Chapter 13
Sufficiency and Unbiased Estimation

1. **Conditional Probability and Expectation**
2. **Sufficiency**
3. **Exponential families and sufficiency**
4. **Uses of sufficiency**
5. **Ancillarity and completeness**
6. **Unbiased estimation**
7. **Nonparametric unbiased estimation: U - statistics**

Chapter 13

Sufficiency and Unbiased Estimation

1 Conditional probability and expectation

References:

- Sections 2.3, 2.4, and 2.5, Lehmann and Romano, TSH; pages 34 - 46.
- Billingsley, 1986, pages 419 - 479;
- Williams, 1991, pages 83 - 92.

Basic Notation. Suppose that X is an integrable random variable on a probability space $(\Omega, \mathcal{A}, P)$, and that $\mathcal{A}_0 \subset \mathcal{A}$ is a sub-sigma field. Typically $\mathcal{A}_0 = T^{-1}(\mathcal{B})$ where T is another random variable on $(\Omega, \mathcal{A}, P)$, $T : (\Omega, \mathcal{A}) \to (\mathbb{T}, \mathcal{B})$

Definition 1.1 A *conditional expectation* of X given $\mathcal{A}_0$, denoted $E(X|\mathcal{A}_0)$, is an integrable $\mathcal{A}_0-$measurable random variable satisfying

$$(1) \qquad \int_{A_0} E(X|\mathcal{A}_0) dP = \int_{A_0} X dP \qquad \text{for all } \ A_0 \in \mathcal{A}_0.$$

Proposition 1.1 $E(X|\mathcal{A}_0)$ exists.

Proof. Consider $X \geq 0$. Define ν on $\mathcal{A}_0$ by

$$nu(A_0) = \int_{A_0} X dP \qquad \text{for } \ A_0 \in \mathcal{A}_0.$$

The measure ν is finite since X is integrable, and ν is absolutely continuous with respect to $P|_{\mathcal{A}_0}$. Hence by the Radon - Nikodym theorem there is an $\mathcal{A}_0-$measurable function f such that

$$\int_{A_0} X dP = \nu(A_0) = \int_{A_0} f dP.$$

This function f has the desired properties; i.e. $f = E(X|\mathcal{A}_0)$. if $X = X^+ - X^-$, then $E(X^+|\mathcal{A}_0) - E(X^-|\mathcal{A}_0)$ works. $\square$

Theorem 1.1 (Properties of conditional expectations). Let X, Y, Y_n be integrable random variables on $(\Omega, \mathcal{A}, P)$. Let $\mathcal{D}$ be a sub-sigma field of $\mathcal{A}$. Let g be measurable. Then for *any* versions of the conditional expectations the following hold:

(i) (Linearity) $E(aX + bY|\mathcal{D}) = aE(X|\mathcal{D} + bE(Y|\mathcal{D})$.

(ii) $EY = E[E(Y|\mathcal{D})]$.

(iii) (Monotonicity) $X \leq Y$ a.s. P implies $E(X|\mathcal{D}) \leq E(Y|\mathcal{D})$ a.s.

(iv) (MCT) If $0 \leq Y_n \uparrow Y$ a.s. P, then $E(Y_n|\mathcal{D}) \uparrow E(Y|\mathcal{D})$ a.s.

(v) (Fatou) If $0 \leq Y_n$ a.s. P, then $E(\underline{\lim} Y_n|\mathcal{D}) \leq \underline{\lim} E(Y_n|\mathcal{D})$ a.s.

(vi) (DCT) If $|Y_n| \leq X$ for all n and $Y_n \rightarrow_{a.s.} Y$ a.s. P, then $E(Y_n|\mathcal{D}) \rightarrow E(Y|\mathcal{D})$ a.s.

(vii) If Y is $\mathcal{D}-$measurable and XY is integrable, then $E(XY|\mathcal{D}) = YE(X|\mathcal{D})$ a.s.

(viii) If $\mathcal{F}(Y)$ and $\mathcal{D}$ are independent, then $E(Y|\mathcal{D}) = E(Y)$ a.s.

(ix) (Stepwise smoothing). If $\mathcal{D} \subset \mathcal{E} \subset \mathcal{A}$, then $E[E(Y|\mathcal{D})] = E[Y|\mathcal{D}]$ a.s.

(x) If $\mathcal{F}(Y, X_1)$ is independent of $\mathcal{F}(X_2)$, then $E(Y|X_1, X_2) = E(Y|X_1)$ a.s.

(xi) c_r, Hölder, Liapunov, Minkowski, and Jensen inequaltieis hold for $E(\cdot|\mathcal{D})$. Jensen: $g(E(Y|\mathcal{D})) \leq E[g(Y)|\mathcal{D})$ a.s. $P|_{\mathcal{D}}$ for g convex and $g(Y)$ integrable.

(xii) If $Y_n \rightarrow_r Y$ for $r \geq 1$, then $E(Y_n|\mathcal{D}) \rightarrow_r E(Y|\mathcal{D})$.

(xiii) g is a version of $E(Y|\mathcal{D})$ if and only if $E(XY) = E(Xg)$ for all bounded $\mathcal{D}-$measurable random variables X.

(xiv) If $P(D) = 0$ or 1 for all $D \in \mathcal{D}$, then $E(Y|\mathcal{D}) = EY$ a.s.

In the case $\mathcal{A}_0 = T^{-1}(\mathcal{B})$ where $T : (\Omega, \mathcal{A}) \rightarrow (\mathbb{T}, \mathcal{B})$, the assertion that $f = E(X|\mathcal{A}_0)$ is $\mathcal{A}_0$-measurable is equivalent to stating that $f(\omega) = g(T(\omega))$ for all $\omega \in \Omega$ where g is a $\mathcal{B}$-measurable function on $\mathbb{T}$; see lemma 2.3.1, TSH, page 35. Thus for $A_0 = T^{-1}(B)$ with $B \in \mathcal{B}$, the change of variable theorem (lemma 2.3.2) TSH page 36 yields

$$\int_{A_0} f dP = \int_{T^{-1}(B)} f dP = \int_{T^{-1}(B)} g(T) dP = \int_B g dP_T$$

where P_T is the measure induced on $(\mathbb{T}, \mathcal{B})$ by $P_T(B) \equiv P(T^{-1}(B))$. We may write

$$f(\omega) = E(X|\mathcal{A}_0)(\omega) = E(X|T(\omega)), \qquad \mathcal{A}_0 - \text{measurable},$$

or vies it as the $\mathcal{B}-$measurable function g on $\mathbb{T}$

$$g(t) \equiv E(X|t), \qquad \mathcal{B} - \text{measurable}.$$

For $X = 1_A$, $A \in \mathcal{A}$, the conditional expectation is called conditional probability. Its defining equation is thus

$$P(A_0 \cap A) = \int_{A_0} f dP \qquad \text{for all } \ A_0 \in \mathcal{A}_0,$$

and we denote it by $P(A|\mathcal{A}_0)$ on Ω and by $P(A|t)$ on $\mathbb{T}$ when $\mathcal{A}_0 \equiv T^{-1}(\mathcal{B})$ where $T : (\Omega, \mathcal{A}) \to (\mathbb{T}, \mathcal{B})$. Thus for each fixed set $A \in \mathcal{A}$, we have defined uniquely a.s. P_T a function $P(A|t)$. But in elementary classes we think of $P(A|t)$ as a distribution on $(\Omega, \mathcal{A})$ for each fixed t. The following theorem says that this is usually justified.

Theorem 1.2 (Existence of regular conditional probabilities). If $(\Omega, \mathcal{A})$ is Euclidean, then there exist determinations of the functions $P(A|t)$ on $\mathbb{T}$ such that for each fixed t the function $P(\cdot|t)$ from $\mathcal{A}$ to $[0, 1]$ is a probability measure over $\mathcal{A}$. We denote them by $P_{X|t}(A)$, $A \in \mathcal{A}$. (These are called *regular conditional probabilities*.)

Theorem 1.3 If X is a random vector and $f(X)$ is integrable, then

$$E\{f(X)|t\} = \int_{\mathcal{X}} f(x) dP_{X|t}(x)$$

for all t except possibly in some set B having $P_T(B) = 0$.

