is the identity, part (c) of Assumption 3.4.2 holds trivially.

□

We next show that $\mathbf{T}$ induces a transformation group on the parameter space Θ.

Theorem 3.4.1 *The relation* $P_\theta \circ g^{-1} = P_{\bar{g}\theta}$ *defines a map* $\bar{g} : \Theta \to \Theta$*, and* $\bar{\mathbf{T}} = \{\bar{g} : g \in \mathbf{T}\}$ *is a group of transformations on* Θ*. Furthermore,*

$$\text{for any } \theta_1, \theta_2 \in \Theta, \text{ there is } \bar{g} \in \bar{\mathbf{T}} \text{ such that } \bar{g}\theta_1 \;=\; \theta_2\;. \tag{3.94}$$

Proof. Given $\theta \in \Theta$, $P_\theta \circ g^{-1} \in \mathbf{P}$, so $P_\theta \circ g^{-1} = P_{\theta'}$ for some $\theta' \in \Theta$. Since we are assuming an identifiable parameterization, θ' is uniquely determined by θ and g, so if we write $\theta' = \bar{g}\theta$, the mapping $\bar{g}$ is well defined. Note that

$$P_\theta[gX \in A] \;=\; P_{\bar{g}\theta}[X \in A] \tag{3.95}$$

for all $A \in \mathcal{G}$. One sees that $\bar{g} : \Theta \to \Theta$.

Note that p has the following property,

$$\bar{g}p(h) \;=\; p(hg)\;. \tag{3.96}$$

To see this,

$$P_{\bar{g}p(h)}[X \in C] \;=\; P_{p(h)}[gX \in C] \;=\; Q \circ h^{-1}[gX \in C]$$

$$= Q[gX \in h^{-1}C] \ = \ Q[hgX \in C] \ = \ Q \circ (hg)^{-1}[X \in C] \ .$$

so $p(hg) = \bar{g}p(h)$, by identifiability.

Next we show that each $\bar{g}$ is a measurable map $(\Theta, \mathcal{H}) \to (\Theta, \mathcal{H})$. For $H \in \mathcal{H}$,

$$(\bar{g})^{-1}(H) \ = \ \{ \, \theta \in \Theta \ : \ \bar{g}\theta \in H \, \} \ .$$

Since p is surjective, as h ranges over all of $\mathbf{T}$, $p(h)$ ranges over all of Θ. Therefore, we may replace θ by $p(h)$, $h \in \mathbf{T}$ in the last display, and we obtian

$$(\bar{g})^{-1}(H) \ = \ \{ \, p(h) \ : \ h \in \mathbf{T} \,\&\, \bar{g}p(h) \in H \, \} \ = \ \{ \, p(h) \ : \ h \in \mathbf{T} \,\&\, p(hg) \in H \, \} \ ,$$

where the last equation follows from (3.96). Now let $f = hg$ in the above, i.e. $h = fg^{-1}$, then

$$(\bar{g})^{-1}(H) \ = \ \{ \, p(fg^{-1}) \ : \ f \in \mathbf{T} \,\&\, p(f) \in H \, \} \ ,$$

since as f ranges over all of $\mathbf{T}$, $h = fg^{-1}$ ranges over all of $\mathbf{T}g^{-1} = \mathbf{T}$ (see Exercise 2.5.1) Thus,

$$(\bar{g})^{-1}(H) \ = \ \{ \, p(fg^{-1}) \ : \ f \in \mathbf{T} \,\&\, f \in p^{-1}(H) \, \} \ = \ p(p^{-1}(H)g^{-1}) \ .$$

Now $p^{-1}(H) \in \mathbf{E}$ by the assumed measurability of p, and $p^{-1}(H)g^{-1} \in \mathbf{E}$ by compatibility of $\mathbf{E}$ with the group structure on $\mathbf{T}$. Finally, $p(p^{-1}(H)g^{-1}) \in \mathcal{H}$ since it is assumed that forward images under p of measurable sets are also measurable. This shows $\bar{g}$ is measurable.

Now we show that $\bar{\mathbf{T}}$ is closed under composition. Let $g_1, g_2 \in \mathbf{T}$, then

$$P_\theta[g_1 g_2 X \in A] \ = \ P_{\bar{g}_1\theta}[g_2 X \in A] \ = \ P_{\bar{g}_2\bar{g}_1\theta}[X \in A]$$

which shows that $\overline{g_1 g_2}$ is in $\bar{\mathbf{T}}$, and in fact

$$\overline{g_1 g_2} \ = \ \bar{g}_2 \bar{g}_1 \ . \tag{3.97}$$

This completes the verification of property (ii) of Definition 3.2.5(a).

Now we show that $\bar{g}$ is one to one and onto (bijective), and that the inverse $(\bar{g})^{-1} = \bar{h}$ for some $h \in \mathbf{T}$, namely $h = g^{-1}$. This will show property (iii) of Definition 3.2.5(a). Suppose $\bar{g}\theta_1 = \bar{g}\theta_2$ for some θ_1 and θ_2, i.e. $P_{\theta_1} \circ g^{-1} = P_{\theta_2} \circ g^{-1}$. This means $P_{\theta_1}[X \in g^{-1}A] = P_{\theta_2}[X \in g^{-1}A]$ for all $A \in \mathcal{G}$. Since g is bijective (and bimeasurable), letting $B = g^{-1}A$, as A ranges over $\mathcal{G}$, B also ranges over all of $\mathcal{G}$. Hence, $P_{\theta_1}[X \in B] = P_{\theta_2}[X \in B]$ for all $B \in \mathcal{G}$, i.e. $P_{\theta_1} = P_{\theta_2}$, and since we are assuming identifiability, $\theta_1 = \theta_2$. This completes the proof that $\bar{g}$ is one to one (injective).

Now we show that $\bar{g}$ is onto, i.e. surjective. By (3.92), given $\theta_1, \theta_2 \in \Theta$ there are a $g_1, g_2 \in \mathbf{T}$ such that $P_{\theta_1} = Q \circ g_1^{-1}$ and $P_{\theta_2} = Q \circ g_2^{-1}$, i.e. $p(g_i) = \theta_i$ for $i = 1, 2$. Now let

$$g \ = \ g_2 \circ g_1^{-1} \quad ,$$

which is in $\mathbf{T}$ by the group properties. We claim that $\bar{g}\theta_1 = \theta_2$, which incidentally establishes (3.94). To see this,

$$P_{\theta_1}[gX \in A] \;=\; P_{\theta_1}[X \in g^{-1}A] \;=\; P_{\theta_1}[X \in (g_2 g_1^{-1})^{-1}A]$$
$$=\; P_{\theta_1}[X \in g_1 g_2^{-1}A] \;=\; (Q \circ g_1^{-1})[X \in g_1 g_2^{-1}A] \;=\; Q[g_1 X \in g_1 g_2^{-1}A]$$
$$=\; Q[X \in g_2^{-1}A] \;=\; P_{\theta_2}[X \in A]\;.$$

Thus, looking back at the defining relation (3.95), we see that $\bar{g}\theta_1 = \theta_2$, as claimed. This completes the proof that $\bar{g} : \Theta \to \Theta$ is surjective, as well as the claim (3.94).

Since $\bar{g}$ is both injective and surjective, it is bijective, and its inverse map exists. Now let $h = g^{-1}$, and note that

$$P_\theta[X \in A] \;=\; P_\theta[hgX \in A] \;=\; P_{\bar{h}\theta}[gX \in A] \;=\; P_{\bar{g}\bar{h}\theta}[X \in A]\;,$$

so $P_\theta = P_{\bar{g}\bar{h}\theta}$ and again using identifiability, we have that $\bar{g}\bar{h}\theta = \theta$, i.e.

$$(\bar{g})^{-1} \;=\; \bar{h} \text{ with } h \;=\; g^{-1}\;. \tag{3.98}$$

This shows $(\bar{g})^{-1} \in \bar{\mathbf{T}}$ and completes the verification of property (iii) of Definition 3.2.5(a). Hence, $\bar{\mathbf{T}}$ is a group of transformations on Θ.

□

Remarks 3.4.1 (a) The property (3.94) is called *transitivity* of the transformation group $\bar{\mathbf{T}}$.

(b) As a consequence of (3.95), we have

$$E_\theta[h(gX)] \;=\; E_{\bar{g}\theta}[h(X)] \tag{3.99}$$

for any function h for which the expectation is defined. See Exercise 3.4.2.

(c) If G_1 and G_2 are groups, then a mapping $\alpha : G_1 \to G_2$ is called a *homomorphism* iff $\alpha(g_1 g_2) = \alpha(g_1)\alpha(g_2)$ and $\alpha(g_1^{-1}) = \alpha(g_1)^{-1}$,for all $g_1, g_2 \in G_1$. α is called an *antihomomorphism* iff $\alpha(g_1 g_2) = \alpha(g_2)\alpha(g_1)$ and $\alpha(g_1^{-1}) = \alpha(g_1)^{-1}$,for all $g_1, g_2 \in G_1$. Note that if G_2 is commutative, then an antihomomorphism is a homomorphism. In the proof of the last proposition, (3.97) and (3.98) show in fact that the mapping $\alpha : \mathbf{T} \to \bar{\mathbf{T}}$ is an antihomomorphism.

□

Example 3.4.3 (This is a continuation of Example 3.4.1.) Let us consider the application of the previous result to the location example. Given a and b in $I\!R$ $= \Theta$ with $g_a(\underline{x}) = \underline{x} + a\underline{1}$,

$$P_b[g_a\underline{X} \in B] \;=\; P_b[\underline{X} \in B - a\underline{1}] \;=\; Q[\underline{X} \in B - (a+b)\underline{1}]$$

so that

$$\bar{g}_a(b) \;=\; b + a \;,$$

i.e. $\bar{g}_a$ is just translation by a on $\Theta = I\!\!R$. Note that each $\bar{g}_a$ is obviously measurable w.r.t. $\mathcal{B}$. This is typical for much of the area: we work hard to prove measurability in a general setting as in Theorem 3.4.1, but in most concrete examples measurability is trivially obvious.

□

Next we consider the properties required of the estimand in order that we can utilize the principle of equivariance. Suppose we wish to estimate $u(\theta)$ where $u : \Theta \to A$, and A is the action space. We will say that the estimand u is *compatible* with $\mathbf{T}$ (or $\bar{\mathbf{T}}$, or simply "with the group structure") if

$$u(\theta_1) \;=\; u(\theta_2) \;\Rightarrow\; u(\bar{g}\theta_1) \;=\; u(\bar{g}\theta_2) \text{ for all } \; g \in \mathbf{T} \;.$$

Assumption 3.4.3 *Assume*
(i) $u(\Theta) = A$, i.e. u is onto or surjective.
(ii) $u : (\Theta, \mathcal{H}) \to (A, \mathbf{D})$ is measurable, and for every measurable $H \in \mathcal{H}$, $u(H)$ is measurable (i.e. $u(H) \in \mathbf{D}$).
(iii) u is compatible with $\mathbf{T}$.

□

One consequence of this last assumption is that we can obtain still another group of transformations. The proof of the following is left as Exercise 3.4.7.

Theorem 3.4.2 *The map $g^\star : A \to A$ given by*

$$g^\star(u(\theta)) \;=\; u(\bar{g}\theta) \textit{ for all } \theta \in \Theta \;,$$

is well defined, and $\mathbf{T}^\star = \{g^\star : g \in \mathbf{T}\}$ is a group of transformations on A. Furthermore,

$$\textit{for any } d_1, d_2 \in A, \textit{ there is } g^\star \in \mathbf{T}^\star \textit{ such that } g^\star d_1 \;=\; d_2 \;. \qquad (3.100)$$

□

Assumption 3.4.4 *Assume $\mathbf{T}^\star$ is commutative.*

□

Lemma 3.4.3 *The map $g^\star$ in (3.100) is unique for each d_1, d_2.*

Proof. Suppose that for some $d \in A$ and $g_1^\star, g_2^\star \in \mathbf{T}^\star$, $g_1^\star d = g_2^\star d$. One must be careful here: this equation is only assumed for a single $d \in A$. We will show $g_1^\star d' = g_2^\star d'$ for all $d' \in A$, and hence that $g_1^\star = g_2^\star$.

By part (i) of Assumption 3.4.3, $d = u(\theta)$ for some θ, and for any $d' \in A$, $d' = u(\theta')$ for some θ'. By (3.94), $\theta' = \bar{g}'\theta$ for some $\bar{g}'$, so $d' = u(\bar{g}'\theta) = g'^\star u(\theta) = g'^\star d$. Hence,

$$g_1^\star d' \;=\; g_1^\star g'^{star} d \;=\; g'^{star} g_1^\star d \;=\; g'^{star} g_2^\star d \;=\; g_2^\star g'\, stard \;=\; g_2^\star d' \quad ,$$

where commutativity was used in the second and fourth equalities.

□

Example 3.4.4 (This is a continuation of Example 3.4.1.) In the location example, the estimand most likely of interest is the identity, i.e. $u(b) = b$. For this estimand, $g^\star = \bar{g}$, so the conclusions of Theorem 3.4.2 holds trivially. Also, $\mathbf{T}^\star$ is a group of shifts or translations on $I\!R$ and is clearly commutative, so the conclusions of Lemma 3.4.3 hold trivially.

For another estimand which satisfies Assumptions 3.4.3, consider the "fractional part",

$$u(b) \;=\; b \;-\; \lfloor b \rfloor \;, \tag{3.101}$$

where $\lfloor b \rfloor$ is the largest integer $\leq b$. See Exercise 3.4.4.

□

Finally, we are ready for the following definitions.

Definition 3.4.1 *(a) A function $v : (\Xi, \mathcal{G}) \to (\Omega, \mathcal{F})$ is* invariant *iff*

$$v(g\underline{x}) \;=\; v(\underline{x}) \textit{ for all } g \in \mathbf{T},\ \underline{x} \in \Xi \;.$$

(b) A function $L : \Theta \times A \to I\!R$ for which $L(\theta, \cdot) : (A, \mathbf{D}) \to (I\!R, \mathcal{B})$ (such as a loss function $L(\theta, d)$) is invariant *iff*

$$L(\bar{g}\theta, g^\star d) \;=\; L(\theta, d) \textit{ for all } g \in \mathbf{T},\ \theta \in \Theta,\ \textit{and } d \in A \;.$$

(c) An estimator $\delta : (\Xi, \mathcal{G}) \to (A, \mathbf{D})$ is called equivariant *iff*

$$\delta(g\underline{x}) \;=\; g^\star \delta(\underline{x}) \textit{ for all } g \in \mathbf{T} \textit{ and } \underline{x} \in \Xi \;.$$

□

Equivariance of an estimator δ is a property of δ as a function and the estimand $u(\theta)$ is irrelevant other than that u determines the range of δ, i.e. the action space.

Example 3.4.5 (This is a continuation of Example 3.4.1.) For instance, with the estimand $u(b) = b$, a Borel function $v : \mathbb{R}^n \to \mathbb{R}^m$ is location invariant iff given any $a \in \mathbb{R}$,

$$v(\underline{x} - a\underline{1}) \; = \; v(\underline{x}) \quad , \quad \text{for all } \underline{x} \in \mathbb{R}^n \; .$$

Note that the function

$$T_1(\underline{x}) \; = \; (x_1 - x_n, x_2 - x_n, ..., x_{n-1} - x_n) \tag{3.102}$$

is location invariant.

We have the following characterization of location invariant functions: *a Borel function $v : \mathbb{R}^n \to \mathbb{R}^m$ is location invariant iff there is a Borel function $w : \mathbb{R}^{n-1} \to \mathbb{R}^m$ such that $v = w \circ T_1$, where T_1 is given in (3.102).* Pictorially, this can be represented as follows:

$$\begin{tikzcd}
& (\mathbb{R}^{n-1}, \mathbf{B}_{n-1}) \arrow[dd, dashed, "w \text{ exists iff } v \text{ is invariant.}"] \\
(\mathbb{R}^n, \mathbf{B}_n) \arrow[ur, "T_1"] \arrow[dr, "v"'] & \\
& (\mathbb{R}^m, \mathcal{B}_m)
\end{tikzcd}$$

(3.103)

To prove this claim, notice that a function of a location invariant function is location invariant, i.e. if $v = w \circ T_1$ then v is location invariant since T_1 is location invariant. Conversely, if v is invariant, let

$$w(t_1, t_2, ..., t_{n-1}) \; = \; v(0, t_1, t_2, ..., t_{n-1}) \; .$$

Then by location invariance of v,

$$v(x_1, x_2, x_3, ..., x_n) \; = \; v(x_1 - x_1, x_2 - x_1, x_3 - x_1, ..., x_n - x_1)$$

$$= \; w(x_2 - x_1, x_3 - x_1, ..., x_n - x_1) \; = \; w \circ T_1(x_1, x_2, x_3, ..., x_n) \; .$$

A loss function is location invariant if

$$L(\theta + a, d + a) \; = \; L(\theta, d)$$

$$\text{for all } a \in I\!\!R,\ \theta \in I\!\!R,\ \text{and } d \in I\!\!R\ .$$

Note that squared error loss,

$$L(\theta, d) \ = \ (\theta - d)^2$$

satisfies this requirement. More generally, any loss of the form

$$L(\theta, d) \ = \ \lambda(\theta - d) \tag{3.104}$$

for some function $\lambda : I\!\!R \rightarrow [0, \infty]$ is location invariant (Exercise 3.4.1). One can also show the converse, i.e. that any location invariant loss function can be put in the form of (3.104) (Exercise 3.4.1).

An estimator $\delta : I\!\!R^n \rightarrow I\!\!R$ is location equivariant iff

$$\delta(\underline{x} + b\underline{1}) \ = \ \delta(\underline{x}) \ + \ b \text{ for all } b \in I\!\!R \text{ and } \underline{x} \in I\!\!R^n\ . \tag{3.105}$$

For instance, the estimator $\bar{X}$ is location equivariant, as is the estimator X_1. More generally, any estimator of the form $\sum a_i X_i + c$ where $\sum a_i = 1$ is location equivariant in this setup (Exercise 3.4.1). One can give a simple characterization of location equivariant estimators: given any location equivariant estimator δ_0, an estimator δ is location equivariant if and only if

$$\delta(\underline{x}) \ = \ \delta_0(\underline{x}) \ - \ v(\underline{x})\ , \tag{3.106}$$

where v is location invariant. If v is location invariant, then δ as given in (3.106) is location equivariant since

$$\delta(\underline{x} + a\underline{1}) \ = \ \delta_0(\underline{x} + a\underline{1}) \ - \ v(\underline{x} + a\underline{1})$$

$$= \ \delta_0(\underline{x}) \ + \ a \ - \ v(\underline{x}) \ = \ \delta(\underline{x}) \ + \ a\ .$$

Conversely, if δ is location equivariant, then let

$$v(\underline{x}) \ = \ \delta_0(\underline{x}) \ - \ \delta(\underline{x})$$

and one can easily check that v is location invariant.

□

Our immediate goals are to give a useful characterizations of invariant and equivariant functions in general, similar to the ones in (3.103), (3.104), and (3.106) for the location model. First, we need some more definitions. Given $x \in \Xi$, the *orbit* of x under the transformation group $\mathbf{T}$ is

$$\mathbf{T}x \ = \ \{\, gx \,:\, g \in \mathbf{T} \,\} \subset \Xi \quad .$$

Note that two orbits either overlap or are disjoint, i.e. either $\mathbf{T}x_1 = \mathbf{T}x_2$ or $\mathbf{T}x_1 \cap \mathbf{T}x_2 = \emptyset$. To see this, suppose $x \in \mathbf{T}x_1 \cap \mathbf{T}x_2$, then $x = g_1 x_1$ and $x =$

g_2x_2, but then $x_1 = g_1^{-1}g_2x_2$, so $gx_1 = gg_1^{-1}g_2x_2$, which shows $\mathbf{T}x_1 \subset \mathbf{T}x_2$. The reverse inclusion holds by symmetry (i.e. interchange x_1 and x_2).

Now denote the collection of all orbits by

$$\mathbf{Orb}(\Xi, \mathbf{T}) \; = \; \{ \, \mathbf{T}x \, : x \in \Xi \, \} \, . \tag{3.107}$$

Note that $\mathbf{Orb}(\Xi, \mathbf{T})$ is a collection of subsets of Ξ. We define a σ-field $\mathbf{J}$ on $\mathbf{Orb}(\Xi, \mathbf{T})$ by

$$J \in \mathbf{J} \text{ iff } \bigcup_{\mathbf{T}x \in J} \mathbf{T}x \; \in \; \mathcal{G} \, . \tag{3.108}$$

That is, a subset J of $\mathbf{Orb}(\Xi, \mathbf{T})$ (which is a collection of subsets of Ξ) is measurable iff the union of the subsets of Ξ in J is a measurable subset of Ξ. It is left to the student to check that $\mathbf{J}$ so defined is indeed a σ-field (Exercise 3.4.5). With these definitions, the following result is easy to show.

Theorem 3.4.4 *Let $T : \Xi \to \mathbf{Orb}(\Xi, \mathbf{T})$ be given by $T(x) = \mathbf{T}x$. Then T is measurable as a map $(\Xi, \mathcal{G}) \to (\mathbf{Orb}(\Xi, \mathbf{T}), \mathbf{J})$, and T is invariant.*

A function $v : (\Xi, \mathcal{G}) \to (\Omega, \mathcal{F})$ is invariant if and only if there is a function $w : \mathbf{Orb}(\Xi, \mathbf{T}) \to (\Omega, \mathcal{F})$ such that $v = w \circ T$.

Proof. Let $J \in \mathbf{J}$. First we claim that $\mathbf{T}x \in J$ iff $x \in \mathbf{T}x_0$ for some $\mathbf{T}x_0 \in J$. Clearly $x \in \mathbf{T}x$ since $x = \mathit{ident} x$ where ident is the identity map in $\mathbf{T}$. Thus, $\mathbf{T}x \in J$ implies $x \in \mathbf{T}x_0$ for some $\mathbf{T}x_0 \in J$. Conversely, if $x \in \mathbf{T}x_0$ then $x = g_0x_0$ for some g_0 so $x_0 = g_0^{-1}x$ and hence $gx_0 = gg_0^{-1}x$ and $\mathbf{T}x_0 = \mathbf{T}x$, and thus if $\mathbf{T}x_0 \in J$ and $x \in \mathbf{T}x_0$, then $\mathbf{T}x \in J$.

Assuming $J \in \mathbf{J}$, then

$$T^{-1}(J) \; = \; \{ \, x \in \Xi \, : \, \mathbf{T}x \in J \, \} \; = \; \bigcup_{\mathbf{T}x \in J} \mathbf{T}x \, ,$$

where the last equality follows from our first claim. This shows the measurability of T.

To show invariance of T, note that $T(gx) = (\mathbf{T}g)x$. But if $y \in (\mathbf{T}g)x$, then $y = hgx$ for some $h \in \mathbf{T}$, and then since $hg \in \mathbf{T}$ (property (ii) of Definition 3.2.5), we have $y \in \mathbf{T}x$, so $\mathbf{T}gx \subset \mathbf{T}x$. Since two elements of $\mathbf{Orb}(\Xi, \mathbf{T})$ are either equal or disjoint, we have $\mathbf{T}gx = \mathbf{T}x$, i.e. $T(gx) = T(x)$, and hence T is invariant.

Now suppose $v : (\Xi, \mathcal{G}) \to (\Omega, \mathcal{F})$ is invariant. Given any $\mathbf{T}x \in \mathbf{Orb}(\Xi, \mathbf{T})$, define $w(\mathbf{T}x) = v(x)$. We need to show that w is well defined, i.e. that if $\mathbf{T}x = \mathbf{T}x'$ then $w(\mathbf{T}x) = w(\mathbf{T}x')$. If $\mathbf{T}x = \mathbf{T}x'$, then $x' \in T(x)$ by an argument given above, and thus $x' = gx$ for some $g \in \mathbf{T}$, so $v(x') = v(gx) = v(x)$. Now we need to show that w is measurable. If $F \in \mathcal{F}$, then

$$w^{-1}(F) \; = \; \{ \, \mathbf{T}x \, : \, w(\mathbf{T}x) \in F \, \} \; = \; \{ \, \mathbf{T}x \, : \, v(x) \in F \, \} \, .$$

Hence,

$$\bigcup_{\mathbf{T}x \in w^{-1}(F)} \mathbf{T}x \; = \; \{ \, x \, : \, v(x) \in F \, \} \; = \; v^{-1}(F)$$

and hence $w^{-1}(F) \in \mathbf{J}$ by definition of $\mathbf{J}$ and measurability of v. Finally, we need to show $v = w \circ T$, but this is easy since $w \circ T(x) = w(\mathbf{T}x) = v(x)$, for all $x \in \Xi$, by definition of w.

Now to prove the converse, assume that $v = w \circ T$ where $w : \mathbf{Orb}(\Xi, \mathbf{T}) \to (\Omega, \mathcal{F})$. Then for any $x \in \Xi$ and $g \in \mathbf{T}$, by invariance of T we have $v(gx) = w(T(gx)) = w((Tx)) = v(x)$ so v is invariant.

□

Remarks 3.4.2 The function T of the last Theorem is called a (the) *maximal invariant*. A function of an invariant function is invariant, and the result above says that every invariant function is a function of the maximal invariant. In general, if $T_1 : (\Xi, \mathcal{G}) \to (\Omega, \mathcal{F})$ is such that there is a bijective bimeasurable map h such that $T_1 = h \circ T$, then T_1 is also a maximal invariant. For instance, referring back to Example 3.4.1, we see that T_1 given in (3.102) is not the same map as T in Theorem 3.4.4, but we proved that the T_1 in (3.102) has the maximal invariance property, so there is a bijective bimeasurable function h such that $T_1 = h \circ T$ (Exercise 3.4.1).

□

Next we indicate the generalization of (3.104). Fix $\theta_0 \in \Theta$, and for each $\theta \in \Theta$, let $g_\theta \in \mathbf{T}$ be such that $\bar{g}_\theta \theta_0 = \theta$. (Such a g_θ exists by (3.94).) Suppose $L : \Theta \times A \to I\!R$ is invariant as in Definition 3.4.1(b). Then there exists $\lambda : (A, \mathbf{D}) \to (I\!R, \mathcal{B})$ such that

$$L(\theta, d) \; = \; \lambda((g_\theta^\star)^{-1} d) \; ,$$

namely,

$$\lambda(d) \; = \; L(\theta_0, d) \; .$$

See Exercise 3.4.6.

Finally, we turn to the generalization of (3.106).

Theorem 3.4.5 *Let δ_0 be a fixed equivariant estimator. Then $\delta : (\Xi, \mathcal{G}) \to (A, \mathbf{D})$ is equivariant if and only if there is a function $\gamma^\star : \mathbf{Orb}(\Xi, \mathbf{T}) \to \mathbf{T}^\star$ such that $\delta(x) = \gamma^\star(T(x))\delta_0(x)$ for all $x \in \Xi$.*

Proof. By Lemma 3.4.3, for each x there is a $\Gamma^\star(x) \in \mathbf{T}^\star$ such that

$$\delta(x) \; = \; \Gamma^\star(x)\delta_0(x) \; .$$

Assume δ is equivariant. We will show that $\Gamma^\star(gx) = \Gamma^\star(x)$ for all g and all x. This will imply that $\Gamma^\star = \gamma^\star \circ T$ for some $\gamma^\star : \mathbf{Orb}(\Xi, \mathbf{T}) \to \mathbf{T}^\star$ by the proof of of Theorem 3.4.4. Note that we are not showing measurability of $\Gamma^\star$ nor of $\gamma^\star$, so the measurability part of the proof of Theorem 3.4.4 doesn't apply.

To show the claim, note that

$$\delta(gx) \;=\; \Gamma^\star(gx)\delta_0(gx)$$

and by equivariance, this is the same as

$$g^\star\delta(x) \;=\; \Gamma^\star(gx)g^\star\delta_0(x) \;=\; g^\star\Gamma^\star(gx)\delta_0(x)$$

where Assumption 3.4.4 was used at the last step. Multiplying both sides of the latter by $(g^\star)^{-1}$ gives

$$\delta(x) \;=\; \Gamma^\star(gx)\delta_0(x) \;.$$

Since this is the same as the defining equation for $\Gamma^\star(x)$, it follows that $\Gamma^\star(gx)$ $= \Gamma^\star(x)$.

Conversely, assume $\delta(x) = \gamma^\star(T(x))\delta_0(x)$ for some $\gamma^\star : \mathbf{Orb}(\Xi, \mathbf{T}) \to \mathbf{T}^\star$. Then using invariance of T and Assumption 3.4.4 again,

$$\begin{aligned}\delta(gx) \;&=\; \gamma^\star(T(gx))\delta_0(gx) \;=\; \gamma^\star(T(x))g^\star\delta_0(x) \\ &=\; g^\star\gamma^\star(T(x))\delta_0(x) \;=\; g^\star\delta_0(x) \;=\; g^\star\delta(x) \quad ,\end{aligned}$$

which shows δ is equivariant.

□

Now we introduce the final important ingredient that makes the principle of equivariance work. Given a decision rule $\delta(\underline{X})$ and loss $L(\theta, d)$, the corresponding risk is

$$R(\theta, \delta) \;=\; E_\theta[L(\theta, \delta(\underline{X})] \;.$$

Now by (3.99), for any $\theta \in \Theta$ and any $g \in \mathbf{T}$,

$$\begin{aligned}R(\bar{g}\theta, \delta) \;&=\; E_{\bar{g}\theta}[L(\bar{g}\theta, \delta(\underline{X})] \;=\; E_\theta[L(\bar{g}\theta, \delta(g\underline{X})] \qquad (3.109) \\ &=\; E_\theta[L(\bar{g}\theta, g^\star\delta(\underline{X})] \;=\; E_\theta[L(\theta, \delta(\underline{X})] \;=\; R(\theta, \delta) \;,\end{aligned}$$

where in the second to last equation, we assume δ is equivariant, and in the last equation that L is invariant. Now, given $\theta, \theta' \in \Theta$, by (3.94) there is a g such that $\bar{g}\theta = \theta'$, and then

$$R(\theta', \delta) \;=\; R(\bar{g}\theta, \delta) \;=\; R(\theta, \delta) \;,$$

where the last equation follows from (3.109). This proves the next result.

Theorem 3.4.6 *Under an invariant loss, the risk of any equivariant estimator is constant.*

□

Example 3.4.6 (This is a continuation of Example 3.4.1.) Consider equivariant estimators of location under a loss of the form (3.104). If δ is location equivariant, we have

$$R(b,\delta) \;=\; \int \lambda(b-\delta(\underline{x}))q(\underline{x}-b\underline{1})\,d\underline{x} \quad .$$

Now make the change of variables $\underline{y} = \underline{x} - b\underline{1}$, and then the latter is

$$= \int \lambda(b-\delta(\underline{y}+b\underline{1})q(\underline{y})\,d\underline{y} \;=\; \int \lambda(b-[\delta(\underline{y})+b])q(\underline{y})\,d\underline{y}$$

where the last equation follows by location equivariance of δ. Cancelling the b's inside λ we have

$$R(b,\delta) \;=\; \int \lambda(-\delta(\underline{y}))q(\underline{y})\,d\underline{y} \tag{3.110}$$

which is a constant independent of b.

□

Recall that in Section 1 of Chapter 3, we introduced methods for comparing estimators (or more general decision rules) which were based on comparing their corresponding risk functions. One of the problems we encountered was that the risk is a function of θ, so one estimator can be better than a second estimator at some values of θ but worse at others. However, for the setup of Theorem 3.4.2, if an equivariant estimator has smaller risk than a second equivariant estimator at one value of θ, then it has smaller risk at all values. Thus, if we constrain ourselves to equivariant estimators, then we see that the problem of finding a minimum risk estimator is considerably simplified since the risk doesn't change from one parameter value to another.

Now we are ready to show how "easy" it is to obtain MRE estimators. To minimize $R(\theta,\delta)$ over equivariant estimators δ, we need only minimize $R(\theta_0,\delta)$ by Theorem 3.4.6, and using Theorem 3.4.5,

$$R(\theta_0,\delta) \;=\; E_{\theta_0}[L(\theta_0,\delta(X))] \tag{3.111}$$

$$= E_{\theta_0}[L(\theta_0,\gamma^\star(T(X))\delta_0(X))]$$

$$= E_{\theta_0}\,\{\,E_{\theta_0}\,[\,L(\theta_0,\gamma^\star(t)\delta_0(X))\,|\,T(X)=t\,]\,\}\ .$$

Now suppose for fixed $t \in \mathbf{Orb}(\Xi,\mathbf{T})$ we can find $\gamma^\star(t) \in \mathbf{T}^\star$ to minimize over $g^\star$ the function

$$\rho(t,g^\star) \;:=\; E_{\theta_0}\,[\,L(\theta_0,g^\star\delta_0(X))\,|T(X)=t\,]\ ,$$

which is only a function of t and $g^\star$ since θ_0 and δ_0 are fixed, and X is "averaged" out w.r.t. the distribution $\mathrm{Law}_{\theta_0}[X|T(X)=t]$. Provided $\delta(x)=\gamma^\star(T(x))\delta_0(x)$ is measurable, then δ will minimize risk among equivariant estimators by (3.111).

Example 3.4.7 (This is a continuation of Example 3.4.1.) Let δ_0 be a fixed location equivariant estimator. We will work through the details of the previous paragraph. Using (3.110) with $\lambda(-x) = x^2$, we have for any equivariant estimator δ,

$$R(b, \delta) \;=\; \int (\delta(\underline{x}))^2 q(\underline{x})\, d\underline{x} \;=\; \int (\delta_0(\underline{x}) - v(\underline{x}))^2 q(\underline{x})\, d\underline{x}$$

where $v(\underline{x})$ is invariant, and by (3.103),

$$\begin{aligned} R(b, \delta) \;&=\; \int (\delta_0(\underline{x}) - w(x_2 - x_1, x_3 - x_1, ..., x_n - x_1))^2 q(\underline{x})\, d\underline{x} \\ &=\; E_0[(\delta_0(\underline{X}) - w(X_2 - X_1, X_3 - X_1, ..., X_n - X_1))^2] \\ &=\; E_0 \left\{ E_0 \left[(\delta_0(\underline{X}) - w(\underline{t}))^2 \,|(X_2 - X_1, X_3 - X_1, ..., X_n - X_1) = \underline{t} \right] \right\} \; . \end{aligned}$$

Note that $Q = P_0$. To minimize

$$\rho(w) \;:=\; E_0 \left[(\delta_0(\underline{X}) - w)^2 \,|\, (X_2 - X_1, X_3 - X_1, ..., X_n - X_1) = \underline{t} \right] \quad ,$$

we take

$$w \;=\; E_0 \left[\delta_0(\underline{X}) \,|\, (X_2 - X_1, X_3 - X_1, ..., X_n - X_1) = \underline{t} \right] \quad .$$

Now

$$\delta^\star(\underline{X}) \;=\; \delta_0(\underline{X}) \;-\; E_0 \left[\delta_0(\underline{X}) \,|\, (X_2 - X_1, X_3 - X_1, ..., X_n - X_1) \right] \quad , \qquad (3.112)$$

is clearly measurable, so $\delta^\star$ is the minimum mean squared error location equivariant for location estimation in the location model.

Now we give a more concrete representation of the estimator in (3.112). Take $\delta_0(\underline{X}) = X_n$. The joint density under Q of $\underline{T} = T_1(\underline{X})$ and X_n (where T_1 is given by (3.102)) is $q(\underline{t} + x_n \underline{1}, x_n)$. Thus, the conditional density of X_n given $\underline{T} = \underline{t}$ is

$$q(x_n | t) \;=\; \frac{q(\underline{t} + x_n \underline{1}, x_n)}{\int q(\underline{t} + \zeta \underline{1}, \zeta)\, d\zeta} \qquad (3.113)$$

where $\underline{1}$ here is an $n-1$-dimensional vector. Thus, the estimator in (3.112) is

$$X_n \;-\; \frac{\int \zeta q(\underline{T} + \zeta \underline{1}, \zeta)\, d\zeta}{\int q(\underline{T} + \zeta \underline{1}, \zeta)\, d\zeta} \; .$$

In the last expression, make the change of variables $\zeta = X_n - \xi$, and we obtain

$$\delta^\star(\underline{X}) \;=\; \frac{\int \xi q(\underline{X} - \xi \underline{1})\, d\xi}{\int q(\underline{X} - \xi \underline{1})\, d\xi} \; . \qquad (3.114)$$

The estimator above is known as the *Pitman estimator of location* after the famous statistician who first discovered this formula.