2 Sufficiency

References:

- Section 1.6, Lehmann and Casella, TPE;
- Sections 1.9 and 2.6, Lehmann and Romano, TSH.

Notation. The typical statistical setup is often

$$Prob(X \in A) = P_\theta(A) \qquad \text{when} \quad \theta \in \Theta \quad \text{is true}$$

where $(\mathcal{X}, \mathcal{A}, P_\theta)$ is a probability space for each $\theta \in \Theta$.

Definition 2.1 $T : (\mathcal{X}, \mathcal{A}) \to (\mathbb{T}, \mathcal{B})$ is *sufficient for* θ (or for $\mathcal{P} \equiv \{P_\theta : \ \theta \in \Theta\}$) if there exist versions of $P_\theta(A|t)$ or of their densities $p_\theta(x|t)$ which do not depend on θ.

Example 2.1 Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ) with $0 < \theta < 1$; let $T \equiv \sum_1^n X_i$. Then T is sufficient for θ since

$$p_\theta(\underline{x}|t) = \frac{p_\theta(\underline{x}, t)}{p_\theta(t)} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

for all θ and all $\underline{x}$ having $p_\theta(t) > 0$.

Example 2.2 Let $X_1, \ldots, X_n$ be i.i.d. Poisson(θ) with $0 < \theta < \infty$; let $T = \sum_1^n X_i$. Then T is sufficient for θ since

$$p_\theta(\underline{x}|t) = \frac{p_\theta(\underline{x}, t)}{p_\theta(t)} = \frac{e^{-n\theta}\theta^{\sum x_i}/\prod x_i!}{e^{-n\theta}(n\theta)^t/t!} = \binom{t}{x_1 \cdots x_n}\left(\frac{1}{n}\right)^t.$$

Example 2.3 Let $X_1, \ldots, X_n$ be i.i.d. with continuous d.f. F. Then the order statistics $(X_{(1)}, \ldots, X_{(n)}$ are sufficient for F; equivalently $\mathbb{F}_n \equiv n^{-1}1\{X_i \le \cdot\}$ is sufficient for F. (See TSH pages 37 - 176.)

Example 2.4 Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, and let $(\mathcal{X}^n, \mathcal{A}^n)$ be its $n-$fold product space. For any $P \in \mathcal{M} \equiv \{$all probability measures on $\mathcal{A}\}$, let P^n denote the distribution of $X_1, \ldots, X_n$ i.i.d. P. Let $\mathbb{P}_n \equiv n^{-1}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure. Then $\mathbb{P}_n$ is sufficient for $P \in \mathcal{M}$. (For the proof, see the end of this section.)

Theorem 2.1 (Neyman - Fisher - Halmos - Savage factorization theorem). If the distributions $\{P_\theta : \ \theta \in \Theta\}$ have densities p_θ with respect to a $\sigma-$finite measure μ, then T is sufficient for θ if and only if there exist nonnegative $\mathcal{B}-$measurable functions g_θ on $\mathbb{T}$ and a non-negative $\mathcal{A}-$measurable function h on $\mathcal{X}$ such that

$$p_\theta(x) = g_\theta(T(x))h(x) \qquad \text{a.e.} \quad (\mathcal{X}, \mathcal{A}, \mu).$$

Proof. TSH, theorem 2.6.2 and corollary 2.6.1, pages 45 and 46. □

Example 2.5 (Markov dependent Bernoulli trials). Suppose that $X_i \sim$ Bernoulli(p), $i = 1, \ldots, n$ as in example 2.1, but now suppose that the X_i form a Markov chain with

$$P(X_i = 1 | X_{i-1}) = \lambda, \qquad i = 2, 3, \ldots, n.$$

Then the remaining transitions probabilities are all determined and

$$\begin{aligned} P(X_i = 1 | X_{i-1} = 0) &= (1 - \lambda)p/q, \\ P(X_i = 0 | X_{i-1} = 1) &= 1 - \lambda, \\ P(X_i = 0 | X_{i-1} = 0) &= (1 - 2p + \lambda p)/q, \end{aligned}$$

and

$$\Theta = \{(p, \lambda) : \ (2p - 1)/p \ \vee 0 \leq \lambda \leq 1, \ 0 \leq p \leq 1\}.$$

Then

$$P_\theta(\underline{X} = \underline{x}) = \frac{(1 - 2p + \lambda p)}{q^{n-2}} a^r b^s c^t$$

where

$$r = \sum_{i=2}^n x_{i-1} x_i, \qquad s = \sum_{i=1}^n x_i, \qquad t = x_1 + x_n,$$

and

$$\begin{aligned} a &= \frac{\lambda(1 - 2p + \lambda p)}{p(1 - \lambda)^2}, \\ b &= \frac{(1 - \lambda)^2 pq}{(1 - 2p + \lambda p)^2}, \\ c &= \frac{(1 - 2p + \lambda p)}{q(1 - \lambda)}. \end{aligned}$$

Thus $(R, S, T) \equiv (\sum_2^n X_{i-1} X_i, \sum_1^n X_i, X_1 + X_n)$ is sufficient for Θ by the factorization theorem; see Klotz (1973).

Example 2.6 (Univariate normal). Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then $(\sum X_i, \sum X_i^2)$ or $(\overline{X}, S^2)$ (with $S^2 \equiv \sum (X_i - \overline{X})^2/(n-1)$ is sufficient for (μ, σ^2) by the factorization theorem.

Example 2.7 (Multivariate normal). Let $\underline{X}_1, \ldots, \underline{X}_n$ be i.i.d. $N_k(\underline{\mu}, \Sigma)$. Then $(\sum \underline{X}_i, \sum \underline{X}_i \underline{X}_i^T)$ or $(\underline{\overline{X}}, \hat{\Sigma})$ (with $\hat{\Sigma} \equiv n^{-1} \sum \underline{X}_i \underline{X}_i^T - \underline{\overline{X}}\,\underline{\overline{X}}^T$) is sufficient for (μ, Σ) by the factorization theorem.

Example 2.8 Suppose that $X_1, \ldots, X_n$ are i.i.d. Exponential(μ, σ):

$$p_\theta(x) = \sigma^{-1} \exp(-(x - \mu)/\sigma) 1_{[\mu, \infty)}(x)$$

where $\theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$. Then $(\min X_i, \sum_{i=1}^n (X_i - \min X_j))$ is sufficient for $\theta = (\mu, \sigma)$ by the factorization theorem.

Example 2.9 if $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$ in $\mathbb{R}^n$ where $\underline{\epsilon} \sim N_n(0, \sigma^2 I)$, then $\underline{\hat{\beta}}_n \equiv (X^T X)^{-1} X^T \underline{Y}$ and $SS_E \equiv \|\underline{Y} - X\underline{\hat{\beta}}\|^2$ are sufficient for $(\underline{\beta}, \sigma^2)$ by the factorization theorem.

Example 2.10 If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, c^2\mu^2)$ with c^2 known, then $(\overline{X}, S^2)$ is sufficient for μ.

Example 2.11 Let $X_1, \ldots, X_n$ be i.i.d. Exponential(θ), $p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$. Let $x_0 > 0$ be a fixed number, and suppose we observe only $Y_i \equiv X_i \wedge x_0$, $\delta_i \equiv 1\{X_i \le x_0\}$, $i = 1, \ldots, n$. Then

$$p_\theta(\underline{y}, \underline{\delta}) = \prod_{i=1}^n \{\theta e^{-\theta y_i}\}^{\delta_i} \{e^{-\theta x_0}\}^{1-\delta_i} = \theta^N \exp(-\theta T)$$

where $N \equiv \sum_1^n \delta_i =$ the number of observations failed by time x_0 and

$$T \equiv \sum_{i=1}^n Y_i \delta_i + x_0(n - N) = \text{total time on test.}$$

Thus (N, T) is sufficient for θ by the factorization theorem.

Example 2.12 (Buffon's needle problem). Perlman and Wichura (1975) give a very nice series of examples of the use of sufficiency in variants of the classical "Buffon's needle problem".

Proof. **for example 2.5**: First, let

$$\mathcal{S}_n \equiv \sigma\{A \in \mathcal{A}^n : \ \pi A = A \ \text{ for all } \ \pi \in \Pi_n\};$$

here Π_n is the collection of all permutations of $\{1, \ldots, n\}$ and $\pi \underline{x} = (x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$. We claim that

$$P^n(A|\mathcal{S}_n) = \frac{1}{n!} \sum_{\pi \in \Pi} 1_A(\pi \underline{X}) \qquad \text{a.s.} \ \ P^n.$$

To see this, for any integrable function $f : \mathcal{X}^n \to \mathbb{R}$, let $Y = f(\underline{X})$, and set

$$f_0(\underline{x}) = \frac{1}{n!} \sum_{\pi \in \Pi} f(\pi \underline{x}) = \frac{1}{n!} \#\text{of permutations of } \ \underline{x} \ \text{ in } A$$

if $f = 1_A$. Then f_0 is $\mathcal{S}_n$-measurable, since it is a symmetric function of its arguments. Also, since the X's are identically distributed, for $A_0 \in \mathcal{S}_n \equiv \mathcal{A}_0$,

$$\int_{A_0} f(\underline{x}) dP^n(\underline{x}) = \int_{A_0} f(\pi \underline{x}) dP^n(\underline{x})$$

for all $\pi \in \Pi_n$. Summing across this equality on π and dividing by $n!$ yields

$$\int_{A_0} f(\underline{x}) dP^n(\underline{x}) = \int_{A_0} f_0(\underline{x}) dP^n(\underline{x}).$$

and this implies that

$$E(f(\underline{X})|\mathcal{S}_n) = E(Y|\mathcal{S}_n) = f_0(\underline{X}) \in m\mathcal{S}_n.$$

To get from $\mathcal{S}_n$ to $\mathbb{P}_n$ see Dudley (1999), theorem 5.1.9, page 177.

3 Exponential Families and Sufficiency

Definition 3.1 Suppose that $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^m, \mathcal{B}^m)$ for some $m \geq 1$, and that $X \sim P_\theta$ has density

$$p_\theta(\underline{x}) = c(\theta)\exp(\sum_{j=1}^k Q_j(\theta)T_j(\underline{x}))h(\underline{x})$$

with respect to a $\sigma-$finite measure μ on some subset of $\mathbb{R}^m$. Then $\{p_\theta : \theta \in \Theta\}$ is called a *k-parameter exponential family*.

Example 3.1 (Bernoulli). If $X = (X_1, \ldots, X_n)$ are i.i.d. Bernoulli(θ)

$$p_\theta(\underline{x}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = e^{n\log(1-\theta)}\exp(\log(\theta/(1-\theta))\sum_1^n x_i)$$

on $\{0,1\}^n$ is an exponential family, and, by the factorization theorem $T = \sum_1^n X_i$ is sufficient.

Example 3.2 If $X = (X_1, \ldots, X_n)$ are i.i.d. $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, then

$$\begin{aligned} p_\theta(\underline{x}) &= (2\pi\sigma^2)^{-n/2}\exp(-(2\sigma^2)^{-1}\sum_1^n (x_i-\mu)^2) \\ &= (2\pi\sigma^2)^{-n/2}e^{-n\mu/2\sigma^2}\exp((-1/2\sigma^2)\sum_1^n x_i^2 + (\mu/\sigma^2)\sum_1^n x_i) \end{aligned}$$

is an exponential family, and by the factorization theorem $(\sum_1^n X_i, \sum_1^n X_i^2)$ is sufficient.

Example 3.3 (Counterexample: shifted exponential distributions). Suppose that $X_1, \ldots, X_n$ are i.i.d. with the shifted Exponential(μ, σ) distribution

$$p_\theta(x) = \sigma^{-1}\exp(-(x-\mu)/\sigma)1_{[\mu,\infty)}(x).$$

Then

$$p_\theta(\underline{x}) = \sigma^{-n}\exp\left\{-\sum_{i=1}^n (X_i-\mu)/\sigma\right\}1_{[\mu,\infty)}(\min x_i)$$

is *not* an exponential family. As noted in section 2, the factorization theorem still works and shows that $(\sum(X_i - \min X_j), \min X_i)$ are sufficient. Note that a support set depending on θ is *not* allowed for an exponential family.

Example 3.4 (Inverse Gaussian). This distribution is given by the density

$$p(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} x^{-3/2}\exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)1_{(0,\infty)}(x).$$

Here μ is the mean and λ is a precision parameter. It sometimes is useful to reparametrize using $\alpha = \lambda/\mu^2$, yielding

$$p(x;\alpha,\lambda) = (2\pi x^3)^{1/2}\exp\left((\alpha\lambda)^{1/2} - \frac{1}{2}\log\lambda - \frac{1}{2}\alpha x - \frac{\lambda}{2}x^{-1}\right),$$

so that for a sample of size n, $(\sum_1^n X_i, \sum_1^n X_i^{-1})$ is sufficient for the natural parameter $(\alpha/2, \lambda/2)$.

Theorem 3.1 For the $k-$parameter exponential family $T = (T_1(X), \ldots, T_k(X))$ is sufficient.

Proof. This follows immediately from the factorization theorem. $\Box$

Remark 3.1 If

$$p_\theta(x) = c(\theta) \exp\left(\sum_{j=1}^k \theta_j T_j(x)\right) h(x), \qquad \theta \in \Theta \subset \mathbb{R}^k,$$

with respect to μ, then p_θ is said to have its *natural parametrization*. Note that Θ is convex in this parametrization since, for $0 < \lambda < 1$, $\overline{\lambda} \equiv 1 - \lambda$, $\theta, \theta^* \in \Theta \subset \mathbb{R}^k$,

$$\begin{aligned}
&\int \exp\left(\sum_{j=1}^k (\lambda\theta_j + \overline{\lambda}\theta_j^*) T_j(x)\right) h(x) d\mu(x) \\
&= \int \left\{\exp\left(\sum_{j=1}^k \theta_j T_j(x)\right)\right\}^\lambda \left\{\exp\left(\sum_{j=1}^k \theta_j^* T_j(x)\right)\right\}^{\overline{\lambda}} h(x) d\mu(x) \\
&\leq \left\{\int \exp\left(\sum_{j=1}^k \theta_j T_j(x)\right) h(x) d\mu(x)\right\}^\lambda \left\{\int \exp\left(\sum_{j=1}^k \theta_j^* T_j(x)\right) h(x) d\mu(x)\right\}^{\overline{\lambda}} \\
&< \infty
\end{aligned}$$

by Hölder's inequality with $p = 1/\lambda$, $q = 1/\overline{\lambda}$.

Theorem 3.2 If X has the $k = r + s$ parameter exponential family density

$$p_{\theta,\xi}(x) = c(\theta, \xi) \exp\left\{\sum_{i=1}^r \theta_i U_i(x) + \sum_{j=1}^s \xi_j T_j(x)\right\} h(x)$$

with respect to μ, then the marginal distribution of T is the exponential family

$$p_{\theta,\xi}(t) = c(\theta, \xi) \exp\left\{\sum_{j=1}^s \xi_j t_j\right\} H_\theta(t),$$

and the conditional distribution of U given $T = t$ is of the exponential family form

$$p_\theta(u|t) = c_t(\theta) \exp\left\{\sum_{i=1}^r \theta_i u_i\right\} \tilde{H}_t(u).$$

Proof. See TSH, page 48, lemma 2.7.2. $\Box$

4 Applications of Sufficiency

Our first application of sufficiency is to show quite generally that nothing is lost in terms of risk if we base decisions on a sufficient statistic.

Theorem 4.1 Let $X \sim P_\theta \in \mathcal{P}$, $\theta \in \Theta$, and let $T = T(X)$ be sufficient for $\mathcal{P}$. Suppose the loss function is $L : \Theta \times \mathcal{A} \to \mathbb{R}^+$. Then for any procedure $d = d(\cdot|X) \in \mathcal{D}$ there exists a (possibly randomized) procedure $d^*(\cdot|T)$ depending on X only through $T(X)$ which has the same risk function as $d(\cdot|X)$: $R(\theta, d^*) = R(\theta, d)$ for all $\theta \in \Theta$.