We consider some special cases. First, assume the X_i's are i.i.d. $N(\mu, \sigma^2)$ where μ is the unknown location parameter. Then

$$q(\underline{X} - \xi\underline{1}) = (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \xi)^2\right]$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\xi - \bar{X})^2\right\}\right]$$

$$= \left\{(2\pi\sigma^2/n)(2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]\right\} \times$$

$$\left\{(2\pi\sigma^2/n)^{-1} \exp\left[\frac{-n}{2\sigma^2}(\xi - \bar{X})^2\right]\right\} .$$

In the last expression, note that the first factor is independent of ξ and the second factor is the $N(\bar{X}, \sigma^2/n)$ density in ξ. Thus, the first factor is the denominator integral in (3.114) and will cancel from the numerator. Thus,

$$\delta^\star(\underline{X}) = \int \xi\,(2\pi\sigma^2/n)^{-1} \exp\left[\frac{-n}{2\sigma^2}(\xi - \bar{X})^2\right] d\xi = \bar{X} ,$$

i.e. $\bar{X}$ is the minimum mean squared error location equivariant estimator of a normal mean. Note that σ^2 doesn't enter into the formula for $\delta^\star$, so whether it is known or not is immaterial.

Next consider the family of shifted exponential distributions with known scale parameter $\{Exp(a, b) : b \in I\!R\}$, $a > 0$ known (see Example 2.3.5). Then

$$q(\underline{x}) = \frac{1}{a^n} \exp\left[-\frac{1}{a}\sum_{i=1}^{n} x_i\right] \prod_{i=1}^{n} I_{(0,\infty)}(x_i)$$

$$= \frac{1}{a^n} \exp\left[-\frac{1}{a}\sum_{i=1}^{n} x_i\right] I_{(0,\infty)}(x_{(1)}) \quad ,$$

where $x_{(1)}$ is the minimum of the x_i. Thus,

$$q(\underline{X} - \xi\underline{1}) = \frac{1}{a^n} \exp\left[-\frac{1}{a}\sum_{i=1}^{n}(X_i - \xi)\right] I_{(0,\infty)}(X_{(1)} - \xi)$$

$$= \left\{\frac{1}{na^{n-1}} \exp\left[-\frac{1}{a}\sum_{i=1}^{n}(X_i - X_{(1)})\right]\right\} \left\{\frac{n}{a} \exp\left[-\frac{n}{a}(X_{(1)} - \xi)\right] I_{(-\infty, X_{(1)})}(\xi)\right\} .$$

In the last expression, note that the first factor is independent of ξ and the second factor is the density of $-Y$, where Y is an $Exp(a/n, -X_{(1)})$ random variable. Thus,

$$\delta^\star(\underline{X}) = \int_{-\infty}^{X_{(1)}} \xi\,\frac{n}{a} \exp\left[-\frac{n}{a}(X_{(1)} - \xi)\right] dxi$$

$$= X_{(1)} - a/n .$$

Unlike the normal case, here the Pitman estimator depends on the scale parameter. However, similarly to the normal case, the Pitman estimator is also the UMVUE.

3.4.2 Location-Scale Families.

In this subsection, we summarize and review the previous section and apply the methodology developed there to location and scale families. As in the example of i.i.d. $N(\mu, \sigma^2)$ observations, rarely is σ^2 known, so the problems in which both location and scale are simulaneously unknown are more realistic.

Let $\underline{X}$ be a random n-vector with Lebesgue density

$$f_{ab}(\underline{x}) \;=\; a^{-n} q(a^{-1}(\underline{x} - b\underline{1})) \quad ,$$

where q is a given Lebesgue probability density and the parameter is $\theta = (a, b)$ with $a \in (0, \infty)$ and $b \in (-\infty, \infty)$, i.e. $\Theta = (0, \infty) \times (-\infty, \infty)$. This is a group family generated by the transformation group of location-scale transformations

$$\mathbf{T} \;=\; \{\, g_{\alpha,\beta} \,:\, (\alpha, \beta) \in (0, \infty) \times (-\infty, \infty)\,\}$$

where we define

$$g_{\alpha,\beta}(\underline{x}) \;=\; \alpha\underline{x} \,+\, \beta\underline{1} \quad .$$

To finish verification of Assumption 3.4.1, we need the following.

Proposition 3.4.7 *The location-scale family above is identifiable.*

Proof. Assume $P_{\theta_1} = P_{\theta_2}$. Letting Q denote the p.m. with Lebesgue density q, we have

$$\mathrm{Law}_Q[a_1\underline{X} + b_1\underline{1}] \;=\; \mathrm{Law}_Q[a_2\underline{X} + b_2\underline{1}]\,. \tag{3.115}$$

Assuming say $a_2 \leq a_1$ and writing $\underline{Y} = a_2\underline{X} + b_2\underline{1}$, we have

$$\mathrm{Law}_Q[\underline{Y}] \;=\; \mathrm{Law}_Q[c\underline{Y} + d\underline{1}]$$

where

$$c \;=\; \frac{a_1}{a_2} \;\geq\; 1$$

$$d \;=\; b_1 - b_2/a_2 \quad .$$

We consider the two cases $c = 1$ and $c > 1$ separately.

Under $c = 1$, we have

$$Q[\,\underline{Y} \in A\,] \;=\; Q[\,\underline{Y} + d\underline{1} \in A\,] \;=\; Q[\,\underline{Y} \in A - d\underline{1}\,] \quad ,$$

$$\text{for all } A \in \mathcal{B}_n\,.$$

Since $A \in \mathcal{B}_n$ implies $A - d\underline{1} \in \mathcal{B}_n$, we can substitute $A - d\underline{1}$ for A and obtain

$$Q[\,\underline{Y} \in A - d\underline{1}\,] \;=\; Q[\,\underline{Y} \in A - d\underline{1} - d\underline{1}\,] \;=\; Q[\,\underline{Y} \in A - 2d\underline{1}\,] \quad .$$

Combining the two previous displays gives

$$Q[\,\underline{Y} \in A\,] \;=\; Q[\,\underline{Y} \in A - 2d\underline{1}\,] \quad .$$

By an induction argument, one can show

$$Q[\,\underline{Y} \in A\,] \;=\; Q[\,\underline{Y} \in A - md\underline{1}\,] \quad , \tag{3.116}$$

$$\text{for all } A \in \mathcal{B}_n \;\text{ and all } m \in \mathit{naturals} \quad .$$

Now we can find a bounded set, say the n-dimensional cube $[-M, M]^n$ which has sidelength $2M$, such that

$$Q([-M, M]^n) \;\geq\; 3/4\,. \tag{3.117}$$

To see this, let $h_k = I_{[-k,k]^n}$ and note that $0 \leq h_k \uparrow 1$ so by the monotone convergence theorem $Q([-k, k]^n) = \int h_k \, dQ \to \int 1 \, dQ = 1$, so for all k large enough, $Q([-k, k]^n) \geq 3/4$. Now, if $d \neq 0$, then we can take m large enough that

$$m|d| \;>\; 2M$$

which implies that

$$(\,[-M, M]^n\,) \cap (\,[-M, M]^n \;-\; md\underline{1}\,) \;=\; \emptyset\,.$$

To see this, note that $\underline{x} \in [-M, M]^n$ iff $|x_i| \leq M$ for $i = 1, 2, ..., n$, but then $|x_i - md| \geq m|d| - |x_i| > 2M - M > M$, which implies $\underline{x} - md\underline{1} \notin [-M, M]^n$. Thus,

$$\begin{aligned} & Q\;[\,(\,[-M, M]^n\,) \cup (\,[-M, M]^n \;-\; md\underline{1}\,)\,] \\ &= \; Q\;(\,[-M, M]^n\,) \;+\; Q\;(\,[-M, M]^n \;-\; md\underline{1}\,) \\ &= \; Q([-M, M]^n) \;+\; Q([-M, M]^n) \;\geq\; 3/2 \quad , \end{aligned}$$

where (3.116) and (3.117) were used in the last line. But this is a contradiction since Q is a p.m., so our assumption $d \neq 0$ must be wrong, i.e. $d = 0$ and since we are assuming $c = 1$, it follows that $(a_1, b_1) = (a_2, b_2)$, i.e. that $\theta_1 = \theta_2$ in (3.115). (NOTE: This also shows identifiability of the location model in Example 3.4.1 of the previous section.)

Turning to the case $c > 1$, we have

$$Q[\,\underline{Y} \in A\,] \;=\; Q[\,c\underline{Y} + d\underline{1} \in A\,] \;=\; Q[\,\underline{Y} \in c^{-1}(A - d\underline{1})\,] \quad ,$$

$$\text{for all } A \in \mathcal{B}_n \;\text{ and all } m \in \mathit{naturals} \quad ,$$

where $cA = \{c\underline{x} : \underline{x} \in A\}$. Using a similar argument to the one used for (3.116), one can show (Exercise 3.4.12)

$$Q[\,\underline{Y} \in A\,] \;=\; Q[\,\underline{Y} \in c^{-}mA - d\sum_{i=1}^{m-1} c^{-i}\,\underline{1}\,] \tag{3.118}$$

$$= \; Q[\,\underline{Y} \in c^{-}mA - [dc^{-1}(1 - c^{-m})/(1 - c^{-1})]\underline{1}\,] \quad ,$$

$$\text{for all } A \in \mathcal{B}_n \ \text{ and all } m \in naturals \quad .$$

Now take

$$h_m \ = \ I_{A_m} \quad , \quad A_m \ = \ c^{-m}A - [dc^{-1}(1-c^{-m})/(1-c^{-1})]\underline{1}\,] \quad .$$

It is easy to see that

$$A \text{ bounded } \Rightarrow \ h_m(\underline{y}) \to 0 \text{ for all } \underline{y} \neq [d/(1-c^{-1})]\underline{1} \quad , \tag{3.119}$$

(Exercise 3.4.13). Hence, since $Q \ll m^n$, $h_m \to 0$ Q-a.s. Since h_m is bounded in absolute value by 1, which is Q integrable, we have by the dominated convergence theorem that $\int h_m \, dQ \to 0$. But, $\int h_m \, dQ = Q[\underline{Y} \in A_m] = Q[\underline{Y} \in A]$ by (3.118), and thus $Q[\underline{Y} \in A] = 0$ for all bounded Borel sets $A \subset \mathbb{R}^n$, which is a contradiction since Q is a p.m. (see Exercise 3.4.14). Hence, $c > 1$ is not possible. This completes the proof.

□

Continuing on with Assumption 3.4.2 (a), since $\mathbf{T}$ is in one to one correspondence with $(0,\infty) \times \mathbb{R}$ because each $g \in \mathbf{T}$ is associated with (a,b) where a is the scale change and b is the location shift, we have a natural σ-field on $\mathbf{T}$, namely the image of $\mathbf{B}_2$ under the one to one correspondence. More explicitly, given $B \in \mathbf{B}_2$, with $B \subset (0,\infty) \times \mathbb{R}$, let $B' = \{g_{(a,b)} : (a,b) \in B\}$. Then the collection of all such B' is $\mathbf{E}$, the σ-field on $\mathbf{T}$. Furthermore, this is compatible with the group structure since if $B' \in \mathbf{E}$ then $g_{(a,b)}B' = B'g_{(a,b)}$ corresponds with $aB + b\underline{1}$ which is clearly a Borel set contained in $(0,\infty) \times \mathbb{R}$.

Now for Assumption 3.4.2 (b), since $\Theta = (0,\infty) \times \mathbb{R}$, we will use for $\mathcal{H}$ the Borel subsets of $(0,\infty) \times \mathbb{R}$. Since p is the map $p(g_{(a,b)}) = (a,b)$ which was used above to construct $\mathbf{E}$, measurability of p and the fact that forward images under p of measurable sets are measurable follows trivially, i.e. Assumption 3.4.2 (c) holds. The transformation group $\bar{\mathbf{T}}$ on Θ is easily identified, viz. if $\theta = (\alpha, \beta) \in \Theta$ then

$$P_\theta[\, g_{(a,b)}\underline{X} \in A\,] \ = \ Q[\, a(\alpha\underline{X} + \beta) + b \in A\,]$$
$$= \ Q[\,(a\alpha)\underline{X} + (a\beta + b) \in A\,] \ = \ P_{(a\alpha, a\beta+b)}[\,\underline{X} \in A\,] \quad .$$

Thus,

$$\bar{g}_{(a,b)}(\alpha,\beta) \ = \ (a\alpha, a\beta + b) \quad .$$

One can check directly that $\{\bar{g}_{(a,b)} : 0 < a < \infty, \text{ and } -\infty < b < \infty \ \}$ is a transformation group on Θ and satisfies (3.94) without using Theorem 3.4.1 (Exercise 3.4.4).

We are mainly interested in two different estimands, the scale parameter and the location parameter:

$$u_s(\alpha,\beta) \ = \ \alpha \quad , \ A_s \ = \ (0,\infty) \quad ,$$

$$u_l(\alpha, \beta) \;=\; \beta \quad , \; A_l \;=\; (-\infty, \infty) \quad .$$

Another estimand of interest which can be treated is

$$u_{s^m}(\alpha, \beta) \;=\; \alpha^m \quad , \; A_{s^m} \;=\; (0, \infty) \quad .$$

For instance, in the context of the i.i.d. normal model, $m = 2$ corresponds to variance estimation whereas $m = 1$ corresponds to standard deviation estimation. See Example 3.4.8 below. This estimand will be treated in Exercise 3.4.28.

We will see that each of the estimands u_s and u_l requires its own kind of invariant loss function. Clearly part (i) of Assumption 3.4.3 holds for each estimand. Measurability of each estimand is also clear since it is a projection (see Theorem 1.3.5 and exercise 1.3.13). The fact that forward images of measurable sets under the estimands are measurable is not so clear. However, this assumption can be avoided as follows: measurability of forward images is used in only one step of the proof to show that each $g^\star$ is measurable $(A, \mathbf{D}) \to (A, \mathbf{D})$. Thus, if we simply show each $g^\star$ is so measurable after identifying what a $g^\star$ looks like, then it is not necessary to check the second part of Assumption 3.4.3 (ii). Finally, to check Assumption 3.4.3 (iii), suppose we are given (α_1, β_1) and (α_2, β_2) for which

$$u(\alpha_1, \beta_1) \;=\; u(\alpha_2, \beta_2) \tag{3.120}$$

where u is either u_s or u_l. Then, for any $g_{(a,b)}$, we have

$$\begin{aligned} u(\bar{g}_{(a,b)}(\alpha_i, \beta_i)) \;&=\; u(a\alpha_i, a\beta_i + b) \\ &=\; \begin{cases} a\alpha_i & \text{if } u = u_s, \\ & \\ a\beta_i + b & \text{if } u = u_l. \end{cases} \end{aligned}$$

For the $u = u_s$ case, (3.120) is equivalent to $\alpha_1 = \alpha_2$, so we see that $a\alpha_1 = a\alpha_2$, i.e. $u(\bar{g}_{(a,b)}(\alpha_1, \beta_1)) = u(\bar{g}_{(a,b)}(\alpha_2, \beta_2))$. A similar simple argument applies if $u = u_l$.

Now we can identify $\mathbf{T}^\star$, and we also need to check measurability of the elements in it. Using the defining relation in Theorem 3.4.2, we compute

$$u_s(\bar{g}_{(a,b)}(\alpha, \beta)) \;=\; a\alpha$$

so for the scale estimand

$$g^\star_{(a,b)}, s(\alpha) \;=\; a\alpha \quad ,$$

i.e. $\mathbf{T}^\star_s$ is the group of scale transformations on $(0, \infty)$. Clearly the elements of this transformation group are measurable. Similarly, for the location estimand

$$g^\star_{(a,b)}, l(\beta) \;=\; a\beta \,+\, b \quad .$$

Here, $\mathbf{T}^\star_1$ is the group of affine transformations on $I\!R$. Again, measurability is clear. This illustrates that as usual, the measurability assumptions are immediate in practice, and we will not bother to check them in further applications of the principle of equivariance.

Turning to Assumption 3.4.4 , commutativity of $\mathbf{T}_2^\star$ follows immediately from commutativity of multiplication of real numbers. However, $\mathbf{T}_1^\star$ is not commutative (Exercise 3.4.16), and so Assumption 3.4.4 doesn't hold. The only places where Assumption 3.4.4 is used is in the proof of Lemma 3.4.3 (which is only used for the proof of Theorem 3.4.5) and in the proof of Theorem 3.4.5, which gives a characterization of of equivariant estimators. Thus, if we can give another characterization of equivariant estimators for the location estimand in the location scale model which can be used for finding the MRE estimator of location, then we will not need Assumption 3.4.4. In Theorem 3.4.8 below we give an alternative characterization of location equivariance in the setup of a location-scale family.

Now we turn to investigation of invariant loss functions. For the scale estimation problem, $L_s(\theta, \alpha)$ is invariant by definition iff for all $g_{(a,b)} \in \mathbf{T}$, all $\theta \in \Theta$, and all $\alpha \in A_s$,

$$L_s(\bar{g}_{(a,b)}(\theta_1, \theta_2), g^\star_{(a,b)}, s\alpha) \;=\; L_s((a\theta_1, a\theta_2 + b), a\alpha) \;=\; L_s(\theta, \alpha) \quad . \tag{3.121}$$

One can easily see that this is equivalent to L_s being of the form

$$L_s((\theta_1, \theta_2), \alpha) \;=\; \lambda_s(\alpha/\theta_1) \tag{3.122}$$

for some $\lambda_s : (0, \infty) \to (0, \infty)$. See Exercise 3.4.17. In general, we would want the ratio α/θ_1 to be close to 1 in say absolute value, so an example of a reasonable scale invariant loss function is

$$L_s((\theta_1, \theta_2), \alpha) \;=\; |1 - \alpha/\theta_1| \;=\; \left| \frac{\theta_1 - \alpha}{\theta_1} \right|^p \tag{3.123}$$

where $p > 0$. Taking $p = 2$ gives *relative squared error loss* since the quantity inside absolute values in (3.123) is the relative error, i.e. the error divided by the true quantity. (Generally, relative error is only meaningful when we are estimating a positive quantity so there is no chance of dividing by 0.)

For location estimation, $L_l(\theta, \beta)$ is invariant by definition iff for all $g_{(a,b)} \in \mathbf{T}$, all $\theta \in \Theta$, and all $\beta \in A_l$,

$$L_l((a\theta_1, a\theta_2 + b), a\beta + b) \;=\; L_l(\theta, \beta) \quad . \tag{3.124}$$

A necessary and sufficient condition for (3.124) to hold is that L_l have the form

$$L_l((\theta_1, \theta_2), \beta) \;=\; \lambda_l\left(\frac{\theta_2 - \beta}{\theta_1} \right) \quad . \tag{3.125}$$

See Exercise 3.4.17. An example is

$$L_l((\theta_1, \theta_2), \beta) \;=\; \left| \frac{\theta_2 - \beta}{\theta_1} \right|^p \quad .$$

The case $p = 2$ is what we might call *normalized squared error* since the error is normalized by dividing by the scale parameter.

Now Theorem 3.4.6 (constancy of the risk function) holds for both the location and scale estimation problems. Note that this theorem doesn't depend on the characterization of equivariant estimators given in Theorem 3.4.5, which only applies to scale estimation in the current setup. However, the derivation of the Minimum Risk Equivariant (MRE) estimator in (3.111) does depend on the characterization of Theorem 3.4.5, so at this point we will restrict attention to the scale estimation problem, returning to location estimation after we give a characterization of the location equivariant estimators in the setting of location-scale families.

Now to use the derivation in (3.111), we need to (i) find a fixed scale equivariant estimator $\delta_0(\underline{X})$; (ii) find a maximal invariant T; and (iii) for each fixed value t of the maximal invariant T, minimize over $a \in (0, \infty)$

$$\rho(t, a) \;=\; E_{(1,0)}[\,\lambda(a\delta_0(\underline{X}))\,|\,T(\underline{X}) = t\,] \quad . \tag{3.126}$$

The last equation is derived from the display following (3.111) as follows. First we take θ_0 in (3.111) equal to $(1, 0)$. Next, notice that $g^\star_{(a,b)}\delta_0(\underline{X}) = a\delta_0(\underline{X})$, so the value of b is irrelevant and it is only necessary to optimize over a. Finally, we used the characterization of a scale invariant loss in (3.122) in the r.h.s. of (3.126). Once we find $a^\star(t)$ which minimizes the r.h.s of (3.126) (we fix t in (3.126) and find the a which minimizes the r.h.s., so this optimal a depends on t), then the MRE scale estimator for this setting is given by

$$\delta^\star_s(\underline{X}) \;=\; a^\star(T(\underline{X}))\delta_0(\underline{X}) \; . \tag{3.127}$$

For example, consider relative squared error loss, i.e. L_s as given in (3.122) with $\lambda(a) = (1 - a)^2$. Then

$$\rho(t, a) \;=\; E_{(1,0)}[\,(1 - a\delta_0(\underline{X}))^2\,|\,T(\underline{X}) = t\,]$$

$$1 \;-\; 2aE_{(1,0)}[\,\delta_0(\underline{X})\,|\,T(\underline{X}) = t\,] \;+\; a^2E_{(1,0)}[\,\delta_0(\underline{X})^2\,|\,T(\underline{X}) = t\,] \quad .$$

It is easy to see that this is minimized when

$$a^\star \;=\; \frac{E_{(1,0)}[\,\delta_0(\underline{X})\,|\,T(\underline{X}) = t\,]}{E_{(1,0)}[\,\delta_0(\underline{X})^2\,|\,T(\underline{X}) = t\,]}$$

so the scale equivariant estimator which minimizes mean squared relative error is

$$\delta^\star_s(\underline{X}) \;=\; \frac{\delta_0(\underline{X})E_{(1,0)}[\,\delta_0(\underline{X})\,|\,T(\underline{X}) = t\,]}{E_{(1,0)}[\,\delta_0(\underline{X})^2\,|\,T(\underline{X}) = t\,]} \quad . \tag{3.128}$$

Of course, we must find a fixed equivariant estimator δ_0 first (this will be easy), then the maximal invariant T (this will be a little harder), and finally $\mathrm{Law}_{(1,0)}[\,\delta_0(\underline{X})\,|\,T(\underline{X}) = t\,]$ (this will involve some computational difficulty, but is straightforward).

Now we implement the recipe of the previous paragraph. Part (i), finding a fixed scale equivariant estimator is easy. Now for δ to be scale equivariant for the location-scale family, by definition it must satisfy

$$\delta(g_{(a,b)}\underline{X}) \;=\; g^{\star}_{(a,b)}\delta(\underline{X})$$

which is the same as

$$\delta(a\underline{X} + b\underline{1}) \;=\; a\delta(\underline{X}) \quad .$$

Note that when $a = 1$, δ is invariant of location so it must be a function of any maximal invariant of the location transformation group we found in the previous section, say δ is a function of $(X_2 - X_1, X_3 - X_1, ..., X_n - X_1)$. Now $\delta(\underline{X}) = X_2 - X_1$ satisfies the equivariance property in that

$$\delta(a\underline{X} + b\underline{1}) \;=\; a(X_2 - X_1) \;=\; a\delta(\underline{X})$$

but this δ isn't positive (i.e. doesn't take values in A_s, the action space). We may take

$$\delta_0(\underline{X}) \;=\; |X_2 - X_1| \quad , \tag{3.129}$$

and this is scale equivarian and positive provided $X_2 - X_1 \neq 0$. Of course, $X_2 - X_1 = 0$ happens with P_θ-probability 0 for all θ, so we may exclude the set $\{\underline{x} : x_2 - x_1 = 0\}$ from the observation space Ξ. In the above, it was implicitly assumed that

$$n \;\geq\; 2\;.$$

In fact, there are no scale equivariant estimators if $n = 1$ (Exercise 3.4.27(a)), so this trivial case can be discarded. See Exercise 3.4.18 for related examples of scale equivariant estimators.

Next, we need to find a maximal invariant T. As a general rule, if one characterizes the orbit space in an algebraic fashion, there will be an obvious formula for a maximal invariant. To characterize a general orbit $\mathbf{T}\underline{x}$, suppose we have $\underline{y}$ $= g_{(a,b)}\underline{x}$ for some (a, b) with $a > 0$, i.e. $yi = ax_i + b$ for each $1 \leq i \leq n$ for some (a, b) with $a > 0$. Thus, each component y_i of $\underline{y}$ is the same "linear" function of the corresponding component x_i of $\underline{x}$. We can use two pairs (x_i, y_i) to determine the slope a and intercept b of this "linear" function, and then $\underline{y} \in \mathbf{T}\underline{x}$ if and only if both the slope is positive and all other pairs (x_i, y_i) for $i > 2$ are given by the same "linear" function. Using the pairs $(x_1, y1)$ and $(x_2, y2)$, we have

$$a \;=\; \frac{y2 - y1}{x_2 - x_1} \text{ and } b \;=\; y1 \;-\; ax_1 \quad . \tag{3.130}$$

Then $\underline{y} \in \mathbf{T}\underline{x}$ if and only if $a > 0$ and $yi = ax_i + b$ for $i > 2$ where a and b are given in (3.130). This can be simplified algebraically to

$$\frac{y_i - y_1}{x_i - x_1} \;=\; \frac{y_2 - y_1}{x_2 - x_1} \;>\; 0 \quad , \text{ for } 3 \leq i \leq n \quad . \tag{3.131}$$

Because we are assuming $x_2 - x_1 \neq 0$ and $y2 - y1 \neq 0$, the inequality $a = (y_2 - y_1)/(x_2 - x_1) > 0$ is equivalent to

$$sgn(y_2 - y_1) \;=\; sgn(x_2 - x_1) \tag{3.132}$$

where the *signum* function is defined by

$$sgn(x) \;=\; \begin{cases} x/|x| & \text{if } x \neq 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

(It is called the "signum" function instead of the "sign" function so it is not confused with the "sine" function.) By cross multiplying in (3.131), we see that it is equivalent to

$$\frac{y_i - y_1}{|y_2 - y_1|} \;=\; \frac{x_i - x_1}{|x_2 - x_1|} \;>\; 0 \quad , \text{ for } 2 \leq i \leq n \quad . \tag{3.133}$$

We have used one trick here: the case $n = 2$ of (3.133) is equivalent to (3.132).

The form of (3.133) is motivated by the choice of δ_0 in (3.129). For instance, if instead $\delta_0(\underline{X}) = |X_n - X_1|$, it would be more convenient to use this in choosing the denominator of (3.133).

While we "derived" (3.133) as necessary and sufficient condition for $\underline{y} \in \mathbf{T}\underline{x}$, it is a good idea to check that (3.133) is equivalent to $\underline{y} = g_{(a,b)}\underline{x}$ for some (a, b) with $a > 0$. (In general, one can probably guess the form of the maximal invariant, and then simply check its correctness as we are about to do.) Assuming $\underline{y} = a\underline{x} + b\underline{1}$ gives

$$y_i - y_1 \;=\; a(x_i - x_1) \quad , \text{ for } 2 \leq i \leq n \quad . \tag{3.134}$$

Since $a > 0$, taking absolute values of the case $i = 2$ of (3.134) gives

$$|y_2 - y_1| \;=\; a|x_2 - x_1|$$

and because both $x_2 - x_1 \neq 0$ and $y2 - y1 \neq 0$, we may divide the last equation into (3.134) to obtain (3.133). Conversely, assuming (3.133), let

$$a \;=\; \frac{|y_2 - y_1|}{|x_2 - x_1|} \quad .$$

Note that a is defined and positive since both $x_2 - x_1 \neq 0$ and $y_2 - y_1 \neq 0$. Multiplying both sides of (3.133) by $|y2 - y1|$ gives

$$y_i - y_1 \;=\; a(x_i - x_1) \quad , \text{ for } 2 \leq i \leq n \quad ,$$

and taking

$$b \;=\; y1 - ax_1$$

we have

$$y_i \;=\; ax_i \,+\, b \quad , \text{ for } 1 \leq i \leq n \quad ,$$

(note that the case $i = 1$ is trivial), which is what we wanted to show, i.e. that $\underline{y} = g_{(a,b)}\underline{x}$ for some (a, b) with $a > 0$.

Now we seek a simple and obvious maximal invariant (other than the abstract maximal invariant of Theorem 3.4.4). Notice that (3.133) says that $\underline{y} \in \mathbf{T}\underline{x}$ if and only if $T(\underline{y}) = T(\underline{x})$ where $T : \Xi \to \{-1, 1\} \times I\!\!R^{n-2}$ is given by

$$T(\underline{x}) \;=\; \left(sgn(x_2 - x_1)\,,\, \frac{x_3 - x_1}{|x_2 - x_1|}\,,\, \ldots\,,\, \frac{x_n - x_1}{|x_2 - x_1|} \right) \quad , \tag{3.135}$$

It is easy to see that this means T is a maximal invariant (Exercise 3.4.19). Note that T is constant on any orbit, and that T "separates" orbits (i.e. if $\mathbf{T}x_1 \neq \mathbf{T}x_2$, then $T(x_1) \neq T(x_2)$).

Finally, we need to compute the conditional distribution of $\delta_0 = |X_2 - X_1|$ given $\underline{T} = \underline{t}$ under $\theta = (1, 0)$ in order to plug into (3.126) or (3.128). Let

$$\underline{W} \;=\; (|X_2 - X_1|, \underline{T})$$

which is a random vector taking values in $(0, \infty)$ *times* $\{-1, 1\}$ *times* $I\!\!R^{n-2}$. If we can get a joint density for $\text{Law}_{(1,0)}[\underline{W}]$ (w.r.t. m *times* $\#$ m^{n-2}), then it will be easy to get the desired conditional density. For the time being, all calculations are with the parameter value $\theta = (1, 0)$, so we do not mention it further. Letting

$$\underline{U} \;=\; (\, X_2 - X_1\,,\, \frac{X_3 - X_1}{X_2 - X_1}\,,\, \ldots\,,\, \frac{X_n - X_1}{X_2 - X_1}\,)$$

$$=\; W_2\,(\,W_1\,,\, W_3\,,\, W_4\,,\, \ldots,\, W_n\,) \quad ,$$

it is clear that we can obtain the density of $\underline{U}$ (w.r.t. m^{n-2}) by a Jacobian argument, so we need only figure out how to get the density for $\underline{W}$ from that of $\underline{U}$, by somehow accounting for the signs. If $A \subset (0, \infty) \times \{-1, 1\} \times I\!\!R^{n-2}$ is measurable, let

$$A_1 \;=\; \{\underline{w} \in A : w_2 = 1\}$$

and

$$A_2 \;=\; \{\underline{w} \in A : w_2 = -1\} \quad .$$

Then

$$\begin{aligned} P_{(1,0)}[\underline{W} \in A] \;&=\; Q[\underline{W} \in A_1] \,+\, Q[\underline{W} \in A_2] \\ &=\; Q[\,U_1 > 0\,,\ \&\ (U_1, 1, U_2, ..., U_{n-1}) \in A_1\,] \\ &+\, Q[\,U_1 < 0\,,\ \&\ (-U_1, 1, -U_2, ..., -U_{n-1}) \in A_2\,] \quad . \end{aligned} \tag{3.136}$$

Now let

$$B_1 \;=\; \{\,\underline{u} \in I\!\!R^{n-1} \;:\; u_1 > 0\,,\ \&\ (u_1, 1, u_2, ..., u_{n-1}) \in A_1\,]$$

so that the first probability in the last expression in (3.136) equals

$$\int_{B_1} f_{\underline{U}}(\underline{u})\, d\underline{u}$$

$$= \int_{A_1} f_{\underline{U}}(w_2w_1, w_2w_3, w_2w_4, ..., w_2w_n)\, d(m \times \# \times m^{n-2})(\underline{w}) \quad .$$

Similarly for the second term, so

$$f_{\underline{W}}(\underline{w}) \;=\; f_{\underline{U}}(w_2w_1, w_2w_3, w_2w_4, ..., w_2w_n)\ . \tag{3.137}$$

Now to derive $f_{\underline{U}}$, if

$$\underline{V} \;=\; (X_2 - X_1, X_3 - X_1, ..., X_n - X_1)$$

then it's density is

$$f_{\underline{V}}(\underline{v}) \;=\; \int q(\xi, \underline{v} + \xi\underline{1})\, dxi$$

by the derivation of (3.113). Now

$$\underline{V} \;=\; (\, U_1,\, U_1U_2\,,\, U_1U_3\,,\, ...\,, U_1U_{n-1}\,) \quad ,$$

so the Jacobian determinant is

$det\ [\,matrixccol1aboveu_2above.above.above.aboveu_{n-1}ccol0aboveu_1above.above.above.above0ccol.$

and hence

$$f_{\underline{U}}(\underline{u}) \;=\; |u_1|^{n-2}\, f_{\underline{V}}\,(\,u_1\,,\, u_1u_2\,,\, u_1u_3\,,\, ...\,,\, u_1u_{n-1}\,)$$

$$= \int |u_1|^{n-2}\, q(\,\xi\,,\, u_1 + \xi\,,\, u_1u_2 + \xi\,,\, ...\,,\, u_1u_{n-1} + \xi\,)\, d\xi \quad .$$

Plugging this into (3.137) and using that $w_1 = |u_1|$ and $w_2^2 = 1$ gives

$$f_{\underline{W}}(\underline{w}) \;=\; w_1^{n-2} \int q(\,\xi\,,\, w_1w_2 + \xi\,,\, w_1w_3 + \xi\,,\, w_1w_4 + \xi\,,\, ...\,,\, w_1w_n + \xi\,)\, d\xi \quad .$$

Now taking conditional densities (recall the definition of $\underline{W}$),

$$f_{W_1|\underline{T}}(w_1|\underline{t}) \;=\; \frac{w_1^{n-2} \int\, q(\,\xi\,,\, w_1t_1 + \xi\,,\, w_1t_2 + \xi\,,\, ...\,,\, w_1t_{n-1} + \xi\,)\, d\xi}{\int_0^\infty\, \omega^{n-2} \int\, q(\,\xi\,,\, \omega t_1 + \xi\,,\, \omega t_2 + \xi\,,\, ...\,,\, \omega t_{n-1} + \xi\,)\, d\xi\, d\omega}$$

Now we have for any integrable function $h(|X_2 - X_1|, \underline{T})$,

$$E_{(1,0)}[\, h(|X_2 - X_1|, \underline{T})\, |\, \underline{T} = \underline{t}\,] \;= \tag{3.138}$$

$$\frac{\int_0^\infty\, h(\omega, \underline{t})\omega^{n-2} \int\, q(\,\xi\,,\, \omega t_1 + \xi\,,\, \omega t_2 + \xi\,,\, ...\,,\, \omega t_{n-1} + \xi\,)\, d\xi\, d\omega}{\int_0^\infty\, \omega^{n-2} \int\, q(\,\xi\,,\, \omega t_1 + \xi\,,\, \omega t_2 + \xi\,,\, ...\,,\, \omega t_{n-1} + \xi\,)\, d\xi\, d\omega} \quad .$$

To obtain a more aesthetically pleasing formula, make the change of variables

$$\alpha \;=\; \frac{|X_2 - X_1|}{\omega} \quad , \quad \beta \;=\; X_1 - \frac{|X_2 - X_1|}{\omega}\xi \,. \tag{3.139}$$

This gives (Exercise 3.4.20)

$$E_{(1,0)}[\, h(|X_2 - X_1|, \underline{T}) \,|\, \underline{T} = \underline{t}\,] \;= \tag{3.140}$$

$$\frac{\int_0^\infty \; h(|X_2 - X_1|/\alpha, \underline{t})\alpha^{-n} \int \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta\, d\alpha}{\int_0^\infty \; \alpha^{-n} \int \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta\, d\alpha} \quad .$$

Plugging this into (3.128) gives the *minimum mean squared relative error location-scale equivariant estimator of scale* as

$$\hat{a} \;=\; \frac{|X_2 - X_1| \; \int_0^\infty \; (|X_2 - X_1|/\alpha)\alpha^{-n} \; \int_{-\infty}^\infty \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta\, d\alpha}{\int_0^\infty \; (|X_2 - X_1|/\alpha)^2\alpha^{-n} \; \int_{-\infty}^\infty \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta\, d\alpha}$$

$$= \frac{\int_0^\infty \; \alpha^{-(n+1)} \; \int_{-\infty}^\infty \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta\, d\alpha}{\int_0^\infty \; \alpha^{-(n+2)} \; \int_{-\infty}^\infty \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta\, d\alpha} \quad . \tag{3.141}$$

Example 3.4.8 We consider the i.i.d. $N(\mu, \sigma^2)$ family. Then, by some elementary algebra (Exercise 3.4.21),

$$q((\underline{X} - \beta\underline{1})/\alpha) \;=\; n^{-1/2}(2\pi)^{-(n-1)/2}\alpha \;\exp\left[\frac{-1}{2\alpha^2}\sum_{i=1}^n (X_i - \bar{X})^2\right] \times$$

$$\left\{ \frac{1}{\sqrt{2\pi(\alpha^2/n)}} \;\exp\left[\frac{-1}{2(\alpha^2/n)}\,(\beta - \bar{X})^2\right]\right\} \tag{3.142}$$

Hence

$$\int_{-\infty}^\infty \; q((\underline{X} - \beta\underline{1})/\alpha)\; d\beta \;=\; n^{-1/2}(2\pi)^{-(n-1)/2}\alpha \;\exp\left[\frac{-1}{2\alpha^2}\sum_{i=1}^n (X_i - \bar{X})^2\right] \quad .$$

If we write

$$S^2 \;=\; \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

and substitute this into (3.141) there results

$$\hat{\sigma} \;=\; \frac{\int_0^\infty \; \alpha^{-n} \;\exp\left[-\frac{nS^2}{2\alpha^2}\right]\, d\alpha}{\int_0^\infty \; \alpha^{-(n+1)} \;\exp\left[-\frac{nS^2}{2\alpha^2}\right]\, d\alpha} \quad . \tag{3.143}$$

Now make a change of variables

$$v \;=\; \frac{nS^2}{2\alpha^2}$$

and we obtain (Exercise 3.4.22)

$$\hat{\sigma} \;=\; \frac{\Gamma((n-1)/2)}{\Gamma(n/2)} \left(\frac{n}{2}\right)^{1/2} S \quad . \tag{3.144}$$

In Exercise 3.4.28, the student is asked to derive the minimum mean relative squared error location-scale equivariant estimator of variance σ^2.