Proof. First we give the proof for a finite action space $\mathcal{A} = \{a_1, \ldots, a_k\}$. Define a new rule d^* by

$$d^*(a_i|T) \equiv E\{d(a_i|X)|T\};$$

by sufficiency d^* does not depend on θ. Then

$$\begin{aligned}
R(\theta, d) &= E_\theta L(\theta, d(\cdot|X)) \\
&= \int_{\mathcal{X}} \sum_{i=1}^k L(\theta, a_i) d(a_i|x) dP_\theta(x) \\
&= \sum_{i=1}^k L(\theta, a_i) E_\theta\{d(a_i|X)\} \\
&= \sum_{i=1}^k L(\theta, a_i) E_\theta\{E[d(a_i|X)|T]\} \\
&= \sum_{i=1}^k L(\theta, a_i) E_\theta\{d^*(a_i|T)\} \\
&= \int_{\mathbb{T}} \sum_{i=1}^k L(\theta, a_i) d^*(a_i|t) dP_\theta^T(t) \\
&= R(\theta, d^*),
\end{aligned}$$

completing the proof in the case that $\mathcal{A}$ is finite. $\square$

Now we prove the statement for a general action space $\mathcal{A}$ under the assumption that regular conditional expectations exist. Our proof will use the following lemma:

Lemma 4.1 if $f \geq 0$, then

$$\int f d\mu = \int_0^\infty \mu(\{x : f(x) > h\}) dh.$$

Proof. This is almost exactly the same as in the case of a probability measure μ:

$$\int f d\mu = \int_{\mathcal{X}} \int_{(0, f(x))} dh d\mu(x) = \int_0^\infty \int_{\{x: f(x) > h\}} d\mu(x) dh.$$

$\square$

Proof. (continued). Now for the general proof: for a.e. (P_T) fixed value of $T=t$, $P(\cdot|T=t)$ is a probability distribution that does not depend on θ (since T is sufficient). Thus for $B \in \mathcal{B}_{\mathcal{A}}$ (a sigma-field of subsets of the actions space $\mathcal{A}$) we may define

$$d^*(B|t) = \int_{\mathcal{X}} d(B|x) dP_{X|T}(x|t) = E\{d(B|X)|T=t\};$$

here $d(B|x)$ is a bounded measurable function of x. Thus $d^* : \mathcal{B}_{\mathcal{A}} \times \mathbb{T} \to [0,1]$ is a decision rule. Then by using the lemma (at the second and seventh equalities), Fubini (at the third and sixth equalities), and by computing conditionally (at the fourth equality),

$$\begin{aligned}
R(\theta, d) &= \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) d(da|x) dP_\theta(x) \\
&= \int_{\mathcal{X}} \int_0^\infty d(\{a : L(\theta, a) > h\}|x) dh dP_\theta(x) \\
&= \int_0^\infty \int_{\mathcal{X}} d(\{a : L(\theta, a) > h\}|x) dP_\theta(x) dh \\
&= \int_0^\infty \int_{\mathbb{T}} \int_{\mathcal{X}} d(\{a : L(\theta, a) > h\}|x) dP_{X|T}(x|t) dP_\theta^T(t) dh \\
&= \int_0^\infty \int_{\mathbb{T}} d^*(\{a : L(\theta, a) > h\}|t) dP_\theta^T(t) dh \\
&= \int_{\mathbb{T}} \int_0^\infty d^*(\{a : L(\theta, a) > h\}|t) dh dP_\theta^T(t) \\
&= \int_{\mathbb{T}} \int_{\mathcal{A}} L(\theta, a) d^*(da|t) dP_\theta^T(t) \\
&= R(\theta, d^*);
\end{aligned}$$

where we used the definition of d^* in the fifth equality. □

Here is a related result which does not involve sufficiency per se, but illustrates the role of convexity of the loss function $L(\theta, a)$.

Proposition 4.1 if $L(\theta, \cdot)$ is convex for each $\theta \in \Theta$ and if $\mathcal{A}$ is convex, then for any rule $\phi \in \mathcal{D}$ there is a nonrandomized rule ϕ^* which is at least as good: $R(\theta, \phi^*) \le R(\theta, \phi)$ for all θ.

Proof. This is a straightforward application of Jensen's inequality:

$$\begin{aligned}
R(\theta, \phi) &= \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) \phi(da|x) dP_\theta(x) \\
&\ge \int_{\mathcal{X}} L(\theta, \int_{\mathcal{A}} a \phi(da|x)) dP_\theta(x) \qquad \text{by Jensen's inequality since } L \text{ is convex} \\
&\equiv \int_{\mathcal{X}} L(\theta, \phi^*(x)) dP_\theta(x) \\
&= R(\theta, \phi^*).
\end{aligned}$$

□

Note that we can think of ϕ^* as either

$$\phi^*(B|x) = \delta_{\int_{\mathcal{A}} a\phi(da|x)}(B), \qquad \phi^* : \mathcal{B}_{\mathcal{A}} \times \mathcal{X} \to [0,1],$$

or as

$$\phi^*(x) = \int_{\mathcal{A}} a\phi(da|x), \qquad \phi^* : \mathcal{X} \to \mathcal{A}.$$

The following result shows that by conditioning on a sufficient statistic in the presence of a convex loss function we always yields smaller risk.

Theorem 4.2 (Rao-Blackwell theorem). Let $X \sim P_\theta \in \mathcal{P} \equiv \{P_\theta : \ \theta \in \Theta\}$ and let T be sufficient for $\mathcal{P}$. Suppose that $L(\theta, a)$ is a convex function of a for each $\theta \in \Theta$ and that S is an estimator of $g(\theta)$ (possibly randomized, $S = \phi(\cdot|X)$) with finite risk

$$R(\theta, S) = E_\theta L(\theta, S) \qquad \text{for all} \ \ \theta \in \Theta.$$

Let $S^* \equiv E(S|T)$. Then

$$R(\theta, S^*) \leq R(\theta, S) \qquad \text{for all} \ \ \theta \in \Theta. \tag{1}$$

If $L(\theta, a)$ is a strictly convex function of a, then strict inequality holds in (1) unless $S = S^*$.

Proof. By Jensen's inequality for conditional expectations we have

$$E[L(\theta, S)|T] \geq L(\theta, E(S|T))] \qquad \text{a.s.}$$

Hence

$$\begin{aligned} R(\theta, S) &= E_\theta L(\theta, S) = E_\theta\{E[L(\theta, S)|T]\} \\ &\geq E_\theta\{L(\theta, E(S|T))\} \\ &= E_\theta L(\theta, S^*) = R(\theta, S^*). \end{aligned}$$

if L is strictly convex, then the inequality is strict unless $S = S^*$ a.s. $\square$

5 Ancillarity and completeness

The notion of sufficiency involves lack of dependence on θ of a *conditional distribution*. But it is also of interest to know what functions $V \equiv V(X)$ have *unconditional* distributions which do not depend on θ.

Definition 5.1 Let $X \sim P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$. A statistic $V = V(X)$ ($V : (\mathcal{X}, \mathcal{A}) \to (\mathcal{V}, \mathcal{C})$) is *ancillary* if $P_\theta(V(X) \in C)$ does not depend on θ for all $C \in \mathcal{C}$. V is *first - order ancillary* if $E_\theta V(X)$ does not depend on θ.

Definition 5.2 Let $X \sim P_\theta$ and suppose that T is sufficient . Then $\mathcal{P}_T \equiv \{P_\theta^T : \theta \in \Theta\}$ is *complete* (or T is *complete*) if $E_\theta h(T) = 0$ for all $\theta \in \Theta$ implies $h(T) = 0$ a.s. $\mathcal{P}_T$. Equivalently, T is complete if no non-constant function $h(T)$ is first order ancillary.

Theorem 5.1 (Completeness of an exponential family). Suppose that X has the exponential family distribution with its natural parametrization as in remark 3.1, $T = (T_1, \ldots, T_k)$ and $\mathcal{P}_T = \{P_\theta^T : \theta \in \Theta_0\}$. The family $\mathcal{P}_T$ is complete provided Θ_0 contains a $k-$dimensional rectangle.