□

Next, we return to the location estimand. Recall that Theorem 3.4.5 does not apply here so we seek a characterization of location-scale equivariant estimators of location which can be used to derive an optimal such estimator. To this end, let δ_0 be a given location-scale equivariant estimators of location, and suppose δ is any other such estimator. Then $\Delta(\underline{X}) = \delta(\underline{X}) - \delta_0(\underline{X})$ satisfies

$$\Delta(a\underline{X}+b\underline{1}) \;=\; \delta(a\underline{X}+b\underline{1}) \,-\, \delta_0(a\underline{X}+b\underline{1}) \;=\; [\,a\delta(\underline{X})+b\,] - [\,a\delta_0(\underline{X})+b\,]$$
$$=\; a\,[\,\delta(\underline{X}) \,-\, \delta_0(\underline{X})\,] \;=\; a\,\Delta(\underline{X}) \quad .$$

Thus, Δ acts like a location-scale equivariant scale estimator, except of course it need not be positive. However, if δ_1 is a (positive) location-scale equivariant scale estimator, then consider Δ/δ_1. We have

$$\frac{\Delta(a\underline{X}+b\underline{1})}{\delta_1(a\underline{X}+b\underline{1})} \;=\; \frac{a\Delta(\underline{X})}{a\delta_1(\underline{X})} \;=\; \frac{\Delta(\underline{X})}{\delta_1(\underline{X})} \quad .$$

Thus, Δ/δ_1 is location-scale invariant, and so is a function of the maximal invariant. We have just proved one direction of the following.

Theorem 3.4.8 *Let δ_0 be a fixed location-scale equivariant estimator of location, and let δ_1 be a fixed (positive) location-scale equivariant scale estimator. Then $\delta : \Xi \longrightarrow I\!R$ is a location-scale equivariant estimator of location if and only if there is a (Borel) function*

$$v \;:\; \{-1,1\} \times I\!R^{n-2} \;\rightarrow\; I\!R$$

such that

$$\delta(\underline{x}) \;=\; \delta_0(\underline{x}) \,+\, v(T(\underline{x}))\delta_1(\underline{x}) \quad , \quad \textit{for all } \underline{x} \in \Xi \quad , \tag{3.145}$$

where T is given in (3.135).

Proof. The necessary (⇒) direction was shown in the preceding discussion. Assume that δ has the form given in (3.145), and we will prove it is a location-scale equivariant estimator of location. We have

$$\delta(a\underline{x}+b\underline{1}) \;=\; \delta_0(a\underline{x}+b\underline{1}) \,+\, v(T(a\underline{x}+b\underline{1}))\delta_1(a\underline{x}+b\underline{1})$$
$$=\; [a\delta_0(\underline{x})+b] \,+\, v(T(\underline{x}))[a\delta_1(\underline{x})] \;=\; a[\,\delta_0(\underline{x}) \,+\, v(T(\underline{x}))\delta_1(\underline{x})\,] \,+\, b$$
$$=\; a\delta(\underline{x})+b \quad ,$$

which completes the proof.

□

Now we consider whether or not this characterization can help us in the quest for a minimum risk location-scale equivariant estimator of location. The risk of any such estimator is of course a constant given by

$$R(\theta,\delta) \;=\; E_{(1,0)}[L_l((1,0),\delta(\underline{X}))]$$

$$=\; E_{(1,0)}[L_l((1,0),\delta_0(\underline{X})+v(T(\underline{X}))\delta_1(\underline{X})]$$

$$=\; E_{(1,0)}\left\{ E_{(1,0)}\,[\,L_l(\,(1,0)\,,\,\delta_0(\underline{X})\;+\;v(\underline{t})\delta_1(\underline{X})\,|\,T(\underline{X})=\underline{t}\,]\right\} \quad .$$

Thus, we need only minimize over $v \in I\!R$ the function

$$\rho(tu,v) \;=\; E_{(1,0)}\,[\,L_l(\,(1,0)\,,\,\delta_0(\underline{X})\;+\;v\delta_1(\underline{X})\,|\,T(\underline{X})=\underline{t}\,]$$

for $tu \in \{-1,1\} \times I\!R^{n-2}$ fixed, and then as long as the resulting $v^\star(tu)$ is measurable in tu, the sought after optimal estimator is

$$\delta^\star(\underline{X}) \;=\; \delta_0(\underline{X}) \;+\; v^\star(T(\underline{X}))\delta_1(\underline{X}) \quad .$$

In particular, if we use normalized squared error loss, the optimal estimator can be given by the formula

$$\delta^\star(\underline{X}) \;=\; \delta_0(\underline{X}) \;-\; \frac{E_{(1,0)}[\delta_0(\underline{X})\delta_1(\underline{X})|T(\underline{X})]}{E_{(1,0)}[\delta_1(\underline{X})^2|T(\underline{X})]}\,\delta_1(\underline{X}) \quad . \tag{3.146}$$

See Exercise 3.4.23. Note that to carry out this recipe, we will have to find a fixed location-scale equivariant estimator of location δ_0 (we already have δ_1, e.g. as given in (3.129), and of course we already have T), and then we need to find the joint conditional distribution of ($\delta_0(\underline{X})$, $\delta_1(\underline{X})$) given $T(\underline{X}) = tu$.

Finding a fixed location-scale equivariant estimator of location δ_0 is easy, as usual. Consider $\delta_0(\underline{X}) = X_1$. We have

$$\delta_0(a\underline{X}+b\underline{1}) \;=\; \delta_0(\,(\,aX_1+b\,,\,aX_2+b\,,\,\ldots\,,aX_n+b\,)\,)$$

$$=\; aX_1+b \;=\; a\delta_0(\underline{X}) \;+\; b \quad ,$$

which is all we need to verify to show that this δ_0 is a location-scale equivariant estimator of location.

It is not difficult (Exercise 3.4.24) to modify the derivation (3.140) to obtain

$$E_{(1,0)}[\,h(X_1,|X_2-X_1|,\underline{T})\,|\,\underline{T}\,] \;= \tag{3.147}$$

$$\frac{\int_0^\infty \; h((X_1-\beta)/\alpha,|X_2-X_1|/\alpha,\underline{T})\alpha^{-n}\,\int\; q((\underline{X}-\beta\underline{1})/\alpha)\;d\beta\,d\alpha}{\int_0^\infty \;\alpha^{-n}\,\int\; q((\underline{X}-\beta\underline{1})/\alpha)\;d\beta\,d\alpha} \quad .$$

Plugging this into (3.146), we obtain that the *minimum mean normalized squared error location-scale equivariant estimator of location* is given by

$$\hat{b} \;=\; \frac{\int_0^\infty \int_{-\infty}^\infty \;\beta\,\alpha^{-(n+2)}\,q((\underline{X}-\beta\underline{1})/\alpha)\;d\beta\,d\alpha}{\int_0^\infty \int_{-\infty}^\infty \;\alpha^{-(n+2)}\,q((\underline{X}-\beta\underline{1})/\alpha)\;d\beta\,d\alpha} \quad . \tag{3.148}$$

Example 3.4.9 Again we consider the i.i.d. $N(\mu,\sigma^2)$ family. Using (3.142) we see tha that

$$\int_{-\infty}^{\infty} \beta\,q((\underline{X}-\beta\underline{1})/\alpha)\;d\beta \;=\; n^{-1/2}(2\pi)^{-(n-1)/2}\bar{X}\,\alpha\,\exp\left[\frac{-1}{2\alpha^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \quad .$$

Substituting this into (3.148), we see that $\bar{X}$ factors out in the numerator and the same integrals in α appear in the numerator and denominator, so after cancelling them we obtain that $\bar{X}$ is the minimum mean normalized squared error location-scale equivariant estimator of location for this family. One can arrive at this conclusion without using Theorem 3.4.8 and the subsequent derivations; see Exercise 3.4.25.

□

Exercises for Section 5.4.

3.4.1 Consider the location model of Example 3.4.1. Answer the following questions.

(a) What are gB and Bg in Assumption 3.4.2 (a), for a general g and B? Verify that $(\mathbf{T}, \mathbf{E})$ is compatible with the group structure as defined in Assumption 3.4.2 (a).

(b) Show that a necessary and sufficient condition for a loss to be location invariant is that (3.104) hold for some nonnegative extended Borel function λ.

(c) Suppose $\delta(\underline{X}) = \sum_i a_i X_i + c$ where $\sum_i a_i = 1$. Show that δ is location invariant by verifying (3.105). Also, put δ in the form of (3.106) with both $\delta_0(\underline{X})$ $= \bar{X}$ and $\delta_0(\underline{X}) = X_1$.

(d) Identify $\mathbf{T}x$ and $\mathcal{O}rb(\Xi, \mathbf{T})$ as defined in (3.107). Show that T_1 of (3.102) is a bimeasurable transformation of T as given in Theorem 3.4.4.

(e) Give detailed verifications of (3.113) and (3.114).

3.4.2 Verify equation (3.99)

3.4.3 Show that any bijective bimeasurable map of a maximal invariant also is a maximal invariant (i.e. any invariant function is a function of the maximal invariant).

3.4.4 In the framework of the location model, consider the estimand given in (3.101).

(a) What is the group $\mathbf{T}^\star$?

(b) Show that Assumptions 3.4.3 and 3.4.4 hold.

3.4.5 Check that $\mathbf{J}$ in (3.108) is a σ-field.

3.4.6 Suppose $L(\theta, d)$ is an invariant loss function. Fix $\theta_0 \in \Theta$. Let g_θ be such that $\bar{g}_\theta \theta_0 = \theta$, which exists by (3.94). Show that there exists a function $\lambda : (A, \mathbf{D}) \to (I\!R, \mathcal{B})$ such that $L(\theta, d) = \lambda((g_\theta^\star)^{-1} d)$.

3.4.7 Prove Theorem 3.4.2. In the proof, show that the mapping $\alpha : \mathbf{T} \to \mathbf{T}^\star$ given by $\alpha(g) = gstar$ is a homomorphism.

3.4.8 Show that the Pitman estimator is a function of any sufficient statistic.

3.4.9 Compute the Pitman estimators of location for the following families assuming any other parameters are known.

(a) i.i.d. double exponential:

$$f_b(x) \;=\; \frac{1}{2a} \exp[-|x - b|/a] \quad .$$

In this case, your formula will be rather complicated, but describe it as best you can. You may use the order statistics in place of $\underline{X}$ for the evaluation of the formula.

(b) i.i.d. $Unif[b-a, b+a]$:

$$f(x) \ = \ \frac{1}{2a} I_{(b-a,b+a)}(x) \quad .$$

Does the Pitman estimator depend on a?

(c) $N(b\underline{1}, \sigma^2 D)$ where D is a diagonal matrix with positive diagonal entries. Does the Pitman estimator depend on σ^2?

(d) $N(b\underline{1}, \sigma^2 V)$ where V is a positive definite matrix. Does the Pitman estimator depend on σ^2?

(e)

3.4.10 Consider the *two sample location model* where $\underline{X}$ and $\underline{Y}$ are random n and m dimensional vectors, respectively. Let $q(\underline{x}, \underline{y})$ be a Lebesgue probability density on $I\!R^{(n+m)}$, and let

$$f_{ab}(\underline{x}, \underline{y}) \ = \ q(\underline{x} - a\underline{1}, \underline{y} - b\underline{1}) \quad .$$

The parameter is $\theta = (a, b)$, and the parameter space is $I\!R^2$. Let $\mathbf{T}$ be the group of transformations of the form

$$g_{\alpha,\beta}(\underline{x}, \underline{y}) \ = \ (\underline{x} - \alpha\underline{1}, \underline{y} - \beta\underline{1}) \quad .$$

(a) Give "natural" σ-fields and verify Assumptions 3.4.1 and 3.4.2. What is the parameter map p?

(b) Determine $\bar{\mathbf{T}}$.

(c) Let $u(\theta) = a - b$ be the estimand. Determine $\mathbf{T}^\star$ and verify Assumptions 3.4.3 and 3.4.4.

(d) Determine a maximal invariant and give characterizations similar to (3.103), (3.104), and (3.106) of invariant v, invariant L, and equivariant δ, respectively.

(e) Find the analogue of the Pitman estimator in (3.114).

3.4.11 Evaluate the estimator of Exercise 3.4.10 (e) for the following models.

(a) $\underline{X}$ and $\underline{Y}$ independent $N(a\underline{1}, \sigma_X^2 I)$ and $N(b\underline{1}, \sigma_Y^2 I)$, respectively. Show that the estimator depends on σ_X^2 and σ_Y^2 only through the ratio σ_X^2/σ_Y^2.

3.4.12 Verify equation (3.118).

3.4.13 Verify equation (3.119).

3.4.14 Suppose Q is a p.m. on $I\!R^n$. Show that there exists some bounded Borel set A with $Q(A) > 0$.

3.4.15 Verify directly from the definition (without using Theorem 3.4.1) that $\bar{\mathbf{T}} = \{\bar{g}_{(a,b)} : 0 < a < \infty, \text{ and } -\infty < b < \infty\ \}$ is a transformation group on Θ and satisfies (3.94).

3.4.16 Show that $\mathbf{T}_1^\star$ is not commutative.

3.4.17 (a) Show that a scale estimation loss function L_s is invariant (i.e. satisfies (3.121)) if and only if it is of the form (3.122).

(b) Show that a location estimation loss function L_l is invariant (i.e. satisfies (3.124)) if and only if it is of the form (3.125).

3.4.18 Show that any estimator of the form

$$\delta(\underline{X}) \ = \ \left\{ \sum_j a_j \left| \sum_k b_{jk} X_k \right|^p \right\}^1 /p$$

is scale equivariant (for the location-scale family) if the following hold

$$\begin{aligned} p \neq 0 \quad &, \\ \sum_k b_{jk} \ = \ 0 \text{ for all } j \quad &, \\ a_j \ \geq \ 0 \text{ for all } j \quad &, \\ \sum_j a_j \sum_k |b_{jk}| \ \neq \ 0 \quad &. \end{aligned}$$

You will need to exclude a null set from Ξ to guarantee that $\delta(\underline{X}) > 0$.. (Hint: The first, third, and fourth conditions are needed to show that δ is defined and positive. The second condition is needed because a scale equivariant estimator in the location-scale family must be location invariant.)

3.4.19 (a) Assuming $\mathbf{T}$ is a general transformation group on some observation space Ξ, suppose that T is a function on Ξ (to some range space) which (i) is constant on orbits of $\mathbf{T}$, i.e. $x_1 \in \mathbf{T}x$ and $x_2 \in \mathbf{T}x$ implies $T(x_1) = T(x_2)$; and (ii) T separates orbits, i.e. if $x_1 \in \mathbf{T}x$ and $x_2 \notin \mathbf{T}x$ then $T(x_1) \neq T(x_2)$. Then T is a maximal invariant.

(b) Now consider the location-scale group of transformation on $I\!R^n$. Show that T given in (3.135) satisfies the properties (i) and (ii) above.

3.4.20 Verify that the change of variables in (3.139) transforms (3.138) into (3.140).

3.4.21 Verify equation (3.142).

3.4.22 Verify that (3.144) follows from (3.143).

3.4.23 Verify that the minimum mean squared normalized error location-scale equivariant estimator of location is given by (3.146).

3.4.24 Verify (3.147) and (3.148).

3.4.25 Consider the general location-scale family as in the section.

(a) Show that if the scale parameter is fixed at a particular value, then the resulting subfamily is a "pure" location family.

(b) Suppose that for each fixed value of the scale parameter, the minimum mean squared error location equivariant estimator of location is δ which is functionally independent of the (fixed value of) the scale parameter. Show that δ is then the minimum mean normalized squared error location-scale equivariant estimator of location.

(c) Use the result in (b) to give an alternative derivation that $\bar{X}$ is the optimal location estimator for the i.i.d. normal family as in Example Examp5.4.1.4.

3.4.26 (a) Find the minimum mean relative (normalized) squared error location-scale equivariant estimator of scale (location) in the i.i.d. $Unif(b - a/2, b + a/2)$ family.

(b) Same as (a) but the $Exp(a, b)$ family.

3.4.27 Show the following in the context of a location-scale family on $I\!R$, i.e. when $n = 1$.

(a) There exists no scale equivariant estimators.

(b) The only location equivariant estimator is $\delta(x) = x$.

3.4.28 Consider the estimand $u_{s^m}(\alpha, \beta) = \alpha^m$.

(a) Check that Assumption 3.4.3 holds. Determine the corresponding induced group $\mathbf{T}^\star_{s^m}$.

(b) Characterize invariant loss functions similarly to (3.122). Show that

$$L((\theta_1, \theta_2), \alpha) = \left| \frac{\theta_1^m - \alpha^m}{\theta_1^m} \right|^p$$

where $p > 0$ is invariant.

(c) Characterize location-scale equivariant estimators of u_{s^m}.

(d) Give a description as in (3.126) and (3.127) of the minimum risk location-scale equivariant estimator of u_{s^m}.

(e) Give a formula similar to (3.141) for the minimum mean squared relative error location-scale equivariant estimator of u_{s^m}.

(f) For the i.i.d. $N(\mu, \sigma^2)$ family, find the minimum mean squared relative error location-scale equivariant estimator of σ^2. Compare with the UMVUE.

3.4.29 Consider the "pure" scale family

$$f_a(\underline{x}) \;=\; a^{-n} q(\underline{x}/a) \quad , \quad a > 0 \quad ,$$

where q is a fixed Lebesgue density on $I\!R^n$. We wish to estimate

$$u_m(a) \;=\; a^m$$

for some $m > 0$.

(a) Determine the induced groups $\bar{\mathbf{T}}$ and $\mathbf{T}^\star$, assuring that they exist by checking whatever assumptions need be checked.

(b) Give characterizations of invariant loss functions and scale equivariant estimators of the estimand.

(c) Show how a minimum risk scale equivariant estimator of $u_m(a)$ may be obtained, analogously to (3.126) and (3.127).

(d) Give a formula for the minimum mean relative squared error scale equivariant estimator of $u_m(a)$.

(e) Specialize (d) to the case of i.i.d. $N(0, \sigma^2)$ observations and $m = 1, 2$, obtaining explicit formulae similar to (3.144).

(f) Specialize (d) to the case of i.i.d. $Unif(0, a)$, obtaining a simple formula for the estimator.

(g) Specialize (d) to the case of i.i.d. $Exp(a, 0)$.

3.4.30 (a) Consider the location-scale family as in the text of this section. Show that the family of densities for

$$\underline{Y} \;=\; (X_2 - X_1, X_3 - X_1, ..., X_n - X_1)$$

form a pure scale family on $I\!R^{n-1}$.

(b) Using (a) and Exercise 3.4.29, give an alternative derivation of the minimum risk location-scale equivariant estimator of scale in (3.126) and (3.127).

3.4.31 Suppose $\underline{X}$ and $\underline{Y}$ are random n and m-vectors, respectively, with joint Lebesgue density

$$f_{(a,b,c)}(\underline{x}, \underline{y}) \;=\; a^{-(n+m)} q((\underline{x} - b\underline{1})/a, (\underline{y} - c\underline{1})/a)$$

where $a > 0$, $b \in I\!R$, and $c \in I\!R$. Here, q is a known Lebesgue density.

(a) Extend the theory of the current section to obtain general formulae for optimal appropriately equivariant estimators of (i) a_m and (ii) $b - c$.

(b) Specialize to the i.i.d. normal case.

3.5 Bayesian Estimation.

The Bayesian approach to statistics requires the statistician to make one additional assumption, namely that the unknown parameter is a random element taking values in the parameter space (which now is required to have a measurable structure, i.e. there is a σ–field $\mathcal{H}$ on Θ), and that the distribution of the parameter (call it Π) is known. Π is called the *prior distribution*. Once this leap of faith is made, one arrives at a state of heavenly bliss wherein virtually all problems can be solved (optimally, if there is a decision theoretic structure such as a loss function) subject to the practical problem of computing the result. Furthermore, there is a certain mathematical elegance and simplicity in the theory which is unusual in Statistics, and the mathematical results have applications in non-Bayesian theory. Many stastisticians and scientists are opposed to the use of Bayesian Statistics in practical problems for mostly "philosophical" reasons which we discuss below.

3.5.1 Bayesian Decision Theory.

Suppose we have a Euclidean observation space $(\Xi, \mathcal{G})$ and family of possible distributions $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ for the random observable X. Under measurability assumptions (see (i) of Theorem 1.5.10, the "Two Stage Experiment Theorem"), we may regard $\mathbf{P}$ as a family of conditional distributions for X given the parameter value θ. We may then compute an "average" risk which is the loss of an estimated value averaged over *both* Ξ *and* Θ (the usual averaging is only over Ξ). This extra averaging operation gives the the so called *Bayes Risk*

$$r(\Pi, \delta) \; = \; \int_\Theta R(\theta, \delta) \, d\Pi(\theta) \; = \; \int_\Theta \int_\Xi L(\theta, \delta(x)) \, dP_\theta(x) \, d\Pi(\theta) \quad .$$

The second equation follows from "Furthermore, ..." conclusion of the Two Stage Experiment Theorem and equation (1.71) of Theorem 1.5.6. Note that the ordinary risk $R(\theta, \delta)$ is the Bayes risk $r(\delta_\theta, \delta)$ w.r.t. the prior δ_θ which is a unit point mass at θ. Alternatively, $R(\theta, \delta)$ is the conditional expectation of the loss given θ, i.e.

$$R(\theta, \delta) \; = \; E[\, L(\theta, \delta(\underline{X})) \,|\, \theta \,] \quad . \tag{3.149}$$

(Notational remark: we are using θ both to denote the random element $\theta : (\Omega, \mathcal{F}, P) \longrightarrow (\Theta, \mathcal{H})$ and to denote a particular value $\theta \in \Theta$. Since we have already used the upper case "Θ" to denote the parameter space, it is not available to denote the random element. We hope the reader has sufficient understanding of the difference at this point that he can recognize which meaning of "θ" is intended and act appropriately. The above equation (3.149) would more appropriately be written

$$R(\theta_0, \delta) \; = \; E[\, L(\theta, \delta(\underline{X})) \,|\, \theta = \theta_0 \,] \quad ,$$

so as to avoid confusion between the random element θ and the specific value θ_0. We will continue to abuse the notation of conditional distributions and conditional expectations as in (3.149) for simplicity.) A *Bayes rule* is one which minimizes the Bayes risk, i.e. $\delta^\star_\Pi$ is a Bayes rule iff

$$r(\Pi, \delta^\star_\Pi) \;\leq\; r(\Pi, \delta)$$

for any other decision rule δ. Notice that the Bayes rule depends on the prior (and the loss). Once Π is given, then we are in the business of trying to optimize over δ the *real valued* function $r(\Pi, \delta)$ instead of somehow trying to manage all values of θ simultaneously when we try to uniformly minimize $R(\theta, \delta)$. For this reason, it is not necessary to restrict the class of decision rules in Bayesian decision theory (e.g. requiring unbiasedness of an estimator), although it is sometimes done for convenience. Recall that a main reason for considering restrictions on the class of estimators was to rule out "dumb" estimators like $\hat{\theta} \equiv 5$, which have smaller risk than any reasonable estimator (i.e. one which actually uses the data) if in fact the true value of θ is 5. But the Bayes risk of $\hat{\theta} \equiv 5$ will not be that good as long as the prior is not concentrated in the neighborhood of $\theta = 5$. Of course, if we use $\Pi = \delta_5$, then $\hat{\theta} \equiv 5$ is the Bayes estimator, but this results from a "dumb" choice of prior.

Finding a Bayes rule is in general quite easy, from a formal point of view in the sense that one can write down formulae involving conditional expectations. Under the measurability conditions of the Two Stage Experiment Theorem (i.e. that $P_\theta(B)$ is Borel measurable in θ for each fixed B), then there is a joint distribution $\mathrm{Law}[\underline{X}, \theta]$ on $\Xi \times \Theta$, and $P_\theta = \mathrm{Law}[\underline{X}|\theta]$. *In all of the following, we are taking expectations and conditional expectations w.r.t. the joint distribution of $\underline{X}$ and θ.* If we assume that Θ is a Euclidean space (which is good enough for most applications), then by Theorem 1.5.6, there exists a conditional distribution $\mathrm{Law}[\theta|\underline{X} = \underline{x}]$, which is called the *posterior distribution* for θ. By successive conditioning,

$$\begin{aligned} r(\Pi, \delta) \;&=\; E\,\{\, E[\, L(\theta, \delta(\underline{X}))\,|\,\theta\,]\,\} \;=\; E[\, R(\theta, \delta)\,] \\ &=\; E\,\{\, E[\, L(\theta, \delta(\underline{X}))\,|\,\underline{X}\,]\,\} \;=\; E[\,\bar{\rho}(\underline{X}, \delta(\underline{X}))\,] \end{aligned} \tag{3.150}$$

where

$$\bar{\rho}(\underline{x}, d) \;=\; E[\, L(\theta, d)\,|\,\underline{X} = \underline{x}\,] \tag{3.151}$$

is the *posterior expected loss.* Note that $\bar{\rho}$ may be computed by integration w.r.t. the posterior distribution. If we can find for fixed $\underline{x}$ a value $\delta^\star(\underline{x})$ in action space to minimize $\bar{\rho}(\underline{x}, .)$, then assuming measurability of $\delta^\star$, for any other decision rule δ,

$$\bar{\rho}(\underline{X}, \delta^\star(\underline{X})) \;\leq\; \bar{\rho}(\underline{X}, \delta(\underline{X}))$$

and plugging this into (3.150) we obtain

$$r(\Pi, \delta^\star) \;\leq\; r(\Pi, \delta) \quad ,$$

i.e. $\delta^\star$ is the Bayes rule. This gives a simple recipe for finding the Bayes rule.

As an example, suppose we wish to estimate a real valued estimand $g(\theta)$ under squared error loss, so

$$\bar{\rho}(\underline{x}, d) \;=\; E[\,(g(\theta) - d)^2\,|\,\underline{X} = \underline{x}\,] \qquad (3.152)$$

$$=\; E[\,g(\theta)^2\,|\,\underline{X} = \underline{x}\,] \;-\; 2dE[\,g(\theta)\,|\,\underline{X} = \underline{x}\,] \;+\; d^2 \quad .$$

Minimizing over d gives

$$\delta^\star(\underline{x}) \;=\; E[\,g(\theta)\,|\,\underline{X} = \underline{x}\,] \quad , \qquad (3.153)$$

which we would have expected anyway. Note here that we have implicitly assumed that $E[g(\theta)^2] < \infty$ in order that the Bayes risk be finite. (If $E[g(\theta)^2] = \infty$ then $r(\Pi, \delta) = \infty$ for any δ by Exercise 3.5.7, so estimators have the same Bayes risk, namely ∞, so all estimators are equally "optimal"). It is only necessary of course that $E[|g(\theta)|] < \infty$ for the estimator to be defined and finite.

If we make the additional assumption of a dominated family, i.e. $\mathbf{P} \ll \mu$ where μ is σ–finite, then we can derive a density for the posterior distribution (even if Θ is not a Euclidean space) and some of the general results above can be simplified. Let $f(x|\theta)$ denote the density of P_θ w.r.t. μ. We have written it as a conditional density as it is the conditional density of X given θ by the Two Stage Experiment Theorem. Then we have

$$r(\Pi, \delta) \;=\; \int_\Theta \int_\Xi L(\theta, \delta(x))\, f(x|\theta)\, d\mu(x)\, d\Pi(\theta) \qquad (3.154)$$

$$=\; \int_\Xi \left[\int_\Theta L(\theta, \delta(x))\, f(x|\theta)\, d\Pi(\theta) \right] d\mu(x) \quad .$$

Fubini's theorem was used at the last step, of course. Let A denote the action space. Now suppose that for each fixed value of x we can find $\delta^\star$ to minimize over $d \in A$

$$\rho(x, d) \;=\; \int_\Theta L(\theta, d)\, f(x|\theta)\, d\Pi(\theta) \quad . \qquad (3.155)$$

Notice that the inner integral in (3.154) is $\rho(x, \delta(x))$, so $\delta^\star(x)$ will minimize the double integral in (3.154), i.e. $\delta^\star(x)$ will be the Bayes rule, provided it is measurable. In some sense then, once we observe a value x of X, it is a computing problem to minimize $\rho(x, d)$ for that value of x. It may be necessary to numerically compute the integral in (3.155), and to use this within a numerical optimizer to minimize $\rho(x, .)$.

The difference between ρ in (3.155) and $\bar{\rho}$ in (3.151) is simply a normalization constant, which we now discuss. For a function $h : \Theta \longrightarrow I\!R$ define

$$J(h) \;=\; \int_\Theta h(\theta)\, f(x|\theta)\, d\Pi(\theta) \quad ,$$

provided the integral exists. Note that $J(h) : \Xi \longrightarrow I\!R$, but we will tend to think of x as fixed. Then of course $\rho(x, d) = J(L(., d))\,(x)$. It follows from Exercise 1.5.12 that the posterior distribution Law$[\theta|X = x] \ll \Pi$ has density

$$f(\theta|x) \;=\; \frac{f(x|\theta)}{J(1)} \quad . \tag{3.156}$$

In Bayesian statistics, when $f(x|\theta)$ is thought of as a function of x for fixed θ, it is sometimes called the *sampling density*, and when it is thought of as a function of θ for fixed x, it is called (a version of) the *unnormalized posterior density*, since it only needs to be divided by the normalizing constant $J(1)$ to become the normalized posterior density. Actually, one can use any function $h(\theta|x)$ as the unnormalized posterior density provided $h(\theta|x)$ has the property that dividing by its integral w.r.t. $d\Pi(\theta)$ gives $f(\theta|x)$. Note that we used an unnormalized posterior in (3.155), so we can refer to $\rho(x, d)$ as unnormalized posterior expected loss, and the posterior expected loss is given by

$$\bar{\rho}(x, d) \;=\; \frac{\rho(x, d)}{J(1)} \quad .$$

Bayesian statisticians often write

$$f(\theta|x) \;\propto\; f(x|\theta)$$

to indicate that the posterior density is proportional to an unnormalized posterior density. This is useful since the normalization constant is frequently unneeded as in (3.155).

We shall see that the posterior density has further uses, e.g. for constructing interval or set estimates. Much of the current research in Bayesian statistics centers on finding useful approximations or practical computing methods for evaluating expressions such as (3.153), (3.155), and (3.156).

Now we consider a standard example of Bayesian estimation.

Example 3.5.1 Suppose X_1, X_2, ..., X_n are i.i.d. $N(\mu, \sigma^2)$. For the time being, consider a fairly general prior distribution for (μ, σ^2) which has Lebesgue density denoted

$$\frac{d\Pi}{dm^2}(\mu, \sigma^2) \;=\; \pi(\mu, \sigma^2)$$

where $\pi(\mu, \sigma^2) = 0$ if $\sigma^2 \leq 0$. Letting $\pi(\sigma^2)$ denote the marginal prior density of σ^2 and $\pi(\mu|\sigma^2)$ the conditional prior density of μ given σ^2, we can write the prior joint density as

$$\pi(\mu, \sigma^2) \;=\; \pi(\mu|\sigma^2)\,\pi(\sigma^2) \tag{3.157}$$

(For simplicity, we are adopting a notational convention common in Engineering: if the distribution of a random vector has a density then simply use the arguments of densities to indicate the appropriate random variable. Also, note that we are using the variable σ^2, and to avoid confusion the student may want to replace it with a simpler expression such as v.) Then the posterior density is

$$f(\mu,\sigma^2|\underline{x}) \;\propto\; (\sigma^2)^{-n/2}\,\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\pi(\mu|\sigma^2)\pi(\sigma^2)$$

$$\propto\; \left\{(n^{-1}\sigma^2)^{-1/2}\,\exp\left[\frac{-1}{2(n^{-1}\sigma^2)}(\mu-\bar{x})^2\right]\pi(\mu|\sigma^2)\right\}\times$$

$$\left\{(\sigma^2)^{-(n-1)/2}\,\exp\left[-\frac{ns^2}{2\sigma^2}\right]\pi(\sigma^2)\right\} \qquad (3.158)$$

where

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i \quad , \quad s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad .$$

Note that in (3.158) we have not kept factors which are functionally independent of μ and σ^2.

A simple calculation (see Exercise 3.5.4) shows that

$$f(\mu,\sigma^2|\underline{x}) \;=\; f(\mu|\sigma^2,\underline{x})f(\sigma^2|\underline{x}) \quad . \qquad (3.159)$$

Using this, one can see by inspection of the last expression in (3.158) that

$$f(\mu|\sigma^2,\underline{x}) \;\propto\; (n^{-1}\sigma^2)^{-1/2}\,\exp\left[-\frac{1}{2(n^{-1}\sigma^2)}(\mu-\bar{x})^2\right]\pi(\mu|\sigma^2) \qquad (3.160)$$

since the only place μ appears in the r.h.s. of (3.159) is in $f(\mu|\sigma^2,\underline{x})$.