Proof. Uniqueness of Laplace transforms. See TSH pages 49 and 116. □

Here are some examples of ancillarity and completeness.

Example 5.1 (Bernoulli) If $(X_1, \ldots, X_n)$ are i.i.d. Bernoulli(θ), then $T = \sum_1^n X_i$ is sufficient and complete by theorem 5.1.

Example 5.2 (Normal, one sample). If $X = (X_1, \ldots, X_n)$ are i.i.d. $N(\mu, \sigma^2)$, then $(\overline{X}_n, S^2)$ is sufficient and complete by theorem 5.1.

Example 5.3 (Normal, two samples). If $X = (X_1, \ldots, X_m)$ are i.i.d. $N(\mu, \sigma^2)$, $Y = (Y_1, \ldots, Y_n)$ are i.i.d. $N(\nu, \tau^2)$ and independent of the X_j's, then $(\overline{X}, \overline{y}, S_X^2, S_Y^2)$ is sufficient and complete by theorem 13.5.3.

Example 5.4 (Normal; two samples with equal means). If the model is as in example 5.3, but $\mu = \nu$, then $(\overline{X}, \overline{Y}, S_X^2, S_Y^2)$ is sufficient, but it is *not* complete since $\overline{X} - \overline{Y} = \mu - \mu = 0$, but $h(T) \equiv \overline{X} - \overline{Y} \neq 0$. [A consequence is that there is no UMVUE of $g(\theta) = \mu$ in this model. Question: what is the MLE and what is its asymptotic behavior?]

Example 5.5 (Uniform$(0, \theta)$). If $X = (X_1, \ldots, X_n)$ are i.i.d. Uniform$(0, \theta)$ for all θ, then $T \equiv \max_{1 \le i \le n} X_i = X_{(n)}$ is sufficient and complete:

$$E_\theta h(T) = \int_0^\theta h(t) \frac{n}{\theta} t^{n-1} dt = 0 \qquad \text{for all } \theta;$$

which implies that

$$\int_0^\theta h(t) t^{n-1} dt = 0 \qquad \text{for all } \theta;$$

which implies, since $h = h^+ - h^-$, that

$$\int_0^\theta h^+(t) t^{n-1} dt = \int_0^\theta h^-(t) t^{n-1} dt = 0 \qquad \text{for all } \theta;$$

which implies, by taking differences over θ and then passing to the sigma-field of sets generated by the intervals (the Borel sigma - field), that

$$\int_A h^+(t)t^{n-1}dt = \int_A h^- t^{n-1}dt = 0 \qquad \text{for all Borel sets } A.$$

By taking $A = \{t : \ h(t) > 0\}$ in this last equality we find that

$$\int_{[t:h(t)>0]} h^+(t)t^{n-1}dt = 0$$

which implies that $h^+(t) = 0$ a.e. Lebesgue. Choosing $A = \{t : \ h(t) < 0\}$ yields $h^-(t) = 0$ a.e. Lebesgue, and hence $h = 0$ a.e. Lebesgue. Thus we conclude that

$$P_\theta(h(T) = 0) = 1 \qquad \text{for all } \theta$$

or $h(T) = 0$ a.s. $\mathcal{P}_T$.

Example 5.6 (Uniform($\theta - 1/2, \theta + 1/2$)). If $X_1, \ldots, X_n$ are i.i.d. Uniform($\theta - 1/2, \theta + 1/2$), $\theta \in \mathbb{R}$, then $T = (X_{(1)}, X_{(n)})$ is sufficient, but $V(X) = X_{(n)} - X_{(1)}$ is ancillary and hence T is *not* complete:

$$E_\theta \left\{ X_{(n)} - X_{(1)} - \frac{n-1}{n+1} \right\} = \left\{ \left(\frac{n}{n+1} + \theta - 1/2 \right) - \left(\frac{1}{n+1} + \theta - 1/2 \right) - \frac{n-1}{n+1} \right\} = 0$$

for all θ.

Example 5.7 (Normal location ancillary). If $X = (X_1, \ldots, X_n)$ are i.i.d. $N(\theta, 1)$, then $V(X) = (X_1 - \overline{X}, \ldots, X_n - \overline{X})^T \sim N_n(0, I - n^{-1}\underline{1}\underline{1}^T)$ is ancillary. [Note that $\underline{X}$ is equivalent to $(\overline{X}, V(X))$.]

Example 5.8 If $X = (X_1, \ldots, X_n)$ are i.i.d Logistic($\theta, 1$), then $T(X) = (X_{(1)}, \ldots, X_{(n)})$ is sufficient for θ (in fact T is minimal sufficient; see Lehmann TPE page 43), but $V(X) = v(T(X)) = (X_{(n)} - X_{(1)}, \ldots, X_{(n)} - X_{(n-1)})$ has a distribution which is not a function of θ and hence V is ancillary; thus T is not complete.

Example 5.9 (Nonparametric family; sufficiency of order statistics; ancillarity of the ranks). If $X = (X_1, \ldots, X_n)$ are i.i.d. $F \in \mathcal{F}_c \equiv \{\text{all continuous df's}\}$, then $T(X) = (X_{(1)}, \ldots, X_{(n)})$ is sufficient for F from example 2.3. As will be seen below, T is complete for $F \in \mathcal{F}_{ac} \equiv \{\text{all df's } F \text{ with a density function } f \text{ w.r.t. Lebesgue measure } \lambda\} \subset \mathcal{F}_c$. If $R = (R_1, \ldots, R_n) \equiv V(X)$ with $R_i \equiv \{\text{number of } X_j's \le X_i\}$, then X is equivalent to (T, R), and $V(X) = R$ is ancillary: $P_F(R = r) = 1/n!$ for all $r \in \Pi \equiv \{\text{all permutations of } \{1, \ldots, n\}\}$. In fact, T and R are independent:

$$P_F(T \in A, \ R = r) = \frac{1}{n!} \int_A n!\, dF^n(x), \qquad A \in \mathcal{B}^n(\text{ordered}), \quad r \in \Pi.$$

The phenomenon exhibited in the last example is quite general, as is shown by the following theorem.

Theorem 5.2 (Basu). if T is complete and sufficient for the family $\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$, then any ancillary statistic V is independent of T.

Proof. Since V is ancillary, $P_\theta(V \in A) \equiv p_A$ does not depend on θ for all A. Since T is sufficient for $\mathcal{P}$, $P(V \in A|T)$ does not depend on θ, and $E_\theta P(V \in A|T) = P(V \in A) \equiv p_A$, or

$$E_\theta\{P(V \in A|T) - p_A\} = 0 \qquad \text{for all} \quad \theta.$$

Hence by completeness $P(V \in A|T) = p_A$ almost surely $\mathcal{P}$. Hence V is independent of T. □

Now we will prove the completeness of the order statistics claimed in example 5.9 above.

Theorem 5.3 (Completeness of the order statistics). Let $\mathcal{F}$ be a convex class of absolutely continuous df's which contains all uniform densities. If $X = (X_1, \ldots, X_n)$ are i.i.d. $F \in \mathcal{F}$, then $T(\underline{X}) = (X_{(1)}, \ldots, X_{(n)})$ is a complete statistic for $F \in \mathcal{F}$.