Now, we will make a choice for a *conditional prior* for μ. We will select $\pi(\mu|\sigma^2)$ purely on the basis of mathematical convenience so that (3.160) can be put in a nice form. The posterior for μ conditional on σ^2 will be simple, namely a normal distribution, provided $\pi(\mu|\sigma^2)$ is also a normal density, say

$$\mathrm{Law}[\mu|\sigma^2] \;=\; N(\nu,\tau^2) \quad , \quad \nu \;=\; \nu(\sigma^2) \quad , \quad \tau^2 \;=\; \tau^2(\sigma^2) \quad . \qquad (3.161)$$

(Note that one way to recognize a normal density, say $g(y)$, is to see that y only appears in a quadratic form in an exponential. Since the sum of quadratic forms is also a quadratic form, by choosing such a conditional normal prior for μ we guarantee a conditional normal posterior.) To figure out the mean and variance for $f(\mu|\sigma^2,\underline{x})$ we carry out the usual elementary calculations of completing the square in the exponent, which gives (Exercise 3.5.5)

$$\exp\left\{-\frac{1}{2}\left[\frac{(\mu-\bar{x})^2}{(n^{-1}\sigma^2)}+\frac{(\mu-\nu)^2}{\tau^2}\right]\right\}$$

$$= \exp\left[\frac{-1}{2[(n^{-1}\sigma^2)^{-1} + (\tau^2)^{-1}]^{-1}}\left(\mu - \frac{\bar{x}/(n^{-1}\sigma^2) + \nu/\tau^2}{(n^{-1}\sigma^2)^{-1} + (\tau^2)^{-1}}\right)^2 \right.$$
$$\left. - \frac{1}{2}\frac{(\bar{x}-\nu)^2}{n^{-1}\sigma^2 + \tau^2}\right] \quad . \tag{3.162}$$

There are several algebraic "simplifications" of the above expression which we did not implement so as to permit some of the following interpretations.

(i) If σ^2 is known, we are basically done at this point, i.e. $f(\mu|\sigma^2, \underline{x})$ is the posterior for μ.

(ii) For any random variable Y, define the *precision* as the reciprocal of its variance, i.e.

$$\text{Precision}[Y] \; = \; \frac{1}{\text{Var}[Y]} \quad .$$

From (3.162) one sees that

$$\text{Var}[\mu|\sigma^2, \underline{X}] \; = \; [(\sigma^2/n)^{-1} + (\tau^2)^{-1}]^{-1} \quad , \tag{3.163}$$

which can be restated as

$$\text{Precision}[\mu|\sigma^2, \underline{x}] \; = \; \text{Precision}[\bar{X}|\mu, \sigma^2] \; + \; \text{Precision}[\mu|\sigma^2] \quad . \tag{3.164}$$

(A conditional precision Precision$[Y|Z]$ is defined in the obvious way with the conditional variance.) *Assuming σ^2 is known, the posterior precision of the mean μ is the sum of the sampling precision of the sample mean $\bar{X}$ and the prior precision of μ.* This provides at least an easy way of remembering the formulae. An entirely analogous result holds for multivariate observations. See Exercise 3.5.8.

(iii) One can also read off from (3.162) that

$$E[\mu|\sigma^2, \underline{X}] \; = \; \frac{\bar{X}/(\sigma^2/n) + \nu/\tau^2}{(\sigma^2/n)^{-1} + (\tau^2)^{-1}} \tag{3.165}$$

$$= \; \frac{\text{Precision}[\bar{X}|\mu, \sigma^2]\, \bar{X} \; + \; \text{Precision}[\mu|\sigma^2]\, \nu}{\text{Precision}[\bar{X}|\mu, \sigma^2] \; + \; \text{Precision}[\mu|\sigma^2]} \quad .$$

Thus, *if σ^2 is known then the posterior mean of μ is a weighted average of the sample mean $\bar{X}$ and the prior mean ν with weights proportional to their precisions as individual estimates of μ.* To explain this last statement, if we use sample mean $\bar{X}$ to estimate μ ignoring the prior information, then its precision is $(n^{-1}\sigma^2)^{-1}$, and if we use the prior mean ν to estimate μ ignoring the sample then its precision is $(\tau^2)^{-1}$. Again, besides some aesthetic appeal this provides an easy way to remember the formulae.

(iv) Note that putting $n = 0$ in (3.163) and (3.165) gives τ^2 and ν, respectively. Thus, if we had no data, then the posterior would be the prior, which only makes sense. This also provides a useful check on Bayesian calculations.

Now we return to (3.158) and (3.159) to investigate the marginal posterior of σ^2, i.e. $f(\sigma^2|\underline{x})$. If we substitute for $f(\mu|\sigma^2, \underline{x})$ (which is normal with parameters given in (3.163) and (3.165)) back into (3.159) and integrate out μ (which can be accomplished by "juggling" the normalizing constants), we obtain (Exercise 3.5.5)

$$f(\sigma^2|\underline{x}) \;\propto \tag{3.166}$$

$$(\tau^2)^{-1/2} \left[(n^{-1}\sigma^2)^{-1} \,+\, (\tau^2)^{-1} \right]^{-1/2} (\sigma^2)^{-n/2} \times$$

$$\exp\left\{ -\frac{1}{2} \left[\frac{(\bar{x} - \nu)^2}{n^{-1}\sigma^2 + \tau^2} + \frac{ns^2}{\sigma^2} \right] \right\} \pi(\sigma^2) \quad .$$

(One error to watch out for here is dropping the factor of $(\tau^2)^{-1/2}$. We have to keep track of this because τ^2 depends on σ^2.) We see that there are two rather ugly parts of the last expression: the first pair of factors,

$$(\tau^2)^{-1/2} \left[(n^{-1}\sigma^2)^{-1} \,+\, (\tau^2)^{-1} \right]^{-1/2}$$

and the first term within brackets in the exponent

$$\frac{(\bar{x} - \nu)^2}{n^{-1}\sigma^2 + \tau^2} \quad .$$

Were these two expressions each multiples of $1/\sigma^2$, then we could obtain a power of $1/\sigma^2$ times an exponential of a negative constant times $1/\sigma^2$. We could then transform to the precision parameter $\lambda = 1/\sigma^2$ get a *Gamma* density except for the factor of the prior density. Recall that we may allow τ^2 and ν to depend on σ^2 (see (3.161)), so if we choose

$$\nu(\sigma^2) \equiv \nu \quad , \quad \tau^2(\sigma^2) \;=\; \sigma^2/c \tag{3.167}$$

for some fixed $\nu \in I\!R$ and $c > 0$, then (3.166) becomes

$$f(\sigma^2|\underline{x}) \;\propto \tag{3.168}$$

$$(\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2} \left[\frac{(\bar{x} - \nu)^2}{n^{-1} + c^{-1}} + ns^2 \right] \frac{1}{\sigma^2} \right\} \pi(\sigma^2) \quad .$$

Now make the substitution $\lambda = 1/\sigma^2$ to obtain the posterior for λ as

$$f(\lambda|\underline{x}) \;\propto\; \lambda^n/2 \,\exp\left\{ -\frac{1}{2}\left[\frac{(\bar{x}-\nu)^2}{n^{-1}+c^{-1}} + ns^2 \right] \lambda \right\} \pi(\lambda) \quad . \tag{3.169}$$

(One may be tempted to think that a "Jacobian" factor from the transformation is missing here, but that is absorbed in the prior $\pi(\lambda)$. Note that putting $n = 0$ in the above causes the posterior to reduce to the prior, as it should, but if one incorrectly multiplies by the Jacobian $1/\lambda^2$, then this check will not apply.) Note that the meaning of "$f(.|\underline{x})$" and "$\pi(\cdot)$" is different between (3.168) and (3.169). Choosing a *Gamma* prior for λ, say

$$\text{Law}[1/\sigma^2] \;=\; \text{Law}[\lambda] \;=\; Gamma(\alpha, \beta) \tag{3.170}$$

for some $\alpha > 0$ and $\beta > 0$, then

$$f(\lambda|\underline{x}) \;\propto\; \lambda^{n/2+\alpha-1} \,\exp\left\{ -\frac{1}{2}\left[\frac{(\bar{x}-\nu)^2}{n^{-1}+c^{-1}} + ns^2 + \frac{2}{\beta} \right] \lambda \right\} \quad , \tag{3.171}$$

and we see that the posterior is

$$\text{Law}[\lambda|\underline{X} = \underline{x}] \;=\; Gamma(n/2+\alpha, \eta(\underline{x})) \tag{3.172}$$

where

$$(\eta(\underline{x}))^{-1} \;=\; \frac{1}{2}\left[\frac{(\bar{x}-\nu)^2}{n^{-1}+c^{-1}} + ns^2 \right] + \frac{1}{\beta} \quad . \tag{3.173}$$

Some remarks on this:

(v) Since in (3.170) we specified a *Gamma* prior on the reciprocal or inverse of σ^2, the prior on σ^2 is called an *inverse Gamma* distribution. The special case where $2\alpha = m$ is an integer and $\beta = 2$ is called an inverse χ^2. If $2\alpha = m$ is an integer and $\beta > 0$ is arbitrary, then the prior for σ^2 is a scaled inverse χ^2 with scale parameter $2\beta^{-1}$ and degrees of freedom m. Under this prior, the posterior for σ^2 is also an inverse χ^2 but the posterior scale parameter $2\eta^{-1}$ is the prior scale parameter plus a data dependent term (twice the first term on the r.h.s. of (3.173)), and the posterior degrees of freedom is sample size plus prior degrees of freedom. This provides a convenient way of remembering the posterior, except for the term added into the scale parameter.

(vi) The posterior mean of λ is

$$E[\,\lambda\,|\,\underline{X}\,] \;=\; [\,n + 2\alpha\,]\left[\frac{(\bar{X}-\nu)^2}{n^{-1}+c^{-1}} + nS^2 + \frac{2}{\beta} \right]^{-1} \quad . \tag{3.174}$$

(vii) The posterior mean of σ^2 is

$$E[\sigma^2 \,|\, \underline{x}] \;=\; E[1/\lambda \,|\, \underline{x}] \;=\; [\, n + 2\alpha - 2\,]^{-1} \left[\frac{(\bar{X} - \nu)^2}{n^{-1} + c^{-1}} + nS^2 + \frac{2}{\beta} \right] \quad . \tag{3.175}$$

For this to be defined and finite, we need $n + 2\alpha - 2 > 0$. The above is then the Bayes estimator of σ^2 under squared error loss with the prior specified in (3.161), (3.167), and (3.170) (provided that $\alpha > 2$ so that $E[1/(\sigma^2)^2] < \infty$, and hence that the Bayes risk under squared error loss is finite, for otherwise every estimator is Bayes because they all have infinite Bayes risk).

Even though we have done much work, we are not yet finished. We would still like to find the unconditional posterior distribution of μ. For convenience, let

$$\hat{\mu} \;=\; E[\mu|\sigma^2, \underline{X}] \;=\; \frac{n\bar{X} + c\nu}{n + c} \tag{3.176}$$

where the last expression is easily derivable from (3.165) and (3.167). Since the latter expression is independent of σ^2, in fact we have that $\hat{\mu} = E[\mu|\underline{X}]$ by a successive conditioning argument: $E[\mu|\underline{X}] = E[E[\mu|\sigma^2, \underline{X}]|\underline{X}] = E[\hat{\mu}|\underline{X}] = \hat{\mu}$ since $\hat{\mu}$ is $\sigma(\underline{X})$ measurable. But, to press forward and get the full marginal posterior distribution, note that by (3.159) and (3.171),

$$f(\mu, \lambda|\underline{x}) \;\propto\; f(\mu|\lambda, \underline{x})\, \lambda^{n/2+\alpha-1}\, e^{-\lambda/\eta}$$

and plugging in (3.160) we obtain

$$f(\mu, \lambda|\underline{x}) \;\propto\; \lambda^{n/2+\alpha-1/2} \,\exp\left\{ - \left[\eta^{-1} + (n + c)(\mu - \hat{\mu})^2/2 \right] \lambda \right] \quad . \tag{3.177}$$

Integrating out λ gives

$$f(\mu|\underline{x}) \;\propto\; \left[\eta^{-1} + (n + c)(\mu - \hat{\mu})^2/2 \right]^{-(n+2\alpha+1)/2} \quad . \tag{3.178}$$

A final remark:

(viii) If $2\alpha = m$ is an integer, then

$$\mathrm{Law}[\,[\eta(n + c)(n + m)/2]^{1/2}(\mu - \hat{\mu}) \,|\, \underline{X} = \underline{x}\,] \;=\; t_{(n+m)} \tag{3.179}$$

where t_r denotes Student's t distribution with r degrees of freedom.

Of course, this is all based on our choice of the "convenience prior" given in (3.161), (3.167), and (3.170). Note that our prior has four parameters: ν (the prior mean for μ), $c^{-1/2}$ (the prior scale parameter for μ w.r.t. σ), α (prior shape parameter for the inverse *Gamma* on σ^2), and β^{-1} (prior scale parameter for σ^2). The Bayes estimators are not determined until these are specified, but we can note the following:

$$\text{as } c \to 0 \quad , \quad \hat{\mu} \to \bar{X} \quad , \tag{3.180}$$

i.e. as prior precision for μ tends to 0, our posterior mean approaches the sample mean. Also,

$$\text{as } c \to 0 \quad , \beta \to \infty \quad , \text{ and } \alpha = 1/2 \quad , \quad E[\sigma^2|\underline{X}] \to \frac{n}{n-1} S^2 \quad , \tag{3.181}$$

i.e. under these limiting values of the prior for σ^2, we obtain the UMVUE as the limiting Bayes estimator.

□

The prior chosen in Example 3.5.1 is an example of a *conjugate prior family.* This is a family of prior distributions corresponding to a given sampling distribution Law$[X|\theta]$ wherein the posterior belongs back in this family. For our example above, the prior is sometimes called a (conditional) normal (given σ^2)-inverse *Gamma*, and the posterior is also a normal-inverse *Gamma.* Several examples of conjugate prior families corresponding to Binomial, Poisson, and Gamma sampling distributions are given in Exercises 3.5.9, 3.5.10, and 3.5.11. See also DeGroot *Optimal Statistical Decisions* (1970).

3.5.2 Some Philosophical Aspects of Bayesian Statistics.

Bayesian Statistics is a worthwhile subject of study for many reasons, but it is very controversial. Many stastisticians and scientists are opposed to the use of Bayesian Statistics in practical problems for "philosophical" reasons. The main objections usually involve the assumption of a random parameter and the choice of its prior distribution. Many critics of the Bayesian approach claim that the parameter is simply not random. In their view, it is a fixed but unknown constant. However, consider the following. Suppose a consulting client approaches a statistician and tells him he will measure the weights of several mice in a laboratory. Prior to weighing the mice, their weights are unknown constants, which become revealed upon the act of weighing, and then they become known constants (ignoring measurement error). However, any competent statistician will be willing to regard the weights of the mice as *realizations of some random variable.* The main difference between say the weight of an individual mouse and the mean weight of the population of mice is that when doing inference on the latter it is mathematically convenient and productive to assume that the individual represents a "randomly" chosen mouse from the population. But if it is also mathematically convenient and productive to assume that the population mean weight is a random variable, then why not? Of course, the critic of Bayesian Statistics will probably hasten to point out that the experiment of selecting a mouse at random is one that can be repeated more or less indefinitely, and in the sense of long run frequencies one could in principle obtain a reasonable estimate of the whole distribution of random mouse weights. This is basically the *long run frequency interpretation of probability.* Because of the prevalence of this criticism of Bayesianism, the battle between Bayesians and other statisticians is often formulated as Bayesianism vs. Frequentism (and non-Bayesian statisticians are sometimes called frequentists, even though not all non-Bayesians would accept this designation). The frequentist will only allow that quantities that are obtained from more or less indefinitely repeatable experiments may be considered as realized values of random variables. Since there is only one population mean mouse weight, and one cannot repeat the experiment of getting a population of mice and getting a mean weight for that population, the mean population weight can not be so considered as a realized value of a random variable according to the frequentist.

The author believes that such thinking (that only quantities obtained from repeatable experiments are candidates for randomness) is overly constraining and narrow minded. "Randomness" is a construction of the human mind. Even such "obviously" random experiments as flipping a coin are not so obviously random if one thinks about them for long. Many people might claim that a coin flip is perfectly predictable if one knows the initial position of the coin, the force exerted on it by the thumb and fingers, and the local state of the atmosphere. Therefore, such people would claim, a coin flip is not random, but simply difficult to predict because of the huge amount of information required. Similar remarks hold for other games of chance. It is well known that computer generated random numbers are not "truly random." (Reference ??) The author's belief is that "randomness" has proven to be a useful mathematical model for many phenomena, and to constrain its application with some philosophical dogmas is like the Pope telling Galileo that the Copernican system cannot be used as a model for planetary motions since it conflicts with the official teachings of the Church. The only way to determine if the Bayesian approach of assuming a random parameter is a good one is to learn about it and try it out on practical problems, and see if it leads to good and useful practical results.

Another objection against Bayesian Statistics that is often raised is that one cannot know the prior distribution of the parameter, and by assuming one or another prior distribution one inevitably biases the results in some sense towards the prior. For instance, considering the simple framework of estimating a normal mean from i.i.d. observations with known variance (see remark (iii) in Example 3.5.1 above), one has that the Bayes estimate under squared error loss with a normal prior as in (3.161) has bias

$$E[\hat{\mu}|\mu] - \mu \;=\; \frac{1}{n\tau^2/\sigma^2 \,+\, 1}\,(\nu - \mu) \quad . \tag{3.182}$$

So the bias depends in a complicated way on the prior and σ^2, but bias is proportional to $\nu - \mu$, the difference in the prior mean and the unknown true value of the parameter. More generally, Bayesian methods are "biased" in some way towards the "likely" values of the parameter according to the prior. This is not necessarily so bad for at least two reasons. Firstly, one can investigate the extent of the bias and often times show that it is negligible. For instance, in (3.182) we see that the bias tends to 0 as $n \to \infty$ or $\tau^2 \to \infty$. In the next chapter, we will see that this first effect happens quite generally: the bias from the prior diminishes as the sample size increases. Thus, with a large sample, which is when any kind of statistics works best, the "biasing" effect from the prior becomes negligible. Also, we will discuss "noninformative" priors in the next section (e.g. the prior in the normal mean example wherein $\nu^2 \to \infty$) and see that one can generally choose the prior so as to make any biasing from the prior small. See Exercise 3.5.6. Secondly, one inevitably has such biases anyway. For instance, in the example of mouse weights above, if the client reported that the estimate of the mean mouse weight was 10 tons, any competent statistician would object (of course, any competent client would know something was wrong in the first place). Clearly, we think that 10 tons is a highly improbable value for a mean mouse weight, given our prior experience with weights of mice and other animals. Offhand, the author can confidently predict that the mean weight of any population of adult mice would be between .01 and 100 grams, and one can certainly choose a prior for which the bias of the estimated mean is negligibly small in this range (Exercise 3.5.6).

Of course, one should be well informed about this possible problem of "bias" with Bayesian methods and be somewhat careful with any practical application, but it is not any more difficult to check this than many other aspects of statistical models, such as assumed normality of the distribution of the observations (given the parameter). Indeed, in the mouse weight example, a single extremely obese mouse can be much more of a detriment to the estimate than any bias resulting from the Bayesian prior assumptions. Indeed, in such a case it may be necessary to consider the prior distribution of such extreme obesity so as to adjust the estimator and get a reasonable estimate of the population mean. Scientists almost always discard discrepant observations from their data sets without further thought, indicating that they definitely have prior beliefs about the distribution of their data. There have been various proposals on how to "elicit" priors, meaning to turn these vague prior beliefs into a probability distribution, but these methods are probably seldom if ever used in practice, and it is usually the statistician's responsibility to somehow make good use of these prior beliefs.

The fact that the prior is in most cases indeterminate from a scientific or objective viewpoint seems not so serious when one recalls that many statistical models are "overachievers" (Efron, ??). For instance, we often time use calculations based on a normal distribution for analysis of variance, and the results have been shown to be somewhat insensitive to the assumed normality, both theoretically and experimentally. Many practical results in agriculture have been obtained with the normal distribution when it is obviously not accurate in a narrow sense. Probability models often have a way of bending to accomodate the data without breaking. This is not universally true and a good applied statistician knows what the shortcomings are of any statistical procedure, Bayesian or not.

The criticism that the prior is not objectively determinable is not always true in the sense that sometimes the prior can be inferred from past experience, for instance if in the mouse weight example we consider our population not the population of all mice but rather the population of laboratory mice supplied by whatever company supplied them (which is probably a more reasonable definition of the population), then we may have data available from several suppliers which can be used (if we think of our supplier as random supplier) to estimate a prior distribution. In fact, we really have a two stage statistical model (randomly selecting a supplier, then randomly selecting mice from that supplier), and of course this should be taken into account. The Bayesian calculations then just become a formal way of relating the first stage model (distribution of mean weight of the mice from a random supplier) to the second stage (individual mice selected at random from a given supplier), and one must still decide whether to be Bayesian or not in the estimation problem (i.e. whether or not to use a prior distribution on the supplier parameters). Such two stage models sometimes are called "random effects" models.

Another approach that is sometimes used is to try to estimate the parameters of the prior from the data. In the random effects type models above, if we think of the supplier parameters as prior parameters, then this makes some sense if we have several samples from different suppliers, but the technique is often applied in the more classical Bayesian framework where we have one sample which provides inference about a single realization of a parameter value from the prior. Such an approach is sometimes referred to as "Empirical Bayesian" since the "prior" is determined empirically. However, one can be a Bayesian Empirical Bayesian (i.e. put a prior on the parameters of the distribution of the parameter; this gives rise to so–called "heirarchical" Bayesianism, and one can obviously extend the heirarchy *ad nauseum*, putting priors on top of priors), or a non-Bayesian Empirical Bayesian (i.e. assume the parameters of the prior are unknown constants and estimate them by a non–Bayesian method, most commonly the method of maximum likelihood, discussed in Section 1).

In some cases, a client may be quite happy to use a subjectively determined prior, in which case the statistician can simply make the computations whether or not he or she believes in the client's prior. In many real world situations, one is confronted with the problem of making a decision where there are two kinds of data: (i) quantifiable data, and (ii) nonquantifiable data. For instance, if we wish to predict interest rates for next year, there is plenty of historical data on interest rates which is quantified and can be used with other quantified data such as new housing starts to help in building a statistical model (probably one involving time series analysis) to make quantified predictions, including probability distributions. (Note: in some sense, next year's interest rate is a fixed but unknown constant. After the year is over, it will be a known constant. But it may nonetheless be useful to regard it as random, just as the (now known) values of present and past interest rates may be regarded as observed values of random variables for the purpose of performing a statistical analysis.) However, there are many items which defy quantification in the usual sense, such as consumer optimism and guesses about the likelihood of various options available to the Chairman of the Federal Reserve Board (who has a big impact on interest rates). An expert on such matters may be quite happy to come up with subjectively determined probabilities for the various contingencies, and then Bayesian methods provide a neat mathematical framework for combining the "quantifiable" data with the data wherein quantification is necessarily subjective. This feature of Bayesian methods has attracted some attention in the "Artificial Intelligence" community in recent years (References??). Exercise 3.5.12 (e) gives an example where such ideas may be useful in the Law.

Just as there are non-Bayesian statisticians who close their mind to Bayesian methods on the basis of some rather airy philosophical basis, there are dogmatic Bayesians who are critical of all statistical methods that do not fall within their realm of approval. Their complaints about others' methods often involve labels such as "incoherent" and have generally not attracted a wide following. The author finds their arguments even less convincing than the arguments against Bayesian Statistics, and it is certainly true that many clients would refuse to listen to a statistician who told them that they must use Bayesian methods. For these reasons, we don't consider them here.

Our goal is to explain Bayesian theory so as to bring out its inherent mathematical beauty and provide some understanding of it so that the student who so chooses will be able to apply it wisely in the future. It is the author's opinion that one should not pay too much attention to any brand of philosophy, as in modern times philosophy has mostly proved an impediment to science. Science is a creative activity and should not be subject to prior restrictions because of some abstract arguments.

3.5.3 Extensions of Bayesian Theory.

In this subsection, we discuss non-informative and improper priors. We also discuss various aspects of decision rules obtained from Bayes rules through limiting operations and their application to admissability in the next subsection. Similar ideas will appear in the next section when we discuss least favorable priors and minimax rules.

In Example 3.5.1, we saw that the UMVUE's of μ and σ^2 could be obtained as the *limits of Bayes estimators*. In the next example, we consider what happens to the posterior distribution in this limit.

Example 3.5.2 Suppose X_1, X_2, ..., X_n are i.i.d. $N(\mu, \sigma^2)$, as in Example 3.5.1.

Recall there that we assumed a prior with Lebesgue density

$$\pi(\mu, \sigma^2) \ =$$

$$[c/(2\pi\sigma^2)]^{1/2} \exp\left[\frac{-c}{2\sigma^2}(\mu-\nu)^2\right] \frac{(\sigma^2)^{-(\alpha+1)}}{\Gamma(\alpha)\beta^\alpha} \exp\left[\frac{-1}{\beta\sigma^2}\right] \quad , \quad \sigma^2 > 0 \quad .$$

The prior was most conveniently thought of as (i) μ conditional on σ^2 is $N(\nu, c^{-1}\sigma^2)$, and (ii) the marginal on $1/\sigma^2$ is a *Gamma*(α, β). We may ignore the normalizing constants and write

$$\pi(\,\mu, \sigma^2\,) \ \propto \tag{3.183}$$

$$(\sigma^2)^{-1/2} \exp\left[\frac{-c}{2\sigma^2}(\mu-\nu)^2\right] (\sigma^2)^{-(\alpha+1)} \exp\left[\frac{-1}{\beta\sigma^2}\right] \quad , \quad \sigma^2 > 0 \quad .$$

Using the r.h.s. of (3.183) as the prior density would yield the same posterior when formally plugged into (3.192), even though it is not normalized to be a probability density function. (**Important Remark:** We are using the proportionality calculations with $\propto$ introduced in Example 3.5.1. These can be very useful for saving work in Bayesian calculations, but one must be careful to avoid errors. In calculating the posterior density, one can always "throw away" factors that do not depend on the parameter, even though they may depend on the data. Very often (especially when using conjugate priors), after obtaining the unnormalized posterior density using proportionality calculations one can perform the normalization as in equation (3.156) "by inspection," as one can recognize the form of the unnormalized posterior density.) We already saw in (3.180) and (3.181) that if we let

$$c \to 0 \quad , \beta \to \infty, \text{ and } \alpha = 1/2 \quad ,$$

then we recover the UMVUE's as the limit of the Bayes estimators. Applying these same limiting values to the unnormalized prior on the r.h.s. of (3.183), we obtain

$$\pi(\mu, \sigma^2) \ \propto \ (\sigma^2)^{-1} \quad , \quad \sigma^2 > 0 \quad . \tag{3.184}$$

Note that this is a density w.r.t. Lebesgue measure m^2 of a σ–finite measure. (One can show that this density corresponds to Lebesgue measure m^2 on $(\mu, \log(\sigma^2))$; Exercise 3.5.14). One can formally do the calculations in Example 3.5.1 and obtain the posterior Lebesgue density

$$f(\mu|\sigma^2, \underline{X}) \ = \ (2\pi n^{-1}\sigma^2)^{-1/2} \exp\left[\frac{-1}{2n^{-1}\sigma^2}(\mu-\bar{X})^2\right] \tag{3.185}$$

$$f(\sigma^2|\underline{X}) \ = \ \frac{(\sigma^2)^{-[(n-1)/2+1]}}{\Gamma((n-1)/2)(nS^2)^{-(n-1)/2}} \exp\left[-(\sigma^2)^{-1}/(nS^2)^{-1}\right] \ . \tag{3.186}$$

Note that this is a Lebesgue probability density if and only if $n \geq 2$ (Exercise 3.5.14). Let us use Post$[.|X = x]$ to denote the posterior distribution. This basically means the same thing as Law$[.|X = x]$, but we want to remind ourselves that the usual conditional probability calculations may not be valid because we start out with a prior Π (with Lebesgue density given in (3.184)) which is not a probability measure (so in particular Theorems 1.5.6 and 1.5.10 are no longer valid as they were in Example 3.5.1, and the results of Theorem 1.5.5 are not valid). Then we may write equations (3.185) and (3.186) as

$$\text{Post}[\mu - \bar{x}|\sigma^2, \underline{X} = \underline{x}] \;=\; N(0, \sigma^2/n) \tag{3.187}$$

and

$$\text{Post}[ns^2/\sigma^2|\underline{X} = \underline{x}] \;=\; \chi^2_{n-1} \quad . \tag{3.188}$$

Note that these posterior distributions are formally the same as the sampling distributions Law$[\mu - \bar{X}|\sigma^2, \mu]$ and Law$[nS^2/\sigma^2|\sigma^2, \mu]$. We hasten to add that the interpretations are very different, although a Bayesian who uses the prior of (3.184), which is not a probability measure, would obtain the same results for many inferences as a non-Bayesian. Also, one can show (Exercise 3.5.14)

$$\text{Post}[(\mu - \bar{x})/(s/\sqrt{n-1})|\underline{X} = \underline{x}] \;=\; t_{n-1} \quad , \tag{3.189}$$

which is the same as the sampling distribution Law$[(\bar{X} - \mu)/(S/\sqrt{n-1})|\mu, \sigma^2]$. One should note however that

$$\begin{aligned} &\text{Post}[\,(\mu - \bar{x})/(s/\sqrt{n-1})\,,\, ns^2/\sigma^2\,|\,\underline{X} = \underline{x}\,] \;\neq \\ &\text{Post}[(\mu - \bar{x})/(s/\sqrt{n-1})|\underline{X} = \underline{x}] \;\times\; \text{Post}[ns^2/\sigma^2|\underline{X} = \underline{x}] \quad , \end{aligned} \tag{3.190}$$

whereas the analogous sampling distributions do factor (Exercise 3.5.14).

□

Before this last example, we assumed the prior Π was a probability measure. Now, assuming a dominated family of sampling distributions $\mathbf{P} \ll \mu$ where μ is σ–finite, we see that the expressions in (3.154), (3.155), and (3.156) are still meaningful as mathematical expressions provided Π is σ–finite and

$$J(1)(x) \;=\; \int_\Theta f(x|\theta)\, d\Pi(\theta) \;<\; \infty \quad . \tag{3.191}$$

(Note that we need Π to be σ–finite in order to apply Fubini's theorem (e.g. at (3.154)) and Radon-Nikodym theory.) If (3.191) holds, then as in (3.156) we have a posterior probability density w.r.t. the prior Π, namely

$$f(\theta|x) \;=\; \frac{f(x|\theta)}{J(1)} \quad , \tag{3.192}$$

and this is a probability density. That is, even if we don't require that Π be a probability measure, we may still be able to obtain a probability distribution for the posterior, meaning that $f(\theta|x)$ is the density w.r.t. Π of a probability measure. Note that in Example 3.5.1 we required that $n \geq 2$, and in some cases (3.191) holds for some x and not others (Exercise 3.5.20(d)). If $\Pi(\Theta) = \infty$ then it is called an *improper prior*. If $\Pi(\Theta) < \infty$ (and Π is nonzero), then it is called a *proper prior*. A *generalized prior* is simply a σ–finite prior, which may be proper or improper. Of course a proper prior can be normalized to be a probability measure and the normalizing constant will cancel out in the formula in (3.192). Thus, without loss of generality a proper prior may be assumed to be a probability measure. (Indeed, any two priors Π_1 and Π_2 are equivalent (in that they give the same posterior) if they are proportional, i.e. $\Pi_1(A) = c\Pi_2(A)$ for all measurable $A \subset \Theta$, where c is some positive, finite, constant.) If (3.191) holds, then we say the *posterior is proper for that* x. It follows from Theorem 1.5.6 that the posterior is always proper if the prior is proper.

It can be argued that the improper prior we chose in (3.184) in "*noninformative*". This is not a formal terminology, but intuitively speaking it seems reasonable in that for instance the bias in the Bayes estimators under squared error loss has disappeared, so we are no longer seemingly getting any "information" from the prior. The notion of a "noninformative prior" is not a well defined mathematical concept, although in many circumstances there are one or more "reasonable" choices for a noninformative prior. In Example 3.5.1, the prior density $\pi(\mu, \sigma^2) = 1$ on $I\!R \times (0, \infty)$ may seem like the most obvious choice for a noninformative prior, although it will not agree with (3.184) of course.

Whether the posterior is proper or not, the unnormalized posterior expected loss may still be finite, i.e. we may have

$$\rho(x, d) \; = \; \int_\Theta L(\theta, d)\, f(x|\theta)\, d\Pi(\theta) \; < \; \infty \quad .$$

Again, this could happen for some x and not others. Anyway, the integral is always defined (possibly $+\infty$), and we can define a *generalized Bayes rule* $\delta^\star(X)$ as a decision rule such that

$$\rho(x, \delta^\star(x)) \;\leq\; \rho(x, d) \qquad \text{for all } \; d \in A \quad .$$

Otherwise said, a generalized Bayes rule minimizes unnormalized posterior expected loss under a generalized prior. We are implicitly assuming a dominated family $\mathbf{P} \ll \mu$, σ–finite. A *strict (or proper) Bayes rule* is a Bayes rule for a proper prior. Thus, we have seen that for i.i.d. $N(\mu, \sigma^2)$ observations, the UMVUE's of μ and σ^2 are generalized Bayes estimators. In fact, they are not strict Bayes estimators, and in general an unbiased estimator will never be a proper Bayes estimator except in trivial situations (Exercise 3.5.21).

Improper priors and generalized Bayes rules have some intuitive appeal and are used often in practice by Bayesians. We will see one theoretical application of generalized Bayes rules in Theorem 3.5.1 below. Another mathematically useful concept is the following.

Definition 3.5.1 *A decision rule $\delta^\star$ is called* extended Bayes *iff there is a sequence of proper priors Π_k such that*

$$\lim_n \left[r(\Pi_n, \delta^\star) \, - \, \inf_\delta \, r(\Pi_n, \delta) \right] \; = \; 0 \quad .$$

In the above, the infimum is taken over all decision rules δ.

□

Intuitively, a decision rule is extended Bayes for a sequence of priors Π_k iff it minimizes the Bayes risk "in the limit". Exercise 3.5.15 gives a simple characterization. When computing a Bayes risk, one can in principle use either formula

$$r(\Pi, \delta) \; = \; \int_\Theta \int_\Xi L(\theta, \delta(x)) \, dP_\theta(x) \, d\Pi(\theta) \tag{3.193}$$

$$= \; \int_\Theta R(\theta, \delta) \, d\Pi(\theta) \quad ,$$

or

$$r(\Pi, \delta) \; = \; \int_\Xi \int_\Theta L(\theta, \delta(x)) \, dP(\theta|x) dP_X(x) \quad . \tag{3.194}$$

The first formula is in general easier since we usually already "have" the measures $P_\theta = \text{Law}[X|\theta]$ and Π, meaning usually we know the formulae for their densities. The second formula requires computation of the posterior $\text{Law}[\theta|X = x]$ and also of the unconditional density of X, i.e. $P_X = \text{Law}[X]$. Finding the posterior usually involves an extra integration (to compute $J(1)$), and finding P_X involves an extra integration. For instance, if $\mathbf{P} \ll \mu$ σ–finite, then

$$\frac{dP_X}{d\mu}(x) \; = \; f_X(x) \; = \; \int_\Theta f(x|\theta) \, d\Pi(\theta) \quad . \tag{3.195}$$

Thus, in general we recommend (3.193) as the easier way to calculate $r(\Pi, \delta)$, and in fact to write out all iterated integrals involved in (3.193) and see which integrals are easier to perform and use Fubini's theorem to contemplate various orderings of the iterated integrals. But it is worth thinking about (3.194) since one can often perform the integrations "by inspection" anyway.

3.5.4 Admissability.

Now, we turn to another important notion in decision theory, and then apply Bayesian methods.

Definition 3.5.2 *A decision rule δ_0 is called* admissable *w.r.t. the given loss iff there is no decision rule δ which is better than δ_0.*

□

Admissability is not a very strong condition on a decision rule. For instance, in estimation of $g(\theta)$ with squared error loss and a situation where $\mathbf{P} \ll P_{\theta_0}$ the degenerate estimator $\delta_0(X) = g(\theta_0)$ is admissable. To see this, note that if δ is better then $(\Omega, \mathcal{F}, \mu)E(\theta_0, \delta) = E_{\theta_0}[(g(\theta_0) - \delta(X))^2] \leq 0 = (\Omega, \mathcal{F}, \mu)E(\theta_0, \delta_0)$, so $\delta(X) = g(\theta_0)$, P_{θ_0}-a.s., which is the same as $\delta(X) = g(\theta_0)$, $\mathbf{P}$-a.s., i.e. $\delta(X) = \delta_0(X)$, $\mathbf{P}$-a.s., and hence $(\Omega, \mathcal{F}, \mu)E(\theta, \delta) = (\Omega, \mathcal{F}, \mu)E(\theta, \delta_0)$ for all θ, and in particular δ is not better than δ_0. Of course, this situation wherein $\mathbf{P} \ll P_{\theta_0}$ is very common (e.g. exponential families), so it will typically be the case that these trivial estimators are admissable.

Nonetheless, from a purely decision theoretic point of view, admissability is a minimal requirement of a decision rule. If it is not admissable, then there is one which is better, so (from a purely decision theoretic point of view) why not use the better one? Of course, decision theory must not be taken too seriously in applications (e.g. the model is never really known exactly, practical computability of the procedure is important, etc.), but it is still of interest to enquire as to the admissability of a decision rule. The following provides one sufficient condition.

Theorem 3.5.1 *(a) Let Π be a generalized prior and suppose δ_0 is an essentially unique rule minimizing the generalized Bayes risk, then δ_0 is admissable.*

(b) Suppose:

(i) *A dominated family, i.e. $\mathbf{P} \ll \mu$, σ–finite;*

(ii) *The loss function is strictly convex;*

(iii) *The generalized prior Π satisfies*

$$\Pi(\{\,\theta \,:\, f(x|\theta) > 0\,\}) \;>\; 0 \quad , \quad \textit{for } \mathbf{P} -- \textit{ almost all } x \quad ;$$

(iv) *δ_0 is a generalized Bayes rule with finite generalized Bayes risk.*

Then δ_0 is admissable.

Remarks 3.5.1 (1) The generalized Bayes risk of a decision rule δ is

$$r(\Pi, \delta) \;=\; \int_\Theta \int_\Xi L(\theta, \delta(x))\, dP_\theta(x)\, d\Pi(\theta) \;=\; \int_\Theta R(\theta, \delta)\, d\Pi(\theta) \quad .$$

The generalized Bayes risk is always defined (since a loss is a nonnegative function), but it may be $+\infty$. Assuming a dominated family, one can always show that a generalized Bayes rule is a minimizer of the generalized Bayes risk, but it is often the case that the generalized Bayes risk is infinite for all rules. Thus, an essentially unique generalized Bayes rule is not necessarily the essentially unique minimizer of the generalized Bayes risk. For example, consider the case of i.i.d. $N(\mu, \sigma^2)$ observations with σ^2 known and μ unknown. $\bar{X}$ is the essentially unique generalized Bayes estimato$\underline{\text{r}}$ squared error loss w.r.t. the improper Lebesgue prior, but its risk is constant (σ^2/n) and its Bayes risk is ∞ (as is the Bayes risk of all decision rules in this setup, by Remark (5), below).

(2) The hypothesis in part (a) that δ_0 is the essentially unique rule which minimizes the generalized Bayes risk means the following: if δ_1 is any decision rule satisfying

$$r(\Pi, \delta_1) \ \le \ r(\Pi, \delta) \quad , \qquad \text{for all } \delta, \tag{3.196}$$

then $\delta_1 = \delta_0$, $\mathbf{P}$-a.s.

(3) Condition (iii) in part (b) is a little tricky to state rigorously because of our notational abuses. Letting θ denote the random parameter element (a mapping from the underlying probability space into the parameter space), and θ_0 denote a specific parameter value, condition (iii) may be written as

$$P_{\theta_0}\left[\{\, x \in \Xi \, : \, \Pi(\{\, \theta \, : \, f(x|\theta) > 0 \,\}) \ > \ 0 \,\}\right] \ = \ 1 \quad , \qquad \text{for all } \ \theta_0 \in \Theta \quad .$$

Note that when we fix $x \in \Xi$, $\{\ \theta : f(x|\theta) > 0\ \} \subset \Theta$, so the Π measure of this set makes sense.

(4) Condition (iii) in part (b) holds in all "usual" families and priors.

(5) The requirement of finite generalized Bayes risk in (iv) cannot be deleted. See Exercise 3.5.22. Note that if any decision rule has finite generalized Bayes risk, then the generalized Bayes rule has finite generalized Bayes risk (Exercise 3.5.17).

□

Proof. For part (a), suppose some other estimator δ_1 is as good as δ_0. Then $R(\theta, \delta_1) \le R(\theta, \delta_0)$, which implies that $r(\Pi, \delta_1) \le r(\Pi, \delta_0)$. But since δ_0 is the essentially unique rule which minimizes $r(\Pi, \delta)$ over δ, it follows that $\delta_1(X) = \delta_0(X)$, $\mathbf{P}$-a.s. Hence, δ_1 is not better than δ_0 (in fact, it has the same risk function). Thus, there is no rule better than δ_0, so it is admissable.

For part (b), we will show that the conditions on the prior and the loss imply that any generalized Bayes rule is essentially unique, and if it has finite generalized Bayes risk then it is the essentially unique rule minimizing the generalized Bayes risk. (Note that we are not proving the existence of a generalized Bayes rule, just showing that it is unique if it exists.)

We first note the following useful fact.

Lemma 3.5.2 *If λ is a strictly convex function defined on $A \subset \mathbb{R}^d$ (where A is convex, by definition of a convex function), then there is at most one point $a \in A$ where λ takes its minimum value.*

Proof. Suppose there are two points $a_1 \neq a_2$ in A which minimize λ. By strict convexity, if $0 < p < 1$, then

$$\lambda(pa_1 + (1-p)a_2) \ < \ p\lambda(a_1) \ + \ (1-p)\lambda(a_2) \ = \ \lambda(a_1) \quad .$$

The last equality follows since $\lambda(a_1) = \lambda(a_2) =$ the minimum value of λ. But the inequality above shows that λ had a smaller value at the point $pa_1 + (1-p)a_2$ than at a_1, which contradicts the assumption that λ achieves its minimum at a_1.

□

Using this lemma, we show that under the hypotheses of part (b), the unnormalized posterior expected loss is strictly convex in d for fixed x. To this end, let $0 < p < 1$, then because of strict convexity of L as a function of d,

$$L(\theta, pd_1 + (1-p)d_2) \;<\; p\,L(\theta, d_1) \;+\; (1-p)\,L(\theta, d_2) \quad . \tag{3.197}$$

Now,

$$[\,p\,\rho(\,x\,,\,d_1\,) \,+\, (1-p)\,\rho(\,x\,,\,d_2\,)\,] \;-\; \rho(\,x\,,\,pd_1 \,+\, (1-p)d_2\,) \;=$$

$$\int_\Theta \{\,[\,p\,L(\,\theta\,,\,d_1\,) \,+\, (1-p)\,L(\,\theta\,,\,d_2\,)\,] \;-\; L(\theta, pd_1 + (1-p)d_2)\,\}\; f(x|\theta)\,d\Pi(\theta)$$

$$\geq\; 0 \quad .$$

If the last expression $= 0$, then since the integrand is nonnegative, it follows that the integrand is 0, Π-a.e. (Proposition 1.2.7(b)). This implies

$$\Pi(\{\,\theta\,:\; f(x|\theta) > 0 \text{ and}$$

$$(\,[\,p\,L(\theta, d_1) \,+\, (1-p)\,L(\theta, d_2)\,] \;-\; L(\theta,\, pd_1 \,+\, (1-p)d_2\,) \,>0\,\}$$

$$=\; 0 \quad .$$

In view of (iii), this means there is a set of θ's with positive Π measure on which

$$[\,p\,L(\theta, d_1) \,+\, (1-p)\,L(\theta, d_2)\,] \;-\; L(\theta,\, pd_1 \,+\, (1-p)d_2\,) \,>0\,\} \;=\; 0 \quad .$$

But a set of positive measure is nonempty, so this contradicts the strict convexity of $L(\theta, .)$. Hence, $\rho(x, .)$ is strictly convex.

Now by the lemma, if for each fixed x, $\delta_0(x)$ is a minimizer over d of $\rho(x, d)$, then it is unique for each fixed x. Thus, the generalized Bayes rule is unique. Now we show that such a δ_0 is the essentially unique minimizer ofthe generalized Bayes risk. Using (i) and Fubini's theorem, as in equation (3.154), we obtain for any rule δ

$$r(\Pi, \delta) \;=\; \int_\Theta \int_\Xi L(\theta, \delta(x))\,dP_\theta(x)\,d\Pi(\theta) \;=\; \int_\Xi \rho(x, \delta(x))\,d\mu(x) \quad .$$

If δ_1 has a generalized Bayes risk no larger than δ_0, i.e.

$$\int_\Xi \rho(x, \delta_1(x))\, d\mu(x) \ \le\ \int_\Xi \rho(x, \delta_0(x))\, d\mu(x) \quad .$$

Since both integrals are nonnegative, and the one on the right is finite (hypothesis (iv)), we may subtract and obtain

$$\int_\Xi [\, \rho(x, \delta_1(x)) \,-\, \rho(x, \delta_0(x))\,]\; d\mu(x) \ \le\ 0 \quad .$$

Since δ_0 is unique generalized Bayes, $\rho(x, \delta_0(x)) < \rho(x, \delta_1(x))$, so the integrand above is nonnegative, and hence by Proposition 1.2.7(b) $\rho(x, \delta_0(x)) = \rho(x, \delta_1(x))$ μ-a.s. In particular, there is a set of x's with positive μ measure (and hence a nonempty set) where $\rho(x, \delta_1(x))\ .\ \rho(x, \delta_0(x))$, contradicting the fact proved above that δ_0 is unique generalized Bayes.

Thus, we have shown that δ_0 is the essentially unique minimizer of the generalized Bayes risk (we can always obtain another minimizer by changing δ_0 on a **P**-null set), so it is admissable by part (a).

□

This result does not apply to many examples, such as the case of i.i.d. $N(\mu, \sigma^2)$ observations with σ^2 known and μ unknown. $\bar{X}$ is generalized Bayes under squared error loss w.r.t. the improper Lebesgue prior, but its risk is constant (σ^2/n) and its Bayes risk is ∞ (as is the Bayes risk of all decision rules in this setup, by Remark (5) after the Theorem). The next result can be applied to this case, however. It is similar in spirit to the previous theorem, but deals with extended Bayesian ideas rather than generalized Bayesian ideas.

Theorem 3.5.3 *Suppose:*

(i) *The parameter space Θ is Euclidean, say $\Theta \subset I\!R^p$;*

(ii) *The risk function of any decision with finite risk for all θ is continuous;*

(iii) *δ_0 is a decision rule such that for some sequence of priors $\{\ \Pi_k\ \}$ we have:*

- **(a)** *$r(\Pi_k, \delta_0) < \infty$ for all k;*
- **(b)** *For any open set $U \subset I\!R^p$ with $U \cap \Theta \neq \emptyset$, $\liminf_n\ \Pi_n(U \cap \Theta) > 0$;*
- **(c)** $\lim_n\ [\, r(\Pi_n, \delta^\star) \,-\, \inf_\delta\ r(\Pi_n, \delta)\ ] = 0.$

Then δ_0 is admissable.

Proof. Suppose δ_1 is better than δ_0, i.e. $R(\theta, \delta_1) \le R(\theta, \delta_0)$ for all θ with $R(\theta_0, \delta_1) \le R(\theta_0, \delta_0)$ for some θ_0. Let $\epsilon = (1/2)[\ R(\theta_0, \delta_0) \,-\, R(\theta_0, \delta_1)\ ] > 0$, then by continuity of $R(., \delta_i)$ there is an open set U containing θ_0 such that for all

$\theta \in U \cap \Theta$, $R(\theta_0, \delta_1) < R(\theta_0, \delta_0) - \epsilon$. Now by (iii) (b), there is an N and $M > 0$ such that for all $n \geq N$, $\Pi_n(U \cap \Theta) > M$. Then for all $n \geq N$,

$$r(\Pi_n, \delta_0) \; - \; \inf_{\delta} r(\Pi_n, \delta) \;\; \geq \;\; r(\Pi_n, \delta_0) \; - \; r(\Pi_n, \delta_1)$$

by (iii) (a),

$$= \; \int_{\Theta} [\, R(\theta, \delta_0) \; - \; R(\theta, \delta_1) \,] \, d\Pi_n(\theta) \;\; \geq \;\; \int_{U \cap \Theta} [\, R(\theta, \delta_0) \; - \; R(\theta, \delta_1) \,] \, d\Pi_n(\theta)$$

$$\geq \; \epsilon \, \Pi_n(U \cap \Theta) \;\; \geq \;\; \epsilon M \;\; > \;\; 0 \quad .$$

But this contradicts (iii) (c). Hence, no such better rule δ_1 exists, and hence δ_0 is admissable.

□

The student is asked to find where hypothesis (ii) was used in the above proof. We use this last result in the proof of the next one. This theorem for $d \geq 3$ shocked the statistical community when it was discovered by Charles Stein in 1956.

Theorem 3.5.4 *Let $\underline{X}_1$, $\underline{X}_2$, ..., $\underline{X}_n$ be i.i.d. random d-vectors with the $N(\underline{\mu}, \sigma^2 I)$ distribution where σ^2 is known. Consider estimation of $\underline{\mu}$ under Sum of Squared Errors Loss, viz.*

$$L(\underline{\mu}, \underline{d}) \;\; = \;\; \|\underline{\mu} - \underline{d}\|^2 \quad .$$

Then $\bar{\underline{X}} = (1/n) \sum_i \underline{X}_i$ is admissable if and only if $d < 3$.

Partial Proof. We only show admissability when $d = 1$. The proofs for the other dimensions are rather lengthy. We only mention that for $d \geq 3$, one can explicitly produce an estimator for $\underline{\mu}$ which has strictly smaller risk than $\bar{\underline{X}}$, which is known as the Stein estimator. See Berger, *Statistical Decision Theory.*

We may assume w.l.o.g. that $\sigma^2 = 1$ and $n = 1$ (Exercise 3.5.18), so we write X for $\bar{X}$. Since $R(\mu, X) = 1$ it is necessary to use a sequence of finite priors for μ in Theorem 3.5.3 so that (i) holds. Also, the risk function is clearly continuous, infinitely differentiable in fact (Exercise 3.5.19). Let

$$\Pi_n \;\; = \;\; a_n \, N(0, n)$$

where the a_n are to be determined. Note that we must keep track of the normalizing (or "unnormalizing") here as the constants are important in hypotheses (iii) (b) and (c) of Theorem 3.5.3. Then (iii) (a) holds. For (ii), if $U \subset I\!R$ is open, then it contains an interval say (μ_1, μ_2), and

$$\Pi_n(U) \;\; \geq \;\; \Pi_n((\mu_1, \mu_2)) \;\; = \;\; a_n \int_{\mu_1}^{\mu_2} (2\pi n)^{-1/2} \, \exp\left[-\frac{1}{2} n \, \mu^2 \right] \, dmu$$

$$\geq\ a_n n^{-1/2}\,(2\pi)^{-1/2}\int_{\mu_1}^{\mu_2}\exp\left[-\frac{1}{2}\,\mu^2\right]\,dmu \quad ,$$

where in the last inequality we used the elementary monotonicity properties of the exponential. Thus,

$$\Pi_n(U)\ \geq\ C\,a_n n^{-1/2}$$

where C is a positive constant. Condition (iii) (b) of Theorem 3.5.3 will hold as long as

$$\liminf_n\ a_n/n^{1/2}\ >\ 0 \quad . \tag{3.198}$$

For (iii) (c), we can easily derive the Bayes estimator $\delta_n^\star$, and hence the $\inf_\delta$ $r(\Pi_n,\delta)$, and the result is

$$\inf_\delta\ r(\Pi_n,\delta)\ =\ r(\Pi_n,\delta_n^\star)\ =\ a_n\,\frac{1}{1+1/n}$$

and hence

$$r(\Pi_n,\delta^\star)\ -\ \inf_\delta\ r(\Pi_n,\delta)\ =\ a_n[1-(1+n^{-1})^{-1}]\ =\ \frac{a_n}{n+1} \quad .$$

Thus, (iii) (c) of Theorem 3.5.3 holds if

$$\liminf_n\ a_n/n\ =\ 0 \quad . \tag{3.199}$$

We can satisfy both (3.198) and (3.199) by taking

$$a_n\ =\ n^p \quad , \quad 1/2\leq p<1 \quad .$$

Then Theorem 3.5.3 applies to show that X is admissable.

□

3.5.5 The Posterior Mode.

Suppose we have a dominated family with densities $\{f(x|\theta):\theta\in\Theta\}$ with $\Theta\subset I\!R^p$ and a prior $\Pi\ll m^p$ with Lebesgue density $\pi(\theta)$. A *posterior mode* is a value $\hat{\theta}$ which maximizes the posterior density. If it is the unique maximizer, we speak of *the* posterior mode.

Remarks 3.5.2 (a) Maximizing the (normalized) posterior density $\pi(\theta|x)=f(x|\theta)\pi(\theta)/\int f(x|\vartheta)\pi(\vartheta)d\vartheta$ is equivalent to maximizing the unnormalized posterior $f(x|\theta)\pi(\theta)$. Note that it is no harder to compute the unnormalized posterior than to compute the likelihood $f(x|\theta)$ and the prior density $\pi(\theta)$ and multiply the two together, and so the maximization can often be done numerically with ease, in contrast to the computation of a formal Bayes estimator which requires calculation of a p dimensional integral for each evaluation of the objective function.

(b) It is often easier to work with the logarithm of the posterior. For instance, if $\underline{X} = (X_1, \ldots, X_n)$ where the X_i's are i.i.d. with density $f(x|\theta)$, then maximizing the posterior is equivalent to minimizing

$$-\log \pi(\theta|\underline{X}) \ = \ \sum_{i=1}^{n} -\log f(X_i|\theta) \ + \ (-\log \pi(\theta)) \quad .$$

Note that if $-\log f(X_i|\theta)$ is convex in θ (which happens, for instance, if θ is the natural parameter in an exponential family), $-\log \pi(\theta)$ is convex in θ, and one of them is strictly convex, then $-\log \pi(\theta|\underline{X})$ is strictly convex and then there is at most one mode. We will return to this shortly.

(c) The posterior mode is very dependent on the choice of dominating measure for the prior. We only defined it in the case of a Lebesgue density for the prior (or in Exercise 3.5.12, for a prior density w.r.t. counting measure). If we change from m^p to another measure, say μ, then we need to multiply by $dm^p/d\mu$, and after taking negative logarithms, the objective function to be minimized becomes

$$\sum_{i=1}^{n} -\log f(X_i|\theta) \ + \ (-\log \pi(\theta)) \ + \ \left(-\log \frac{dm^p}{d\mu}(\theta)\right) \quad ,$$

which will in general be somewhat different.

(d) Sometimes, the posterior mode is referred to as the *maximum a posteriori* or MAP estimator.

□

In Exercise 3.5.12 it was seen that the posterior mode is the optimal Bayes estimator under $0-1$ loss assuming a discrete parameter space, but here we are interested in a continuous parameter space. We can show that the posterior mode is a limit of Bayes estimators for a family of loss functions that mimic $0-1$ loss. Specifically, we will lose 0 if the estimate is "close" to the true value, and otherwise lose 1, and then let "closeness" tend to 0.

Proposition 3.5.5 *Let π be a Lebesgue density for a generalized prior. Assume the following:*

(P1) *The posterior is proper. Denote the posterior Lebesgue density by $\pi(\theta|X)$.*

(P2) *$\pi(\theta|X)$ is a continuous function of θ.*

(P3) *The posterior mode $\hat{\theta}$ exists and satisfies the following strong uniqueness condition:*

$$\forall \delta > 0, \quad \mu(\delta) \ = \ \sup\{\pi(\theta|X) : \theta \in \Theta \textit{ and } \|\theta - \hat{\theta}\| > \delta\} \ < \ \pi(\hat{\theta}|X) \quad .$$

(P4) *$\hat{\theta}$ is an interior point of Θ.*

Let $A \subset I\!R^p$ be measurable and for $r > 0$ define the loss function

$$L_r(\theta, d) \;=\; 1 \;-\; I_A((\theta - d)/r) \quad . \tag{3.200}$$

Assume the following:

(L1) *A is bounded i.e.*

$$\exists M < \infty \text{ such that } \forall \theta \in A, \quad \|\theta\| \leq M.$$

(L2) *$m^p(A) > 0$.*

Then there exists $r_0 > 0$ such that for all $r > 0$ with $r \leq r_0$, a generalized Bayes estimate under the loss L_r exists. Letting $\hat{\theta}_r$ denote any such Bayes estimator, we have

$$\hat{\theta}_r \to \hat{\theta} \text{ as } r \to 0. \tag{3.201}$$

Proof. Consider the posterior expected loss

$$\rho_r(d) \;=\; \int L_r(\theta, d) \pi(\theta|X)\, d\theta.$$

Here, the value of X is fixed throughout, so we do not bother to show it as an argument of the functions. Note that since $0 \leq L_r(\theta, d) \leq 1$, $\rho_r(d) \leq 1$ by (P1). For convenience, define for $d \in I\!R^p$ and $r \geq 0$ the set

$$\begin{aligned} d \,+\, rA \;&=\; \{\, \theta \in I\!R^p \,:\, \theta = d + r\alpha \text{ for some } \alpha \in A \,\} \\ &=\; \{\, \theta \in I\!R^p \,:\, (\theta - d)/r \in A \,\} \quad . \end{aligned}$$

Note that by the fact that Lebesgue measure is translation invariant

$$m^p(d + rA) \;=\; m^p(rA) \quad .$$

Also, since $m^p(A) > 0$, we have $m^p(rA) > 0$. (In fact, $m^p(rA) = r^p m^p(A)$.)

Let $\lambda > 0$ be given. Using continuity of the posterior density in (P2), find $\eta(\lambda)$ such that

$$\|\theta - \hat{\theta}\| < \eta(\lambda) \;\Rightarrow\; \pi(\theta|X) \;>\; \frac{1}{2}[\pi(\hat{\theta}|X) + \mu(\lambda/2)] \quad .$$

where $\mu(\cdot)$ is given in (P3). Clearly, $\eta(\lambda) < \lambda/2$. Now, if $r < \eta(\lambda)/M$ and $\|d - \hat{\theta}\| > \lambda$, then $\|d - \hat{\theta}\| > \lambda/2 + Mr$ so $(\theta - d)/r \in A$ implies $\|\theta - \hat{\theta}\| > \lambda/2$ and hence

$$\begin{aligned} \rho_r(d) \;&=\; 1 \;-\; \int_{(d+rA)\cap\Theta} \pi(\theta|X)\, d\theta \\ &\geq\; 1 \;-\; \mu(\lambda/2) m^p(rA) \quad . \end{aligned}$$

But if λ is small enough that we may assume $\|\theta - \hat{\theta}\| < \eta(\lambda)/M$ implies $\theta \in \Theta$ (i.e. the $\eta(\lambda)/M$ neighborhood of $\hat{\theta}$ is contained in Θ), which is possible by (P4), then

$$\begin{aligned} \rho_r(\hat{\theta}) &= 1 - \int_{(\hat{\theta}+rA)\cap\Theta} \pi(\theta|X)\, d\theta \\ &\leq 1 - \frac{1}{2}[\pi(\hat{\theta}|X) + \mu(\lambda/2)])m^p(rA) \quad . \end{aligned}$$

Since $\pi(\hat{\theta}|X) > \mu(\lambda/2)$, for such λ we have for all r sufficiently small that $\rho_r(d) > \rho_r(\hat{\theta})$ when $\|d - \hat{\theta}\| > \lambda$, i.e. any minimizer of ρ_r must be in a λ–neighborhood of $\hat{\theta}$. Since λ can be taken arbitrarily small, this shows that any sequence of Bayes estimators must converge to the posterior mode.

To show that a Bayes estimator exists under the loss L_r for all r sufficiently small, one can first show that the posterior expected loss is continuous (use continuity of the posterior density and boundedness of the set A along with the fact that a continuous function on a closed and bounded set in $I\!R^p$ is uniformly continuous), and then use the fact that a continuous function on a closed and bounded set achieves its maximum. By taking r sufficiently small, we can use the argument above to restrict attention to some neighborhood of $\hat{\theta}$.

□

Note that the posterior mode depends on the dominating measure used for the prior density (we specifically used Lebesgue measure for the dominating measure above, but one can consider the posterior mode under a prior density w.r.t. other dominating measures).

Now we present a formal, decision theoretic justification for maximum likelihood in a very simplified setting. Suppose $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ is a finite set, that we use a uniform prior (i.e. the prior probability $\pi(\theta_i) = 1/k$ is the same for each possible value), and we use a 0–1 loss function (i.e. $L(\theta, d) = 0$ if $d = \theta$ and otherwise $L(\theta, d) = 1$). Then we claim the MLE is the Bayes estimator (Exercise 3.5.12). Referring back to Proposition 3.5.5, suppose the MLE exists and satisfies a condition like the strong uniqueness condition in (P3) o that Proposition. Then if we use a uniform prior on a bounded subset of Θ which includes the MLE, then the posterior mode will be the MLE. Thus, we see that the MLE satisfies some optimality "in the limit."

Note that the posterior mode and MLE have an advantage over a optimal (decision theoretic) Bayes estimator as in (3.155) above in the one need only solve an optimization problem in contrast to the numerical integration in high dimensions that would typically be required to find Bayes estimates.

Exercises for Section 4.2.

3.5.1 (a) Verify (3.153) gives the Bayes estimator of $g(\theta)$ under squared error loss.

(b) Assuming $u(\theta) > 0$, find the Bayes estimator for $g(\theta)$ under the *normalized* or *weighted squared error loss*

$$L(\theta, d) = \left[\frac{d - g(\theta)}{u(\theta)} \right]^2 .$$

State what moments must be finite for the Bayes risk to be finite, and what moments must be finite for the estimator to be finite.

(c) Assuming $g(\theta) > 0$, find the Bayes estimator for $g(\theta)$ under the *relative squared error loss* obtained by putting $u(\theta) = g(\theta)$ in (b), and answer the questions regarding finiteness of moments.

(d) Find Bayes estimators in the setting of Example 3.5.1 for the following estimands under the normal–inverse *Gamma* prior and the following loss functions:

(i) Estimate μ under the normalized squared error loss $L((\mu, \sigma^2), d) = (d - \mu)^2/\sigma^2$.

(ii) Estimate μ^2 under squared error loss.

(iii) Estimate σ^2 under relative squared error loss.

In each of (i), (ii), and (iii), state any conditions on the prior parameters necessary for the Bayes risk to be finite and any conditions necessary for the estimators to be finite.

3.5.2 Assuming a dominated family, show that a Bayes rule is a function of any sufficient statistic. Conclude that we can replace the observation with any sufficient statistic.

3.5.3 Suppose $\Pi \ll \nu$ and $\mathbf{P} \ll \mu$ where ν and μ are σ–finite. Show that the posterior Law$[\theta | X = x] \ll \nu$ and give a formula for the posterior density w.r.t. ν.

3.5.4 Suppose X, Y, and Z are random vectors with joint density $f(x, y, z)$ w.r.t. some product of σ–finite measures $\mu \times \nu \times \rho$. Show that the conditional density of (X, Y) given Z can be factored as $f(x, y|z) = f(x|y, z)f(y|z)$.

3.5.5 (a) Verify equations (3.162), (3.166), (3.168), (3.169), (3.171), (3.172), (3.174), (3.175), (3.176), (3.177), (3.178), and (3.179).

(b) Check that setting $n = 0$ in (3.169) and (3.172) causes the posterior to reduce to the prior.

(c) Show that $E[\mu|\underline{X}] = \hat{\mu}$ provided that $(n + 2\alpha)/2 > 1$. What condition is required for $E[\mu^2] < \infty$?

(d) Verify (3.180) and (3.181).

3.5.6 (a) Verify (3.182).

(b) Show that the bias tends to 0 as $n \to \infty$.

(c) Show that the bias tends to 0 as $\nu^2 \to \infty$.

(d) Assuming $\sigma^2 = 1$, $n = 10$, $0 < \nu < 100$, find a lower bound on τ^2 so that the bias in (3.182) is $< 10^{-4}$ for any value of μ in the range $0 < \mu < 100$.

(d) If we want to make the bias $< 10^{-10}$, under the same assumptions as in (c), how large must we take τ^2?

3.5.7 (a) Suppose (X, Y) are jointly distributed random variables, and that $E[X^2] = \infty$. Then $E[(X - Y)^2] = \infty$.

(b) Show that part (a) implies that $E[g(\theta)^2] = \infty$ implies $r(\Pi, \delta) = \infty$ under squared error loss for any estimator δ.

(c) Give results similar to (b) for normalized and relative squared error loss.

3.5.8 For a random vector $\underline{Y}$ with nonsingular covariance matrix, define

$$\text{Precision}[\underline{Y}] \;=\; (\,\text{Cov}[\underline{Y}]\,)^{-1} \quad .$$

Now let $\underline{X}_1$, $\underline{X}_2$ .., $\underline{X}_n$ be i.i.d. random d-vectors with distribuion $N(\underline{\mu}, \sigma^2 V)$ where V is a known positive definite definite matrix, and $\underline{\mu} \in I\!\!R^d$ and $\sigma^2 > 0$ are unknown. Assume a "convenience" prior for $(\underline{\mu}, \sigma^2)$ of the form

$$\text{Law}[\underline{\mu}|\sigma^2] \;=\; N(\underline{\nu}, c^{-1}\sigma^2 V) \quad , \quad \text{Law}[1/\sigma^2] \;=\; \mathit{Gamma}(\alpha, \beta) \quad .$$

Here, $\nu \in I\!\!R^n$, $c > 0$, $\alpha > 0$, and $\beta > 0$ are parameters of the prior. Let $\underline{Y}$ denote the entire $n \times d$ dimensional observation vector $(\underline{X}_1, \underline{X}_2, .., \underline{X}_n)$.

(a) Find the conditional posterior $\text{Law}[\underline{\mu}|\sigma^2, \underline{Y}]$. State and prove analogues of (3.164) and (3.165) along the way. Explicitly check that (i) your posterior is a genuine probability density function, and (ii) if $n = 0$ the posterior reduces to the prior.

(b) Find the marginal posterior $\text{Law}[\sigma^2|\underline{Y}]$. Make the same checks (i) and (ii) as in part (a).

(c) Find Bayes estimates of σ^2 under squared error loss and relative squared error loss. Determine the ranges of the prior parameter values which make these losses meaningful and the ranges which make the estimates meaningful.

(d) Find Bayes estimates for $g(\underline{\mu}, \sigma^2) = \underline{\gamma}'\underline{\mu}$ under squared error loss and the normalized squared error loss

$$L((\underline{\mu}, \sigma^2), d) \;=\; \left(\frac{d - \underline{\gamma}'\underline{\mu}}{\sigma} \right)^2 \quad .$$

Here, $\underline{\gamma} \in I\!\!R^n$ is given.

(e) Find a Bayes estimate for $g(\underline{\mu}, \sigma^2) = \underline{\gamma}'\underline{\mu}/\sigma$ under squared error loss where $\underline{\gamma} \in I\!\!R^n$ is given.

(f) Show that the prior family chosen here is a conjugate family for this sampling distribution.

3.5.9 Suppose X is $B(n,p)$ with p unknown. Assume a $Beta(\alpha,\beta)$ prior for p.

(a) Show that this family of priors is a conjugate family for the given family of sampling distributions.

(b) Find Bayes estimates for p^m under squared error and relative squared error loss where $m \neq 0$. State conditions on the prior parameters necessary for the Bayes risk to be finite and conditions necessary for the estimator to be defined and finite.

(c) Determine, if you can, values of the prior parameters (possibly limiting values) for which the Bayes estimator of p^m under squared error loss equals the UMVUE, when that exists.

3.5.10 (a) Let X be a $Poisson(\mu)$ random variable with $\mu > 0$ unknown. Find a conjugate family of priors on μ. Determine the corresponding posterior. Be sure to perform the "checks" (i) and (ii) in Exercise 3.5.8 (a).

(b) Find Bayes estimators for μ under squared error loss and relative squared error loss. Determine which ranges of the prior parameters give meaningful Bayes risks and estimators.

(c) What if we have i.i.d. observations X_1, X_2, .., X_n which are each $Poisson(\mu)$? (Hint: using Exercise 3.5.2, you should be able to avoid any hard work.)

3.5.11 Suppose X_1, X_2, ..., X_n are i.i.d. $Gamma(\alpha,\beta)$ random variables.

(a) Find a conjugate family of prior distributions for the parameters α and β.

(b) Find Bayes estimates for as general estimands as you can handle under (regular, normalized, or relative) squared error losses.

3.5.12 (a) Suppose $g(\theta)$ is a discrete estimand, i.e. there is a finite or infinite sequence a_1, a_2, .. such that $g(\theta) \in \{a_1, a_2, ..\}$ for all θ. Otherwise said, g may be written in the form

$$g(\theta) \;=\; \sum_i a_i I_{\Theta_i}(\theta)$$

where the a_i are distinct I R numbers and $\{\Theta_i : i = 1, 2, ..\}$ is a finite or infinite sequence of disjoint subsets of Θ whose union is Θ.

Consider the $0-1$ loss

$$L(\theta, d) \;=\; \begin{cases} 0 & \text{if } d = g(\theta) \quad , \\ 1 & \text{if } d \neq g(\theta) \quad . \end{cases}$$

Let

$$\begin{aligned} B_j \;=\; & \{\, x \in \Xi : \\ & P[\,\theta \in \Theta_j \,|\, X = x\,] \;\geq\; P[\,\theta \in \Theta_i \,|\, X = x\,] \text{ for all } i > j, \text{ and} \\ & P[\,\theta \in \Theta_j \,|\, X = x\,] \;>\; P[\,\theta \in \Theta_i \,|\, X = x\,] \text{ for all } i < j. \,\} \quad . \end{aligned}$$

That is, B_j is the set of observations wherein Θ_j has highest posterior probability, and j is minimal among indices of elements of the partition of parameter space with highest posterior probability. Show that

$$\delta(X) \; = \; \sum_i a_i \, I_{B_i}(X)$$

is a Bayes estimator. When is the Bayes estimator unique?

(b) Specialize the part (a) to $g(\theta) = \theta$ when Θ is a discrete set. Show that the Bayes estimator is a "posterior mode".

(c) Specialize (b) to the case where Θ is finite and Π is the discrete uniform distribution on Θ. Show that the Bayes estimator is the "maximum likelihood estimator".

(d) Specialize (c) to the case $\Theta = \{0, 1\}$. Show that in this case, we may express the posterior distribution as follows

$$\frac{P[\,\theta = 1\,|\,X\,]}{P[\,\theta = 0\,|\,X\,]} \; = \; \frac{f(X|1)}{f(X|0)} \times \frac{P[\,\theta = 1\,]}{P[\,\theta = 0\,]} \quad ,$$

assuming no divisions by 0. Note that the probability of an event divided by the probability of the complementary event is called the *odds (ratio)* of the event. Also, the ratio $f(X|1)/f(X|0)$ is called a *likelihood ratio.* The above may be stated as "the posterior odds equals the prior odds times the likelihood ratio."

3.5.13 Consider the following "application"of Exercise 3.5.12. At a murder scene, some tissue of the murderer has been found and a forensic expert says it is type A. It is known that a proportion p of the population has this tissue type. A suspect is arrested and his tissue is tested. At the trial, various evidence is presented regarding motive, alibi, and so forth and a juror decides that based on this (nonquantifiable) evidence alone, there is a prior probability π that the suspect is guilty. Now, the forensic expert presents the "data", which is the tissue type of the suspect. Our parameter space is $\Theta = \{$ "the suspect is guilty", "a random person is guilty" $\}$. Show that the posterior odds of guilt, say $g/(1-g)$ where g denotes the posterior probability of guilt, may be determined from

$$\frac{g}{1-g} \; = \; \begin{cases} \pi/(p[1-\pi]) & \text{if suspect's type is A,} \\ \\ 0 & \text{if suspect's type is not A.} \end{cases}$$

Comment on the importance of various assumptions in the model and possible violations of those assumptions which may cause serious problems. Also, state why it is important that π be determined without reference to the tissue type data.