Proof. We have to show that $E_F h(T) = 0$ for all $F \in \mathcal{F}$ implies $P_F(h(T) = 0) = 1$ for all $F \in \mathcal{F}$.
Step 1: A function $\delta(x)$ (such as $\delta(x) \equiv h(T(x))$ is a function of T only if it is symmetric in its arguments; $\delta(\pi x) = \delta(x)$ with $\pi x \equiv (x_{\pi(1)}, \ldots, x_{\pi(n)})$ for any permutation $\pi = (\pi(1), \ldots, \pi(n))$ of $(1, \ldots, n)$.
Step 2: Let $f_1, \ldots, f_n$ be n densities corresponding to $F_1, \ldots, F_n \in \mathcal{F}$, and let $\alpha_1, \ldots, \alpha_n > 0$. Then $f(x) = \sum_{i=1}^n \alpha_i f_i / \sum_{i=1}^n \alpha_i$ is a density corresponding to $F \in \mathcal{F}$, and $E_F h(T) = 0$ implies

$$\int \cdots \int \delta(x_1, \ldots, x_n) f(x_1) \cdots f(x_n)\, dx = 0$$

or

$$\int \cdots \int \delta(x_1, \ldots, x_n) \prod_{j=1}^n \left(\sum_{i=1}^n \alpha_i f_i(x_j) \right)\, dx = 0$$

for all $\alpha_1, \ldots, \alpha_n > 0$. The left side may be rewritten as a polynomial in $\alpha_1, \ldots, \alpha_n$ which is identically zero, and hence its coefficients must all be zero. In particular, the coefficient of $\alpha_1, \ldots, \alpha_n$ must be zero. This coefficient is

$$\begin{aligned}
C &\equiv \sum_{\pi \in \Pi} \int \cdots \int \delta(x_1, \ldots, x_n) f_1(x_{\pi(1)}) \cdots f_n(x_{\pi(n)}) dx \\
&= \sum_{\pi \in \Pi} \int \cdots \int \delta(\pi x) \prod_{i=1}^n f_i(x_i) dx \\
&= \sum_{\pi \in \Pi} \int \cdots \int \delta(x) \prod_{i=1}^n f_i(x_i) dx \qquad \text{by symmetry of} \quad \delta \\
&= n! \int \cdots \int \delta(x) \prod_{i=1}^n f_i(x_i) dx.
\end{aligned}$$

Now let $f_i(x) = (b_i - a_i)^{-1} 1_{[a_i, b_i]}(x)$, $i = 1, \ldots, n$; i.e. uniform densities on $[a, b]$. Hence $C = 0$ implies

$$\int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \delta(x) dx = 0;$$

that is, the integral of δ over any $n-$dimensional rectangle is 0, and this implies that $\delta = 0$ except on a set of Lebesgue measure 0. Thus $P_F(h(T) = 0) = 1$ for all $F \in \mathcal{F}$. $\square$

For another method of proof, see Lehmann and Romano TPE, example 4.3.4, page 118.

Remarks on Sufficiency and Ancillarity

Suppose we have our choice of two experiments:

(i) We observe $X \sim P_\theta^X$;

(ii) We observe $T \sim P_\theta^T$, and then conditional on $T = t$ we observe $X \sim P_{X|t}$.

Then the distribution of X is P_θ^X in both cases. Thus it seems reasonable that:

A. Inferences about θ should be identical in both models.

B. Only the experiment of observing $T \sim P_\theta^T$ is informative about θ.

We are thus lead to:

Sufficiency principle: If T is sufficient for θ in a given model, then identical conclusions should be drawn from data points x_1 and x_2 having $T(x_1) = T(x_2)$. Thus T partitions the sample space $\mathcal{X}$ into regions on which identical conclusions are to be drawn. The adequacy of this principle in a decision theoretic framework was demonstrated in Theorem 4.1.

Testing the adequacy of the model: The adequacy of the model can be tested by seeing whether the data X given $T = t$ behave in accord with the known (if the model is true) conditional distribution $P_{X|T=t}$.

Recall that a sufficient statistic T induces a partition of the sample space; and in fact it is this partition, rather than the particular statistic inducing the partition, that is the fundamental object.

If no coarser partition of the sample space that retains sufficiency is possible, then T is called *minimal sufficient*. See Lehmann and Casella, TPE, pages 39, 69, and 78.

Consider again the typical statistical setup $X \sim P_\theta$ on $(\mathcal{X}, \mathcal{A})$ for some unknown $\theta \in \Theta$.

Definition 5.3 If $X = (T, V)$ where the distribution of V is independent of θ, then V is called an *ancillary statistic*. Then T is called conditionally sufficient for θ: we have $f_\theta(t, v) = f_\theta(t|v) f(v)$.

More generally, suppose that $\theta = (\theta_1, \theta_2)$ where θ_2 is a nuisance parameter and $\Theta = \Theta_1 \times \Theta_2$. Now suppose that $X = (T, V)$ where $P_\theta^V = P_{\theta_2}^V$ and $P_\theta^{T|V=v} = P_{\theta_1}^{T|V=v}$ for all v. Then V is called *ancillary for θ_1 in the presence of θ_2*: $f_{\theta_1,\theta_2}(t, v) = f_{\theta_1}(t|v) f_{\theta_2}(v)$. This leads to the following *conditionality principle*:

> *Conditionality Principle*: Conclusions about θ_1 are to be drawn as if V were fixed at its observed v. Conditioning on ancillaries leads to partitioning the sample space (just as sufficiency does). The degree to which these sets contain differing amounts of information about θ_1 determines the benefits to be derived from such conditioning.

Examples showing the reasonableness of this principle appear inn Cox and Hinkley (197x), pages 32, 34, 38. They concern:

- Random sample size.

- Mixtures of two normal distributions.
- Conditioning on the independent variables in multiple regression.
- Two measuring instruments.
- Configurations in location - scale models.

Examples showing difficulties with this principle center on non-uniqueness and lack of general methods for constructing them.

6 Unbiased estimation

One of the classical ways of restricting the class of estimators which are to be considered is by imposing the restriction of *unbiasedness*. This is a rather severe restriction, and in fact, if a complete sufficient statistic T is available, then there exists a unique *uniform minimum variance unbiased estimator*, or UMVUE.

Theorem 6.1 (Lehmann - Scheffé). Suppose that T is complete and sufficient for θ. Let S be unbiased for $g(\theta)$ with finite variance. Then $S^* = E(S|T)$ is the unique UMVUE of $g(\theta)$: for any unbiased estimator $d = d(X)$ of $g(\theta)$,

$$R(\theta, S^*) \equiv E_\theta(g(\theta) - S^*)^2 \le E_\theta(g(\theta) - d(X))^2 = R(\theta, d)$$

for all θ.

Proof. First,

$$E_\theta S^* = E_\theta E(S|T) = E_\theta S = g(\theta),$$

by the unbiasedness of S, and hence S^* is also unbiased. Also $Var_\theta[S^*] \le Var_\theta[S]$ by Blackwell - Rao. Moreover, S^* does not depend on the choice of S: if S_1 is unbiased then

$$E_\theta(S^* - S_1^*) = E_\theta\{E(S|T) - E(S_1|T)\} = E_\theta(S) - E_\theta(S_1) = g(\theta) - g(\theta) = 0$$

for all $\theta \in \Theta$. Thus $S^* = S_1^*$ a.s. $\mathcal{P}_T$ by completeness. $\square$

Remark 6.1 Note that by the Rao-Blackwell theorem 4.2, an analogous result for UM(Risk)UE holds when $L(\theta, \cdot)$ is convex for each θ: $S^* \equiv E(S|T)$ in fact minimizes $R(\theta, S) \equiv E_\theta L(\theta, S)$ for all θ in the class of unbiased estimates. See Lehmann and Casella, TPE, theorem 2.1.11, page 88.

For a treatment of the asymptotic efficiency of UMVUE's in parametric problems, see Portnoy (1977).

Methods for finding UMVUE estimators: When T is sufficient and complete we can produce UMVUE's by several different approaches.

A. Produce an estimator of $g(\theta)$ that is a function of T and is unbiased.

B. Find an unbiased estimator S of $g(\theta)$ and compute $E(S|T)$.

C. Solve $E_\theta d(T) = g(\theta)$ for d.

Example 6.1 (Normal(μ, σ^2)). Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Then $(\overline{X}, S^2)$ is sufficient and complete by theorems 3.1 and 5.1.
A. For estimation of $g(\theta) = \mu$, $E_\theta \overline{X} = \mu$, so $\overline{X}$ is the UMVUE of $g(\theta) = \mu$.
B. If $\mu = 0$ is known, then $Y \equiv \sum_1^n X_i^2$ is sufficient and complete, and $Y/\sigma^2 \sim \chi_n^2$, so

$$E\left(\frac{Y}{\sigma^2}\right)^{r/2} = E(\chi_n^2)^{r/2} = K_{n,r}^{-1} \equiv \frac{2^{r/2}\Gamma\left(\frac{n+r}{2}\right)}{\Gamma(n/2)}$$

for $r > -n$ and $K_{n,r} Y^{r/2}$ is the UMVUE of $g(\theta) \equiv \sigma^r$.
C.(i) If $\theta = (\mu, \sigma^2)$ as in A, then $Y \equiv \sum_1^n (X_i - \overline{X})^2$ has $(Y/\sigma^2) \sim \chi_{n-1}^2$, so $K_{n-1,r} Y^{r/2} =$

$K_{n-1,r}(n-1)^{r/2}S^r$ is the UMVUE of σ^r by method A.
C.(ii) If $g(\theta) = \mu/\sigma$, then $\overline{X}K_{n-1,-1}S^{-1}$ is the UMVUE of $g(\theta)$ by independence of $\overline{X}$, S (*under normality only*!), and method A.
C.(iii) If $g(\theta) = \mu + z_p\sigma \equiv x_p$ where $P(X \le x_p) = p$ for a fixed $p \in (0,1)$, then $\overline{X} + z_p K_{n-1,1}S$ is the UMVUE.
C.(iv) If $\sigma = 1$ is known and $g(\theta) \equiv P_\mu(X \le x) = \Phi(x-\mu)$ with $x \in \mathbb{R}$ fixed, then $\Phi((x-\overline{X})/\sqrt{1-1/n})$ is the UMVUE. [Question: what is the UMVUE of this probability if σ is unknown?]

Proof. This goes by method B: $S(\underline{X}) = 1\{X_1 \le x\}$ is an unbiased estimator of $g(\theta) = P_\mu(X \le x)$; and

$$\begin{aligned} S^* \equiv E(S|T) &= P(X_1 \le x|\overline{X}) \\ &= P(X_1 - \overline{X} \le x - \overline{X}|\overline{X}) \\ &= P(X_1 - \overline{X} \le x - \overline{x}|\overline{X} = \overline{x}) \qquad \text{on } [\overline{X} = \overline{x}] \\ &= \Phi((x-\overline{x})/\sqrt{1-1/n}) \qquad \text{on } [\overline{X} = \overline{x}] \end{aligned}$$

by Basu's theorem since the ancillary $X_1 - \overline{X} \sim N(0, 1-1/n)$ is independent of $\overline{X}$. □

Example 6.2 (Bernoulli(θ)). Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ).
(i) Then $\overline{X} = T/n$ has $E_\theta(\overline{X}) = \theta$ and hence is the UMVUE/ of θ.
(ii) If $g(\theta) = \theta(1-\theta)$, $(T/n)(n-T)/(n-1)$ is the UMVUE.
(iii) If $g(\theta) = \theta^r$ with $r \le n$, then

$$S^* = S^*(T) = \frac{T}{n} \cdot \frac{T-1}{n-1} \cdots \frac{T-r+1}{n-r+1}$$

is the UMVUE.

Proof. (ii) Here we use method C: Let $d(t)$ be the estimator; then we want to solve

$$E_\theta d(T) = \sum_{t=0}^n d(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = \theta(1-\theta) \equiv g(\theta).$$

Equivalently,

$$\sum_{t=0}^n d(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t = \frac{\theta(1-\theta)}{(1-\theta)^n} = \frac{\theta}{1-\theta}\left(1 + \frac{\theta}{1-\theta}\right)^{n-2}.$$

By setting $\rho \equiv \theta/(1-\theta)$ and expanding the power on the right side we find that

$$\sum_{t=0}^n d(t)\binom{n}{t}\rho^t = \rho(1+\rho)^{n-2} = \rho\sum_{k=0}^{n-2}\binom{n-2}{k}\rho^k = \sum_{t=1}^{n-1}\binom{n-2}{t-1}\rho^t.$$

Equating coefficients of ρ^t on both sides yields $d(0) = d(n) = 0$,

$$d(t) = \binom{n-2}{t-1}\Big/\binom{n}{t}, \qquad t = 1, \ldots, n-1,$$

or

$$d(t) = \frac{t(n-t)}{n(n-1)}, \qquad t = 0, \ldots, n.$$

This is just the claimed unbiased estimator. □

Example 6.3 (Two normal samples with equal means). Suppose that $X = (X_1, \ldots, X_n)$ are i.i.d. $N(\mu, \sigma^2)$, and that $Y = (Y_1, \ldots, Y_n)$ are i.i.d. $N(\mu, \tau^2)$. A UMVUE estimator of $g(\theta) = \mu$ *does not exist.*

Proof. Suppose that $a = \tau^2/\sigma^2$ is known. Then the joint density is given by

$$C \cdot \exp\left(-\frac{1}{2\tau^2}\left(a\sum_{i=1}^m X_i^2 - \sum_{j=1}^n Y_j^2\right) + \frac{\mu}{\tau^2}\left(a\sum_{i=1}^m X_i + \sum_{j=1}^n Y_j\right)\right)$$

so

$$T_a = \left(a\sum_{i=1}^m X_i^2 - \sum_{j=1}^n Y_j^2, a\sum_{i=1}^m X_i + \sum_{j=1}^n Y_j\right)$$

is a complete sufficient statistic. Since

$$E\left(\sum_{i=1}^m X_i + \frac{1}{a}\sum_{j=1}^n Y_j\right) = \mu\left(m + \frac{1}{a}n\right),$$

$$S_a(X,Y) = \frac{\left(\sum_{i=1}^m X_i + \frac{1}{a}\sum_{j=1}^n Y_j\right)}{\left(m + \frac{1}{a}n\right)}$$

is a UMVU estimator of μ. Note that S_a is unbiased for the original model for any $a > 0$. Suppose there exists a UMVUE of $g(\theta) = \mu$ in the original model, say S^*. Then $Var_\theta(S^*) \le Var_\theta(S_a)$, hence also when $\tau^2 = a\sigma^2$. But then S_a is the unique UMVUE, which implies $S^* = S_a$. But since a can be arbitrary, this is a contradiction; S^* cannot be equal to two different estimators at the same time. $\square$

Also see Lehmann, TPE, example 6.1, page 444. If σ^2 and τ^2 are known, then the estimator

$$\lambda\overline{X} + (1-\lambda)\overline{Y} \qquad \text{with} \qquad \lambda \equiv \frac{\tau^2/n}{\sigma^2/m + \tau^2/n}$$

has minimal variance over convex combinations of $\overline{X}$ and $\overline{Y}$; and if σ^2 and τ^2 are unknown, then a perfectly reasonable estimator is obtained by replacing λ by

$$\hat{\lambda} \equiv \frac{\hat{\tau}^2/n}{\hat{\sigma}^2/m + \hat{\tau}^2/n}$$

where $\hat{\tau}$ and $\hat{\sigma}$ are estimates of τ and σ.

Example 6.4 (An inadmissible UMVUE). Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$, $g(\theta) = \theta^2$. Then

$$E_\theta\{\overline{X}^2 - \frac{1}{n}\} = Var_\theta(\overline{X}) + \{E_\theta(\overline{X})\}^2 - \frac{1}{n} = \theta^2,$$

so $\overline{X}^2 - 1/n$ is the UMVUE of θ^2. But it is *inadmissible* since sometimes $\overline{X}^2 - 1/n \le 0$ whereas $\theta^2 \ge 0$. Thus the estimator $\delta(X) = (\overline{X}^2 - 1/n) \vee 0$ has smaller risk: recall Lemma 5.1, TPE, page 113.

Proof. If $g(\theta) \in [a, b]$ for all $\theta \in \Theta$, then

$$\begin{aligned} R(\theta, d) &= E_\theta L(\theta, d) \\ &= E_\theta L(\theta, d)\{1\{d(X) < a\} + 1\{d(X) > b\} + E_\theta L(\theta, d) 1\{a \leq d(X) \leq b\}. \end{aligned}$$

□

Example 6.5 (Another inadmissible UMVUE). Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$. Then $T = \sum_1^n X_i^2$ is sufficient and complete so the MLE $T/n = n^{-1} \sum_1^n X_i^2$ is the UMVUE of σ^2 since $E_{\sigma^2}(T/n) = \sigma^2$. But T/n is inadmissible for squared error loss: consider estimates of the form $d_c(T) = cT$. Then

$$\begin{aligned} R(\sigma^2, d_c) &= E_{\sigma^2}(\sigma^2 - cT)^2 \\ &= E\{c\sigma^2(T/\sigma^2 - n) + (nc-1)\sigma^2\}^2 \\ &= c^2\sigma^4 2n + (nc-1)^2\sigma^4 \\ &= \sigma^4\{1 - 2nc + n(n+2)c^2\} \end{aligned}$$

which is minimized by $c = (n+2)^{-1}$. Thus $d_{1/(n+2)}(T) = T/(n+2)$ has minimum squared error in the class d_c. It is, in fact, also inadmissible; see TPE page 274, and Ferguson pages 134 - 136.