In American courts, it is required that the suspect be proved guilty beyond a "reasonable shadow of a doubt." Suppose we interpret this to mean that the posterior probability of guilt must be $\geq$ 95%. If the proportion of tissue type A in the population is 20%, find a lower bound on the prior odds of guilt (based on the nonquantifiable evidence) in order that the subject be found guilty by this standard. Translate this lower bound on odds into a lower bound on probability.

For more discussion of these ideas, see Finkelstein and Fairly, "A Bayesian approach to identification evidence," *Harvard Law Review*, volume 83, pages 489-517, 1970.

3.5.14 (a) Suppose we use the improper prior m^2 on $(\mu, \log \sigma^2) \in I\!R^2$ in the framework of Example 3.5.2. Show that this leads to the improper prior with density (3.184) for the ordinary parameterization (μ, σ^2).

(b) Verify that the posterior in (3.186) is a Lebesgue probability density if and only if $n \geq 2$.

(c) Verify (3.189).

(d) Verify (3.190), but show that the analogous inequality with sampling distributions doesn't hold.

3.5.15 Given $\epsilon > 0$, a rule $\delta^\star$ is called ϵ-Bayes if there is a proper prior Π such that

$$r(\Pi, \delta^\star) \;\leq\; \epsilon + \inf_{\delta} r(\Pi_n, \delta) \quad .$$

Show that $\delta^\star$ is extended Bayes if and only if it is ϵ–Bayes for every $\epsilon > 0$.

3.5.16 Show that a strict Bayes rule is extended Bayes.

3.5.17 Show that if any decision rule has finite generalized Bayes risk, then the generalized Bayes rule has finite generalized Bayes risk.

3.5.18 Verify that in the proof of Theorem 3.5.4 for $d = 1$, there is no loss of generality in assuming that $\sigma^2 = 1$ and $n = 1$.

3.5.19 Show that the risk function in the proof of Theorem 3.5.4 is continuous and infinitely differentiable in $\underline{\mu}$.

3.5.20 Let X be $B(n, p)$ with $0 < p < 1$ unknown. Consider priors of the form

$$\pi_{p,J}(p) \;\propto\; [p(1-p)]^r \quad , \quad 0 < p < 1 \quad ,$$

and loss functions of the form

$$L_0(p, d) \;=\; \left[\frac{d - p}{[p(1-p)]^v} \right]^2 \quad .$$

State conditions under which (i) the prior is proper; (ii) the posterior is proper; (iii) the generalized Bayes estimator exists; (iv) the generalized Bayes estimator has finite Bayes risk, and (v) the generalized Bayes estimator is extended Bayes or proper Bayes. Give formulae for the Bayes estimator, ordinary risk, and Bayes risk where possible.

3.5.21 Show that a proper Bayes estimator δ under squared error loss is unbiased if and only its Bayes risk is 0. Explain why this will never happen in a practical statistical problem.

3.5.22 (a) Suppose $\underline{X}$ is a $N(\underline{\mu}, I)$ random d-vector. Show that under sum of squared errors loss (see Theorem 3.5.4), $\underline{X}$ is generalized Bayes, extended Bayes, UMVUE (in the sense that it uniformly minimizes the sum of mean squared errors among unbiased estimators; you will have to reason with ad hoc methods or extend the theory of UMVUE estimators from Chapter 4).

(b) True or false (i.e. give proof or counterexample):

(i) A generalized Bayes rule is admissable.

(ii) An extended Bayes rule is admissable.

(iii) A UMVUE (in the sense above) is admissable.

3.6 Minimax Estimation.

Consider a decision theoretic setup with loss function $L(\theta, d)$. The *minimax risk* is

$$\inf_{\delta} \sup_{\theta} \; R(\theta, \delta) \quad ,$$

where the infimum is over all allowable decision rules δ and the supremum is over all parameter values. One must be careful that inf and sup are taken in the correct order, but the word "minimax" itself gives the order: the "mini" of "minimax" refers to the infimum and the "max" to the supremum. If $\delta^\star$ is an allowable decision rule whose maximum risk equals the minimax risk, then we say $\delta^\star$ is a *minimax rule*. That is, $\delta^\star$ is minimax if and only if

$$\sup_{\theta} \; R(\theta, \delta^\star) \; = \; \inf_{\delta} \sup_{\theta} \; R(\theta, \delta) \quad . \tag{3.202}$$

Note that the minimax risk is always defined, even if a minimax rule does not exist. Why might a minimax rule be desired? For any rule δ the *maximum risk* $\sup_\theta R(\theta, \delta)$ is a quantity which characterizes in some sense the performance of that rule. Thus, a minimax rule is one which optimizes (minimizes) that performance criterion.

Now the problem of finding a minimax rule is one of minimizing over δ a single real valued function, namely the maximum risk $\sup_\theta R(\theta, \delta)$. This is in contrast with the problem of finding a uniform minimum risk rule, i.e. one which simultaneously minimizes risk over all parameter values. Thus, the principle of minimaxity is similar in spirit to the principle of Bayesian decision theory – we don't try to find a rule which is simultaneously optimal for a lot of functions $R(\theta, \cdot)$, but one which is optimal for a single objective – average risk (under the prior) for a Bayes rule and maximum risk for a minimax rule. Similarly to the principle of Bayesian decision theory, in seeking minimax rules we usually don't restrict the class of rules (e.g. by unbiasedness), but search over all possible rules (subject to measurability, of course).

3.6.1 Least Favorable Prior Distributions.

It turns out that Bayesian methods are very useful for finding minimax rules. For one thing, the Bayes risk of a proper Bayes rule always gives a lower bound on minimax risk. To see this, let Π be a proper prior and δ any decision rule, then

$$\begin{aligned} r(\Pi, \delta) &= \int_\Theta R(\theta, \delta)\, d\Pi(\theta) \\ &\leq \int_\Theta \sup_{\vartheta \in \Theta} R(\vartheta, \delta)\, d\Pi(\theta) \qquad (3.203) \\ &= \left[\sup_{\vartheta \in \Theta} R(\vartheta, \delta) \right] \int_\Theta 1 \, d\Pi(\theta) \\ &= \sup_{\vartheta \in \Theta} R(\vartheta, \delta) \quad , \end{aligned}$$

i.e. an average is always less than the maximum. Hence, taking infimums over allowable decision rules we obtain

$$\inf_\delta r(\Pi,\delta) \;\leq\; \inf_\delta \sup_\theta \; R(\theta,\delta) \quad . \tag{3.204}$$

Of course, if δ_Π is a Bayes rule for this prior, then $\inf_\delta r(\Pi,\delta) = r(\Pi,\delta_\Pi)$, so

$$r(\Pi,\delta_\Pi) \;\leq\; \inf_\delta \sup_\theta \; R(\theta,\delta) \quad . \tag{3.205}$$

But it is not necessary for a Bayes rule to exist in order that (3.204) be valid.

In the computation (3.203), we have equality if and only if $\Pi(\{\theta : R(\theta,\delta) < \sup_\vartheta \; R(\vartheta,\delta)\}) = 0$. Applying this to the Bayes rule, we have

$$r(\Pi,\delta_\Pi) \;=\; \sup_\theta \; R(\theta,\delta_\Pi) \tag{3.206}$$

if and only if

$$R(\theta,\delta_\Pi) \;=\; \sup_\vartheta \; R(\vartheta,\delta_\Pi) \quad , \quad \text{for } \Pi\text{–almost all } \theta. \tag{3.207}$$

We claim that if (3.206) holds, then in fact δ_Π is a minimax rule, for if there is a rule δ with strictly smaller maximum risk, it would have better Bayes risk than δ_Π, i.e. if $\sup_\theta \; R(\theta,\delta) < \sup_\theta \; R(\theta,\delta_\Pi)$, then

$$\begin{aligned} r(\Pi,\delta) &\leq \sup_\theta R(\theta,\delta) && \text{(by (3.203))} \\ &< \sup_\theta \; R(\theta,\delta_\Pi) && \\ &= r(\Pi,\delta_\Pi) && \text{(by (3.206))} \end{aligned}$$

which contradicts that δ_Π is a Bayes rule for Π. So, if we can find a proper prior $\Pi^\star$ whose Bayes rule $\delta^\star$ has Bayes risk equal to its maximum risk, then $\delta^\star$ is a minimax rule.

While the calculations above which show this fact are simple and elegant, there is more intuition behind it. Suppose $\Pi^\star$ is a proper prior whose Bayes rule $\delta^\star$ has Bayes risk equal to its maximum risk. Since the Bayes risk of $\delta^\star$ under $\Pi^\star$ equals the minimax risk, we see from (3.204) that the minimal Bayes risk under $\Pi^\star$ (which is after all $r(\Pi^\star,\delta^\star)$) is largest among minimal Bayes risks over all priors, i.e.

$$\sup_\Pi \inf_\delta r(\Pi,\delta) \;=\; \inf_\delta r(\Pi^\star,\delta) \quad . \tag{3.208}$$

Note that the l.h.s. of this last equation is “maximin” Bayes risk. Otherwise said, $\Pi^\star$ has the worst optimal Bayes risk. When (3.208) holds, we say the prior $\Pi^\star$ is *least favorable.* To explain this terminology, think of the decision problem as a game with opponents Nature vs. the Statistician (or Decision Maker). Nature gets to choose a proper prior Π which is revealed to the Statistician, and the

Statistician then gets to choose a decision rule δ, which of course will be the Bayes rule for that prior (since the Statistician is no dummy and wants to minimize his or her risk). The the Statistician will have the highest possible risk if Nature chooses the least favorable prior, so we see that the prior is "least favorable" for the Statistician.

The above then gives us a strategy for finding minimax rules: try to guess a least favorable prior, find its Bayes rule, and then see if (3.206) holds for that rule. Note that by (3.207), the risk of such a rule is almost surely constant under the prior. A stronger notion is that its risk function is constant everywhere, not just Π–almost everywhere. A decision rule with constant risk (over the entire parameter space) is called an *equalizer rule.* Thus, a proper Bayes rule which is an equalizer rule is minimax.

Example 3.6.1 Let X_1, X_2, …, X_n be i.i.d. with the Fisher–von Mises distribution $FM(\alpha, \phi)$, $\alpha > 0$ and $\phi \in [0, 2\pi)$, which has Lebesgue density

$$f_{\alpha,\phi}(x) \;=\; \frac{1}{I_0(\alpha)} \exp\left[\alpha \cos(x - \phi)\right] \quad , \quad 0 \le x < 2\pi.$$

Here

$$I_0(\alpha) \;=\; \int_0^{2\pi} e^{\alpha \cos u}\, du$$

is a so called modified Bessel function of order 0. This is a common model for angular data, i.e. when the X_i's are angles measured in radians. For example, the X_i's might be replicated measurements of the angle between two surveying stations as measured by a surveyor's transit. The density has a mode at $x = \phi$, and α controls the spread of the density about the mode: larger values of α correspond to a tighter (narrower) spread. Geometrically, a value of X_i is best thought of as a point on the unit circle. Note in particular that $x = .05$ radians and $x = 2\pi - .07$ radians are really just .12 radians apart, not $2\pi - .12$ radians apart. We must remember this when specifying a loss function.

Some other simple properties of this family of distribution will be useful below. First, for x and y in $[0, 2\pi)$, define *mod* 2π *addition* and *subtraction* by

$$x \oplus y \;=\; \begin{cases} x + y & \text{if } x + y < 2\pi, \\ x + y - 2\pi & \text{if } x + y \ge 2\pi, \end{cases}$$

and

$$x \ominus y \;=\; \begin{cases} x - y & \text{if } x - y \ge 0, \\ x - y + 2\pi & \text{if } x - y < 0. \end{cases}$$

Note that this is basically addtion of angles, but we "wrap" around the circle to always keep the value between 0 and 2π. Now, we claim that if $X \sim FM(\alpha, \phi)$, then $X \oplus \varphi \sim FM(\alpha, \phi \oplus \varphi)$, and similarly for $X \ominus \varphi$. (Exercise 3.6.2).

We will assume $\alpha > 0$ is known and consider the problem of estimating ϕ under the loss

$$L(\phi, d) \;=\; 1 - \cos(\phi - d) \quad .$$

Note that this loss has the properties that for $\phi - d$ close to 0, it is approximately the same as squared error loss (look at the Taylor series expansion for $\cos u$) and that as $\phi - d$ increases to values bigger than π, it "wraps" around to reflect the fact that ϕ and d are now getting closer.

It is reasonable to conjecture that a least favorable prior for ϕ is $Unif(0, 2\pi)$, which means we don't favor any part of $[0, 2\pi)$ over any other part, i.e. "complete ignorance." We will compute the Bayes estimator under this prior and see if it has the requisite properties, e.g. is an equalizer rule. Now the posterior density is

$$\begin{aligned}
p(\phi|\underline{X}) &\propto \exp\left[\alpha \sum_{i=1}^{n} \cos(X_i - \phi)\right] \\
&= \exp\left[\alpha \sum_{i=1}^{n} (\cos(X_i)\cos(\phi) + \sin(X_i)\sin(\phi))\right] \\
&= \exp\left[\alpha \left(\cos\phi \sum_{i=1}^{n} \cos X_i \;+\; \sin\phi \sum_{i=1}^{n} \sin X_i\right)\right] \\
&= \exp\left[\alpha\beta\left(\cos(\phi)\cos(\hat{\phi}) + \sin(\phi)\sin(\hat{\phi})\right)\right] \\
&= \exp\left[\alpha\beta\cos(\phi - \hat{\phi})\right] \quad ,
\end{aligned}$$

where

$$\begin{aligned}
\hat{\phi} &= \operatorname{Arg}\left(\sum_{i=1}^{n} \cos X_i, \sum_{i=1}^{n} \sin X_i\right) \\
\beta &= \left[\left(\sum_{i=1}^{n} \cos X_i\right)^2 + \left(\sum_{i=1}^{n} \sin X_i\right)^2\right]^{1/2}
\end{aligned}$$

are the argument and length of the vector $(\sum_{i=1}^{n} \cos X_i, \sum_{i=1}^{n} \sin X_i)$, respectively. Here, the argument of a vector is the counter clockwise angle it makes with the positive x–axis, i.e.

$$\operatorname{Arg}(x, y) \;=\; \begin{cases}
\tan^{-1}(y/x) & \text{if } x > 0 \text{ and } y \geq 0, \\
\pi/2 & \text{if } x = 0 \text{ and } y > 0, \\
\tan^{-1}(y/x) + \pi & \text{if } x < 0 \text{ and } y \neq 0, \\
-\pi/2 & \text{if } x = 0 \text{ and } y < 0, \\
\tan^{-1}(y/x) + 2\pi & \text{if } x > 0 \text{ and } y \leq 0, \\
\text{undefined} & \text{if both } x = 0 \text{ and } y = 0.
\end{cases}$$

Thus, we recognize that the posterior of ϕ is $FM(\alpha\beta, \hat{\phi})$. An expression for an unnormalized posterior expected loss is

$$\begin{aligned}
\rho(\underline{x}, d) &= \int_0^{2\pi} [1 - \cos(d - \phi)] \exp\left[\alpha\beta\cos(\phi - \hat{\phi})\right] d\phi \\
&= I_0(\alpha\beta) \;-\; \int_0^{2\pi} \cos(d - \phi) \exp\left[\alpha\beta\cos(\phi - \hat{\phi})\right] d\phi \\
&= I_0(\alpha\beta) \;-\; \cos(d - \hat{\phi}) \int_0^{2\pi} \cos(\hat{\phi} - \phi) \exp\left[\alpha\beta\cos(\phi - \hat{\phi})\right] d\phi \\
&\qquad\qquad\; - \; \sin(d - \hat{\phi}) \int_0^{2\pi} \sin(\hat{\phi} - \phi) \exp\left[\alpha\beta\cos(\phi - \hat{\phi})\right] d\phi \\
&= I_0(\alpha\beta) \;-\; \cos(d - \hat{\phi}) \int_0^{2\pi} \cos\vartheta \exp\left[\alpha\beta\cos\vartheta\right] d\vartheta \\
&\qquad\qquad\; - \; \sin(d - \hat{\phi}) \int_0^{2\pi} \sin\vartheta \exp\left[\alpha\beta\cos\vartheta\right] d\vartheta \\
&= I_0(\alpha\beta) \;-\; \cos(d - \hat{\phi}) \int_0^{2\pi} \cos\vartheta \exp\left[\alpha\beta\cos\vartheta\right] d\vartheta \quad ,
\end{aligned}$$

where the last equation follows from a simple symmetry argument (the integrals from 0 to π and from π to 2π are equal in magnitude and opposite in sign). Now

$$\begin{aligned}
&\int_0^{2\pi} \cos\vartheta \exp\left[\alpha\beta\cos\vartheta\right] d\vartheta \\
&\quad = \quad 2\left[\int_0^{\pi/2} \cos\vartheta \exp\left[\alpha\beta\cos\vartheta\right] d\vartheta \;+\; \int_{\pi/2}^{\pi} \cos\vartheta \exp\left[\alpha\beta\cos\vartheta\right] d\vartheta\right] \\
&\quad = \quad 2\int_0^{\pi/2} \cos\vartheta \;(\exp\left[\alpha\beta\cos\vartheta\right] \;-\; \exp\left[-\alpha\beta\cos\vartheta\right])\; d\vartheta \\
&\quad > \quad 0 \quad ,
\end{aligned}$$

since $\cos\vartheta > 0$ for $0 < \vartheta < \pi/2$ and $e^a > e^{-a}$ for $a > 0$. Thus, we minimize the posterior expected loss over d by maximizing $\cos(d - \hat{\phi})$, i.e. by taking $d = \hat{\phi}$.

Now we will show that $\hat{\phi}$ is an equalizer. To do this, we write its functional dependence on the data and observe that shifting all the data by a fixed amount (mod 2π) shifts $\hat{\phi}$ by the same amount, i.e.

$$\hat{\phi}(X_1 \oplus \varphi, X + 2 \oplus \varphi, \ldots, X_n \oplus \varphi) \;=\; \hat{\phi}(X_1, X_2, \ldots, X_n) \;\oplus\; \varphi \quad .$$

Recalling that $\mathrm{Law}_{\alpha,\phi}[X_i \oplus \varphi] = \mathrm{Law}_{\alpha,\phi\oplus\varphi}[X_i]$, we have for the distribution of $\hat{\phi}$ that

$$\begin{aligned}
\mathrm{Law}_{\alpha,\phi}[\hat{\phi}(X_1, X_2, \ldots, X_n)] &= \mathrm{Law}_{\alpha,0}[\hat{\phi}(X_1 \oplus \phi, X_2 \oplus \phi, \ldots, X_n \oplus \phi) \\
&= \mathrm{Law}_{\alpha,0}[\hat{\phi}(X_1, X_2, \ldots, X_n) \oplus \phi] \quad ,
\end{aligned}$$

and so

$$\begin{aligned}
&E_{\alpha,\phi}[1 - \cos\left(\phi - \hat{\phi}(X_1, X_2, \ldots, X_n)\right)] \\
&\quad = \quad E_{\alpha,0}[1 - \cos\left(\phi - \left\{\hat{\phi}(X_1, X_2, \ldots, X_n) \oplus \phi\right\}\right)] \\
&\quad = \quad E_{\alpha,0}[1 - \cos\hat{\phi}(X_1, X_2, \ldots, X_n)] \quad ,
\end{aligned}$$

which doesn't depend on ϕ, i.e. the risk for $\hat{\phi}$ is a constant (independent of ϕ) and hence $\hat{\phi}$ is an equalizer. Since $\hat{\phi}$ is a Bayes rule, it then follows that it is minimax.

In Exercise 3.6.3 we consider various issues about this problem and other estimators which might seem reasonable at first glance. The estimator $\hat{\phi}$ derived here from minimaxity considerations is in fact probably the "best" estimator for this problem for general use.

□

3.6.2 Least Favorable Sequences of Priors.

Unfortunately, there are few instances where the above theory can be applied, because all of these arguments require that the prior be proper and a prior which is intuitively "least favorable" on a parameter space of infinite extent will typically be improper. However, it is easy enough to produce "in the limit" versions of the notions above and these prove to be very useful.

Theorem 3.6.1 *Suppose $\{\Pi_n : n \in I\!N\}$ is a sequence of proper priors and $\delta^\star$ is a decision rule such that*

$$\sup_{\theta\in\Theta} R(\theta, \delta^\star) \ = \ \lim_{n\to\infty} \inf_{\delta} r(\Pi_n, \delta) \quad . \tag{3.209}$$

Then $\delta^\star$ is minimax.

Proof. Suppose δ_1 is another decision rule with smaller maximum risk than $\delta^\star$, i.e. $\sup_\theta R(\theta, \delta_1) < \sup_\theta R(\theta, \delta^\star)$. Then, by (3.209) we have for some n that

$$\begin{aligned} \sup_\theta R(\theta, \delta_1) \quad &< \quad \inf_\delta r(\Pi_n, \delta) \\ &\leq \quad r(\Pi_n, \delta_1) \quad , \end{aligned}$$

which violates the fact that the average risk for δ_1 cannot exceed its maximum risk (equation (3.203)). Hence, no such δ_1 exists with smaller maximum risk than $\delta^\star$ and thus $\delta^\star$ is minimax.

□

Corollary 3.6.2 *An extended Bayes equalizer rule is minimax.*

□

Remarks 3.6.1 **(a)** Note that when (3.209) holds, no proper prior Π can have minimal Bayes risk $\inf_\delta r(\Pi, \delta)$ smaller than $\lim_n \inf_\delta r(\Pi_n, \delta)$ since then the maximum risk of $\delta^\star$ would be strictly less than the average risk $r(\Pi, \delta^\star)$. This justifies the following terminology: we say a sequence of proper priors $\{\Pi_n : n \in I\!\!N\}$ satisfying (3.209) for some rule $\delta^\star$ is a *least favorable sequence of priors.* This can provide some intuitive guidance in seeking the sequence of priors with which to construct a minimax estimator.

(b) Since $\sup_\theta R(\theta, \delta^\star) \geq r(\Pi_n, \delta^\star) \geq \inf_\delta r(\Pi_n, \delta)$ it follows that $\delta^\star$ is extended Bayes under the given sequence of priors (assuming that it has finite Bayes risk for all priors in the sequence). Since it's maximum risk equals its limit of Bayes risks, it follows as in (3.207) that $\lim_n \Pi_n(\{\theta : R(\theta, \delta^\star) < \sup_\vartheta R(\vartheta, \delta^\star)\}) = 0$, so again $\delta^\star$ is in some sense "almost" an equalizer rule. Very often, it will be an equalizer rule, so Corollary 3.6.2 will apply.

Example 3.6.2 Let X_1, X_2, ..., X_n be i.i.d. $N(\mu, \sigma^2)$ with μ unknown. We will consider first the case σ^2 known and use squared error loss. Then we know that $\bar{X}$ has constant risk σ^2/n. It follows from the a slight modification of the proof of Theorem 3.5.4 that $\bar{X}$ is extended Bayes, so it is minimax by Corollary 3.6.2. We will briefly sketch that modified proof here. First of all, we need to use proper priors (probability measures), so take $a_n = \sigma^2$ in the proof of that theorem, and thus we are using a sequence $\Pi_n = N(0, n\sigma^2)$ of priors. Computing

$$\lim_n \left[r(\Pi_n, \delta^\star) \; - \; \inf_\delta r(\Pi_n, \delta) \right] \; = \; \sigma^2[1 - (1 + 1/n)^{-1}] \; \rightarrow \; 0 \text{ as } n \rightarrow \infty.$$

Thus, $\bar{X}$ is extended Bayes. Note that our least favorable sequence of priors is one which becomes increasingly "noninformative" about μ – i.e. for any fixed n, the $N(0, n\sigma^2)$ has a mode at $\mu = 0$, but as $n \to \infty$, the variance tends to ∞ thus increasing the region in which we are a priori "likely" to find μ. Looked at differently, $\sqrt{n}\Pi_n$ tends to a multiple of Lebesgue measure (like a "uniform" distribution on $I\!\!R$), which puts equal "likelihood" on each interval of the same length, and it seems reasonable that such a state of indefiniteness would be "least favorable" to the statistician.

Now suppose σ^2 is unknown, and then the risk of $\bar{X}$ under squared error loss is no longer constant (because it depends on the unknown parameter σ^2). However, we can fix this in a simple if artificial way: change to the following "weighted" squared error loss,

$$L((\mu, \sigma^2), \delta) \; = \; \frac{(\delta - \mu)^2}{\sigma^2} \quad . \tag{3.210}$$

Then, the risk for $\bar{X}$ under this loss becomes the constant $1/n$, i.e. $\bar{X}$ is an equalizer rule. Showing it is extended Bayes is easy! Take the sequence of priors $N(0, n) \times \delta_1$ on $\Theta = I\!\!R \times (0, \infty) = \{(\mu, \sigma^2) : \mu \in I\!\!R \;\& \sigma^2 > 0\}$. Since this sequence of priors is one that corresponds to $\sigma^2 = 1$ known (i.e. the factor δ_1 in the prior corresponds to knowing for certain that $\sigma^2 = 1$), and since the weighted squared error loss function above with $\sigma^2 = 1$ reduces to squared error loss, we obtain the same result as when it is known that $\sigma^2 = 1$, i.e. that $\bar{X}$ is extended Bayes.

□

It is very convenient in this last example that we can use the result from known σ^2 to take care of the unknown σ^2 as well. In general, one wouldn't expect that in a multiparameter problem such a trick would work. There are at least two special features of the last example to note: the extended Bayes estimator that was derived under known σ^2 does not depend on the known value of σ^2 and its risk does not depend on the known value of σ^2, so that it continues to be an equalizer rule even when σ^2 is unknown. The student should look for this situation in multiparameter (or nonparametric) settings as it can save much labor.

Example 3.6.3 Now suppose X_1, X_2, ..., X_n are i.i.d. $Poisson(\mu)$ with $\mu \geq 0$ unknown. We will consider first the estimation of $b\mu$ for some constant $b > 0$ when $n = 1$, i.e. just consider a single $X \sim Poisson(\mu)$. In the end, we will show how to reduce the case for $n > 1$ to the case when $n = 1$.

We begin with squared error loss. The mean squared error of the UMVUE bX of $b\mu$ is $\mathrm{Var}_\mu[bX] = b^2\mu$, which is not constant. However, we can rectify this situation very easily (if perhaps artificially) by changing to the weighted squared error loss,

$$L(\mu, d) \; = \; \frac{(d - b\mu)^2}{\mu} \quad , \tag{3.211}$$

under which the risk of bX becomes the constant b^2. Now we need only show that bX is extended Bayes (with this loss) to show it is minimax. Under a proper $\Gamma(\alpha, \beta)$ prior with Lebesgue density of the form

$$\pi(\mu) \propto \mu^{\alpha-1} \exp[-\mu/\beta] \quad , \quad \mu > 0,$$

we have the posterior density

$$\pi(\mu|X) \propto \mu^{X+\alpha-1} \exp[-\mu(1 + \beta^{-1})] \quad , \quad \mu > 0,$$

which is $Gamma(X + \alpha, (1 + \beta^{-1})^{-1})$. Now the unnormalized posterior expected loss is

$$\begin{aligned} \rho(X, d) \quad &= \quad \int_0^\infty \frac{(d - b\mu)^2}{\mu} \mu^{X+\alpha-1} \exp[-\mu(1 + \beta^{-1})] \, d\mu \\ &= \quad d^2\Gamma(X + \alpha - 1)(1 + \beta^{-1})^{-(X+\alpha-1)} \\ &\qquad - 2bd\Gamma(X + \alpha)(1 + \beta^{-1})^{-(X+\alpha)} \\ &\qquad + b^2\Gamma(X + \alpha + 1)(1 + \beta^{-1})^{-(X+\alpha+1)} \quad , \end{aligned}$$

which gives the Bayes estimator

$$\begin{aligned}\delta_{\alpha,\beta}(X) &= -\frac{-2b\Gamma(X+\alpha)(1+\beta^{-1})^{-(X+\alpha)}}{2\Gamma(X+\alpha-1)(1+\beta^{-1})^{-(X+\alpha-1)}} \\ &= \frac{b(X+\alpha-1)}{1+\beta^{-1}} \quad .\end{aligned}$$

To compute Bayes risk, we first compute bias

$$\begin{aligned}\text{Bias}_\mu[\delta_{\alpha,\beta}] &= E\left[\frac{b(X+\alpha-1)}{1+\beta^{-1}} \quad \Big| \quad \mu\right] - \; b\mu \\ &= \frac{b(\mu+\alpha-1)}{1+\beta^{-1}} - \; b\mu \\ &= \frac{-b(\mu\beta^{-1}-\alpha+1)}{1+\beta^{-1}}\end{aligned}$$

and variance

$$\begin{aligned}\text{Var}_\mu[\delta_{\alpha,\beta}] &= \frac{b^2\text{Var}[X|\mu]}{(1+\beta^{-1})^2} \\ &= \frac{b^2\mu}{(1+\beta^{-1})^2}\end{aligned}$$

so the mean squared error is

$$\begin{aligned}\text{MSE}_\mu[\delta_{\alpha,\beta}] &= \frac{b^2[\mu+(\mu\beta^{-1}-\alpha+1)^2]}{(1+\beta^{-1})^2} \\ &= \frac{b^2\{\mu^2\beta^{-2}+\mu[1-2\beta^{-1}(\alpha-1)]+(\alpha-1)^2\}}{(1+\beta^{-1})^2}\end{aligned}$$

so under our $Gamma(\alpha,\beta)$ prior (for which $E[\mu]=\alpha\beta$and $E[\mu^{-1}]=1/[(\alpha-1)\beta]$), the Bayes risk is

$$\begin{aligned}&r(Gamma(\alpha,\beta),\delta_{\alpha,\beta}) \\ &= E\left[\frac{\text{MSE}_\mu[\delta_{\alpha,\beta}]}{\mu}\right] \\ &= \frac{b^2\{E[\mu]\beta^{-2} \; + \; [1-2\beta^{-1}(\alpha-1)] \; + \; E[\mu^{-1}](\alpha-1)^2\}}{(1+\beta^{-1})^2} \\ &= \frac{b^2\{\alpha\beta^{-1} \; + \; [1-2\beta^{-1}(\alpha-1)] \; + \; \beta^{-1}(\alpha-1)\}}{(1+\beta^{-1})^2} \\ &= \frac{b^2}{1+\beta^{-1}} \quad .\end{aligned}$$

So, fixing α and letting $\beta \to \infty$ we obtain

$$\begin{aligned}\lim_{\beta\to\infty} [r(Gamma(\alpha,\beta), bX) \;-\; r(Gamma(\alpha,\beta), \delta_{\alpha,\beta})] \\ &= \lim_{\beta\to\infty}\left\{ b^2 - \frac{b^2}{1+\beta^{-1}}\right\} \\ &= \lim_{\beta\to\infty}\left\{ b^2 \frac{\beta^{-1}}{1+\beta^{-1}}\right\} \\ &= 0 \quad ,\end{aligned}$$

and hence, bX is extended Bayes and thus minimax.

To finish up, returning to the case of i.i.d. X_1, X_2, ..., X_n which are $Poisson(\mu)$, we reduce by sufficiency to $T = \sum_i X_i$, which is $Poisson(n\mu)$, and so we want to estimate $n^{-1}E_\mu[T]$ (i.e. replace μ in the previous argument by $n\mu$ and take $b = n^{-1}$), and it follows that the minimax estimator is T/n, the UMVUE. Recall that this is for the weighted squared error loss in (3.211).

□

3.6.3 Discussion of the Principle of Minimaxity.

While minimaxity has a certain appeal because it avoids the arbitrariness of selecting a restricted class of decision rules as in the principle of uniform minimum risk and avoids the arbitrariness of selecting a prior as in Bayesian decision theory, it is not without its own set of problems. The focus of the principle of minimaxity of optimizing the “worst case” (i.e. maximum risk) leaves something to be desired on philosophical grounds. After all, why should we worry so much about the worst case risk if it is in a part of parameter space which is not too important or likely (but here we are already becoming Bayesian)? More to the point, by putting the emphasis on the worst case of risk, we may obtain a procedure that does very poorly at many other parameter values. Examples are given in Exercises 3.6.8 and 3.6.9 below. Also, minimax procedures do not exist for many problems so the principle is useless in those situations. In general, if one can find a minimax estimator for a reasonable loss function, and if the estimator is “reasonable” from various points of view, then its minimaxity provides somewhat of a compelling argument for the use of that procedure. Certainly in all of the examples above, minimaxity a strong justification for use of the estimators so derived. In Examples 3.6.2 and 3.6.3, we already had considerable justification for the use of the particular procedures that were minimax (e.g. both procedures were UMVUE). For Example 3.6.1 we derived a minimax estimator when unbiasedness would really not apply and it is the only non–Bayesian justification we have seen for the procedure so derived. But it would be highly desirable to have other justifications.

Exercises for Section 4.3.

3.6.1 True or false (i.e. give proof or counterexample):

(a) An admissable equalizer rule is minimax.

(b) If δ_0 is a minimax rule and δ_1 is a better rule, then δ_1 is also minimax.

(c) A minimax rule is admissable.

(d) Assuming a complete and sufficient statistic exists, if $\delta^\star$ is a minimax estimator under squared error loss which is also unbiased and $\delta^\star$ has finite maximum MSE which is achieved at some θ, then $\delta^\star$ is also UMVUE.

(e) If for some proper prior Π,

$$\inf_{\delta} r(\Pi, \delta) \ = \ \infty \quad ,$$

then every rule is minimax.

3.6.2 Let $X \sim FM(\alpha, \phi)$ and let $Y = X \hat{\pm} \varphi$ where $\hat{\pm}$ denotes addition or subtraction mod 2π. Verify that Law$[Y] = FM(\alpha, \phi \hat{\pm} \varphi)$.

3.6.3 In the setup of Example 3.6.1 consider the sample mean $\bar{X}$ as an estimator of ϕ. Is it unbiased? (Hint: Look at $\phi = 0$ and $\phi = \pi$.) Does the principle of unbiasedness make sense in this context?

3.6.4 Suppose $g(\theta)$ is an estimand which takes on only finitely many values, i.e. there is a finite sequence of distinct values a_1, a_2, .., a_m such that $g(\theta) \in \{ \ a_1, a_2, ..., a_m \ \}$ for all θ. Find the minimax estimator under the 0-1 loss, given in Exercise 3.5.12.

3.6.5 Prove Corollary 3.6.2.

3.6.6 Let X_1, X_2, ..., X_n be i.i.d. with unknown density and having a finite second moment. Find the minimax estimator of $\mu = E[X_i]$ under normalized squared error loss given in (3.210). (Hint: the problem is easy.)

3.6.7 Find the minimax estimator of p, $0 < p < 1$, in a $B(n, p)$ observation model under (i) squared error loss, and (ii) the loss

$$L_0(p, d) \ = \ \frac{(d - p)^2}{p(1 - p)} \quad .$$

Compare both minimax estimators with the UMVUE.

3.6.8 Let $\Theta = [0, 1] = A$ and consider the loss $L(\theta, d) = (1 - \theta)d + \theta(1 - d) = \theta(1 - \theta) + d(1 - d) + (d - \theta)^2$. Show that $\delta(X) \equiv 1/2$ is a minimax rule. Note that this is independent of the distribution of the observable. Comment.

3.6.9 Let $\Theta = (0, 1]$ and $A = [0, 1]$. Consider the loss $L(\theta, d) = \min \{ \ (\theta - d)^2/\theta, 2 \ \}$, i.e. the smaller of relative squared error and 2. Suppose X is $B(n, \theta)$. Show that the unique minimax rule is $\delta(X) \equiv 0$. Comment.

3.7 Asymptotic Comparison of Estimators.

Because exact computations of risks such as mean squared error in finite samples is so difficult, it has been common practice for some time to compare asymptotic distributions of point estimators and to consider "asymptotic optimality". Unfortunately, this subject is fraught with considerable mathematical difficulty. Here, we will primarily consider the class of M–estimates with regular score functions as introduced below, which simplifies matters considerably, although the student is forgiven for finding even this to be difficult. In the final subsection, we point out some issues that indicate the difficulties faced in a more general theory and provide some caveats to be watchful of.

3.7.1 Solutions of Random Equations: Univariate Case.

The next result is similar in spirit to the δ-method, and also very useful. Many estimators in practice are obtained by solving random equations of the form

$$\hat{\lambda}_n(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \psi(X_i,\theta) \;=\; 0 \quad , \tag{3.212}$$

where X_1, X_2, … are i.i.d. We will also use X to denote a r.v. with the same distribution as the X_i's. Here, ψ is sometimes referred to as the *score function.* The equation is solved for the variable θ, and we will denote a solution by $\hat{\theta}_n$, where the subscript n indicates the sample size. Let us assume for now that θ is 1-dimensional and that ψ is "regular". Let θ_0 be such that

$$E[\psi(X_i,\theta_0)] \;=\; 0 \quad , \tag{3.213}$$

i.e. θ_0 solves the "population" version of (3.212) where the average over the sample is replaced by $E[\cdot]$. Since the l.h.s. of (3.212) tends to the l.h.s. of (3.213) by the law of large numbers, we expect $\hat{\theta}_n$ to tend to θ_0 in some sense. If $\hat{\theta}_n$ is close enough to θ_0 then by first order Taylor expansion about $\theta = \theta_0$ we have

$$0 \;=\; \hat{\lambda}_n(\hat{\theta}_n) \;=\; \hat{\lambda}_n(\theta_0) + D\hat{\lambda}_n(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + \ldots \quad .$$

Here, D is differentiation w.r.t. the θ variable. After some algebraic manipulation we obtain

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \;=\; \left[D\hat{\lambda}_n(\theta_0)\right]^{-1}\left[-\sqrt{n}\hat{\lambda}_n(\theta_0)\right] + \ldots \quad .$$

By the Weak Law of Large Numbers, $D\hat{\lambda}_n(\theta_0) \xrightarrow{P} E[D\psi(X_i,\theta_0)]$. Note that $\hat{\lambda}_n(\theta_0)$ is an average of mean 0 random variables, and assuming they also have finite variance, by the Central Limit Theorem $-\sqrt{n}\hat{\lambda}_n(\theta_0) \xrightarrow{D} N(0, E[\psi(X_i,\theta_0)^2])$, so by Slutsky's Theorem (assuming $E[D\psi(X_i,\theta_0)] \neq 0$),

$$\left[D\hat{\lambda}_n(\theta_0)\right]^{-1}\left[\sqrt{n}\hat{\lambda}_n(\theta_0)\right] \;\xrightarrow{D}\; N(0,\sigma^2)$$

where

$$\sigma^2 \;=\; \frac{E[\psi(X,\theta_0)^2]}{E[D\psi(X,\theta_0)]^2} \quad . \tag{3.214}$$

Assuming we can show the remainder terms are negligible, we will obtain

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \;\xrightarrow{D}\; N(0,\sigma^2) \tag{3.215}$$

as well.

There is one difficulty we have glossed over here. When solving equations, there are always the problems with existence and uniqueness of a solution. The best we will be able to do is show that there exists a sequence of roots of (3.212) with probability tending to 1 for which (3.215) holds. There may be other roots to (3.212) as well. Another difficulty which we alluded to is the need for some "regularity" conditions, e.g. differentiability of ψ as a function of θ and $E[D\psi(X_i,\theta_0)] \neq 0$. Conditions such as these will be stated in the theorem, but the reader should keep in mind that one does not formulate such requirements before constructing the proof, but rather proceeds with the proof and discovers the "right" conditions as the need arises. Now we are prepared to state the first theorem for univariate θ. Some further difficulties arise in the multivariate version of the theorem. The reader may wish to skip the proof on first reading.

Theorem 3.7.1 *Let X, X_1, X_2, ... be i.i.d. random d-vectors taking values in $\mathcal{W}$. Suppose $\Theta \subset I\!R$ and $\theta_0 \in \Theta$. Let $\psi : \mathcal{W} \times \Theta \longrightarrow I\!R$ be a given score function. Assume that the following hold:*

(A1) *θ_0 is an interior point of Θ;*

(A2) *$E[\psi(X,\theta_0)^2] < \infty$;*

(A3) *$E[\psi(X,\theta_0)] = 0$;*

(A4) *For each fixed $x \in \mathcal{W}$, $\psi(x,\cdot)$ is differentiable (as a function of θ) in a neighborhood of θ_0; the derivative is denoted $D\psi(x,\theta)$;*

(A5) *$E[|D\psi(X,\theta_0)|] < \infty$;*

(A6) *$E[D\psi(X,\theta_0)] > 0$.*

(A7) *There exists $M : \mathcal{W} \longrightarrow [0,\infty)$ and a constant $p > 0$ such that for all θ in a neighborhood of θ_0,*

$$\Big|\; D\psi(x,\theta) \;-\; D\psi(x,\theta_0) \Big| \;\;\leq\;\; M(x)|\theta-\theta_0|^p \tag{3.216}$$

and

$$E[M(X)] \;<\; \infty \quad . \tag{3.217}$$

Let $\hat{\lambda}_n(\theta)$ be defined by (3.212). Then there exists a sequence $\hat{\theta}_n$ such that as $n \to \infty$,

$$P\left[\hat{\lambda}_n(\hat{\theta}_n) \,=\, 0\right] \,\to\, 1 \tag{3.218}$$

and

$$\sqrt{n}\left(\hat{\theta}_n \,-\, \theta_0\right) \,\xrightarrow{D}\, N(0,\sigma^2) \tag{3.219}$$

where σ^2 is given in equation (3.214).

Remarks 3.7.1 (a) In (A4) we specifically mention that differentiability is w.r.t. θ. Note that x takes values in the arbitrary measurable space $\mathcal{W}$, so in general it makes no sense to speak of differentiability w.r.t. x. In particular, x may be a discrete variable.

(b) We could have replaced (A6) with $E[D\psi(X,\theta_0)] \neq 0$. Under this asssumption, one may take $E[D\psi(X,\theta_0)] > 0$ with no loss of generality since one can replace ψ by $-\psi$ if $E[D\psi(X,\theta_0)] < 0$.

(c) The condition in (3.216) is called a Hölder condition of order p on $D\psi(x,\cdot)$. Such conditions are widely used in analysis. A Lipschitz condition is a Hölder condition of order 1. If a function is continuously differentiable, then it satisfies a Lipschitz condition (Exercise 3.7.1). Note that (3.217) is something extra which is required for us to control the behavior of the constant in the Hölder condition that depends on the extra variable x.

(d) Note that θ_0 figures prominently in our assumptions. In a situation where $\hat{\theta}_n$ is used to estimate an unknown parameter θ_0, one will have to verify these conditions for arbitrary θ_0 in Θ. In this setting, condition (A1) immediately rules out boundary points of Θ, and in fact the "usual" asymptotic distribution theory does not hold for boundary values of the parameter.

Before giving the proof of the theorem, we need the following lemma.

Lemma 3.7.2 *Under the assumptions of the theorem, the following results hold for all θ in a fixed neighborhood of θ_0.*

(a) $E[|\psi(X,\theta)|] < \infty$. Hence, the function

$$\lambda(\theta) \;:=\; E[\psi(X,\theta)]$$

is defined.

(b) λ is differentiable and the derivative may be computed by interchanging differentiation and integration, i.e.

$$D\lambda(\theta) \;=\; \int_{\mathcal{W}} D\psi(x,\theta)\, dP_X(x) \quad .$$

where $P_X = Law[X]$ denotes the distribution of X.

(c) $D\lambda$ is continuous at θ_0.

(d) $D\lambda(\theta) > 0$.

Proof of Lemma. Simultaneously applying (A1), (A4), and (A7), let $\delta_0 > 0$ be such that the interval $J_0 = (\theta_0 - \delta_0, \theta_0 + \delta_0) \subset \Theta$, $\psi(x, \cdot)$ is differentiable on J_0, and (3.216) holds. By the mean value theorem, for $\theta \in J_0$,

$$\psi(x,\theta) \ = \ \psi(x,\theta_0) + D\psi(x,\theta_0)(\theta-\theta_0) + [D\psi(x,\tilde{\theta}) - D\psi(x,\theta_0)](\theta-\theta_0) \quad (3.220)$$

where $\tilde{\theta}$ is between θ and θ_0. By (3.216)

$$\Big| \ D\psi(x,\tilde{\theta}) \ - \ D\psi(x,\theta_0) \Big| \ \leq \ M(x)|\tilde{\theta} - \theta_0|^p \ \leq \ M(x)|\theta - \theta_0|^p \quad , \qquad (3.221)$$

the last inequality following from the fact that $\tilde{\theta}$ is between θ and θ_0.

Now to prove part (a) of the Lemma, we have by the last two displays that for all $\theta \in J_0$,

$$E[|\psi(X,\theta)|] \ \leq \ E[|\psi(X,\theta_0)|] \ + \ E[M(X)]|\theta - \theta_0|^{1+p} \ < \ \infty$$

where the last inequality follows from (A2) (which implies of course that $E[|\psi(X,\theta_0)|] < \infty$) and (3.217).

For part (b) we have by (3.220) and (3.221) that for all $\theta \in J_0$,

$$\begin{aligned} |D\psi(X,\theta)| \ &\leq \ |D\psi(X,\theta_0)| \ + \ M(X)|\theta - \theta_0|^p \\ &\leq \ |D\psi(X,\theta_0)| \ + \ M(X)\delta_0^p \ := \ G_0(X) \qquad (3.222) \end{aligned}$$

and $EG_0(X) < \infty$ by (A5) and (3.217). Thus, by the theorem on interchange of differentiation and integration (Theorem 1.2.10), λ is differentiable on J_0 and the derivatives may be computed by interchange of differentiation and integration.

Turning to part (c), we have from (3.221) that

$$D\psi(x,\theta) \ - \ D\psi(x,\theta_0) \ \to \ 0 \quad \text{ as } \theta \to \theta_0$$

and

$$\Big| \ D\psi(x,\theta) \ - \ D\psi(x,\theta_0) \Big| \ \leq \ M(x)\delta_0^p \ := \ G_1(x)$$

with $EG_1(X) < \infty$ by (3.217). Thus we may apply dominated convergence to claim that

$$E[D\psi(X,\theta)] \ \to \ E[D\psi(X,\theta_0)] \quad \text{ as } \theta \to \theta_0 \quad .$$

In view of part (b) (i.e. that we can interchange D and $\int \ldots dP_X$), this shows that $D\lambda(\theta) \to D\lambda(\theta_0)$ as $\theta \to \theta_0$, i.e. that $D\lambda$ is continuous at θ_0.

Finally, for part (d), since $D\lambda$ is continuous at θ_0 and positive at θ_0 (assumption (A6)), it follows that it is positive in a neighborhood of θ_0.

□

Proof of Theorem 3.7.1. The first task is to prove the existence of $\hat{\theta}_n$, and to show simultaneously that $\hat{\theta}_n \xrightarrow{P} \theta_0$. Take δ_1 small enough that for all $|\theta - \theta_0| \le \delta_1$, $D\lambda(\theta) > 0$, and in particular, λ is strictly monotone increasing in the interval $J_1 := [\theta_0 - \delta_1, \theta_0 + \delta_1]$. In view of assumption (A3) (which states that $\lambda(\theta_0) = 0$), this means $\lambda(\theta_0 - \delta_1) < 0$ and $\lambda(\theta_0 + \delta_1) > 0$. Let

$$c_1 \;=\; \frac{1}{2}\min\{-\lambda(\theta_0 - \delta_1), \lambda(\theta_0 + \delta_1)\}$$

which is positive. Now by the weak law of large numbers, $\hat{\lambda}_n(\theta) \xrightarrow{P} \lambda(\theta)$ for each fixed $\theta \in J_1$ (note that we are using part (a) of Lemma 3.7.2 here), and so for the two values $\theta_0 \pm \delta_1$, given $\epsilon_1 > 0$ there exists N_1 such that

$$\forall n \ge N_1, \quad P\left[\hat{\lambda}_n(\theta_0 - \delta_1) < -c_1 \text{ and } \hat{\lambda}_n(\theta_0 + \delta_1) > +c_1\right] \;\ge\; 1-\epsilon_1 \quad . \tag{3.223}$$

(See Exercise 3.7.2.) Now let $\delta_k > 0$ and $\epsilon_k > 0$ be sequences that decrease to 0. Then by iterating this argument, we obtain an increasing sequence N_k and a sequence of positive numbers c_k such that

$$\forall n \ge N_k, \quad P\left[\hat{\lambda}_n(\theta_0 - \delta_k) < -c_k \text{ and } \hat{\lambda}_n(\theta_0 + \delta_k) > +c_k\right] \;\ge\; 1-\epsilon_k \quad . \tag{3.224}$$

Since $\hat{\lambda}_n$ is continuous in J_1 (after all, it is differentiable by (A4) so it is certainly also continuous), on the event in the last display, $\hat{\lambda}_n(\theta) = 0$ for some $\theta \in J_k := (\theta_0 - \delta_k, \theta_0 + \delta_k)$ by the intermediate value theorem. Let $\hat{\theta}_n \in J_k$ be such a root of $\hat{\lambda}_n(\theta) = 0$ for $N_k \le n < N_{k+1}$, and otherwise let $\hat{\theta}_n$ be defined arbitrarily (i.e. on the complement of the event in (3.224) where the root may not exist). Then, given any $\delta > 0$ and $\epsilon > 0$, there is a k such $\delta_k < \delta$ and $\epsilon_k \le \epsilon$, and

$$\forall n \ge N_k, \quad P\left[|\hat{\theta}_n - \theta_0| < \delta\right] \;\ge\; 1 - \epsilon \quad . \tag{3.225}$$

This shows that $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Turning now to the asymptotic normality, on the event where $\hat{\lambda}_n(\hat{\theta}_n) = 0$, following the same reasoning as in (3.220) and (3.221), we have by the mean value theorem that for some $\tilde{\theta}_n$ between θ_0 and $\hat{\theta}_n$,

$$0 \;=\; \hat{\lambda}_n(\theta_0) \;+\; D\hat{\lambda}_n(\theta_0)\left(\hat{\theta}_n - \theta_0\right) \;+\; \left[D\hat{\lambda}_n(\tilde{\theta}_n) \;-\; D\hat{\lambda}_n(\theta_0)\right]\left(\hat{\theta}_n - \theta_0\right) \tag{3.226}$$

and

$$\left|\; D\hat{\lambda}_n(\tilde{\theta}_n) - D\hat{\lambda}_n(\theta_0)\right| \;\le\; \left[\frac{1}{n}\sum_{i=1}^{n} M(X_i)\right] |\tilde{\theta}_n - \theta_0|^p \;\le\; \left[\frac{1}{n}\sum_{i=1}^{n} M(X_i)\right] |\hat{\theta}_n - \theta_0|^p \quad .$$

Now by the weak law of large numbers (in conjunction with (3.217))

$$\frac{1}{n}\sum_{i=1}^{n} M(X_i) \;=\; O_P(1)$$

and by the first part of the proof

$$|\hat{\theta}_n - \theta_0|^p \;=\; o_P(1)$$

and so the product is $o_P(1)$, of course. Using these bounds in (3.226) we obtain

$$-\hat{\lambda}_n(\theta_0) \;=\; \left[D\hat{\lambda}_n(\theta_0) \,+\, o_P(1)\right]\left(\hat{\theta}_n - \theta_0\right) \quad .$$

By (A5) and the weak law in conjunction with Lemma 3.7.2 (b),

$$D\hat{\lambda}_n(\theta_0) \;=\; D\lambda(\theta_0) \,+\, o_P(1) \quad ,$$

which when plugged into the previous display yields

$$-\hat{\lambda}_n(\theta_0) \;=\; [D\lambda(\theta_0) \,+\, o_P(1)]\left(\hat{\theta}_n - \theta_0\right) \quad .$$

Hence

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \;=\; \frac{-\sqrt{n}\hat{\lambda}_n(\theta_0)}{D\lambda(\theta_0) \,+\, o_P(1)} \quad .$$

By the Central Limit Theorem (we need (A2) and (A3) for this),

$$\sqrt{n}\hat{\lambda}_n(\theta_0) \;\xrightarrow{D}\; N(0, E[\psi(X,\theta_0)^2]) \quad ,$$

and by Slutsky's theorem (noting that $D\lambda(\theta_0) \neq 0$ by (A6)),

$$\frac{-\sqrt{n}\hat{\lambda}_n(\theta_0)}{D\lambda(\theta_0) \,+\, o_P(1)} \;\xrightarrow{D}\; N(0,\sigma^2)$$

where σ^2 is given by (3.214).

□

Example 3.7.1 Suppose X_1, X_2, ..., X_n are i.i.d. and modelled by a gamma distribution with density

$$f_\alpha(x) \;=\; \frac{x^{\alpha-1}}{\Gamma(\alpha)}e^{-x} \quad , \quad x > 0.$$

Here, $\alpha > 0$ is an unknown parameter to be estimated. The maximum likelihood estimator (to be discussed in more detail in a later chapter) is obtained by maximizing (as a function of α) the log–likelihood

$$\ell(\alpha) \;=\; \sum_{i=1}^{n} \left[\,(\alpha - 1)\log X_i \,-\, \log\Gamma(\alpha) \,-\, X_i\right] \quad .$$

Taking derivatives and setting to 0, then multiplying by $-1/n$, we get an equation of the form (3.212) where the score function is

$$\psi(x, \alpha) \; = \; -\log X_i + D \log \Gamma(\alpha) \quad .$$

Here, $D = d/d\alpha$. Suppose α_0 is a solution of

$$E[\log X] \; = \; D \log \Gamma(\alpha) \quad .$$

We shall see in the a later chapter that the "regularity" conditions (A1), (A2), and (A4) through (A7) hold here. Furthermore, α_0 is unique, and the solution $\hat{\alpha}_n$ to equation (3.212) is unique in this case. Then we obtain that $\sqrt{n}[\hat{\alpha}_n - \alpha_0] \xrightarrow{D} N(0, \sigma^2)$ where

$$\sigma^2 \; = \; \frac{\mathrm{Var}[\log X]}{\left[D^2 \log \Gamma(\alpha_0)\right]^2} \quad .$$

Note that $E[-\log X] = D \log \Gamma(\alpha_0)$, so $E[\psi(X, \alpha_0)^2] = \mathrm{Var}[\log X]$, here. Moreover, it is easy to get an estimate of this "asymptotic variance" (more properly, variance of the asymptotic normal distribution) by replacing $\mathrm{Var}[\log X]$ with the sample variance of the transformed data $\log X_1$, $\log X_2$, ..., $\log X_n$.

□

We close this subsection by mentioning that the "regularity" conditions involving smoothness of ψ are not necessary. For instance, consider for real valued data

$$\psi(x, \theta) \; = \; \begin{cases} -1 & \text{if } x - \theta < 0; \\ 0 & \text{if } x - \theta = 0; \\ +1 & \text{if } x - \theta > 0. \end{cases}$$

In general, the corresponding equation $\hat{\lambda}_n(\theta) = 0$ may not have a solution, but there does exist $\hat{\theta}_n$ such that $\hat{\lambda}_n(\theta) \leq 0$ for $\theta < \hat{\theta}_n$ and $\hat{\lambda}_n(\theta) \geq 0$ for $\theta > \hat{\theta}_n$. Such a $\hat{\theta}_n$ will be a sample median. Now $\lambda(\theta) = 1 - F(\theta) - F(\theta - 0)$, where F is the c.d.f. of the distribution of the X_i's, and if $F' = f$ exists and is positive at θ_0, the median of F (which is unique when the density f exists and is positive at m), then $\sqrt{n}[\hat{\theta}_n - \theta_0] \xrightarrow{D} N(0, 1/[4f(\theta_0)^2])$, which is exactly the result that would be obtained by formally applying the conclusion of the theorem even though ψ does not satisfy the differentiability requirement.

3.7.2 Solutions of Random Equations: Multiparameter Case.

Now we consider the situation in which the parameter θ is a vector. Many of the calculations and arguments of the previous subsection go through with but minor modification, but there is a major problem in the proof. We will give the intuitive discussion here and derive the result, state the theorem, give the proof of the one difficult point, and leave the rest to the reader.

Now consider a random equation of the form

$$\hat{\underline{\lambda}}_n(\underline{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \underline{\psi}(X_i, \underline{\theta}) \;=\; 0 \quad , \tag{3.227}$$

where now $\underline{\theta} \in \Theta \subset I\!\!R^p$ is a p-dimensional vector, and $\underline{\psi} : \mathcal{X} \times \Theta \longrightarrow I\!\!R^p$, so (3.227) is really p equations in p unknowns. As before, we assume X_1, X_2, ... are i.i.d., and we denote a solution by $\hat{\underline{\theta}}_n$. Let $\underline{\theta}_0$ be the solution of the "population" equation, i.e. $E[\underline{\psi}(X, \underline{\theta}_0)] = \underline{0}$ (where $\underline{0}$ is the zero vector). Assuming then that we can show $\hat{\underline{\theta}}_n \xrightarrow{P} \underline{\theta}_0$, we have similar nonrigorous calculations which are

$$\underline{0} \;=\; \hat{\underline{\lambda}}_n(\hat{\underline{\theta}}_n) \;=\; \hat{\underline{\lambda}}_n(\underline{\theta}_0) \;+\; D\hat{\underline{\lambda}}_n(\underline{\theta}_0)\left(\hat{\underline{\theta}}_n - \underline{\theta}_0\right) \;+\; \ldots \quad .$$

Note that $D\underline{\psi}(x, \cdot)$ is a $p \times p$ matrix. Similarly to the case where θ is unidimensional (except there is a matrix inverse), we obtain

$$\sqrt{n}\left(\hat{\underline{\theta}}_n - \underline{\theta}_0\right) \;=\; \left[D\hat{\underline{\lambda}}_n(\underline{\theta}_0)\right]^{-1}\left[-\sqrt{n}\hat{\underline{\lambda}}_n(\underline{\theta}_0)\right] \;+\; \ldots \quad ,$$

provided $D\hat{\underline{\lambda}}_n(\underline{\theta}_0)$ is invertible (which it will be in the limit under our regularity conditions). Again, $D\hat{\underline{\lambda}}_n(\underline{\theta}_0) \xrightarrow{P} E[D\underline{\psi}(X, \underline{\theta}_0)]$, and by the Central Limit Theorem for vector valued random variables, $-\sqrt{n}\hat{\underline{\lambda}}_n(\underline{\theta}_0) \xrightarrow{D} N(0, V)$ where the covariance matrix is given by

$$V \;=\; E[\underline{\psi}(X, \theta_0)\underline{\psi}(X, \theta_0)'].$$

Again, similarly to the previous argument for one dimensional θ, (see Exercise 3.7.3)

$$\left[D\hat{\underline{\lambda}}_n(\underline{\theta}_0)\right]^{-1}\left[\sqrt{n}\hat{\underline{\lambda}}_n(\underline{\theta}_0)\right] \;\xrightarrow{D}\; N(0, \Sigma) \tag{3.228}$$

where

$$\Sigma \;=\; E[D\underline{\psi}(X, \underline{\theta}_0)]^{-1} E[\underline{\psi}(X, \underline{\theta}_0)\underline{\psi}(X, \underline{\theta}_0)'](E[D\underline{\psi}(X, \underline{\theta}_0)]')^{-1} \quad , \tag{3.229}$$

and we expect the same limit in distribution for $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_0)$.

Recall that a significant portion of the proof of Theorem 3.7.1 dealt with establishing the existence of $\hat{\theta}_n$. That was accomplished by using monotonicity of λ in a neighborhood of θ_0 and invoking the intermediate value theorem. None of that machinery will extend to the case when $\underline{\theta}$ is a vector. The following rather elementary theorem from analysis will be very useful in this context.

Theorem 3.7.3 (Contraction Mapping Fixed Point Theorem) *Suppose $F \subset I\!\!R^d$ is a closed and bounded set and $h : F \longrightarrow F$ is a* contraction map, *i.e. there is a positive constant $C < 1$ such that*

$$\|h(x) - h(y)\| \;\leq\; C\|x - y\|$$

for all x and y in F. Then there exists a unique fixed point for h in F, i.e. a unique point $x_0 \in F$ such that $h(x_0) = x_0$.

□

The proof is left as an Exercise 3.7.6. It may be found in a number of books on analysis, e.g. Rudin *Principles of Mathematical Analysis*, Theorem 9.2.3, p. 220.

Theorem 3.7.4 *Let X, X_1, X_2, ... be i.i.d. random d-vectors taking values in $\mathcal{W}$. Suppose $\Theta \subset I\!R^p$ and $\underline{\theta}_0 \in \Theta$. Let $\underline{\psi} : \mathcal{W} \times \Theta \longrightarrow I\!R$ be a given score function. Assume that the following hold:*

(A1) *$\underline{\theta}_0$ is an interior point of Θ;*

(A2) *$E[\|\underline{\psi}(X, \underline{\theta}_0)\|^2] < \infty$;*

(A3) *$E[\underline{\psi}(X, \underline{\theta}_0)] = \underline{0}$;*

(A4) *For each fixed $x \in \mathcal{W}$, $\underline{\psi}(x, \cdot)$ is differentiable (as a function of $\underline{\theta}$) in a neighborhood of $\underline{\theta}_0$; the derivative is denoted $D\underline{\psi}(x, \underline{\theta})$;*

(A5) *$E[\|D\underline{\psi}(X, \underline{\theta}_0)\|] < \infty$;*

(A6) *The $p \times p$ matrix $D\underline{\psi}(X, \underline{\theta}_0)$ is invertible.*

(A7) *There exists $M : \mathcal{W} \longrightarrow [0, \infty)$ and a constant $q > 0$ such that for all $\underline{\theta}$ in a neighborhood of $\underline{\theta}_0$,*

$$\|D\underline{\psi}(x, \underline{\theta}) \;-\; D\underline{\psi}(x, \underline{\theta}_0)\| \;\leq\; M(x)\|\underline{\theta} - \underline{\theta}_0\|^q \tag{3.230}$$

and

$$E[M(X)] \;<\; \infty \quad . \tag{3.231}$$

Let $\hat{\lambda}_n(\underline{\theta})$ be defined by (3.227). Then there exists a sequence $\hat{\underline{\theta}}_n$ such that as $n \to \infty$,

$$P\left[\hat{\lambda}_n(\hat{\underline{\theta}}_n) \;=\; 0\right] \;\to\; 1 \tag{3.232}$$

and

$$\sqrt{n}\left(\hat{\underline{\theta}}_n \;-\; \underline{\theta}_0\right) \;\xrightarrow{D}\; N(0, \Sigma) \tag{3.233}$$

where Σ is given in equation (3.229).

Partial Proof. Many of the details are basically the same as in the proof of Theorem 3.7.1, and therefore we do not give them. Lemma 3.7.2 holds with minor modifications (e.g. in part (d) one can only claim that $\det D\underline{\lambda}(\underline{\theta}) > 0$, or < 0, depending on the sign of $\det D\underline{\lambda}(\underline{\theta}_0)$). Just to sketch the proof of part (b) of that Lemma, the claim in (3.222) is modified as follows: there exists a $\delta_0 > 0$ such that for $\|\underline{\theta} - \underline{\theta}_0\| < \delta_0$,

$$\begin{aligned} \|D\underline{\psi}(X,\underline{\theta})\| &\leq \|D\underline{\psi}(X,\underline{\theta}_0)\| + M(X)\|\underline{\theta} - \underline{\theta}_0\|^p \\ &\leq \|D\underline{\psi}(X,\underline{\theta}_0)\| + M(X)\delta_0^p := G_0(X) \end{aligned}$$

in the same way.

Now we turn to the main difference in the multiparameter case – establishing existence of $\hat{\underline{\theta}}_n$. For convenience, we put

$$A = E[D\underline{\psi}(X,\underline{\theta}_0)] \quad .$$

We first show that there exists δ_0 such that if $0 < \delta \leq \delta_0$, $\|\zeta_1\| \leq \delta$, and $\|\zeta_2\| \leq \delta$, then

$$\hat{\underline{\lambda}}_n(\underline{\theta}_0 + \zeta_2) = \hat{\underline{\lambda}}_n(\underline{\theta}_0 + \zeta_1) + A(\zeta_2 - \zeta_1) + R_n(\zeta_1, \zeta_2) \tag{3.234}$$

where

$$\|R_n(\zeta_1,\zeta_2)\| \leq p\,[U_n + \overline{M}_n\delta^q]\,\|\zeta_2 - \zeta_1\| \quad . \tag{3.235}$$

Here, U_n and $\overline{M}_n$ are random variables that don't depend on either ζ_1 or ζ_2, and satisfy

$$U_n = o_P(1), \quad \textit{and} \quad \overline{M}_n = O_P(1) \quad . \tag{3.236}$$

To this end, applying the mean value theorem to the ith component of $\hat{\underline{\lambda}}_n$ along the line segment between $\underline{\theta}_0 + \zeta_1$ and $\underline{\theta}_0 + \zeta_2$ (note that the mean value theorem only applies to scalar functions of a scalar variable) gives

$$\hat{\lambda}_{ni}(\underline{\theta}_0 + \zeta_2) = \hat{\lambda}_{ni}(\underline{\theta}_0 + \zeta_1) + \bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0 + \tilde{\zeta})'(\zeta_2 - \zeta_1) \tag{3.237}$$

where $\tilde{\zeta}$ is on the line segment between ζ_1 and ζ_2. Since $\|\tilde{\zeta}\| \leq \delta$, we have by (A7) that

$$\| \bigtriangledown \hat{\lambda}_{ni}(\underline{\theta}_0 + \tilde{\zeta}) - \bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0)\| \leq \overline{M}_n\delta^q$$

where

$$\overline{M}_n = p^{1/2}\frac{1}{n}\sum_{i=1}^{n} M(X_i) = O_P(1) \quad ,$$

the last relation following from the weak law of large numbers using (3.231). Similarly, by the weak law in conjunction with (A5),

$$\bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown\psi_i(X_i,\underline{\theta}_0) \xrightarrow{P} E[\psi_i(X,\underline{\theta}_0)] := \underline{a}_i$$

where $\underline{a}_i'$ is the ith row of A. Thus,

$$\| \bigtriangledown \hat{\lambda}_{ni}(\underline{\theta}_0 + \tilde{\zeta}) - \underline{a}_i\| := U_{ni} = o_P(1) \quad .$$

Now we have by Cauchy-Schwarz

$$\Big|\ [\bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0+\tilde{\zeta})\ -\ \underline{a}_i]'\,[\zeta_2-\zeta_1]\ \Big|$$

$$\leq\ \|\bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0+\tilde{\zeta})\ -\ \bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0)\ +\ \bigtriangledown\hat{\lambda}_{ni}(\underline{\theta}_0)\ -\ \underline{a}_i\|\|\zeta_2-\zeta_1\|$$
$$\leq\ [\overline{M}_n\delta^q\ +\ U_n]\,\|\zeta_2-\zeta_1\| \quad ,$$

where $U_n := \max_i U_{ni}$. Plugging the preceding computations back into (3.237) gives (3.234) through (3.236).

We will apply Theorem 3.7.3 to a closed ball of radius δ centered at the origin, viz.

$$F_\delta\ =\ \{\zeta\in I\!\!R^p : \|\zeta\|\leq\delta\} \quad .$$

Here, δ will be determined. Define

$$h(\zeta)\ =\ \zeta - A^{-1}\hat{\underline{\lambda}}_n(\underline{\theta}_0+\zeta) \quad .$$

By the first part of the proof, if $\|\zeta\|\leq\delta$, then

$$\|h(\zeta)\|\ =\ \|\zeta - A^{-1}[\hat{\underline{\lambda}}_n(\underline{\theta}_0)\ +\ A\zeta\ +\ R_n(0,\zeta)]\|$$

$$=\ \|A^{-1}[\hat{\underline{\lambda}}_n(\underline{\theta}_0)\ +\ R_n(0,\zeta)]\|\ \leq\ \|A^{-1}\|\,\|\hat{\underline{\lambda}}_n(\underline{\theta}_0)\ +\ R_n(0,\zeta)\|$$
$$\leq\ \|A^{-1}\|\,\Big\{\|\hat{\underline{\lambda}}_n(\underline{\theta}_0)\|\ +\ [\overline{M}_n\delta^q\ +\ U_n]\delta\Big\} \quad . \qquad (3.238)$$

Note that $\|A^{-1}\|$ is just a constant C_1. Now by assumptions (A2) and (A3) and Chebyshev's weak law of large numbers, $(1/n)\sum_{i=1}^n\psi_j(X_i,\underline{\theta}_0) = O_P(n^{-1/2})$, which implies $\hat{\underline{\lambda}}_n(\underline{\theta}_0) = O_P(n^{-1/2})$. So, given $\epsilon>0$, there exist finite positive C_2 and C_3 such that

$$P\Big[\|\hat{\underline{\lambda}}_n(\underline{\theta}_0)\|\ >\ C_2n^{-1/2}\Big]\ <\ \epsilon/3$$
$$P\Big[|\overline{M}_n|\ >\ C_3\Big]\ <\ \epsilon/3$$
$$P\left[|U_n|\ >\ 1/(4C_1)\right]\ <\ \epsilon/3 \quad ,$$

for all n sufficiently large. The latter two inequalities follow from (3.236), of course. Now take δ_1 such that

$$C_1C_3\delta_1^q\ \leq\ 1/4$$

and given $\delta<\delta_1$, take n large enough that

$$C_2n^{-1/2}\ <\ \delta/(2C_1) \quad .$$

We have that the event

$$\mathcal{E}_n\ :=\ \Big[\|\hat{\underline{\lambda}}_n(\underline{\theta}_0)\|\ <\ C_2n^{-1/2}\,\&\,|\overline{M}_n|\ \leq\ C_3\,\&\,|U_n|\ \leq\ 1/(4C_1)\Big]$$

has probability $P(\mathcal{E}_n) > 1 - \epsilon$ for all n sufficiently large. On this event, if $\delta < \delta_1$, we have by (3.238)

$$\|h(\zeta)\| \;\leq\; C_1 \left\{ \frac{\delta}{2C_1} + \left[\frac{1}{4C_1} + \frac{1}{4C_1} \right] \delta \right\} \;=\; \delta \quad ,$$

i.e., on the event $\mathcal{E}_n$, h maps F_δ into itself for all $\delta < \delta_1$.

Now we show the contraction mapping property. For $\delta < \delta_1$ and for all n sufficiently large, we have on the event $\mathcal{E}_n$ for all ζ_1 and ζ_2 in F_δ,

$$\begin{aligned} \|h(\zeta_1) - h(\zeta_2)\| \;&=\; \|\zeta_1 \;-\; A^{-1}\underline{\hat{\lambda}}_n(\underline{\theta}_0 + \zeta_1) \;-\; \zeta_2 \;+\; A^{-1}\underline{\hat{\lambda}}_n(\underline{\theta}_0 + \zeta_2)\| \\ &=\; \|A^{-1}R_n(\zeta_1, \zeta_2)\| \;\leq\; C_1[U_n + \overline{M}_n\delta^q]\, \|\zeta_1 - \zeta_2\| \\ &\leq\; C_1 \left[\frac{1}{4C_1} + \frac{1}{4C_1} \right] \|\zeta_1 - \zeta_2\| \;=\; \frac{1}{2}\|\zeta_1 - \zeta_2\| \quad . \end{aligned}$$

This shows the contraction mapping property, so we conclude that there is a unique ζ_0 such that $h(\zeta_0) = \zeta_0$, i.e.

$$\zeta_0 \;=\; \zeta_0 \;-\; A^{-1}\underline{\hat{\lambda}}_n(\underline{\theta}_0 + \zeta_0) \;\Longrightarrow\; \underline{\hat{\lambda}}_n(\underline{\theta}_0 + \zeta_0) \;=\; \underline{0} \quad .$$

Setting $\underline{\hat{\theta}}_n = \underline{\theta}_0 + \zeta_0$ on the event $\mathcal{E}_n$ (where we know ζ_0 exists) and defining $\underline{\hat{\theta}}_n$ arbitrarily off of $\mathcal{E}_n$, we have

$$P\left[\underline{\hat{\lambda}}_n(\underline{\hat{\theta}}_n) \;=\; \underline{0} \right] \;\geq\; P(\mathcal{E}_n) \;>\; 1 - \epsilon \quad .$$

This argument will also give that $\underline{\hat{\theta}}_n \xrightarrow{P} \underline{\theta}_0$ with a little extra work. What was shown is that given $\epsilon > 0$ there exists δ_1 (which depends on ϵ) such that for all $\delta < \delta_1$ there exists N (which depends on δ) such that for all $n \geq N$, there is a root $\underline{\hat{\theta}}_n$ of $\underline{\hat{\lambda}}_n(\underline{\theta}) = \underline{0}$ satisfying $\|\underline{\hat{\theta}}_n - \underline{\theta}_0\| \leq \delta$. This latter follows because we took $\underline{\hat{\theta}}_n = \underline{\theta}_0 + \zeta_0$ and $\|\zeta_0\| \leq \delta$. Thus, given a positive sequence $\epsilon_k \downarrow 0$ we can find a positive sequence $\delta_k \downarrow 0$ and a sequence of positive integers $N_k \uparrow \infty$ such that for all $n \geq N_k$, $P[\|\underline{\hat{\theta}}_n - \underline{\theta}_0\| > \delta_k] < \epsilon_k$. As in the proof of Theorem 3.7.1, this suffices to show $\underline{\hat{\theta}}_n \xrightarrow{P} \underline{\theta}_0$.

Now the proof of asymptotic normality follows as in the proof of Theorem 3.7.1.

□

3.7.3 Asymptotic Relative Efficiency.

For an estimand $g(\theta)$ consider an estimator $\hat{\gamma}_n$ based on n observations (where θ denotes the parameter). The estimator (sequence) $\hat{\gamma}_n$ is called *consistent and asymptotically normal* (*CAN* for short) if no matter what the true value θ_0 is, we have $\sqrt{n}[\hat{\gamma}_n - g(\theta_0)] \xrightarrow{D} N(0, V_g(\theta_0))$ where $V_g(\theta_0)$ is a nonsingular covariance matrix of appropriate dimension. (Conceivably, one could have a different normalizing factor than $\sqrt{n}$, but we only consider here $\sqrt{n}$.) Note that consistency of such an estimator follows from the fact that we center by $g(\theta_0)$, which implies $\hat{\gamma}_n \xrightarrow{D} g(\theta_0)$.

We will now restrict attention to real valued estimands $g(\theta)$ so $V_g(\theta)$ is just a positive scalar. Let $\hat{\gamma}_{n1}$ and $\hat{\gamma}_{n2}$ be two CAN estimators of the same estimand $g(\theta)$ with asymptotic variance functions $V_{g1}(\theta)$ and $V_{g2}(\theta)$, respectively. Then the *asymptotic relative efficiency* of $\hat{\gamma}_{n2}$ with respect to $\hat{\gamma}_{n1}$ is

$$ARE(\theta; \hat{\gamma}_{n2}, \hat{\gamma}_{n1}) \ := \ \frac{V_{g1}(\theta)}{V_{g2}(\theta)} \quad .$$

For simplicity, we will often write ARE_{21}, $ARE(\theta)$, $ARE_{\hat{\gamma}_{n2},\hat{\gamma}_{n1}}$, etc. rather then $ARE(\theta; \hat{\gamma}_{n2}, \hat{\gamma}_{n1})$ depending on the context. Usually, we choose the order of the estimators so that the $ARE \le 100\%$. To explain this, and in particular explain why the ratio of the variances has the particular order chosen, suppose we are going to take n_1 observations and estimate $g(\theta)$ with $\hat{\gamma}_{n1}$. Then of course the approximate variance of the estimator is $V_{g1}(\theta)/n_1$. How many observations would we have to take to achieve the same "accuracy" using the estimator $\hat{\gamma}_{n2}$? Well, using the asymptotic variance formula again, we would need n_2 observations where n_2 solved $V_{g2}(\theta)/n_2 = V_{g1}(\theta)/n_1$, which means that $n_2 = n_1/ARE_{21}$. (Of course, this may depend on the parameter θ.) Note that if $ARE_{21} < 100\%$, then we need more observations using $\hat{\gamma}_{n2}$, and we would say $\hat{\gamma}_{n2}$ is *less efficient* than $\hat{\gamma}_{n1}$.

Example 3.7.2 Let X_1, X_2, …, X_n be i.i.d. $N(\mu, \sigma^2)$ random variables. Two possible estimators for μ are the sample mean $\overline{X}_n$ and the sample median, denoted M_n. and Both are CAN estimators:

$$\sqrt{n}(\overline{X}_n - \mu) \ \xrightarrow{D} \ N(0, \sigma^2)$$

and

$$\sqrt{n}(M_n - \mu) \ \xrightarrow{D} \ N(0, 1/[4f(\mu)^2]) \ = \ N(0, 1.570796\sigma^2) \quad .$$

Thus, the asymptotic relative efficiency of the median w.r.t. the mean for $N(\mu, \sigma^2)$ location estimation is

$$ARE_{M_n,\overline{X}_n} \ = \ \frac{\sigma^2}{1.570796\sigma^2} \ = \ 63.7\% \quad .$$

Note that in this case, the ARE doesn't depend on the parameter. This is typical of location–scale problems. See Exercise 3.7.14.

□

In Exercise 3.7.9 it is shown that that sample median is more efficient than the sample mean for location estimation in the Laplace (double exponential) family.

Example 3.7.3 Suppose X_1, X_2, …, X_n are i.i.d. from $Gamma(\alpha, 1)$ distribution. (Recall that the mean of $Gamma(\alpha, \beta)$ is $\alpha\beta$ and the variance is $\alpha\beta^2$.) The Maximum Likelihood estimation of α was already discussed in Example 3.1.2. From our theory above, we have for the asymptotic distribution of the MLE, denoted $\hat{\alpha}_n$,

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) \overset{D}{\to} N(0, 1/\psi_1(\alpha_0))$$

where $\psi_1 = D^2 \log \Gamma$ is the trigamma function. Since $E_\alpha[X] = \alpha$, a method of moment estimator (which is much easier to compute) would be $\bar{X}_n$, which has asymptotic distribution

$$\sqrt{n}(\bar{X}_n - \alpha_0) \overset{D}{\to} N(0, \alpha_0) \quad .$$

Now, the ARE of $\bar{X}_n$ w.r.t. $\hat{\alpha}_n$ is

$$ARE(\alpha) \; = \; 1/(\alpha\psi_1(\alpha)) \quad .$$

In this case, the efficiency depends on the parameter α. It is relatively easy to evaluate the expression for the ARE within Mathematica. In Figure 8.1 is shown a plot of $ARE(\alpha)$. Note that apparently as $\alpha \to 0$, $ARE(\alpha) \to 0$, and as $\alpha \to \infty$, $ARE(\alpha) \to 100\%$. These facts can be shown mathematically. Also, $ARE(1) = 60.8\%$. Thus, we see that although the method of moment estimator is very easy to compute, it has relatively low efficiency w.r.t. the MLE except for large values of α.

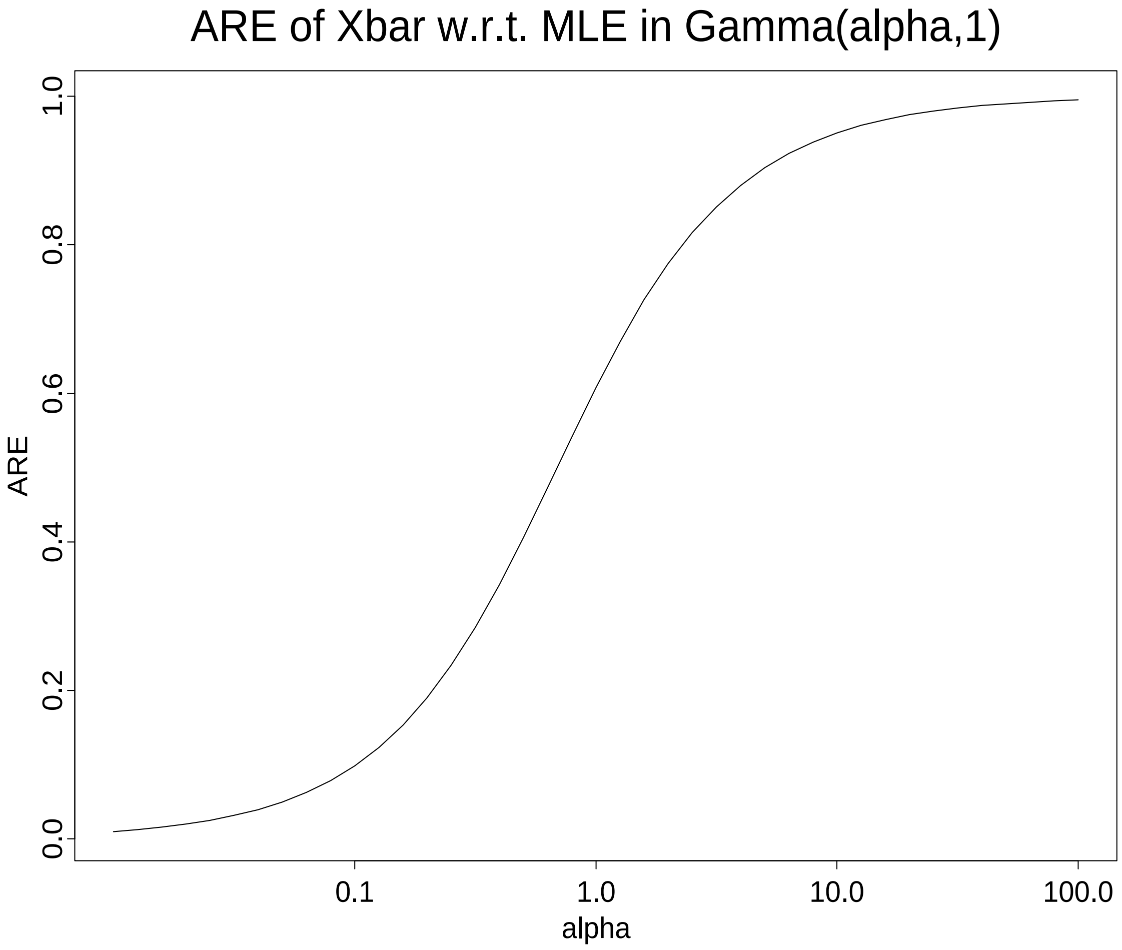

Figure 8.1

□

3.7.4 Asymptotically Optimal M–Estimators.

In each of the examples of the previous subsection, we saw that the MLE was asymptotically more efficient than the other estimator. Here, we will prove in general that under the regularity conditions of Theorem 3.1.2, the MLE will be more efficient than any other M–estimator with score function satisfying the regularity conditions of Theorem 3.7.4 for all $\theta_0 \in \mathrm{int}\Theta$, with one extra regularity condition thrown in. We will also discuss more general results that have been proven about the asymptotic efficiency of the MLE and other estimators. In general, under regularity conditions, we say a CAN estimator is *(asymptotically) (fully) efficient* if the variance of it's limiting normal distribution is the same as that of the MLE. (The words "asymptotically" and "fully" are in parentheses since they are frequently omitted.) Sometimes, a fully efficient estimator is called *Best Asymptotically Normal* or *BAN* for short. That is, if the estimand $g(\theta)$ is real valued, then the estimator sequence $\hat{\gamma}_n$ is asymptotically efficient or BAN iff under any value of θ in the interior of Θ,

$$\sqrt{n}[\hat{\gamma}_n - g(\theta)] \overset{D}{\to} N\left(0, \bigtriangledown g(\theta)^t \mathcal{I}(\theta)^{-1} \bigtriangledown g(\theta)\right) \quad .$$

We do note that when the regularity conditions do not hold, then the MLE may not even have an asymptotically normal distribution. An example is given in Exercise 3.7.8.

Now we consider M–estimators for a real valued estimand $g(\theta)$. The general form of the estimation equation would be

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \gamma) \;=\; 0 \quad , \tag{3.239}$$

where the solution is $\hat{\gamma}_n$. In order to be a consistent estimator of $g(\theta)$, we need

$$E_\theta[\psi(X, g(\theta))] \;=\; 0 \quad , \quad \forall \theta \in \text{int}\Theta. \tag{3.240}$$

This is assumption (A3) of Theorem 3.7.1, except we are estimating $g(\theta)$, not θ. Assume the other regularity conditions of Theorem 3.7.1 hold for all P_θ, $\theta \in \text{int}\Theta$. Then we have under any such θ,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \overset{D}{\to} N\left(0, v_\psi(\theta)\right) \quad , \tag{3.241}$$

where

$$v_\psi(\theta) \;=\; \frac{E_\theta[\psi(X, g(\theta))^2]}{\{E_\theta[D\psi(X, g(\theta))]\}^2} \quad . \tag{3.242}$$

This is just equation (3.214).

Now we will show that the asymptotic variance of any regular M–estimator is larger than that of the MLE. Thus, we are in the setting of a dominated family, so we have densities $dP_\theta/d\mu = f(\cdot;\theta)$. We will also assume that all the regularity conditions of Theorem 3.1.2 hold, and others as needed.

For simplicity, we first consider the case where θ is one dimensional and $g(\theta) = \theta$. First rewrite equation (3.240) as

$$\int \psi(x,\theta) f(x;\theta)\, d\mu(x) \;=\; 0$$

and differentiate under the integral sign (this is the other regularity condition we need), using the product rule, to obtain

$$\int D\psi(x,\theta) f(x;\theta)\, d\mu(x) \;+\; \int \psi(x,\theta) Df(x;\theta)\, d\mu(x) \;=\; 0 \quad .$$

Now applying the usual trick of multiplying and dividing by $f(x;\theta)$ in the second integral, we obtain

$$\int D\psi(x,\theta) f(x;\theta)\, d\mu(x) \;=\; -\int \psi(x,\theta)[D\log f(x;\theta)] f(x;\theta)\, d\mu(x) \tag{3.243}$$

and applying Cauchy–Schwartz to the second integral, we obtain

$$\left\{\int D\psi(x,\theta) f(x;\theta)\, d\mu(x)\right\}^2 \ \le$$

$$\left\{\int \psi(x,\theta)^2 f(x;\theta)\, d\mu(x)\right\}\left\{\int [D\log f(x;\theta)]^2 f(x;\theta)\, d\mu(x)\right\} \quad .$$

Plugging this back into equation (3.242) yields

$$\begin{aligned} v_\psi(\theta) \ &\ge\ \frac{\int \psi(x,\theta)^2 f(x;\theta)\, d\mu(x)}{\{\int \psi(x,\theta)^2 f(x;\theta)\, d\mu(x)\}\{\int [D\log f(x;\theta)]^2 f(x;\theta)\, d\mu(x)\}} \\ &=\ \frac{1}{\int [D\log f(x;\theta)]^2 f(x;\theta)\, d\mu(x)} \ =\ \frac{1}{\mathcal{I}(\theta)} \end{aligned} \qquad (3.244)$$

which is the asymptotic variance of the MLE of θ. This establishes then that in the case of a univariate parameter, no M–estimator is more efficient than the MLE, under regularity.

Now we consider the multiparameter case but with a univariate estimand. To derive the analog of (3.243), differentiating w.r.t. θ and interchanging differentiation and integration in the equation

$$\int \psi(x, g(\theta)) f(x;\theta)\, d\mu(x) \ =\ 0$$

(which is equation (3.240)) gives the following analog of (3.243):

$$\int D\psi(x, g(\theta)) f(x;\theta)\, d\mu(x)\ \triangledown\, g(\theta) \ =\ -\int \psi(x,\theta)[\triangledown \log f(x;\theta)] f(x;\theta)\, d\mu(x) \quad . \qquad (3.245)$$

The factor $\triangledown g(\theta)$ comes from the chain rule:

$$\triangledown_\theta \psi(x, g(\theta)) \ =\ D_\gamma \psi(x,\gamma)\Big|_{\ \gamma=g(\theta)} \triangledown\, g(\theta).$$

In this last display we have been careful to indicate which variables the differentiations are with respect to. Thus, on the r.h.s. one differentiates $\psi(x,\gamma)$ w.r.t. its second argument (remember, we never differentiate w.r.t. x), evaluates that at $g(\theta)$, and multiplies this scalar times the vector $\triangledown g(\theta)$. Of course, one may verify this by computing the appropriate partial derivatives which are the components of the vector on the l.h.s. and applying the chain rule for each of these.

Denote the MLE by $\hat\theta_n$ and recall that the asymptotic variance of the MLE is

$$\triangledown g(\theta)^t \mathcal{I}(\theta)^{-1} \triangledown\, g(\theta)$$

so what we want to show is that

$$\frac{E_\theta[\psi(X, g(\theta))^2]}{\{E_\theta[D\psi(X, g(\theta))]\}^2} \ \ge\ \triangledown g(\theta)^t \mathcal{I}(\theta)^{-1} \triangledown\, g(\theta)^t \quad ,$$

or what is the same

$$E_\theta[\psi(X, g(\theta))^2] \;\geq\; \{E_\theta[D\psi(X, g(\theta))] \bigtriangledown g(\theta)\}^t \mathcal{I}(\theta)^{-1} \left[E_\theta[D\psi(X, g(\theta))] \bigtriangledown g(\theta)\right\} \quad .$$

But in view of (3.245) the last display is equivalent to

$$\begin{gathered} E_\theta[\psi(X, g(\theta))^2] \;\geq \\ \{E_\theta[\psi(X, g(\theta)) \bigtriangledown \log f(X;\theta)]\}^t \mathcal{I}(\theta)^{-1} \{E_\theta[\psi(X, g(\theta)) \bigtriangledown \log f(X;\theta)]\} \quad . \end{gathered}$$

For simplicity, write

$$\begin{aligned} Y &:= \psi(X, g(\theta)) \\ \underline{Z} &:= \bigtriangledown \log f(X;\theta). \end{aligned}$$

Recall that $\mathcal{I}(\theta) = \mathrm{Cov}_\theta[\bigtriangledown \log f(X;\theta)] = \mathrm{Cov}[\underline{Z}]$, and that $E[Y] = 0$ and $E[\underline{Z}] = \underline{0}$. So what we want to show is that

$$\mathrm{Var}[Y] \;\geq\; \mathrm{Cov}[Y, \underline{Z}]\mathrm{Cov}[\underline{Z}]^{-1}\mathrm{Cov}[\underline{Z}, Y] \quad .$$

But this identity is easy. Denote by $\underline{\gamma} = \mathrm{Cov}[\underline{Z}, Y]$ and $V = \mathrm{Cov}[\underline{Z}]$ and let $\underline{a}$ be the vector $\underline{a} = V^{-1}\underline{\gamma}$. Then $\mathrm{Var}[\underline{a}^t\underline{Z}] = \underline{a}^t V \underline{a} = \underline{\gamma}^t V^{-1} V V^{-1} \underline{\gamma} = \underline{\gamma}^t V^{-1} \underline{\gamma}$. We also have $\mathrm{Cov}[Y(\underline{a}^t\underline{Z})] = \mathrm{Cov}[Y, \underline{Z}]\underline{a} = \underline{\gamma}^t V^{-1}\underline{\gamma}$. Therefore, using the correlation inequality (or Cauchy–Schwartz), $[\underline{\gamma}^t V^{-1}\underline{\gamma}]^2 = \{\mathrm{Cov}[Y(\underline{a}^t\underline{Z})]\}^2 \leq \mathrm{Var}[Y]\mathrm{Var}[\underline{a}^t\underline{Z}]$ $= \mathrm{Var}[Y]\underline{\gamma}^t V^{-1}\underline{\gamma}$, which shows that $\mathrm{Var}[Y] \geq \underline{\gamma}^t V^{-1}\underline{\gamma}$, as desired.

We briefly comment on the extra regularity condition needed to derive (3.243) or (3.245), namely that one be able to differentiate under the integral sign in

$$\int \psi(x, g(\theta)) f(x;\theta)\, d\mu(x) \;=\; 0 \quad .$$

In Lemma 3.7.2 it was shown that the assumptions of Theorem 3.7.1 allow one to differentiate under the integral sign in

$$\lambda(\gamma) \;=\; \int \psi(x, \gamma) f(x;\theta_0)\, d\mu(x) \quad .$$

Note that θ_0 is fixed here and we are differentiating w.r.t. the second variable in ψ. In Theorem 3.1.2 we assumed that one can differentiate under the integral sign in

$$\int f(x;\theta)\, d\mu(x) \;=\; 1 \quad .$$

Note that neither of these is enough to permit the differentiation under the integral sign that we needed above.

Certainly the student who is familiar with the derivation of the Rao–Cramér inequality (and all graduate students of statistics should be familiar with that argument) will find similarity with the above argument. It is common to argue that the fact that the MLE is the most efficient regular estimator follows somehow from the Rao–Cramér inequality, but this is false. Recall that the Rao–Cramér inequality says that

$$\frac{\mathrm{Var}_\theta(\hat{\gamma}_n)}{n} \geq \bigtriangledown g(\theta)^t \mathcal{I}(\theta)^{-1} \bigtriangledown g(\theta) \quad ,$$

where $\hat{\gamma}_n$ is any unbiased estimator of $g(\theta)$ based on n i.i.d. observations and $\mathcal{I}(\theta)$ is the information for a single observation. What we have claimed here is that if $\hat{\gamma}_n$ is any regular CAN estimator of $g(\theta)$, then the asymptotic variance $V_g(\theta)$ in

$$\sqrt{n}[\hat{\gamma}_n - g(\theta)] \overset{D}{\to} N(0, V_g(\theta)) \tag{3.246}$$

satisfies

$$V_g(\theta) \geq \bigtriangledown g(\theta)^t \mathcal{I}(\theta)^{-1} \bigtriangledown g(\theta) \quad , \quad \forall \theta \in \mathrm{int}\Theta \quad . \tag{3.247}$$

Now there is a world of difference between the two statements. In the second statement, we are not saying anything about whether the estimator is unbiased or not. Furthermore, $\sqrt{n}[\hat{\gamma}_n - g(\theta)] \overset{D}{\to} N(0, V_g(\theta))$ does not imply $\mathrm{Var}[\sqrt{n}(\hat{\gamma}_n - g(\theta_0)] \to V_g(\theta)$. Recall that convergence in distribution does not imply convergence of moments. Thus, in principle we could have $\mathrm{Var}_\theta[\sqrt{n}(\hat{\gamma}_n - g(\theta))] = \infty$ for all n and $V_g(\theta)$ below the Rao–Cramér lower bound. See Exercises 3.7.12 and 3.7.13 for relevant examples. Thus, even though the statement of the Rao–Cramér inequality is very suggestive about the variance of the limiting normal distribution, it in fact does not apply to give the kind of result we have here.

3.7.5 Counterexamples and Further Topics.

We now show that in fact it is possible for an estimator sequence to be CAN with an asymptotic variance that violates (3.247). The following example is due to Hodges and Le Cam, but one can construct much more general examples.

Example 3.7.4 Let X_1, X_2, …, X_n be i.i.d. $N(\mu, 1)$. Then the information about μ in a single observation is $\mathcal{I}(\mu) = 1$, and the MLE is $\overline{X}_n$ and of course $\mathrm{Law}[\sqrt{n}(\overline{X}_n - \mu)] = N(0, 1)$, $\forall n$. Now fix $a > 0$ and define

$$\delta_n = \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \geq n^{-1/4}, \\ a\overline{X}_n & \text{if } |\overline{X}_n| < n^{-1/4}. \end{cases} \tag{3.248}$$

Given $\epsilon > 0$,

$$\begin{aligned} & P_\mu[\,|\sqrt{n}(\delta_n - \mu) - \sqrt{n}(\overline{X}_n - \mu)| > \epsilon\,] \\ \leq \quad & P_\mu[\delta_n \neq \overline{X}_n] \\ = \quad & P_\mu[-n^{-1/4} < \overline{X}_n < n^{-1/4}] \\ = \quad & P_\mu[\sqrt{n}(-n^{-1/4} - \mu) < \sqrt{n}(\overline{X}_n - \mu) < \sqrt{n}(n^{-1/4} - \mu)] \\ = \quad & \Phi(n^{1/4} - n^{1/2}\mu) - \Phi(-n^{1/4} - n^{1/2}\mu) \quad . \end{aligned}$$

Now if $\mu > 0$, then $\pm n^{1/4} - n^{1/2}\mu \to -\infty$ as $n \to \infty$, and if $\mu < 0$, then $\pm n^{1/4} - n^{1/2}\mu \to \infty$ as $n \to \infty$. Either way, both terms of the r.h.s. of the last display tend to the same value (1 or 0 according as $\mu < 0$ or > 0), and so their difference tends to 0. Thus, we have shown that for $\mu \neq 0$, $\sqrt{n}(\delta_n - \mu)$ and $\sqrt{n}(\overline{X}_n - \mu)$ are convergence equivalent. Since the latter is $N(0,1)$ for all n, it follows that

$$\sqrt{n}(\delta_n - \mu) \overset{D}{\to} N(0,1) \quad , \quad \text{as } n \to \infty, \quad \text{when } \mu \neq 0. \tag{3.249}$$

Similarly, for $\mu = 0$,

$$\begin{array}{rl} & P_0[\,|\sqrt{n}\delta_n \;-\; \sqrt{n}a\overline{X}_n| \;>\; \epsilon\,] \\ \leq & P_0[|\overline{X}_n| \geq n^{-1/4}] \\ = & \Phi(-n^{1/4}) \,+\, 1 \,-\, \Phi(n^{1/4}) \\ \to & 0 \quad , \quad \text{as } n \to \infty \quad , \end{array}$$

which shows that $\sqrt{n}\delta_n$ and $\sqrt{n}a\overline{X}_n$ are convergence equivalent when $\mu = 0$. Thus,

$$\sqrt{n}(\delta_n - \mu) \overset{D}{\to} N(0,a^2) \quad , \quad \text{as } n \to \infty, \quad \text{when } \mu = 0. \tag{3.250}$$

For this example, the Fisher Information for a single observation is $\mathcal{I}(\mu) = 1$, so if we take $a < 1$ then we can make the limiting normal distribution of $\sqrt{n}(\delta_n - \mu)$ have a variance less that of the MLE at the point $\mu = 0$, although at every other value of μ it has a limiting normal distribution with variance equal to the variance of the MLE. We say that such an estimator is *superefficient* at $\mu = 0$. In fact, we can take $a = 0$ and one has that $\sqrt{n}(\delta_n - \mu) \overset{D}{\to} 0$, a degenerate probability measure at 0. One can also construct estimators which are superefficient at more than one point. See Exercise 3.7.15.

□

Such superefficient estimators provide a counterexample to the conjecture that the variance of the asymptotic normal distribution of any CAN estimator cannot be less than the variance of the MLE. However, it can be shown (under regularity conditions on the family, of course) that the set of points where the variance of a limiting normal distribution is less than the variance of the of the asymptotic normal distribution of the MLE is a set of Lebesgue measure 0, so in some sense the set of superefficiency must be a small one. See Shao [Ref???], Theorem 4.16, or Lehmann [Ref ???] ??? for more details and references. Also, if one restricts attention to "regular" estimators (which in some sense we did in the previous section, but we considered only M–estimators), then the MLE is asymptotically optimal in some sense with respect to a loss function which is almost squared error loss. See ???. Now the MLE is not unique in this regard.

There exist many (infinitely many) estimators which are asymptotically fully efficient. For instance, Bayes estimators with regular priors are asymptotically efficient, in regular families. Nonetheless, Bayesian methods have not enjoyed the popularity of maximum likelihood estimation at least classically, perhaps in part because it is relatively easy to compute MLE's and the associated confidence regions. As modern computers have become more powerful the computational issues associated with the integrals required for the computation of the posterior distribution have become much more tractable and the popularity of Bayesian methods is on the increase.

Exercises for Section 5.7.

3.7.1 Let f be a real valued function defined on a compact subset of $I\!R^d$. Suppose f is continuously differentiable. Show that f satisfies the Lipschitz condition

$$|f(x) - f(y)| \;\leq\; M\|x - y\|$$

for all x and y in the domain of f, where M is a constant that doesn't depend on x or y.

3.7.2 Verify that display (3.223) follows from the fact that $\hat{\underline{\lambda}}_n(\theta_0 \pm \delta_1) \xrightarrow{P} \lambda(\theta_0 \pm \delta_1)$.

3.7.3 Verify (3.228).

3.7.4 Under the assumptions of Theorem 3.7.1, show that there exists $\delta_1 > 0$ such that θ_0 is the unique root of $\lambda(\theta) = 0$ in $\{\theta : |\theta - \theta_0| < \delta_1\}$. (See Exercise 3.7.7 below for a result on the uniqueness of $\hat{\theta}_n$ which however depends on the more complicated proof of Theorem 3.7.4.)

3.7.5 Let $\psi_0 : I\!R \longrightarrow I\!R$ be a bounded twice continuously differentiable function with both ψ_0' and ψ_0'' bounded. Assume $\psi_0' > 0$ and $\psi_0(-x) = -\psi_0(x)$ for all x. Let X_1, X_2, ..., be i.i.d. random variables with distribution symmetric about its median m_0 (i.e. $F(m_0 + x) = 1 - F(m_0 - x)$ where F is the c.d.f. of Law$[X_i]$). Show that there is a sequence of random variables $\hat{m}_n$ such that

(i) For each n, $\hat{m}_n$ is the unique solution to

$$\sum_{i=1}^{n} \psi_0(X_i - m) \;=\; 0 \quad .$$

(ii) $\sqrt{n}[\hat{m}_n - m_0] \xrightarrow{D} N(0, \sigma^2)$, and find the asymptotic variance σ^2.

3.7.6 Prove Theorem 3.7.3.

Hints: Let x_1 be any point of F and recursively define $x_{n+1} = h(x_n)$. Show that for $m > n$,

$$\|x_m - x_n\| \;\leq\; \sum_{i=n}^{m-1} \|x_{i+1} - x_i\| \;\leq\; MC^n$$

where M is a finite constant that doesn't depend on n or m. Argue then that $\{x_n\}$ is a Cauchy sequence, that $x_0 = \lim x_n \in F$, and that x_0 is a fixed point.

3.7.7 Under the assumptions of Theorem 3.7.4, show the following:

(a) There exists $\delta_1 > 0$ such that $\underline{\theta}_0$ is the unique root of $\underline{\lambda}(\underline{\theta}) = \underline{0}$ in $\{\underline{\theta} : \|\underline{\theta} - \underline{\theta}_0\| < \delta_1\}$.

(b) Given $\epsilon > 0$, there is a $\delta > 0$ and N such that for all $n \geq N$,

$$P\left[\hat{\lambda}_n(\underline{\theta}) = 0 \quad \text{has a unique root in} \quad \{\underline{\theta} : \|\underline{\theta} - \underline{\theta}_0\| < \delta\}\right] \ > \ 1 - \epsilon \quad .$$

3.7.8 Let $X_{(n)}$ be the maximal order statistic from n i.i.d. $U(0, \theta)$ distribution. Show that $n[\theta_0 - X_{(n)}]$ converges in distribution to an exponential r.v. and determine the mean of the limiting exponential distribution. Hint: $P_{\theta_0}[n(\theta_0 - X_{(n)}) > y] = P_{\theta_0}[\text{all } X_i < \theta_0 - y/n] = \{P_{\theta_0}[X_1 < \theta_0 - y/n]\}^n = [1 - y/(n\theta_0)]^n \to \exp[-y/\theta_0]$ as $n \to \infty$.

3.7.9 Find the *ARE* of the sample mean w.r.t. the sample median as an estimator of location in the location family generated by the Laplace distribution, i.e. the density is

$$f_a(x) \ = \ \frac{1}{2} \exp\left[-|x - a|\right] \quad , \quad x \in \mathbb{R} \quad ,$$

where $-\infty < a < \infty$.

What is the result if we include a scale parameter as well?

3.7.10 (a) Let X_1, X_2, ..., X_n be i.i.d. with $Expo(\mu)$ distribution. Consider the following estimators for $g(\mu) = P_\mu[X_i > x_0] = \exp[-x_0/\mu]$:

$$\begin{aligned} \delta_{1n} &= \frac{1}{n} \sum_{i=1}^{n} I_{(x_0,\infty)}(X_i) \\ \delta_{2n} &= \exp[-x_0/\overline{X}_n] \end{aligned}$$

Thus, δ_{1n} is the fraction of observations exceeding x_0. Find the ARE of the less efficient estimator w.r.t. the more efficient estimator.

(b) Let $r = x_0/\mu$. Find the limits of ARE as $r \to 0$ and $r \to \infty$.

(c) Plot the ARE over a reasonable range of values of r and find graphically the maximum value of ARE and the value of r at which this occurs (approximately).

(d) Suppose the data are not truly from an exponential distribution. Are both estimators still CAN?

3.7.11 Consider i.i.d. $Unif(0, \theta)$ observations and the two estimators

$$\begin{aligned} \delta_{1n} &= \max\{X_i : 1 \leq i \leq n\} \\ \delta_{2n} &= \frac{n+1}{n} \delta_{1n} \end{aligned}$$

Here, δ_{1n} is the MLE and δ_{2n} is the UMVUE. Of course, neither estimator is asymptotically normal, but we will make an asymptotic comparison as best as possible.

(a) Show that the variance of the UMVUE is larger than that of the MLE, but the ratio of variance tends to 1 as $n \to \infty$.

(b) Show that asymptotically, the MSE of the UMVUE is half that of the MLE. Is it reasonable to claim that the MLE is asymptotically efficient in this case?

(Hint: the density of δ_{1n} is $n\theta^{-n}x^{n-1}$ for $0 \le x \le \theta$.)

3.7.12 (a) Suppose $\sqrt{n}[\hat{\gamma}_n - g(\theta)] \overset{D}{\to} N(0, V_g(\theta))$. Let V_n, Y_n, and $\hat{\gamma}_n$ be mutually idependent with Y_n having a finite mean but infinite variance and V_n Bernoulli with success probability $p_n \to 1$ as $n \to \infty$. Put

$$\tilde{\gamma}_n \; := \; V_n\hat{\gamma}_n \; + \; (1 - V_n)Y_n \quad .$$

Show that $Var[\tilde{\gamma}_n] = \infty$ for all n but $\sqrt{n}[\tilde{\gamma}_n - g(\theta)] \overset{D}{\to} N(0, V_g(\theta))$.

(b) Give a distribution for a random variable Y which has finite mean but infinite variance.

3.7.13 Show that the MLE S_n^2 of σ^2 based on i.i.d. $N(\mu, \sigma^2)$ observations (both parameters unknown) has a variance which is strictly smaller than the Rao–Cramér lower bound for unbiased estimates. How can this be?

Advanced Exercises.

3.7.14 Consider a location–scale family

$$f_{ab}(x) \; = \; \frac{1}{b} f_{01}\left(\frac{x-a}{b}\right)$$

(a) Suppose $\hat{a}_1$ and $\hat{a}_2$ are CAN estimators of location which are location–scale equivariant (i.e. location estimators). Show that ARE_{21} is independent of the parameters.

(b) Same as (a) but for CAN estimators of scale which are location invariant and scale equivariant (i.e. scale estimators).

3.7.15 Let X_1, X_2, ..., X_n be i.i.d. $N(\mu, 1)$. Construct estimators of μ which are superefficient at a set of points $\Lambda = \{\mu_i : 1 \le i \le \infty\}$ where

(a) $\inf\{|\mu_i - \mu_j| : i \ne j\} > 0$.

(b) Λ is the set of dyadic rationals, i.e. the set of all points of the form $k/2^{-j}$ where k is an arbitrary integer and j is a nonnegative integer. (Hint: choose sequences $j_n \to \infty$ at an appropriate rate and $\alpha_n \to 0$ at an appropriate rate and put $\delta_n = k/2^{-j}$ for a maximal j with $j \le j_n$ provided $|\overline{X}_n - k/2^{-j}| \le \alpha_n$ and otherwise $\delta_n = \overline{X}_n$.)

3.7.16 Obtain formulae the ARE's of the method of moment estimators of α and β w.r.t. the MLE's in the $Gamma(\alpha, \beta)$ family. If you have access to Mathematica or a similar system, plot graphs of these ARE's.