See Ferguson pages 123 -124 for a nice description of a bioassay problem in which sufficiency was used to good advantage.

7 Nonparametric Unbiased Estimation; U - statistics

Suppose that P is a probability distribution on some sample space $(\mathcal{X}, \mathcal{A})$ and suppose that $X_1, \ldots, X_m$ are i.i.d. P. Let $h : \mathcal{X}^m \to \mathbb{R}$ be a symmetric "kernel" function:

$$h(\pi \underline{x}) \equiv h(x_{\pi(1)}, \ldots, x_{\pi(m)}) = h(x_1, \ldots, x_m) = h(x).$$

for all $x \in \mathcal{X}^m$ and $\pi \in \Pi_m$; if h is not symmetric we can symmetrize it: replace h by

$$h_s(\underline{x}) \equiv \frac{1}{m!} \sum_{\pi \in \Pi_m} h(\pi \underline{x}).$$

Note that

$$E_P h(X_1, \ldots, X_m) = \int \cdots \int h(x_1, \ldots, x_m) dP(x_1) \cdots dP(x_m) \equiv g(P).$$

Now suppose that $X_1, \ldots, X_n$ are i.i.d. P with $n \geq m$, write $\underline{X} = (X_1, \ldots, X_n)$, and let

$$U_n \equiv U_n(\underline{X}) \equiv \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \ldots, X_{i_m})$$

where $\sum_c$ denotes summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ of $\{1, \ldots, n\}$. U_n is called an *$m-$th order $U-$statistic*. Clearly U_n is an unbiased estimator of $g(P)$:

$$E_P U_n = g(P).$$

Moreover, U_n is a symmetric function of the data:

$$U_n(\underline{X}) = U_n(\pi \underline{X})$$

for all $\pi \in \Pi_n$. All this becomes more explicit when $\mathcal{X} = \mathbb{R}$, and we write F instead of P for the probability measure described in terms of its distribution function. Then we write the empirical measure $\mathbb{P}_n$ in terms of the empirical distribution $\mathbb{F}_n$, and this is equivalent to the order statistics $T \equiv \underline{X}_{(\cdot)}$, and $U_n = U_n(\underline{X}_{(\cdot)})$. In fact, if $S = h(X_1, \ldots, X_m)$ so that $E_F S = g(F)$, then

$$S^* \equiv E\{S|T\} = U_n.$$

Example 7.1 Let $\mathcal{F}_2 \equiv \{F \in F_c : \; E_F X^2 < \infty\}$, $g(F) \equiv E_F X = \int x dF(x)$. Then

$$\overline{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \sum_{i=1}^n X_{(i)}$$

is the UMVUE of $g(F)$ (since $T(\underline{X}) = (X_{(1)}, \ldots, X_{(n)})$ is sufficient and complete for $\mathcal{F}_2$. Note that $E(X_1 | T(\underline{X}) = \overline{X}$.

Example 7.2 If $\mathcal{F}_2 \equiv \{F \in F_c : \; E_F X^2 < \infty\}$ and $g(F) \equiv (E_F X)^2 = E_F(X_1 X_2)$, then

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} X_i X_j$$

is the UMVUE of $g(F) = (E_F X)^2$.

Example 7.3 If $\mathcal{F}_4 \equiv \{F \in F_c : \ E_F X^4 < \infty\}$ and

$$g(F) \equiv Var_F(X) = \frac{1}{2} E_F(X_1 - X_2),$$

then

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \frac{1}{2}(X_i - X_j)^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2 = S^2$$

is the unique UMVUE of $g(F) = Var_F(X)$.

Example 7.4 If $F \in \mathcal{F}_c$

$$g(F) \equiv F(x_0) = \int 1_{(-\infty, x_0]}(x) dF(x),$$

then $U_n = n^{-1} \sum_{i=1}^n 1_{(-\infty, x_0]}(X_i) = \mathbb{F}_n(x_0)$ is the unique UMVUE of $g(F)$.

Example 7.5 Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. F on $\mathbb{R}^2$. Set

$$g(F) \equiv \int \int [F(x,y) - F(x,\infty)F(\infty,y)]^2 dF(x,y) = \int \cdots \int h(z_1, \ldots, z_5) dF(z_1) \cdots dF(z_5)$$

where $z_i = (x_i, y_i)$, $i = 1, \ldots, 5$, and

$$h(z_1, \ldots, z_5) = \frac{1}{4} \psi(x_1, x_2, x_3) \psi(x_1, x_4, x_5) \psi(y_1, y_2, y_3) \psi(y_1, y_4, y_5)$$

where

$$\psi(u_1, u_2, u_3) = 1_{[u_2 \le u_1]} - 1_{[u_3 \le u_1]}.$$

Remark 7.1 Note that U_n has a close relative, the $m-$th order $V-$statistic V_n defined by

$$V_n \equiv \frac{1}{n^m} \sum_{i_1=1}^n \cdots \sum_{i_m=1}^n h(X_{i_1}, \ldots, X_{i_m}) = \int \cdots \int_{\mathcal{X}^m} h(x_1, \ldots, x_m) d\mathbb{P}_n(x_1) \cdots d\mathbb{P}_n(x_m)$$

in which the sum is extended to include all of the diagonal terms.

Remark 7.2 A necessary condition for $g(P)$ to have an unbiased estimate is that $g(\alpha P_1 + (1-\alpha)P_2)$ be a polynomial (in α) of degree $m \le n$.
Proof: If $g(P) = \int \cdots \int h(\underline{x}) dP^m(x)$, then

$$g(\alpha P_1 + (1-\alpha)P_2) = \int \cdots \int h(\underline{x}) d\{\alpha P_1(x_1) + (1-\alpha)P_2(x_1)\} \cdots \{\alpha P_1(x_m) + (1-\alpha)P_2(x_m)\}$$

is a polynomial of degree m.

Remark 7.3 There is a lot of theory and probability tools available for $U-$statistics. See Serfling (1980), chapter 5, and Lehmann (1975), appendix 5. For some very interesting work on U-processes, see e.g. Arcones and Giné (1993), a topic which was apparently initiated by Silverman (1983).

Bibliography

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *Ann. Probab.* **21** 1494–1542.

Arcones, M. A. and Giné, E. (1994). U-processes indexed by Vapnik-Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *Stochastic Process. Appl.* **52** 17–38.

Billingsley, P. (1995). *Probability and measure.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

Brown, T. C. and Silverman, B. W. (1979). Rates of Poisson convergence for U-statistics. *J. Appl. Probab.* **16** 428–432.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics.* Chapman and Hall, London.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach.* Probability and Mathematical Statistics, Vol. 1, Academic Press, New York.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *Ann. Statist.* **1** 373–379.

Lehmann, E. L. (1975). *Nonparametrics: statistical methods based on ranks.* Holden-Day Inc., San Francisco, Calif.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation.* Springer Texts in Statistics, Springer-Verlag, New York.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* Springer Texts in Statistics, Springer, New York.

Perlman, M. D. and Wichura, M. J. (1975). Sharpening Buffon's needle. *Amer. Statist.* **29** 157–163.

Portnoy, S. (1977). Asymptotic efficiency of minimum variance unbiased estimators. *Ann. Statist.* **5** 522–529.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* John Wiley & Sons Inc., New York.

Silverman, B. and Brown, T. (1978). Short distances, flat triangles and Poisson limits. *J. Appl. Probab.* **15** 815–825.

Williams, D. (1991). *Probability with martingales.* Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge.