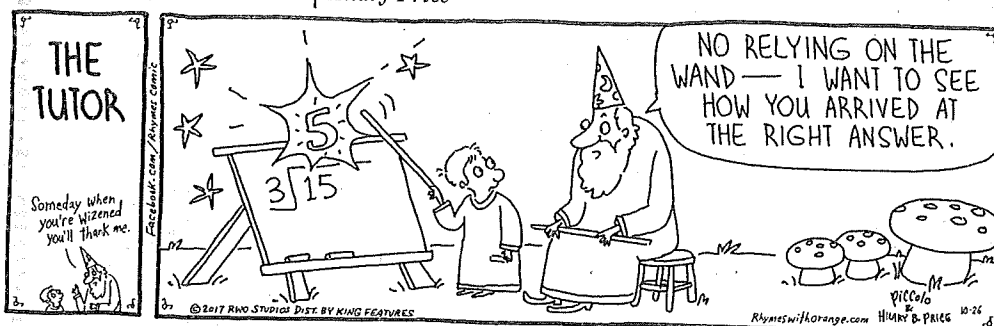


PROBABILITY and MATHEMATICAL STATISTICS II.

CLASS NOTES FOR STAT 513

Michael D. Perlman
Department of Statistics
University of Washington
Seattle, Washington 98195
michael@stat.washington.edu

RHYMES WITH ORANGE | Hilary Price



11. Statistical Models and Sufficient Statistics.....	161
11.1. Data reduction by sufficiency.....	161
11.2. The Fisher-Neyman Factorization Criterion for sufficiency.....	173
11.3. The Factorization Criterion and the likelihood ratio.....	174
11.4. Minimal sufficiency.....	176
12. Ancillarity and Invariance; Sufficiency and Completeness; Minimum-Variance Unbiased Estimation.....	183
12.1. Ancillary statistics and group-invariant families.....	183
12.2. Completeness, sufficiency, and ancillarity.....	186
12.3. Minimum-variance unbiased estimation via a complete sufficient statistic.....	198
12.4. Extension to general convex loss functions.....	205
13. The Information Inequality.....	207
13.1. Variance bounds.....	207
13.2. The role of nuisance parameters.....	215
13.3. Information and sufficiency.....	218
13.4. A rigorous proof of the variance bound.....	220
14. The Role of the Likelihood Ratio in Statistical Inference; the Method of Maximum Likelihood.....	223
14.1. The maximum likelihood estimator.....	230
14.2. Strong consistency of the MLE.....	232
14.3. Asymptotic normality and asymptotic efficiency of the MLE.....	234
14.4. The possibility of multiple roots of the LEQ.....	238
14.5. The multiparameter case.....	245
14.6. The effect of nuisance parameters on asymptotic efficiency.....	250
15. The EM Algorithm for the MLE when Data is Missing.....	261
16. Bayes Estimators.....	272
16.1. Prior distributions: proper vs. improper, informative vs. uninformative.....	276

17. The Elements of Statistical Decision Theory.....	280
17.1. Bayes decision rules.....	282
17.2. Admissible Bayes estimators.....	285
18. Testing Statistical Hypotheses.....	289
18.1. Testing a simple hypothesis vs. a simple alternative.....	289
18.2. Testing composite hypotheses and/or alternatives.....	299
18.3. One-parameter testing problems with one-sided alternatives.....	300
18.4. One-parameter testing problems with two-sided alternatives.....	302
18.5. One-parameter testing problems with nuisance parameters.....	304
18.6. Testing composite hypotheses; the general LRT.....	306
18.7. Relation between $-2 \log \lambda_n$ and Pearson's χ^2 for multinomial hypotheses.....	312
18.8. Proof of Wilks' Theorem; consistency of the Case II LRT.....	314
18.9. Properties of the MLE when the model is incorrect.....	317
18.10. Case III: Non-separated but non-smooth hypotheses.....	320
18.11. Properties of p -values.....	321
19. Sequential Tests and Estimators.....	322
19.1. The sequential probability ratio test.....	322
19.2. Uniform consistency of tests; need for sequential sampling.....	331
Supplement 1. Censored Data exercise with solution.....	342
20. Estimation and Hypothesis Testing with Normal Data	
21. Invariant Tests and Equivariant Estimators	
22. The James-Stein Estimator	
23. How Likely is Simpson's Paradox?	
24. Sharpening Buffon's Needle.	
25. Estimating the Face Probabilities of Shaved Dice.	
26. Circular and Spherical Copulas.	
27. The Emperor's New Tests.	
28. The Role of Reversals in Order-restricted Inference.	
29. Predicting Extinction or Explosion in a Galton-Watson Branching Process.	
30. Variance-Stabilizing Transformations for a Normal Correlation Coefficient.	

11. Statistical Models and Sufficient Statistics.

A *statistical model* $(\mathcal{X}, \mathcal{P})$ consists of a sample space \mathcal{X} and a family

$$(11.1) \quad \mathcal{P} \equiv \{P_\theta \mid \theta \in \Omega\}$$

of possible probability distributions on \mathcal{X} , where θ is an unknown parameter. The model is called *parametric* if θ is finite-dimensional; otherwise the model is called *nonparametric*. Given an observed data vector $X \in \mathcal{X}$, our goals are two-fold:¹⁶

- Make inferences about the unknown P_θ that gave rise to X ;
- Assess the accuracy of our inferences.

11.1. Data reduction by sufficiency.

Often the data can be reduced to a simpler *sufficient statistic* $T(X)$ without losing any information relevant to the goals of inference.

Definition 11.1 (broad). $T(X)$ is a *sufficient statistic* for \mathcal{P} (i.e., for θ) if, for any inference procedure based on X , there exists an equivalent procedure based on $T(X)$. That is, $T(X)$ contains all the relevant information that X provides about the unknown P_θ .

Definition 11.2 (precise). $T(X)$ is a *sufficient statistic* for \mathcal{P} (i.e., for θ) if the conditional distribution of X given T does not depend on θ , i.e., if for every event $A \subseteq \mathcal{X}$, $P_\theta[X \in A \mid T]$ does not depend on θ .

To see that Def. 11.2 \Rightarrow Def. 11.1, suppose that You observe X while I only observe $T(X)$. But I can then generate a “pseudo-observation” X^* according to the conditional distribution $P_\theta[\cdot \mid T]$, which is known to me because, by the sufficiency of T , it does not depend on the unknown θ . Thus, regardless of θ , $X^* \stackrel{d}{=} X$. Therefore, for every possible inference procedure that You can apply to X , I can apply the same procedure to X^* , thereby producing a probabilistically equivalent inference. Since $X \stackrel{d}{=} X^*$, I lose no inferential ability knowing only T rather than X .

¹⁶ We do not discuss here the important question of how to select the model.

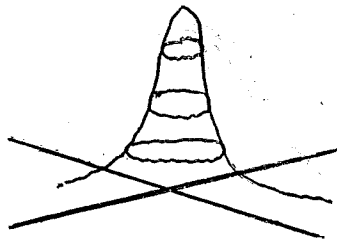
[In fact, the additional information in X that is not already contained in $T(X)$ is extraneous and may be detrimental for inferences about P_θ . An example is provided by the Rao-Blackwell theorem – see the Improvement Lemma 12.2.]

Definition 11.3 (Bayesian sufficiency). Suppose that θ is itself random with prior distribution π on \mathcal{P} (or Ω). $T(X)$ is a *sufficient statistic* for \mathcal{P} (i.e., for θ) if, for every π , the conditional (posterior) distribution of $\theta | X$ depends on X only through the value of $T(X)$.

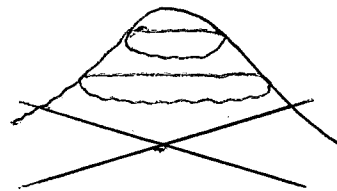
Theorem 11.1. *Definitions 11.1, 11.2, and 11.3 are (usually) equivalent. (This holds in all common cases where \mathcal{P} is a family of pdfs or pmfs, but may not hold in very large nonparametric families, such as the family of all distributions on \mathcal{X} .) (Proof omitted.)*

Example 11.1. (*Normal*($0, \sigma^2$)) Let X_1, \dots, X_n be i.i.d. $N_1(0, \theta)$ rvs with $\theta \equiv \sigma^2$ unknown. The joint pdf of $X \equiv (X_1, \dots, X_n)$ is

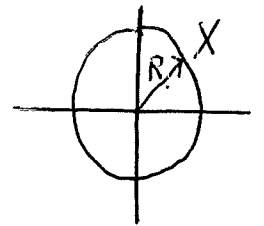
$$(11.2) \quad f_\theta(x) = \prod_{i=1}^n \left[\frac{1}{(2\pi\theta)^{1/2}} e^{-x_i^2/2\theta} \right] = \frac{1}{(2\pi\theta)^{n/2}} e^{-\|x\|^2/2\theta}, \quad x \in \mathbf{R}^n.$$



θ small



θ large.



Thus $f_\theta(x)$ is radial, i.e., has spherical contours, and the distribution of X grows more dispersed as θ increases. If we represent X by its “polar coordinates” $(R, \vec{X}) \equiv (\|X\|, \frac{X}{\|X\|})$, it follows from (6.37) – (6.42) that

$$(11.3) \quad R \perp \vec{X}; \quad R^2 \sim \theta \chi_n^2; \quad \vec{X} \sim \text{Uniform on } \mathcal{S}_n \text{ (the unit sphere in } \mathbf{R}^n).$$

This suggests that all the information about θ in X is contained in $T(X) \equiv R$, and that the unit direction vector \vec{X} contains no information about θ , i.e., this suggests that R is a sufficient statistic for θ .

To verify this according to Definition 11.2, we must show that the conditional distribution of $X \mid R$ does not depend on θ . But $X = R \cdot \vec{X}$, so it follows from (11.3) that $X \mid R$ is uniformly distributed over the sphere of radius R . Since this conditional distribution does not depend on θ , we conclude that R is sufficient.

To verify that R is sufficient according to Definition 11.1, note that if You observe X while I only observe R , I can generate a pseudo-observation $X^* \equiv R \cdot U \stackrel{d}{=} X$ by choosing U uniformly on \mathcal{S}_n , independent of R . \square

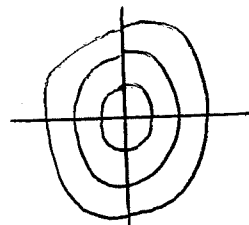
Remark 11.1. The sufficiency of R extends to the much larger nonparametric model \mathcal{P} consisting of *all radial pdfs*

$$(11.4) \quad f_g(x) = g(\|x\|^2), \quad x \in \mathbf{R}^n.$$

Here we can think of “ g ” playing the role of the unknown parameter θ . This holds [verify!] because (11.3) remains valid with “ $R^2 \sim \theta \chi_n^2$ ” replaced by (recall (6.39))

$$(11.5) \quad R \text{ has pdf } c_n \cdot r^{n-1} g(r^2).$$

Remark 11.2. Since $R \xleftrightarrow{1-1} R^2$, R^2 is also sufficient for θ . The statistics $R(X)$ and $R^2(X)$ induce the same partitioning of the sample space \mathcal{X} , that is, have the same contour lines. Whether we are given R or R^2 simply specifies the sphere (centered at 0) on which X lies:



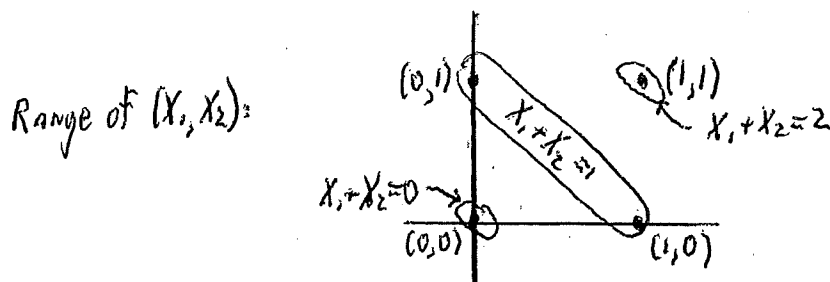
Note: In general, if T is a sufficient statistic and $T \xleftrightarrow{1-1} V$, then V is also sufficient.

Example 11.2. (*Binomial*(n, p)) Toss a coin twice ($n = 2$) with $\theta \equiv p = P[\text{Heads}]$ unknown. Let X_1, X_2 be the Bernoulli rvs indicating the two outcomes, i.e., $X_i = 1(0)$ if H (T) occurs on the i th toss. Then $X \equiv (X_1, X_2)$ has pmf

$$(11.6) \quad \begin{aligned} f_\theta(x_1, x_2) &= \theta^{x_1} (1 - \theta)^{1-x_1} \cdot \theta^{x_2} (1 - \theta)^{1-x_2} \\ &= \theta^{x_1+x_2} (1 - \theta)^{2-(x_1+x_2)} \end{aligned}$$

for $(x_1, x_2) \in \{0, 1\} \times \{0, 1\} \equiv \{0, 1\}^2$. This suggests that $T(X) \equiv X_1 + X_2$ is a sufficient statistic for θ .

To verify this according to Definition 11.2, consider the conditional distribution of $(X_1, X_2) \mid (X_1 + X_2)$. The range $\{0, 1\}^2$ of (X_1, X_2) consists of 4 points, and $X_1 + X_2$ partitions this range into 3 subsets:



The conditional distribution of $(X_1, X_2) \mid (X_1 + X_2)$ is as follows:

$$\begin{aligned}
 P_\theta[(X_1, X_2) = (0, 0) \mid X_1 + X_2 = 0] &= 1, \\
 P_\theta[(X_1, X_2) = (1, 1) \mid X_1 + X_2 = 2] &= 1, \\
 P_\theta[(X_1, X_2) = (1, 0) \mid X_1 + X_2 = 1] \\
 &= \frac{P_\theta[(X_1, X_2) = (1, 0)]}{P_\theta[(X_1, X_2) = (1, 0)] + P_\theta[(X_1, X_2) = (0, 1)]} \\
 &= \frac{\theta(1 - \theta)}{\theta(1 - \theta) + (1 - \theta)\theta} \\
 &= \frac{1}{2}.
 \end{aligned}$$

Because this conditional distribution *does not depend on* θ , $T \equiv X_1 + X_2$ is sufficient for θ .

[Given the value of $X_1 + X_2$, how can we generate the pseudo-observation $X^* \stackrel{\text{distn}}{=} X$ without knowing θ ?]. \square

This example extends to the case of n independent Bernoulli rvs X_1, \dots, X_n , where $T(X_1, \dots, X_n) \equiv X_1 + \dots + X_n$ is sufficient for θ . Here

$$(11.7) \quad f_\theta(x_1, \dots, x_n) = \theta^{x_1 + \dots + x_n} (1 - \theta)^{n - (x_1 + \dots + x_n)}$$

for $(x_1, \dots, x_n) \in \{0, 1\}^n$. Then $(X_1, \dots, X_n) \mid (X_1 + \dots + X_n)$ has the following conditional distribution: for $t = 0, 1, \dots, n$,

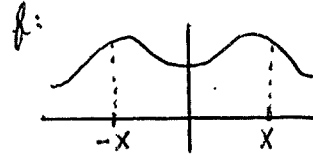
$$\begin{aligned}
P_\theta[(X_1, \dots, X_n) = (x_1, \dots, x_n) \mid X_1 + \dots + X_n = t] \\
= \begin{cases} 0 & \text{if } x_1 + \dots + x_n \neq t; \\ \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \text{if } x_1 + \dots + x_n = t \end{cases} \\
= \begin{cases} 0 & \text{if } x_1 + \dots + x_n \neq t; \\ \frac{1}{\binom{n}{t}} & \text{if } x_1 + \dots + x_n = t \end{cases} .
\end{aligned}$$

Because this does not depend on θ , T is a sufficient statistic.

[Given that $X_1 + \dots + X_n = t$, how can we generate the pseudo-observation $X^* \stackrel{\text{distn}}{=} X$ without knowing θ ?].

Example 11.3. Let \mathcal{P} be the nonparametric family of all *symmetric* pdfs f on \mathbf{R}^1 , that is,

$$(11.8) \quad f(-x) = f(x) \quad \forall x \in \mathbf{R}^1.$$



Let X be a *single* observation from an unknown $f \in \mathcal{P}$. Note that

$$(11.9) \quad f(x) = f_+(|x|) \quad \forall x \in \mathbf{R}^1 \setminus \{0\},$$



where f_+ is the restriction of f to $\mathbf{R}_+^1 := (0, \infty)$, so “ f_+ ” plays the role of θ . We will show that $T(X) = |X|$ is a sufficient statistic for \mathcal{P} (i.e., for f_+). For this we will show that the conditional distribution of $X \mid |X|$ does not depend on f_+ .

Represent X as $|X| \cdot \Psi$, where $(|X|, \Psi \equiv \text{sign}(X))$ are the one-dimensional “polar coordinates” of X . The conditional distribution of $X \mid |X|$ is equivalent to that of $\Psi \mid |X|$. Here $|X|$ and Ψ are independent: for all $t > 0$,

$$\begin{aligned}
P[0 < |X| \leq t \mid \Psi = 1] &= \frac{P[0 < X \leq t]}{P[X > 0]} \\
&= 2P[0 < X \leq t] \\
&= P[0 < |X| \leq t] \quad \text{by symmetry,}
\end{aligned}$$

and similarly $P[0 < |X| \leq t \mid \Psi = -1] = P[0 < |X| \leq t]$. Thus the conditional distribution of $\Psi \mid |X|$ is the same as the unconditional distribution of Ψ , so

$$(11.10) \quad P[\Psi = \pm 1 \mid |X|] = P[\Psi = \pm 1] = \frac{1}{2} \quad \text{by symmetry.}$$

Since this conditional distribution does not depend on f_+ , $|X|$ is a sufficient statistic for f_+ .

[Given that $|X| = t$, how can we generate the pseudo-observation $X^* \stackrel{\text{distr}}{=} X$ without knowing f ?]. \square

Exercise 11.1. More generally, suppose that X_1, \dots, X_n are i.i.d. rvs with a symmetric pdf f on \mathbf{R}^1 . Show that $(|X_1|, \dots, |X_n|)$ is sufficient for f_+ .

Example 11.4. Contrary to appearance, it is not the symmetry (11.8) of the pdfs f in Example 11.3 that leads to the sufficiency of $|X|$, but rather the fact that the ratio $\frac{f(-x)}{f(x)}$ does not depend on f . To see this, generalize Example 11.3 as follows. Let \mathcal{P} be the nonparametric family of all pdfs f on \mathbf{R}^1 that satisfy

$$(11.11) \quad \frac{f(-x)}{f(x)} = h_0(x) > 0 \quad \forall x \in \mathbf{R}_+^1,$$

where h_0 is a known function on \mathbf{R}_+^1 . Note that

$$(11.12) \quad f(x) = f_+(|x|) \cdot h(x) \quad \forall x \in \mathbf{R}^1 \setminus \{0\},$$

where

$$h(x) = I_{(0,\infty)}(x) + h_0(-x)I_{(-\infty,0)}(x),$$

so “ f_+ ” again plays the role of θ . We show that $T(X) \equiv |X|$ is a sufficient statistic for f_+ by finding the conditional distribution of $\Psi \mid T$.

Note that T is continuous and Ψ is discrete. First we find the conditional distribution of $T \mid \Psi$: for all $t > 0$,

$$P[0 < T \leq t \mid \Psi = 1] = \frac{P[0 < X \leq t]}{P[\Psi = 1]},$$

$$P[0 < T \leq t \mid \Psi = -1] = \frac{P[-t \leq X < 0]}{P[\Psi = -1]},$$

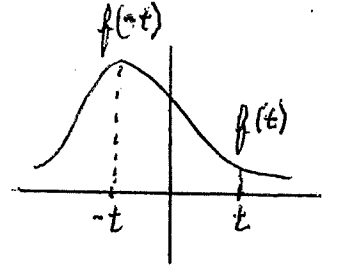
so

$$f_T(t \mid \Psi = 1) = \frac{f(t)}{P[\Psi = 1]},$$

$$f_T(t \mid \Psi = -1) = \frac{f(-t)}{P[\Psi = -1]}.$$

Then by Bayes formula (4.14) for the mixed case, for $t > 0$ we have,

$$(11.13) \quad \begin{aligned} P(\Psi = 1 \mid T = t) &= \frac{f_T(t \mid \Psi = 1) P[\Psi = 1]}{f_T(t)} \\ &= \frac{f(t)}{f(t) + f(-t)} \quad [\text{verify!}] \end{aligned}$$



$$(11.14) \quad \begin{aligned} &= \frac{f(t)}{f(t) + f(t)h_0(t)} \quad [\text{by (11.11)}] \\ &= \frac{1}{1 + h_0(t)}; \end{aligned}$$

$$(11.15) \quad \begin{aligned} P(\Psi = -1 \mid T = t) &= \frac{f_T(t \mid \Psi = -1) P[\Psi = -1]}{f_T(t)} \\ &= \frac{f(-t)}{f(t) + f(-t)} \\ &= \frac{f(t)h_0(t)}{f(t) + f(t)h_0(t)} \\ &= \frac{h_0(t)}{1 + h_0(t)}. \end{aligned}$$

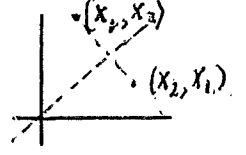
Since this conditional distribution does not depend on f_+ , $T \equiv |X|$ is a sufficient statistic for f_+ .

[Given that $|X| = t$, how can we generate the pseudo-observation $X^* \stackrel{\text{distn}}{=} X$ without knowing f ?]. \square

Remark 11.3. Note that (11.13) was derived without any symmetry assumptions on f whatsoever.

Example 11.5. Let \mathcal{P} be the nonparametric family of all *exchangeable* \equiv *symmetric* \equiv *permutation-invariant* pdfs f on \mathbf{R}^2 , that is,

$$(11.16) \quad f(x_2, x_1) = f(x_1, x_2) \quad \forall (x_1, x_2) \in \mathbf{R}^2.$$

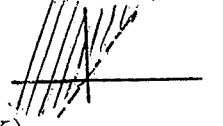


Let $X \equiv (X_1, X_2)$ be a *single* observation from some $f \in \mathcal{P}$. Note that

$$(11.17) \quad f(x_1, x_2) = f_{<}(x_{(1)}, x_{(2)}) \quad \forall (x_1, x_2) \in \mathbf{R}^2 \setminus \{x_1 = x_2\},$$

where $X_{(1)} < X_{(2)}$ are the order statistics and $f_{<}$ is the restriction of f to

$$\mathbf{R}_{<}^2 \equiv \{(x_1, x_2) \mid -\infty < x_1 < x_2 < \infty\},$$



so " $f_{<}$ " plays the role of θ . We will show that the order statistic $T(X) \equiv (X_{(1)}, X_{(2)})$ is a sufficient statistic for $f_{<}$.

To show that the conditional distribution of $(X_1, X_2) \mid (X_{(1)}, X_{(2)})$ does not depend on $f_{<}$, represent (X_1, X_2) in terms of $(X_{(1)}, X_{(2)})$ and the random permutation $\Pi \equiv \Pi(X) \equiv (\Pi_1, \Pi_2)$ defined by

$$(11.18) \quad (X_1, X_2) = (X_{(\Pi_1)}, X_{(\Pi_2)}),$$

that is,

$$(11.19) \quad (\Pi_1, \Pi_2) = \begin{cases} (1, 2) & \text{if } X_1 < X_2; \\ (2, 1) & \text{if } X_2 < X_1. \end{cases}$$

The conditional distribution of $(X_1, X_2) \mid (X_{(1)}, X_{(2)})$ is equivalent to that of $\Pi \mid (X_{(1)}, X_{(2)})$. But $(X_{(1)}, X_{(2)})$ and Π are independent: for $B \subseteq \mathbf{R}_{<}^2$,

$$\begin{aligned} P[(X_{(1)}, X_{(2)}) \in B \mid \Pi = (1, 2)] &= \frac{P[(X_1, X_2) \in B]}{P[\Pi = (1, 2)]} \\ &= 2P[(X_1, X_2) \in B] \\ &= P[(X_{(1)}, X_{(2)}) \in B] \quad \text{by symmetry,} \end{aligned}$$

and similarly $P[(X_{(1)}, X_{(2)}) \in B \mid \Pi = (2, 1)] = P[(X_{(1)}, X_{(2)}) \in B]$. Thus the conditional distribution of $\Pi \mid (X_{(1)}, X_{(2)})$ is the same as the unconditional distribution of Π , so

$$(11.20) \quad P[\Pi = (1, 2) \mid (X_{(1)}, X_{(2)})] = P[\Pi = (1, 2)] = \frac{1}{2} \quad \text{by symmetry.}$$

Since this conditional distribution does not depend on $f_{<}$, $(X_{(1)}, X_{(2)})$ is a sufficient statistic for $f_{<}$.

[Given $(X_{(1)}, X_{(2)}) \in \mathbf{R}_{<}^2$, how can we generate the pseudo-observation $(X_1^*, X_2^*) \stackrel{\text{distn}}{=} (X_1, X_2)$ without knowing f ?]. \square

Exercise 11.2. As in Example 11.4, it is not the symmetry (11.16) of the pdfs f in Example 11.5 that leads to the sufficiency of the order statistic $(X_{(1)}, X_{(2)})$, but rather the fact that the ratio $\frac{f(x_2, x_1)}{f(x_1, x_2)}$ does not depend on f . To see this, generalize Example 11.5 as follows. Let \mathcal{P} be the nonparametric family of all pdfs f on \mathbf{R}^2 that satisfy

$$(11.21) \quad \frac{f(x_2, x_1)}{f(x_1, x_2)} = h_0(x_1, x_2) > 0 \quad \forall (x_1, x_2) \in \mathbf{R}_{<}^2,$$

where h_0 is a known function on $\mathbf{R}_{<}^2$. Note that

$$(11.22) \quad f(x_1, x_2) = f_{<}(x_{(1)}, x_{(2)}) \cdot h(x_1, x_2) \quad \forall (x_1, x_2) \in \mathbf{R}^2 \setminus \{x_1 = x_2\},$$

where, for $x_1 \neq x_2$,

$$h(x_1, x_2) = I_{\mathbf{R}_{<}^2}(x_1, x_2) + h_0(x_2, x_1)I_{\mathbf{R}_{>}^2}(x_1, x_2),$$

so “ $f_{<}$ ” again plays the role of θ . Show that the order statistic $T(X) \equiv (X_{(1)}, X_{(2)})$ is a sufficient statistic for $f_{<}$.

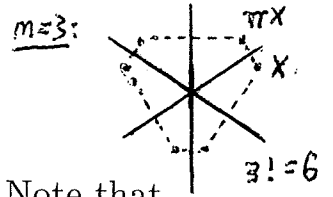
[Given $(X_{(1)}, X_{(2)}) \in \mathbf{R}_{<}^2$, how can we generate the pseudo-observation $(X_1^*, X_2^*) \stackrel{\text{distn}}{=} (X_1, X_2)$ without knowing f ?]. \square

Example 11.6. (Extension of Example 11.5 to exchangeable pdfs on \mathbf{R}^n .) Let \mathcal{P} be the nonparametric family of all *exchangeable* \equiv *symmetric* \equiv *permutation-invariant* pdfs f on \mathbf{R}^n , that is,

$$(11.23) \quad f(\pi x) = f(x) \quad \forall x \in \mathbf{R}^n \text{ and } \forall \text{ permutations } \pi.$$

Let $X \equiv (X_1, \dots, X_n)$ be a *single* observation from some $f \in \mathcal{P}$. Note that

$$(11.24) \quad f(x_1, \dots, x_n) = f_{<}(x_{(1)}, \dots, x_{(n)}) \quad \forall (x_1, \dots, x_n) \in \mathbf{R}^n \setminus \{x_1 = \dots = x_n\},$$



where $X_{(1)} < \dots < X_{(n)}$ are the order statistics and $f_{<}$ is the restriction of f to

$$\mathbf{R}_{<}^n \equiv \{(x_1, \dots, x_n) \mid -\infty < x_1 < \dots < x_n < \infty\},$$

so " $f_{<}$ " plays the role of θ . We will show that the order statistic $T(X) \equiv (X_{(1)}, \dots, X_{(n)})$ is a sufficient statistic for $f_{<}$.

To show that the conditional distribution of $X \mid T$ does not depend on $f_{<}$, represent X in terms of T and the random permutation $\Pi \equiv \Pi(X) \equiv (\Pi_1, \dots, \Pi_n)$ defined by

$$(11.25) \quad X \equiv (X_1, \dots, X_n) = (X_{(\Pi_1)}, \dots, X_{(\Pi_n)}) \equiv \Pi T.$$

(Note that Π_i is simply the *rank* of X_i among X_1, \dots, X_n .) Thus the conditional distribution of $X \mid T$ is equivalent to that of $\Pi \mid T$. But T and Π are independent: for $B \subseteq \mathbf{R}_{<}^n$ and any permutation π ,

$$\begin{aligned} P[T \in B \mid \Pi = \pi] &\equiv P[\pi^{-1}X \in B \mid \Pi = \pi] && \text{[by (11.25)]} \\ &= \frac{P[\pi^{-1}X \in B]}{P[\Pi = \pi]} && \text{[since } B \subseteq \mathbf{R}_{<}^n] \\ &= n!P[\pi^{-1}X \in B] \\ &= P[T \in B] && \text{[by symmetry].} \end{aligned}$$

Thus the conditional distribution of $\Pi \mid T$ is the same as the unconditional distribution of Π , so for any $t \in \mathbf{R}_{<}^n$,

$$(11.26) \quad P[\Pi = \pi \mid T = t] = P[\Pi = \pi] = \frac{1}{n!} \quad \text{by symmetry.}$$

Since this conditional distribution does not depend on $f_{<}$, the order statistic T is a sufficient statistic for $f_{<}$.

[Given $T \equiv (X_{(1)}, \dots, X_{(n)}) \in \mathbf{R}_{<}^n$, how can we generate the pseudo-observation $(X_1^*, \dots, X_n^*) \stackrel{d}{=} (X_1, \dots, X_n)$ without knowing f ?]. \square

Exercise 11.3. (Recall Remark 11.3) If f is any pdf on \mathbf{R}^n , use Bayes' formula for the mixed case to extend (11.26) as follows:

$$(11.27) \quad P[\Pi = \pi \mid T = t] = \frac{f(\pi t)}{\sum_{\pi'} f(\pi' t)},$$

where the summation extends over all $n!$ permutations π' .

Remark 11.4. In Example 11.6, suppose we replace the exchangeable family \mathcal{P} by the *smaller* (still nonparametric) family \mathcal{P}_1 consisting of all pdfs f on \mathbf{R}^n of the form

$$(11.28) \quad f(x_1, \dots, x_n) = g(x_1) \cdots g(x_n).$$

That is, assume that X_1, \dots, X_n is an i.i.d. sample from some unknown pdf g on \mathbf{R}^1 . Note that

$$(11.29) \quad f(x_1, \dots, x_n) = g(x_{(1)}) \cdots g(x_{(n)}) \quad \forall (x_1, \dots, x_n) \in \mathbf{R}^n,$$

so “ g ” plays the role of θ . Because the order statistics are sufficient for \mathcal{P} and $\mathcal{P}_1 \subset \mathcal{P}$, clearly the order statistics are also sufficient for \mathcal{P}_1 , i.e., for g , by the following trivial lemma.

Lemma 11.1. If $T(X)$ is a sufficient statistic for \mathcal{P} and $\mathcal{P}_1 \subset \mathcal{P}$, then $T(X)$ is sufficient for \mathcal{P}_1 . \square

Example 11.7. (*Uniform* $(0, \theta]$) Let X_1, \dots, X_n be i.i.d. $\text{Uniform}(0, \theta]$ rvs, where $\theta \in (0, \infty)$ is an unknown scale parameter. The joint pdf of $X \equiv (X_1, \dots, X_n)$ is

$$(11.30) \quad \begin{aligned} f_\theta(x) &= \prod_{i=1}^n \left[\frac{1}{\theta} I_{(0, \theta]}(x_i) \right] \\ &= \frac{1}{\theta^n} I_{(0, \theta]}(x_{(n)}) \cdot I_{(0, \infty)}(x_{(1)}). \end{aligned}$$

Because the family of pdfs in (11.30) is a subfamily of (11.28), it follows from Remark 11.4 that the order statistic $T(X) \equiv (X_{(1)}, \dots, X_{(n)})$ is a sufficient statistic for θ . However, T is not *minimal sufficient*, for (11.30) suggests that further reduction to $S(X) \equiv X_{(n)}$ still preserves sufficiency.

To verify this directly via Definition 11.2 requires finding the conditional distribution of $X \mid S$. This can be done, but it is slightly complicated, being a discrete mixture of several continuous distributions [think about this]. Instead we can proceed in steps, using the following lemma.

Lemma 11.2. Let \mathcal{P} be a family of distributions for X and suppose that $T \equiv T(X)$ is a sufficient statistic for \mathcal{P} . Let \mathcal{Q} be the family of distributions of T induced by \mathcal{P} and suppose that $S \equiv S(T)$ is sufficient for \mathcal{Q} . Then $S \equiv S(T(X))$ is also sufficient for \mathcal{P} .

Proof. For $P \in \mathcal{P}$ let Q be the induced distribution of T . Then for $A \subseteq \mathcal{X}$,

$$\begin{aligned}
 P[X \in A \mid S] &= E_P\{P[X \in A \mid S, T] \mid S\} \\
 &= E_P\{P[X \in A \mid T] \mid S\} && [S = S(T)] \\
 &\equiv E_P\{g_A(T) \mid S\} && [T \text{ is sufficient for } \mathcal{P}] \\
 &= E_Q\{g_A(T) \mid S\} && [\text{definition of } Q] \\
 &= E\{g_A(T) \mid S\}. && [S \text{ is sufficient for } \mathcal{Q}]
 \end{aligned}$$

Because this does not depend on $P \in \mathcal{P}$, S is sufficient for \mathcal{P} . \square

Continuation of Example 11.7: Thus to show that S is sufficient for θ w.r.to X , it suffices to show that S is sufficient for θ w.r.to T .

The joint pdf of $T \equiv (T_1, \dots, T_{n-1}, T_n) \equiv (X_{(1)}, \dots, X_{(n-1)}, X_{(n)})$ is obtained from (9.5) and (11.30):

$$(11.31) \quad f_\theta(t_1, \dots, t_{n-1}, t_n) = \frac{n!}{\theta^n} I_{(0, \theta]}(t_n) \cdot I_{(0, \infty)}(t_1) \cdot I_{\mathbf{R}_<^n}(t_1, \dots, t_{n-1}, t_n).$$

Next, the pdf of $S \equiv T_n \equiv X_{(n)}$ is given by

$$(11.32) \quad f_\theta(t_n) = \frac{nt_n^{n-1}}{\theta^n} I_{(0, \theta]}(t_n). \quad [\text{verify}]$$

Finally, the conditional distribution of $T \mid S$ is equivalent to the conditional distribution of $(T_1, \dots, T_{n-1}) \mid T_n$, and from (11.31) and (11.32) this conditional pdf is

$$\begin{aligned}
 f(t_1, \dots, t_{n-1} \mid t_n) &= \frac{f_\theta(t_1, \dots, t_{n-1}, t_n)}{f_\theta(t_n)} \\
 (11.33) \quad &= \frac{(n-1)!}{t_n^{n-1}} I_{(0, t_n)}(t_{n-1}) \cdot I_{(0, \infty)}(t_1) \cdot I_{\mathcal{R}_<^{n-1}}(t_1, \dots, t_{n-1}).
 \end{aligned}$$

Because this conditional pdf does not depend on θ , we conclude that S is sufficient for θ w.r.to T , as required. \square

Remark 11.5. From (11.31) and (11.33), the conditional distribution of $(X_{(1)}, \dots, X_{(n-1)}) \mid X_{(n)}$ is the same as the distribution of the order statistics for a sample of size $n - 1$ from the Uniform $(0, X_{(n)})$ distribution. This implies that the $n - 1$ ratios $0 < \frac{X_{(1)}}{X_{(n)}} < \dots < \frac{X_{(n-1)}}{X_{(n)}} < 1$ are independent of $X_{(n)}$ and are distributed as the order statistics for a sample of size $n - 1$ from the Uniform $(0, 1)$ distribution.

11.2. The Fisher-Neyman Factorization Criterion for sufficiency.

Consider a (parametric or nonparametric) statistical model $(\mathcal{X}, \mathcal{P})$, where each $P_\theta \in \mathcal{P}$ is determined by a pdf or pmf $f_\theta(x)$.

Theorem 11.1. *The statistic $T \equiv T(X)$ is sufficient for \mathcal{P} (i.e., for θ) if and only if $f_\theta(x)$ factors as follows:*

$$(11.34) \quad f_\theta(x) = g_\theta(T(x)) \cdot h(x)$$

where $g_\theta(T(x))$ depends on x only through $T(x)$ and $h(x)$ does not depend on θ .

Proof (sketch). CB Theorem 6.2.6 contains a proof for the discrete case. There, (11.34) comes from the following factorization: if $T(x) = t$,

$$(11.35) \quad f_\theta(x) = P_\theta[X = x, T(X) = t] = f_\theta(t) \cdot f_\theta(x \mid t),$$

since if T is sufficient then $f_\theta(x \mid t) \equiv h(x)$ does not depend on θ . [See Bahadur *Ann. Math. Statist.* (1954) for a general proof. \square]

We have already encountered several examples of the factorization (11.34) – **verify** that the following all have this form: (11.2), (11.4), (11.6), (11.7), (11.9), (11.12), (11.17), (11.22), (11.24), (11.29), (11.30), (11.31).

Example 11.8. (*Uniform* $[\theta_1, \theta_2]$) Let X_1, \dots, X_n be i.i.d. Uniform $[\theta_1, \theta_2]$ rvs, where $-\infty < \theta_1 < \theta_2 < \infty$ are unknown. The joint pdf of $X \equiv (X_1, \dots, X_n)$ is

$$(11.36) \quad \begin{aligned} f_\theta(x) &= \prod_{i=1}^n \left[\frac{1}{\theta_2 - \theta_1} I_{[\theta_1, \theta_2]}(x_i) \right] \\ &= \frac{1}{(\theta_2 - \theta_1)^n} I_{[\theta_1, \infty)}(x_{(1)}) \cdot I_{(-\infty, \theta_2]}(x_{(n)}) \cdot 1. \end{aligned}$$

This has the form (11.34) with $\theta = (\theta_1, \theta_2)$, $T(X) = (X_{(1)}, X_{(n)})$, and $h(x) \equiv 1$, so the pair of order statistics $(X_{(1)}, X_{(n)})$ is sufficient for (θ_1, θ_2) .

Example 11.9. (*Normal* (μ, σ^2)) Let X_1, \dots, X_n be i.i.d. $N_1(\mu, \sigma^2)$ rvs with $\theta \equiv (\mu, \sigma^2)$ unknown. The joint pdf of $X \equiv (X_1, \dots, X_n)$ is

$$\begin{aligned} f_{\mu, \sigma^2}(x_1, \dots, x_n) &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum x_i^2 + \frac{\mu}{\sigma^2} \sum x_i - \frac{n\mu^2}{2\sigma^2}} \cdot 1. \end{aligned}$$

This has the form (11.34) with $T(X) = (\sum X_i, \sum X_i^2)$ and $h(x) \equiv 1$, so the pair $(\sum X_i, \sum X_i^2)$ is sufficient for (μ, σ^2) . Because

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right) \xleftrightarrow{1-1} (\bar{X}_n, s_n^2),$$

the sample mean and sample variance are also a pair of sufficient statistics.

11.3. The Factorization Criterion and the likelihood ratio.

As in §11.2, let $\mathcal{P} \equiv \{f_\theta(x) \mid \theta \in \Omega\}$ be a general statistical model specified by a family of pdfs/pmf. Suppose we know that $\theta = \theta_1$ or θ_2 and wish to decide between these two possibilities based on the data X . It is appropriate [see p.223(a)] to base our decision on the value of the *likelihood ratio* (*LR*)

$$(11.37) \quad L_{\theta_1, \theta_2}(x) \equiv \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$$

according to a decision rule of the following form : for a fixed constant c ,

$$(11.38) \quad \text{decide } \theta = \begin{cases} \theta_2 & \text{if } L_{\theta_1, \theta_2}(x) > c; \\ \theta_1 & \text{if } L_{\theta_1, \theta_2}(x) < c; \\ \text{arbitrary} & \text{if } L_{\theta_1, \theta_2}(x) = c. \end{cases}$$

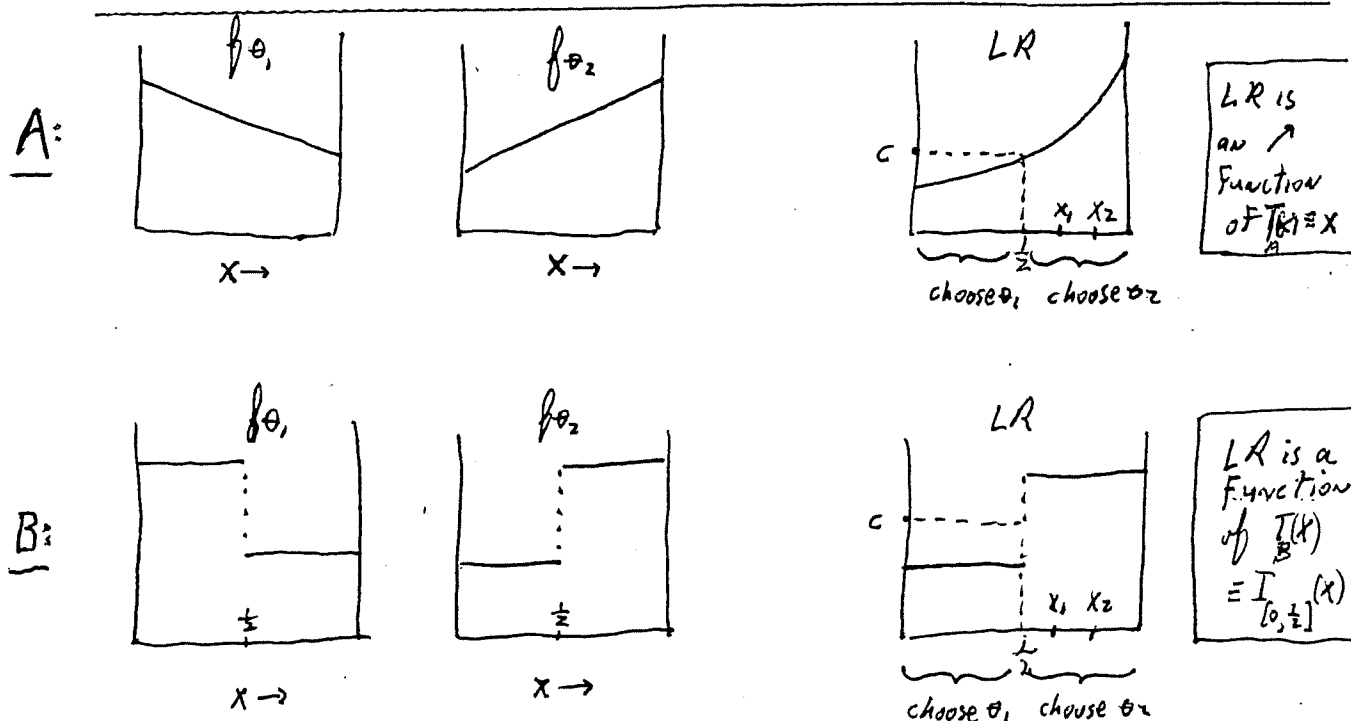
Such a rule is optimal for deciding between f_{θ_1} and f_{θ_2} , both in the Neyman-Pearson sense (cf. Theorem 18.6) of maximizing the probability of selecting

θ_2 when θ_2 is true while controlling the probability of selecting θ_2 when θ_1 is true, and in the Bayesian sense of minimizing a weighted average of the two error probabilities (cf. (17.11)). Here we emphasize the fundamental importance of the LR by its dependence on any sufficient statistic $T(X)$.

This dependence is easily demonstrated by expressing the LR (11.37) according to the factorization criterion (11.34) as

$$(11.39) \quad L_{\theta_1, \theta_2}(x) = \frac{g_{\theta_2}(T(x))h(x)}{g_{\theta_1}(T(x))h(x)} = \frac{g_{\theta_2}(T(x))}{g_{\theta_1}(T(x))}.$$

Thus, the LR depends on X only through the value of the sufficient statistic $T(X)$. Therefore the optimal decision rules (11.38) depend on X only through $T(X)$. This is a further indication of the role of a sufficient statistic.



Example 11.10. In Case A, the observation x_2 provides stronger evidence for θ_2 than does the observation x_1 , because $L_{\theta_1, \theta_2}(x_2) > L_{\theta_1, \theta_2}(x_1)$. In Case B, however, x_1 and x_2 convey the same degree of evidence for θ_2 , since $L_{\theta_1, \theta_2}(x_2) = L_{\theta_1, \theta_2}(x_1)$. Thus, in B we can reduce the data from X to $T(X) \equiv I_{[0, \frac{1}{2}]}(X)$ without losing any relevant information for distinguishing between f_{θ_1} and f_{θ_2} , so $T(X)$ is a sufficient statistic for $\{\theta_1, \theta_2\}$. In A, however, this $T(X)$ is not sufficient, for relevant information would be lost if only $T(X)$ were known, not X .

11.4. Minimal sufficiency.

In both Cases A and B above, the statistic X itself is trivially sufficient for $\{\theta_1, \theta_2\}$. In B, however, X can be reduced further to the sufficient statistic $T_B(X)$, whereas X cannot be reduced in A without losing sufficiency. Thus, X is *minimal sufficient* in Case A but not in Case B.

Definition 11.4. $T^*(X)$ is a *minimal sufficient statistic* for $\mathcal{P} \equiv \{P_\theta\}$ if T^* is sufficient and if, for every other sufficient statistic $T(X)$, T^* is a reduction of T , i.e., $T^*(X) = h(T(X))$ for some function h . \square

In (11.39) we applied the Factorization Criterion to show that

$$T(X) \text{ is sufficient} \Rightarrow L_{\theta_1, \theta_2}(X) \text{ is a function of } T(X) \forall \theta_1, \theta_2.$$

The converse is also true:

$$L_{\theta_1, \theta_2}(X) \text{ is a function of } T(X) \forall \theta_1, \theta_2 \Rightarrow T(X) \text{ is sufficient.}$$

For, by setting $\theta = \theta_2$ and fixing θ_1 , we have

$$f_\theta(x) = L_{\theta_1, \theta}(x) \cdot f_{\theta_1}(x) \equiv g_\theta(T(x)) \cdot h(x),$$

so $T(X)$ satisfies the factorization criterion. Thus:

$$\begin{aligned} (11.40) \quad T(X) \text{ is sufficient} &\iff L_{\theta_1, \theta_2}(X) \text{ is a function of } T(X) \forall \theta_1, \theta_2 \\ &\iff T^{**}(X) \text{ is a function of } T(X), \end{aligned}$$

where

$$(11.41) \quad T^{**}(X) \equiv \{L_{\theta_1, \theta_2}(X) \mid \theta_1, \theta_2 \in \Omega\}$$

is the *entire family of pairwise LR's*. Thus (11.40) and Definition 11.4 show that T^{**} is a *minimal sufficient statistic*. This can be stated as follows: *the set of likelihood ratios is a minimal sufficient statistic*. This again emphasizes the fundamental role of the LR in statistical inference.

Remark 11.6. Together with (11.39), this also suggests that the statistic $T(X)$ appearing in the factorization criterion is (usually) a minimal sufficient statistic.

Theorem 11.2 (Lehmann-Scheffe). Suppose that X has pmf or pdf $f_\theta(x)$, $\theta \in \Omega$ and that $T^*(X)$ satisfies the following property:

for every pair of sample points $x, y \in \mathcal{X}$, the ratio $\frac{f_\theta(y)}{f_\theta(x)}$ is θ -free
(does not depend on θ) iff $T^*(x) = T^*(y)$.

Then T^* is minimal sufficient for θ .

Proof. By hypothesis,

$$\begin{aligned} T^*(x) = T^*(y) &\iff \frac{f_\theta(y)}{f_\theta(x)} \text{ is } \theta\text{-free} \\ &\iff \frac{f_{\theta_1}(y)}{f_{\theta_1}(x)} = \frac{f_{\theta_2}(y)}{f_{\theta_2}(x)} \quad \forall \theta_1, \theta_2 \\ &\iff \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = \frac{f_{\theta_2}(y)}{f_{\theta_1}(y)} \quad \forall \theta_1, \theta_2. \\ &\iff L_{\theta_1, \theta_2}(x) = L_{\theta_1, \theta_2}(y) \quad \forall \theta_1, \theta_2 \\ &\iff T^{**}(x) = T^{**}(y) \quad \text{by definition of } T^{**}. \end{aligned}$$

Thus T^* and T^{**} are equivalent statistics and T^{**} is minimal sufficient, so T^* is also minimal sufficient. \square

Note: This proof is not completely rigorous for it implicitly assumes that $f_\theta(x) > 0 \quad \forall x, \theta$. Bahadur (1954) gives a rigorous proof via measure theory.

Example 11.11. (1-parameter exponential family) Let X_1, \dots, X_n be an i.i.d. sample from a distribution with pdf (continuous) or pmf (discrete) of the exponential form

$$(11.42) \quad f_\theta(x) = a(\theta) \exp[\theta T(x)] \cdot h(x),$$

where $\theta \in \Omega$ is a real parameter. Then $X \equiv (X_1, \dots, X_n)$ has joint pdf

$$(11.43) \quad f_\theta(x) = [a(\theta)]^n \exp \left[\theta \sum_{i=1}^n T(x_i) \right] \cdot \prod_{i=1}^n h(x_i),$$

so $\sum T(X_i)$ is a sufficient statistic by the factorization criterion. To see that it is minimal sufficient, apply the Lehmann-Scheffe Theorem:

$$\frac{f_\theta(y)}{f_\theta(x)} \equiv \exp \left[\theta \left(\sum T(y_i) - \sum T(x_i) \right) \right] \cdot \frac{\prod h(y_i)}{\prod h(x_i)}$$

is θ -free iff $\sum T(y_i) = \sum T(x_i)$ (provided that the parameter space Ω contains at least two points). Note also that the LR

$$(11.44) \quad L_{\theta_1, \theta_2}(X) = \left[\frac{a(\theta_2)}{a(\theta_1)} \right]^n \exp \left[(\theta_2 - \theta_1) \sum T(X_i) \right]$$

is a strictly increasing function of $\sum T(X_i)$ for every pair $\theta_1 < \theta_2$. \square

Many common 1-parameter families are exponential families [verify]:

$$N_1(\mu, 1): \quad \theta = \mu, \quad T(X_i) = X_i;$$

$$N_1(0, \sigma^2): \quad \theta = -\frac{1}{2\sigma^2}, \quad T(X_i) = X_i^2;$$

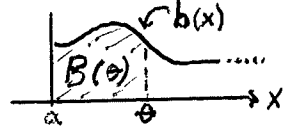
$$\text{Binomial}(n, p): \quad \theta = \log \frac{p}{1-p}, \quad T(X_i) = X_i;$$

$$\text{Poisson}(\lambda): \quad \theta = \log \lambda, \quad T(X_i) = X_i;$$

$$\text{Exponential}(\lambda): \quad \theta = -\lambda, \quad T(X_i) = X_i.$$

Example 11.12. (*1-parameter truncation family*) Let X_1, \dots, X_n be an i.i.d. sample from a distribution with pdf of the *truncation* form

$$(11.45) \quad f_\theta(x) = [B(\theta)]^{-1} I_{(a, \theta]}(x) \cdot b(x), \quad x > a,$$



where $-\infty \leq a < \infty$ is specified, $\theta > a$ is a real parameter, $b(x) > 0$ on (a, ∞) ,¹⁷ and $B(\theta) \equiv \int_a^\theta b(x) dx < \infty \forall \theta > a$. (The Uniform(0, θ) pdf is a special case with $a = 0$, $b(x) \equiv 1$, and $B(\theta) = \theta$.) Here $X \equiv (X_1, \dots, X_n)$ has joint pdf

$$(11.46) \quad f_\theta(x) = [B(\theta)]^{-n} I_{(a, \theta]}(x_{(n)}) \cdot I_{(a, \infty)}(x_{(1)}) \prod_{i=1}^n b(x_i),$$

so $T \equiv X_{(n)}$ is a sufficient statistic by the factorization criterion. To see that $X_{(n)}$ is minimal sufficient, apply the Lehmann-Scheffe Theorem:

$$\frac{f_\theta(y)}{f_\theta(x)} \equiv \frac{I_{(a, \theta]}(y_{(n)})}{I_{(a, \theta]}(x_{(n)})} \cdot \frac{I_{(a, \infty)}(y_{(1)})}{I_{(a, \infty)}(x_{(1)})} \frac{\prod b(y_i)}{\prod b(x_i)}$$

¹⁷ If $b \equiv 0$ on some interval then θ would not be identifiable.

is θ -free iff $x_{(n)} = y_{(n)}$ [verify!]. (We set $\frac{1}{0} = \infty$ and $\frac{0}{0} = 1$.)

Similarly, $T \equiv X_{(1)}$ is minimal sufficient if (11.45) is replaced by

$$(11.47) \quad [B(\theta)]^{-1} I_{[\theta, a)}(x) \cdot b(x), \quad x < a,$$

where $-\infty < a \leq \infty$ is specified, $\theta < a$ is a real parameter, $b(x) > 0$ on (∞, a) , and $B(\theta) \equiv \int_{\theta}^a b(x) dx < \infty \forall \theta < a$. \square

Example 11.13. (*k-parameter exponential family*) Let X_1, \dots, X_n be an i.i.d. sample of rvs or rvtrs from a distribution with pdf (continuous) or pmf (discrete) of the exponential form

$$a(\theta_1, \dots, \theta_k) \exp[\theta_1 T_1(x) + \dots + \theta_k T_k(x)] \cdot h(x),$$

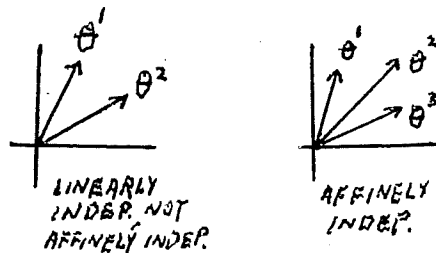
where $\theta \equiv (\theta_1, \dots, \theta_k) \in \Omega$ is a k -dimensional parameter. Then $X \equiv (X_1, \dots, X_n)$ has pdf

$$(11.48) \quad f_{\theta}(x) = [a(\theta)]^n \exp \left[\theta_1 \sum_{i=1}^n T_1(x_i) + \dots + \theta_k \sum_{i=1}^n T_k(x_i) \right] \cdot \prod_{i=1}^n h(x_i),$$

so $(\sum T_1(X_i), \dots, \sum T_k(X_i))$ is a k -dimensional sufficient statistic. To see that it is minimal sufficient, apply the Lehmann-Scheffe Theorem:

$$\frac{f_{\theta}(y)}{f_{\theta}(x)} \equiv \exp \left[\sum_{j=1}^k \theta_j \left(\sum_{i=1}^n T_j(y_i) - \sum_{i=1}^n T_j(x_i) \right) \right] \cdot \frac{\prod h(y_i)}{\prod h(x_i)}$$

is θ -free iff $\sum T_j(y_i) = \sum T_j(x_i)$ for $j = 1, \dots, k$, provided that the parameter space $\Omega \subseteq \mathbf{R}^k$ affinely spans \mathbf{R}^k . That is, Ω must contain a set of $k+1$ affinely independent vectors $\theta^1, \dots, \theta^{k+1}$, i.e., $\theta^1, \dots, \theta^{k+1}$ are not contained in any hyperplane of dimension $\leq k-1$:



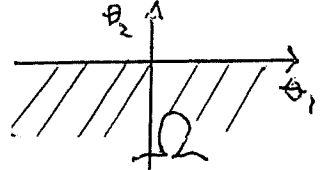
For example, let X_1, \dots, X_n be an i.i.d. sample from $N_1(\mu, \sigma^2)$. Write the pdf of $X \equiv (X_1, \dots, X_n)$ in 2-parameter exponential family form:

$$(11.49) \quad f_{\mu, \sigma}(x) = \frac{e^{-\frac{n\mu^2}{2\sigma^2}}}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left(\frac{\mu}{\sigma^2} \sum x_i - \frac{1}{2\sigma^2} \sum x_i^2\right).$$

Here $k = 2$, while $\theta \equiv (\theta_1, \theta_2) \equiv (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ are the “natural” exponential parameters ((μ, σ) and (μ, σ^2) are not). Thus, in order for the pair $(\sum X_i, \sum X_i^2) \xrightarrow{1-1} (\bar{X}_n, s_n^2)$ to be minimal sufficient, the “proviso” concerning the affine span of Ω must be verified for (θ_1, θ_2) rather than for (μ, σ) or (μ, σ^2) :

(i) If Ω is the entire parameter space $\{(\mu, \sigma) \mid -\infty < \mu < \infty, 0 < \sigma < \infty\}$, then equivalently

$$(11.50) \quad \Omega = \{(\theta_1, \theta_2) \mid -\infty < \theta_1 < \infty, -\infty < \theta_2 < 0\},$$



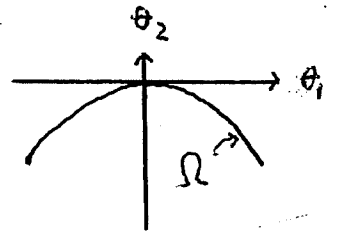
so its affine span is \mathbf{R}^2 , hence the 2-dimensional statistic $(\sum X_i, \sum X_i^2)$ (equivalently, (\bar{X}_n, s_n^2)) is minimal sufficient.

(ii) Now impose the restriction $\sigma^2 = \mu^2$ ($\mu \neq 0$) on Ω , i.e., $X_i \sim N_1(\mu, \mu^2)$, so the parameter space is essentially 1-dimensional. However,

$$(\theta_1, \theta_2) = \left(\frac{\mu}{\mu^2}, -\frac{1}{2\mu^2}\right) = \left(\frac{1}{\mu}, -\frac{1}{2\mu^2}\right),$$

so

$$(11.51) \quad \Omega = \left\{(\theta_1, \theta_2) \mid \theta_2 = -\frac{\theta_1^2}{2}, \theta_1 \neq 0\right\}.$$

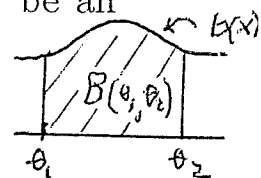


Thus Ω is a parabola in \mathbf{R}^2 so its affine span is again \mathbf{R}^2 , hence the 2-dimensional statistic $(\sum X_i, \sum X_i^2) \xrightarrow{1-1} (\bar{X}_n, s_n^2)$ remains minimal sufficient, although the parameter space is only 1-dimensional! Reduction of the data to either \bar{X}_n or s_n^2 alone will result in a loss of relevant information for inference about μ . (Also see Remark 11.7, p.181.)

Note: Case (ii) is an example of a “curved” exponential family. □

Example 11.14. (*2-parameter truncation family*) Let X_1, \dots, X_n be an i.i.d. sample from a distribution with pdf of the *truncation* form

$$[B(\theta_1, \theta_2)]^{-1} I_{[\theta_1, \theta_2]}(x) \cdot b(x), \quad -\infty < x < \infty,$$



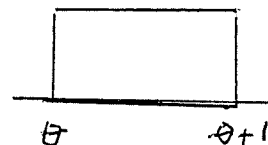
where $-\infty < \theta_1 < \theta_2 < \infty$ are real parameters, $b(x) > 0$ on $(-\infty, \infty)$,¹⁸ and $B(\theta_1, \theta_2) \equiv \int_{\theta_1}^{\theta_2} b(x) dx < \infty \forall \theta_1 < \theta_2$. (The Uniform $[\theta_1, \theta_2]$ pdf is a special case with $b(x) \equiv 1$ and $B(\theta_1, \theta_2) = \theta_2 - \theta_1$.) Here $X \equiv (X_1, \dots, X_n)$ has joint pdf

$$(11.52) \quad f_{\theta}(x) = [B(\theta_1, \theta_2)]^{-n} I_{[\theta_1, \infty)}(x_{(1)}) I_{(-\infty, \theta_2]}(x_{(n)}) \cdot \prod_{i=1}^n b(x_i),$$

so $(X_{(1)}, X_{(n)})$ is a sufficient statistic for (θ_1, θ_2) by the factorization criterion. To see that $(X_{(1)}, X_{(n)})$ is minimal sufficient, apply the Lehmann-Scheffe Theorem:

$$\frac{f_{\theta}(y)}{f_{\theta}(x)} \equiv \frac{I_{[\theta_1, \infty)}(y_{(1)}) I_{(-\infty, \theta_2]}(y_{(n)})}{I_{[\theta_1, \infty)}(x_{(1)}) I_{(-\infty, \theta_2]}(x_{(n)})} \cdot \frac{\prod b(y_i)}{\prod b(x_i)}$$

is (θ_1, θ_2) -free iff $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$ [verify!].



Exercise 11.4. (Uniform $[\theta, \theta + 1]$) In Example 11.14, take $\theta_1 = \theta$ and $\theta_2 = \theta + 1$, where $\theta \in (-\infty, \infty)$ is a real-valued location parameter. For simplicity set $b(x) \equiv 1$, so each $X_i \sim \text{Uniform}[\theta, \theta + 1]$. Show that $(X_{(1)}, X_{(n)})$ remains a 2-dimensional minimal sufficient statistic for the 1-dimensional parameter θ .

Remark 11.7. (*Ancillary statistics and conditional inference*) Example 11.13(ii) and Exercise 11.4 show that we may find a 2-dimensional minimal sufficient statistic (T_1, T_2) for a 1-dimensional parameter θ . In such cases there may exist an equivalent minimal sufficient statistic $(U, V) \xleftrightarrow{1-1} (T_1, T_2)$ such that U, V are each 1-dimensional and V is *ancillary*, i.e., the distribution of V does not depend on θ (see §12.1). In this case the joint pdf/pmf of (U, V) must have the form

$$(11.53) \quad f_{\theta}(u, v) = f_{\theta}(u | v) \cdot f(v),$$

¹⁸ If $b \equiv 0$ on some interval then $\theta \equiv (\theta_1, \theta_2)$ would not be identifiable.

so the likelihood ratio is given by

$$(11.54) \quad L_{\theta_1, \theta_2}(u, v) \equiv \frac{f_{\theta_2}(u, v)}{f_{\theta_1}(u, v)} = \frac{f_{\theta_2}(u | v)}{f_{\theta_1}(u | v)}.$$

This suggests that efficient inference about θ might be obtained from the conditional distribution of $U | V$. However the ancillary statistic V may not be unique, in which case there will be several to choose among and some might work better than others. Furthermore, even if V is unique, this conditional distribution may not be simple (see Footnote 19). \square

Exercise 11.5. For example, in Exercise 11.4, the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ is equivalent to the pair $(X_{(1)}, R_n)$ where $R_n \equiv X_{(n)} - X_{(1)}$ is the sample range. Note that $0 \leq X_{(1)} - \theta \leq 1 - R_n$, so R_n , which is clearly ancillary hence provides no information about θ by itself, nonetheless governs the accuracy of $X_{(1)}$ as an estimator of θ . In fact, because R_n is ancillary, we can base inference about θ on the conditional distribution of $X_{(1)} | R_n$.

(i) Find this conditional distribution. (See CB Example 5.4.7 for a related discussion.) Use this conditional distribution to find an estimator $\tilde{\theta}_n$ that is conditionally unbiased for θ , thus unconditionally unbiased.

(ii) Let $\check{\theta}_n = X_{(1)} - \frac{1}{n+1}$. Show that $\check{\theta}_n$ is unbiased for θ , that $\text{Var}(\tilde{\theta}_n) < \text{Var}(\check{\theta}_n)$ for all n , and that $\lim_{n \rightarrow \infty} \text{Var}(\tilde{\theta}_n) / \text{Var}(\check{\theta}_n) = \frac{1}{2}$.

(iii) Find a confidence interval for θ , centered at $\tilde{\theta}$, whose conditional and unconditional confidence coefficient is $(1 - \alpha)$. \square

Exercise 11.6. In Example 11.13(ii), show that $t_n^2 \equiv \frac{\bar{X}_n^2}{s_n^2}$ is ancillary. Clearly the minimal sufficient statistic $(\bar{X}_n, s_n^2) \xrightarrow{1-1} (\bar{X}_n, t_n^2)$, hence contains the nontrivial ancillary statistic t_n^2 . Thus, as stated in Remark 11.7, inference on μ can be based on the conditional distribution of $\bar{X}_n | t_n^2$.¹⁹ \square

¹⁹ However, this conditional distribution is not simple, see D. V. Hinkley (1977) "Conditional inference about a normal mean with known coefficient of variation," *Biometrika* 64 105-108.

12. Ancillarity and Invariance; Sufficiency and Completeness; Minimum-Variance Unbiased Estimation.

In Remark 11.7 we noted that a minimal sufficient statistic $T \equiv T(X)$ may include an ancillary statistic $V \equiv V(X)$, in which case inference based on the conditional distribution of $T \mid V$ is suggested. In §12.1 we show that ancillary statistics commonly arise as *invariant statistics in a group-invariant statistical model*, i.e., one generated by applying a group of transformations to a single standard distribution.

In §12.2 we focus on the case where T is also *complete*, i.e., contains no ancillary information, so inference should be based on the unconditional distribution of T . Here Basu's Theorem 12.1 implies that $T \perp\!\!\!\perp V$ for any ancillary statistic V , so V is truly irrelevant for inference.

In §12.3 the Rao-Blackwell-Lehmann-Scheffe approach to minimum-variance unbiased estimation based on a complete and (minimal) sufficient statistic T is presented. Here it is shown that use of ancillary information can actually be detrimental. In §12.4 these results are extended to general convex loss functions.

12.1. Ancillary statistics and group-invariant families.

Definition 12.1. A statistic $V \equiv V(x)$ on \mathcal{X} is *ancillary* for $\mathcal{P} \equiv \{P_\theta\}$ if its distribution does not depend on θ ; i.e., for any $A \subset \mathcal{X}$ and any integrable g on \mathcal{X} , $P_\theta[V \in A]$ and $E_\theta[g(V)]$ are θ -free (do not depend on θ). \square

Any location-invariant statistic in a location-parameter family is ancillary, as is any scale-invariant statistic in a scale-parameter family. These are special cases of a *group-invariant statistical model*:

Definition 12.2. Let X_0 (unobservable) have a specified distribution P_0 on \mathcal{X} and let $\Gamma \equiv \{\gamma\}$ be a *group* of 1-1 transformations $\gamma : \mathcal{X} \leftrightarrow \mathcal{X}$. The *group-invariant model* \mathcal{P}_Γ is the set of distributions of the random variates

$$(12.1) \quad \{X_\gamma \equiv \gamma X_0 \mid \gamma \in \Gamma\}.$$

Here $\Gamma (\equiv \Omega)$ is the parameter space, $\gamma (\equiv \theta)$ is the unknown parameter, $P_\gamma (\equiv P_\theta) = P_0 \circ \gamma^{-1}$, and $X \equiv X_\gamma$ is observed. \square

Lemma 12.1. Let $V \equiv V(X)$ be a Γ -invariant statistic on \mathcal{X} , that is,

$$V(\gamma x) = V(x) \quad \forall x \in \mathcal{X}, \quad \forall \gamma \in \Gamma.$$

Then V is ancillary for \mathcal{P}_Γ .

Proof. $V(X) \equiv V(X_\gamma) \equiv V(\gamma X_0) = V(X_0)$, so the distribution of $V(X)$ is γ -free.

Example 12.1. (*location family*) Let P_0 be determined by a pdf f_0 on $\mathcal{X} \equiv \mathbf{R}^1$ and let $\Gamma \equiv \{\mu\} \equiv \mathbf{R}^1$ be the group of all translations μ of \mathbf{R}^1 given by

$$\mu : x \rightarrow x + \mu.$$

Then \mathcal{P}_Γ is the *location-parameter family* of pdfs on \mathbf{R}^1 given by

$$\{f_\mu(x) \equiv f_0(x - \mu) \mid \mu \in \mathbf{R}^1\}.$$

More generally, if X_1, \dots, X_n is an i.i.d. sample from this family of pdfs, then $\mathcal{X} = \mathbf{R}^n$, P_0 is determined by the joint pdf $\prod_{i=1}^n f_0(x_i)$ on \mathbf{R}^n , $\Gamma \equiv \{\mu\} \equiv \mathbf{R}^1$ is the group of all translations of \mathbf{R}^n given by

$$\mu : (x_1, \dots, x_n) \rightarrow (x_1 + \mu, \dots, x_n + \mu),$$

and \mathcal{P}_Γ is the *location-parameter family* determined by the family of pdfs

$$(12.2) \quad \left\{ f_\mu(x_1, \dots, x_n) \equiv \prod_{i=1}^n f_0(x_i - \mu) \mid \mu \in \mathbf{R}^1 \right\}$$

on \mathbf{R}^n . By Lemma 12.1, any location-invariant statistic

$$(12.3) \quad V(x_1, \dots, x_n) = V(x_1 + \mu, \dots, x_n + \mu) \quad \forall \mu \in \mathbf{R}^1$$

is ancillary. Examples include:

- the set of sample spacings $(X_{(2)} - X_{(1)}, \dots, X_{(n)} - X_{(n-1)})$,
- the sample range $X_{(n)} - X_{(1)}$,
- the sample variance $s_n^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. □

Example 12.2 (*scale family*) Let P_0 be determined by a pdf f_0 on $\mathcal{X} \equiv \mathbf{R}^1$ and let $\Gamma \equiv \{\sigma\} \equiv \mathbf{R}_+^1$ be the group of all scale transformations of \mathbf{R}^1 given by

$$\sigma : x \rightarrow \sigma x.$$

Then \mathcal{P}_Γ is the *scale-parameter family* of pdfs on \mathbf{R}^1 given by

$$\{f_\sigma(x) \equiv \sigma^{-1} f_0(\sigma^{-1}x) \mid \sigma \in \mathbf{R}_+^1\}.$$

More generally, if X_1, \dots, X_n is an i.i.d. sample from this family of pdfs, then $\mathcal{X} = \mathbf{R}^n$, P_0 is determined by the joint pdf $\prod_{i=1}^n f_0(x_i)$ on \mathbf{R}^n , $\Gamma \equiv \{\sigma\} \equiv \mathbf{R}_+^1$ is the group of all scale transformations σ of \mathbf{R}^n given by

$$\sigma : (x_1, \dots, x_n) \rightarrow (\sigma x_1, \dots, \sigma x_n),$$

and \mathcal{P}_Γ is the *scale-parameter family* determined by the family of pdfs

$$(12.4) \quad \left\{ f_\sigma(x_1, \dots, x_n) \equiv \sigma^{-n} \prod_{i=1}^n f_0(\sigma^{-1}x_i) \mid \sigma \in \mathbf{R}_+^1 \right\}$$

on \mathbf{R}^n . By Lemma 12.1, any scale-invariant statistic

$$(12.5) \quad V(x_1, \dots, x_n) = V(\sigma x_1, \dots, \sigma x_n) \quad \forall \sigma \in \mathbf{R}_+^1$$

is ancillary. Examples include:

- the set of sample ratios $\left(\frac{X_{(1)}}{X_{(n)}}, \dots, \frac{X_{(n-1)}}{X_{(n)}} \right)$;
- the t -statistic $t \equiv \frac{\bar{X}_n}{s_n}$ or robust $\tilde{t} \equiv \frac{\tilde{X}_n}{X_{(\frac{3n}{4})} - X_{(\frac{n}{4})}}$ ($\tilde{X}_n =$ sample median).

Example 12.3 (*location/scale family*) Let P_0 be determined by a pdf f_0 on $\mathcal{X} \equiv \mathbf{R}^1$ and let $\Gamma \equiv \mathbf{R}^1 \times \mathbf{R}_+^1$ be the group of all location-scale transformations (μ, σ) of \mathbf{R}^1 given by

$$(\mu, \sigma) : x \rightarrow \sigma x + \mu.$$

Then \mathcal{P}_Γ is the *location/scale-parameter family* of pdfs on \mathbf{R}^1 given by

$$\{f_{\mu, \sigma}(x) \equiv \sigma^{-1} f_0(\sigma^{-1}(x - \mu)) \mid (\mu, \sigma) \in \mathbf{R}^1 \times \mathbf{R}_+^1\}.$$

More generally, if X_1, \dots, X_n is an i.i.d. sample from this family of pdfs, then $\mathcal{X} = \mathbf{R}^n$, P_0 is determined by the joint pdf $\prod_{i=1}^n f_0(x_i)$ on \mathbf{R}^n , $\Gamma \equiv \{g\} \equiv \mathbf{R}^1 \times \mathbf{R}_+^1$ is the group of all location-scale transformations of \mathbf{R}^n given by

$$g \equiv (\mu, \sigma) : (x_1, \dots, x_n) \rightarrow (\sigma x_1 + \mu, \dots, \sigma x_n + \mu),$$

and \mathcal{P}_Γ is the *location/scale-parameter family* given by the family of pdfs

$$(12.6) \quad \left\{ f_{\mu, \sigma}(x_1, \dots, x_n) \equiv \sigma^{-n} \prod_{i=1}^n f_0(\sigma^{-1}(x_i - \mu)) \mid (\mu, \sigma) \in \mathbf{R}^1 \times \mathbf{R}_+^1 \right\}$$

on \mathbf{R}^n . By Lemma 12.1, any location/scale-invariant statistic

$$(12.7) \quad V(x_1, \dots, x_n) = V(\sigma x_1 + \mu, \dots, \sigma x_n + \mu) \quad \forall (\mu, \sigma) \in \mathbf{R}^1 \times \mathbf{R}_+^1$$

is ancillary. Examples include:

- the set of normalized sample spacings $\left(\frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}, \dots, \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}} \right)$;
- the sample range/sample s.d. ratio $\frac{X_{(n)} - X_{(1)}}{s_n}$. □

12.2. Completeness, sufficiency, and ancillarity.

Definition 12.3. Let $(\mathcal{X}, \mathcal{P} \equiv \{P_\theta\})$ be a statistical model. A statistic $T \equiv T(X)$ is *complete* for \mathcal{P} if

$$(12.8) \quad E_\theta[g(T)] \text{ is } \theta\text{-free} \implies g(T) \equiv \text{constant (a.e.)}. \quad \square$$

Contrast this with the definition of ancillarity: V is *ancillary* if

$$(12.9) \quad \forall g, E_\theta[g(V)] \text{ is } \theta\text{-free}.$$

Thus *completeness and ancillarity are antithetical properties*. In fact:

Theorem 12.1. (i) If T is complete for \mathcal{P} , then no (non-constant) function of T is ancillary.

(ii) (**Basu**) If T is complete and sufficient for \mathcal{P} , then for every θ , T is independent of any ancillary statistic V .

Proof. (i) $g(T)$ ancillary $\Rightarrow E_\theta[g(T)]$ is θ -free $\Rightarrow g(T) \equiv \text{constant}$ by the completeness of T .

(ii) V ancillary $\Rightarrow P_\theta[V \in B] \equiv P[V \in B]$ is θ -free $\forall B$.

T sufficient $\Rightarrow P_\theta[V \in B | T] \equiv P[V \in B | T]$ is θ -free $\forall B$.

Thus: $g(T) \equiv P[V \in B | T] - P[V \in B]$ is θ -free $\forall B$,

that is, $g(T)$ is an actual statistic (not involving θ). But

$$\begin{aligned} E_\theta[g(T)] &= E_\theta\{P[V \in B | T] - P[V \in B]\} \\ &= P[V \in B] - P[V \in B] \\ &= 0 \quad \forall \theta, \end{aligned}$$

hence $g(T) \equiv 0$ by the completeness of T . Thus $P[V \in B | T] = P[V \in B]$, so V is independent of T . \square

Exercise 12.1*. Prove the following theorem, suggested by Theorem 12.1:

Theorem 12.2. If T is complete and sufficient, it is minimal sufficient.

Usually, sufficiency of T can be verified by the Factorization Criterion (§11.2) while ancillarity of V often can be verified by invariance (§12.1). Conditions for completeness of T are now presented. Note that the completeness condition (12.8) is equivalent to

$$(12.10) \quad E_\theta[g(T)] = 0 \quad \forall \theta \quad \implies \quad g(T) \equiv 0 \quad (\text{a.e.}).$$

Example 12.4. (i) (Example 11.2 contd.) If $T \sim \text{Binomial}(n, \theta)$, $0 < \theta < 1$, then T is complete. To verify this via (12.10) suppose that

$$\begin{aligned} E_\theta[g(T)] &= \sum_{t=0}^n g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} \\ &= (1-\theta)^n \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0 \quad \forall \theta. \end{aligned}$$

Thus the polynomial

$$\sum_{t=0}^n g(t) \binom{n}{t} x^t = 0 \quad \forall 0 < x < \infty,$$

hence its coefficients $g(0) = \cdots g(n) = 0$, so (12.10) holds.

(ii) By a similar argument, T is complete if $T \sim \text{Poisson}(\theta)$, $0 < \theta < \infty$ [verify]. (Both (i) and (ii) are special cases of the next result.) \square

Proposition 12.1. *(i) (1-parameter exponential family.) Let T have pdf (continuous) or pmf (discrete) of the exponential form*

$$(12.11) \quad f_{\theta}(t) = a(\theta)e^{\theta t}h(t), \quad \theta \in \Omega \subseteq \mathbf{R}^1.$$

If Ω contains a nondegenerate interval (a, b) then T is complete.

(ii) (k-parameter exponential family.) Let $T \equiv (T_1, \dots, T_k)$ have pdf (continuous) or pmf (discrete) of the exponential form

$$(12.12) \quad f_{\theta}(t_1, \dots, t_k) = a(\theta)e^{\theta_1 t_1 + \cdots + \theta_k t_k} h(t_1, \dots, t_k), \quad \theta \in \Omega \subseteq \mathbf{R}^k,$$

where $\theta = (\theta_1, \dots, \theta_k)$. If Ω contains a nondegenerate k -dimensional rectangle then T is complete.

Proof. (i) Apply (12.10): if $E_{\theta}[g(T)] \equiv \int g(t)f_{\theta}(t)dt = 0 \quad \forall \theta$ then for any fixed $\theta_0 \in (a, b)$,

$$(12.13) \quad \int e^{(\theta - \theta_0)t} e^{\theta_0 t} g^{+}(t) h(t) dt = \int e^{(\theta - \theta_0)t} e^{\theta_0 t} g^{-}(t) h(t) dt \quad \forall \theta \in \Omega,$$

where $g \equiv g^{+} - g^{-}$. This implies that the moment-generating functions of the two (nonnormalized) pdfs $e^{\theta_0 t} g^{+}(t) h(t)$ and $e^{\theta_0 t} g^{-}(t) h(t)$ agree on $(a - \theta_0, b - \theta_0)$. By the uniqueness of the mgf, this in turn implies that $e^{\theta_0 t} g^{+}(t) h(t) = e^{\theta_0 t} g^{-}(t) h(t)$ a.e., hence $g^{+}(t) = g^{-}(t)$ a.e., hence $g(t) = 0$ a.e., as required. (In the discrete case, replace \int by \sum .)

(ii) The proof is similar, using the uniqueness of the k -dimensional mgf. \square

Proposition 12.2. (General k -parameter exponential family.) Let X have pdf (continuous) or pmf (discrete) of the exponential form

$$(12.14) \quad f_{\theta}(x) = a(\theta)e^{\theta_1 T_1(x) + \cdots + \theta_k T_k(x)} h(x), \quad \theta \in \Omega \subseteq \mathbf{R}^k,$$

where $\theta = (\theta_1, \dots, \theta_k)$. If Ω contains a k -dimensional open rectangle then $T(X) \equiv (T_1(X), \dots, T_k(X))$ is complete.

Example 12.5. (Univariate normal distribution) X_1, \dots, X_n are i.i.d. $\sim N_1(\mu, \sigma^2)$.

Case 1: $-\infty < \mu < \infty$ is unknown, $\sigma^2 = \sigma_0^2$ is known.

This is a location family of the form (12.2) with $f_0(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-x^2/2\sigma_0^2}$. It is also a 1-parameter exponential family (12.14) with $\theta = \mu/\sigma_0^2$ and

$$T(X) = \sum X_i.$$

Thus $\sum X_i$ is complete and sufficient for μ while the vector of residuals

$$V \equiv (X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$$

is location-invariant, hence ancillary. Thus s_n^2 is ancillary, so Basu's Theorem implies that $\bar{X}_n \perp\!\!\!\perp s_n^2$, which was proved by a direct argument in §8.4.2. It also implies that \bar{X}_n is independent of

- the set of sample spacings $(X_{(2)} - X_{(1)}, \dots, X_{(n)} - X_{(n-1)})$,
- the sample range $X_{(n)} - X_{(1)}$,
- $\bar{X}_n - \tilde{X}_n$ (\tilde{X}_n = sample median),
- the set of standardized residuals $\left(\frac{X_1 - \bar{X}_n}{\sigma_0}, \dots, \frac{X_n - \bar{X}_n}{\sigma_0} \right)$,

each of which is location-invariant hence ancillary. These independences would be harder to prove directly.

Because the distribution of the residuals does not depend on the parameter μ and is independent of the statistic $\sum X_i$ used for inference about μ , these residuals can be used to *independently test the model*, i.e., the assumption of normality $N_1(\cdot, \sigma_0^2)$ (e.g., via Q-Q plots).

Case 2: $\mu = \mu_0$ is known, $0 < \sigma^2 < \infty$ is unknown.

The distribution of (X_1, \dots, X_n) is *not* a scale family of the form (12.4) (unless $\mu_0 = 0$), but that of $Y \equiv (Y_1, \dots, Y_n) \equiv (X_1 - \mu_0, \dots, X_n - \mu_0)$ does have this form with $f_0(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$. The latter distribution is also a 1-parameter exponential family (12.14) with $\theta = -1/2\sigma^2$ and

$$(12.15) \quad T(Y) = \sum Y_i^2 \equiv \sum (X_i - \mu_0)^2.$$

Thus $\sum (X_i - \mu_0)^2$ is complete and sufficient for σ while the vector of standardized residuals

$$\left(\frac{X_1 - \mu_0}{s_n}, \dots, \frac{X_n - \mu_0}{s_n} \right) \equiv \left(\frac{Y_1}{s_n}, \dots, \frac{Y_n}{s_n} \right)$$

is a scale-invariant function of (Y_1, \dots, Y_n) [verify] hence ancillary, as is the centered t -statistic

$$(12.16) \quad t_0 \equiv \frac{\bar{X}_n - \mu_0}{s_n} \equiv \frac{\bar{Y}_n}{s_n}.$$

Thus

- $\sum (X_i - \mu_0)^2 \perp\!\!\!\perp \left(\frac{X_1 - \mu_0}{s_n}, \dots, \frac{X_n - \mu_0}{s_n} \right);$
- $\sum (X_i - \mu_0)^2 \perp\!\!\!\perp t_0.$

Because the distribution of these standardized residuals does not depend on σ^2 and is independent of the statistic $\sum (X_i - \mu_0)^2$ used for inference about σ^2 , these residuals can be used to *independently test the model*, i.e., the assumption of normality $N_1(\mu_0, \cdot)$ (e.g., via Q-Q plots).

Note: In Case 2 the sample variance s_n^2 is *not* sufficient for σ^2 because $\bar{X}_n \sim N_1(\mu_0, \sigma^2/n)$ also contains relevant information about σ^2 .

Case 3: $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ are both unknown.

This is a location-scale family of the form (12.6) with $f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. It is also a 2-parameter exponential family (12.14) with

$$(12.17) \quad \theta \equiv (\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right),$$

natural parameter space $\mathbf{R}^1 \times \mathbf{R}_+^1$, and sufficient statistic

$$(12.18) \quad T(X) \equiv (T_1(X), T_2(X)) \equiv \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right).$$

Since the parameter space $\mathbf{R}^1 \times \mathbf{R}_+^1$ contains a nondegenerate rectangle, $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ or, equivalently, (\bar{X}_n, s_n^2) , is complete and sufficient for (μ, σ^2) , while the set of standardized residuals

$$(12.19) \quad \left(\frac{X_1 - \bar{X}_n}{s_n}, \dots, \frac{X_n - \bar{X}_n}{s_n} \right)$$

is location-scale invariant, hence ancillary for (μ, σ) . By Basu's Theorem, therefore,

$$\bullet (\bar{X}_n, s_n^2) \perp\!\!\!\perp \left(\frac{X_1 - \bar{X}_n}{s_n}, \dots, \frac{X_n - \bar{X}_n}{s_n} \right).$$

Also:

$$\begin{aligned} \bullet (\bar{X}_n, s_n^2) &\perp\!\!\!\perp \frac{X_{(n)} - X_{(1)}}{s_n} \text{ (the sample range/sample s.d. ratio);} \\ \bullet (\bar{X}_n, s_n^2) &\perp\!\!\!\perp \left(\frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}, \dots, \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}} \right) \text{ (the set of normalized sample spacings).} \end{aligned}$$

Because the distributions of the standardized residuals and the normalized spacings do not depend on (μ, σ^2) and is independent of the statistic (\bar{X}_n, s_n^2) used for inference about (μ, σ^2) , either set can be used to *independently test the model*, i.e., the assumption of normality $N_1(\cdot, \cdot)$ (e.g., via Q-Q plots). \square

Exercise 12.2. Consider the normal model $N_1(\mu, \mu^2)$ treated in Example 11.13(ii). In Exercise 11.6 it was shown that the minimal sufficient statistic (\bar{X}_n, s_n^2) contains the nontrivial ancillary statistic $t_n^2 \equiv \frac{\bar{X}_n^2}{s_n^2}$, so Theorem 12.1(i) implies that (\bar{X}_n, s_n^2) is not complete. Give an alternate proof of the non-completeness of (\bar{X}_n, s_n^2) by finding another nontrivial function $g(\bar{X}_n, s_n^2)$ (essentially different than t_n^2) such that $E_\mu[g(\bar{X}_n, s_n^2)]$ is μ -free.

Exercise 12.3. (i) Suppose $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$, $X \perp\!\!\!\perp Y$. Use Basu's Theorem to show that

$$(12.20) \quad (X + Y) \perp\!\!\!\perp \frac{X}{X + Y}.$$

Show in turn that

$$(12.21) \quad E\left(\frac{X}{X + Y}\right) = \frac{E(X)}{E(X + Y)}.$$

(ii) Suppose that $X_i \sim \text{Gamma}(\alpha_i, \lambda)$, $i = 1, \dots, n$, and X_1, \dots, X_n are mutually independent. Show that (recall Exercise 6.3)

$$(12.22) \quad (X_1 + \dots + X_n) \perp\!\!\!\perp \left(\frac{X_1}{X_1 + \dots + X_n}, \dots, \frac{X_{n-1}}{X_1 + \dots + X_n} \right).$$

Proposition 12.3. (i) (1-parameter truncation family.) Let X_1, \dots, X_n be an i.i.d. sample from the truncation pdf

$$(12.23) \quad f_\theta(x) \equiv [B(\theta)]^{-1} I_{(a, \theta]}(x) \cdot b(x), \quad x > a,$$

where $a \in [-\infty, \infty)$ is specified, $\theta \in (a, \infty)$ is a real parameter, $b(x) > 0$ on (a, ∞) , and $B(\theta) \equiv \int_a^\theta b(x)dx < \infty \forall \theta > a$ (see Example 11.12). The sufficient statistic $T(X) \equiv X_{(n)}$ is complete, thus minimal sufficient for θ .

Similarly, the sufficient statistic $T(X) \equiv X_{(1)}$ is complete, hence minimal sufficient for θ , if $I_{(a, \theta]}(x)$, $x > a$, is replaced by $I_{[\theta, a)}(x)$, $x < a$.

(ii) (2-parameter truncation family.) Let X_1, \dots, X_n be an i.i.d. sample from the truncation pdf

$$(12.24) \quad f_{\theta_1, \theta_2}(x) \equiv [B(\theta_1, \theta_2)]^{-1} I_{[\theta_1, \theta_2]}(x) \cdot b(x), \quad -\infty < x < \infty,$$

where $-\infty < \theta_1 < \theta_2 < \infty$, $b(x) > 0$, and $B(\theta_1, \theta_2) \equiv \int_{\theta_1}^{\theta_2} b(x)dx < \infty$ (see Example 11.14). The sufficient statistic $T(X) \equiv (X_{(1)}, X_{(n)})$ is complete, hence minimal sufficient for (θ_1, θ_2) .

Proof. (i) The cdf of $T \equiv X_{(n)}$ is

$$F_\theta(t) = P_\theta[X_{(n)} \leq t] = (P_\theta[X_1 \leq t])^n = \left(\frac{B(t)}{B(\theta)} \right)^n, \quad a \leq t \leq \theta,$$

so its pdf is $f_\theta(t) = \frac{n[B(t)]^{n-1}b(t)}{[B(\theta)]^n} I_{(a,\theta]}(t)$. Thus, if

$$E_\theta[g(T)] \equiv \int_a^\theta g(t) \frac{n[B(t)]^{n-1}b(t)}{[B(\theta)]^n} dt = 0 \quad \forall \theta > a,$$

then

$$\int_a^\theta g^+(t)[B(t)]^{n-1}b(t)dt = \int_a^\theta g^-(t)[B(t)]^{n-1}b(t)dt \quad \forall \theta > a.$$

This implies that $g^+(t) = g^-(t)$ for a.e. $t > a$ [why?], i.e., $g = 0$ a.e., so T is complete. \square

Exercise 12.4*. Prove part (ii).

Hint: One method begins by finding the joint pdf of $X_{(1)}, X_{(n)}$. A second method is to find the conditional distribution of $X_{(1)} \mid X_{(n)}$, then iterate $E_{\theta_1, \theta_2}[g(X_{(1)}, X_{(n)})]$ by conditioning on $X_{(n)}$, then apply part (i). \square

Exercise 12.5. (*Uniform* $[\theta, \theta+1]$) In Exercise 11.4, show that the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ is not complete. \square

Example 12.6. (*Uniform* $(0, \theta]$) Suppose that X_1, \dots, X_n are i.i.d. observations from the scale family *Uniform* $(0, \theta]$, $\theta > 0$ unknown. Then $T \equiv X_{(n)}$ is sufficient and complete for θ (Example 11.7 and Proposition 12.3(i)) and

$$(12.25) \quad V \equiv \left(\frac{X_{(1)}}{X_{(n)}}, \dots, \frac{X_{(n-1)}}{X_{(n)}} \right)$$

is scale-invariant, hence ancillary for θ (Example 12.2). Thus the joint distribution of these ratios does not depend on θ , in fact they are distributed as the order statistics based on a sample of size $n-1$ from *Uniform* $(0, 1)$ (see Remark 11.5, p.173). Because

$$(12.26) \quad X_{(n)} \perp\!\!\!\perp \left(\frac{X_{(1)}}{X_{(n)}}, \dots, \frac{X_{(n-1)}}{X_{(n)}} \right)$$

by Basu's Theorem, these ratios can be used to test the assumption of uniformity independently of any inference about θ based on $X_{(n)}$. [How? - justify any reasonable method].

Exercise 12.6. (*Uniform* $[\theta_1, \theta_2]$) Let X_1, \dots, X_n be i.i.d. rvs from the location/scale family [verify!] $\text{Uniform}[\theta_1, \theta_2]$, where $-\infty < \theta_1 < \theta_2 < \infty$ are both unknown. Then $T \equiv (X_{(1)}, X_{(n)})$ is sufficient and complete for (θ_1, θ_2) (Example 11.8, p.173 and Proposition 12.3(ii), p.192) and

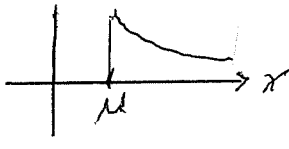
$$(12.27) \quad V \equiv \left(\frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}, \dots, \frac{X_{(n-1)} - X_{(1)}}{X_{(n)} - X_{(1)}} \right)$$

is location/scale-invariant, hence ancillary for (θ_1, θ_2) . Show that they are distributed as the order statistics based on a sample of size $n - 2$ from $\text{Uniform}(0, 1)$. Because

$$(12.28) \quad (X_{(1)}, X_{(n)}) \perp\!\!\!\perp \left(\frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}, \dots, \frac{X_{(n-1)} - X_{(1)}}{X_{(n)} - X_{(1)}} \right)$$

by Basu's Theorem, show how these ratios can be used to test the assumption of uniformity independently of any inference about (θ_1, θ_2) based on $(X_{(1)}, X_{(n)})$ - justify any reasonable method. \square

Example 12.7. (*The location/scale family Exponential* (μ, σ)) The following location-scale family combines the features of a 1-parameter exponential family and a truncation-parameter family: Let $f_0(x) = e^{-x} I_{(0, \infty)}(x)$, so by (12.6) on p.186),

$$(12.29) \quad \begin{aligned} f_{\mu, \sigma}(x_1, \dots, x_n) &= \prod_{i=1}^n \sigma^{-1} e^{-(x_i - \mu)/\sigma} I_{[\mu, \infty)}(x_i) \\ &= \sigma^{-n} e^{n\mu/\sigma} e^{-\sum x_i/\sigma} \cdot I_{[\mu, \infty)}(x_{(1)}). \end{aligned}$$


Then $(X_{(1)}, \sum X_i)$, or equivalently $(X_{(1)}, \sum X_{(i)})$, or equivalently

$$(12.30) \quad T \equiv \left(X_{(1)}, \sum_{i=1}^n X_{(i)} - nX_{(1)} \right) = \left(X_{(1)}, \sum_{i=2}^n (X_{(i)} - X_{(1)}) \right),$$

is sufficient for (μ, σ) (by the factorization criterion). To show that T is complete for (μ, σ) , proceed as follows:

First fix σ , so (12.29) is the joint pdf of a sample from a 1-parameter truncation family with pdf of the form (11.47) with $\theta = \mu$, $a = \infty$, and

$b(x) = e^{-x/\sigma}$. Note that this is also a location family with $\theta = \mu$. From (12.29) $T_1 \equiv X_{(1)}$ is sufficient for μ , while by Proposition 12.3(i), T_1 is complete for μ . Furthermore, $T_2 \equiv \sum_{i=2}^n (X_{(i)} - X_{(1)})$ is location-invariant hence ancillary. Thus by Basu's Theorem, $T_1 \perp\!\!\!\perp T_2 \forall \mu, \sigma$.

Next, for fixed μ , $(Z_1, \dots, Z_n) \equiv (X_1 - \mu, \dots, X_n - \mu)$ is an i.i.d. sample from $\text{Expo}(\frac{1}{\sigma})$ and

$$(12.31) \quad (X_{(1)}, \dots, X_{(n)}) \stackrel{d}{=} (Z_{(1)} + \mu, \dots, Z_{(n)} + \mu).$$

By the memory-free property of $\text{Expo}(\cdot)$, $Z_{(2)} - Z_{(1)}, \dots, Z_{(n)} - Z_{(1)}$ have the same distribution as the order statistics $U_{(1)}, \dots, U_{(n-1)}$ from a sample U_1, \dots, U_{n-1} of size $n-1$ from $\text{Expo}(\frac{1}{\sigma})$, so

$$(12.32) \quad T_2 \equiv \sum_{i=2}^n (Z_{(i)} - Z_{(1)}) \stackrel{d}{=} \sum_{i=1}^{n-1} U_{(i)} = \sum_{i=1}^{n-1} U_i \sim \text{Gamma}(n-1, \frac{1}{\sigma}),$$

a 1-parameter exponential family. Thus T_2 is complete for σ by Proposition 12.1(i), p.188.

Now apply (12.10) to verify the completeness of (T_1, T_2) : suppose that

$$0 = E_{\mu, \sigma}[g(T_1, T_2)] \equiv E_{\mu, \sigma} \{E_{\sigma}[g(T_1, T_2) \mid T_1]\} \quad \forall \mu, \sigma.$$

Since T_1 is complete for μ with σ fixed, this implies that

$$E_{\sigma}[g(t_1, T_2) \mid T_1 = t_1] = 0 \quad \text{for a.e. } t_1$$

for every σ . Since $T_1 \perp\!\!\!\perp T_2$ and T_2 is complete for σ , this implies that for a.e. fixed t_1 , $g(t_1, t_2) = 0$ for a.e. t_2 . Thus $g(t_1, t_2) = 0$ for a.e. (t_1, t_2) , hence (T_1, T_2) is complete for (μ, σ) as claimed.²⁰

Now apply the location/scale-invariance [verify] of

$$(12.33) \quad V \equiv (V_2, \dots, V_{n-1}) \equiv \left(\frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}, \dots, \frac{X_{(n-1)} - X_{(1)}}{X_{(n)} - X_{(1)}} \right)$$

²⁰ A completely rigorous proof would require consideration of exceptional null sets.

to see that V is ancillary for (μ, σ) . By Basu's Theorem, therefore,

$$(12.34) \quad \left(X_{(1)}, \sum_{i=2}^n (X_{(i)} - X_{(1)}) \right) \perp\!\!\!\perp V. \quad \square$$

Exercise 12.7. Since the distribution of $V \equiv (V_2, \dots, V_{n-1})$ in (12.33) is (μ, σ) -free, these ratios can be used to test the model assumption that $f_0(x) = e^{-x} I_{(0, \infty)}(x)$. Show that the joint pdf of (V_2, \dots, V_{n-1}) is given by (12.35)

$$f(v_2, \dots, v_{n-1}) = \frac{(n-1)!(n-2)!}{(v_2 + \dots + v_{n-1} + 1)^{n-1}}, \quad 0 < v_2 < \dots < v_{n-1} < 1.$$

*Justify any reasonable method based on this pdf that uses V to test the model assumption $\text{Exponential}(\cdot, \cdot)$. \square

Remark 12.1. We have seen that for an i.i.d. sample from an exponential family or truncation family, the dimensionality of the minimal sufficient statistic does not vary with the sample size n . The same is true for the combined exponential/truncation family in Example 12.7. Under mild regularity conditions it can be shown that *only* exponential families, truncation families, or a combination of the two, have this property. (E. B. Dynkin, L. D. Brown, O. Barndorff-Nielsen, etc.) In most other cases the order statistics or their multivariate extension are minimal sufficient. \square

Example 12.8. (*Some non-normal location parameter families*). Let X_1, \dots, X_n be i.i.d. r.v.s, where either $X_i \sim \text{Cauchy}(\theta)$, $X_i \sim \text{double exponential}(\theta)$, or $X_i \sim \text{logistic}(\theta)$. Then the order statistics $X_{(1)}, \dots, X_{(n)}$ are minimal sufficient but not complete, since $V := X_{(n)} - X_{(1)}$ is a nontrivial ancillary statistic. (See CB Exercise 6.9.) \square

Exercise 12.8. Let $X_1, \dots, X_m, Y_1, \dots, Y_n$ be independent with $X_i \sim N_1(\mu, \sigma^2)$, $Y_j \sim N_1(\mu, \tau^2)$, $i = 1, \dots, m$, $j = 1, \dots, n$. Show that

$$(12.36) \quad \left(\sum X_i, \sum X_i^2, \sum Y_j, \sum Y_j^2 \right)$$

is a minimal sufficient statistic for (μ, σ^2, τ^2) ; hence so is the equivalent statistic $(\bar{X}_m, s_x^2, \bar{Y}_n, s_y^2)$. However neither is complete: $g := \bar{X} - \bar{Y}$ is an unbiased estimator of 0; recall condition (12.10) for completeness.)

[*Note:* This is the null model for the *Behrens-Fisher problem* of testing $\mu = \nu$ based on samples from two normal populations $N_1(\mu, \sigma^2)$ and $N_1(\nu, \tau^2)$ when both variances are unknown and unrelated. No simple ancillary statistic is known when $\mu = \nu$, so no simple conditional inference procedure for the common mean is available.] \square

Example 12.9. Let $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be an i.i.d. sample from the bivariate normal distribution

$$(12.37) \quad N_2 \left[\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \right].$$

Thus $X_i \perp\!\!\!\perp Y_i$ so this constitutes a 4-parameter exponential family model in which the sample means and sample variances $(\bar{X}_n, s_n^2, \bar{Y}_n, t_n^2)$ together constitute a sufficient and complete statistic for $(\mu, \sigma^2, \nu, \tau^2)$ [verify via Proposition 12.2]. This can also be viewed as the combination of two independent location/scale families, so the sample correlation coefficient

$$(12.38) \quad r_n \equiv \frac{\sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{[\sum (X_i - \bar{X}_n)^2]^{1/2} [\sum (Y_i - \bar{Y}_n)^2]^{1/2}},$$

being location/scale-invariant, is ancillary. Thus by Basu's Theorem,

$$(12.39) \quad r_n \perp\!\!\!\perp (\bar{X}_n, s_n^2, \bar{Y}_n, t_n^2),$$

so the ancillary statistic r_n can be used to test the assumption of independence, assuming normality. [How? Justify any reasonable method.] \square

Exercise 12.9. (*Complete sufficient statistics in nonparametric models*)

(a) $\mathcal{P} = \{\text{all symmetric pdfs } f(-x) = f(x) \text{ on } \mathbf{R}^1\}$ (see Example 11.3). The sufficient statistic $T(X) \equiv |X|$ is complete for f_+ [verify!], hence minimal sufficient, and $\Psi \equiv \text{sign}(X)$ is ancillary by symmetry, so $|X| \perp\!\!\!\perp \Psi$ by Basu.

(b) $\mathcal{P} = \{\text{all exchangeable pdfs } f(\pi x) = f(x) \text{ on } \mathbf{R}^n\}$ (see Example 11.6). The sufficient statistic $T(X) \equiv (X_{(1)}, \dots, X_{(n)})$ is complete for $f_<$ [verify!], hence minimal sufficient, and $\Pi \equiv (\text{rank}(X_1), \dots, \text{rank}(X_n))$ is ancillary by exchangeability, so $T \perp\!\!\!\perp \Pi$ by Basu.

(c) $\mathcal{P} = \{\text{all radial pdfs } f(x) = g(\|x\|) \text{ on } \mathbf{R}^n\}$ (see Remark 11.1). The sufficient statistic $R \equiv \|X\|$ is complete for g [verify!], hence minimal sufficient, and $\vec{X} \equiv \frac{X}{\|X\|}$ is ancillary [verify!], so $R \perp\!\!\!\perp \vec{X}$ by Basu. \square

12.3. Minimum-variance unbiased estimation via a complete sufficient statistic.

Suppose that we wish to estimate a *real-valued function* $\tau \equiv \tau(\theta)$ of the parameter θ based on the observed data $X \sim P_\theta$. For example:

$$(12.40) \quad \tau(\theta) = \theta \text{ or } \frac{1}{\theta}, \quad \tau(\theta_1, \theta_2) = \theta_1 \text{ or } \theta_1 - \theta_2 \text{ or } \frac{\theta_1}{\theta_2}.$$

An *estimator* of τ is any real-valued statistic $\tilde{\tau} \equiv \tilde{\tau}(X)$ with the same range as τ . It is customary and mathematically convenient to evaluate the performance of $\tilde{\tau}$ by its *mean-square error (MSE) function*

$$(12.41) \quad \text{MSE}_\theta(\tilde{\tau}) = \text{E}_\theta[\tilde{\tau}(X) - \tau(\theta)]^2.$$

Notice that this depends on the unknown θ .

As in (5.4) we have the *basic MSE decomposition*

$$(12.42) \quad \begin{aligned} \text{MSE}_\theta(\tilde{\tau}) &= \text{Var}_\theta(\tilde{\tau}) + [\text{E}_\theta(\tilde{\tau}) - \tau(\theta)]^2 \\ &\equiv \text{Var}_\theta(\tilde{\tau}) + [\text{Bias}_\theta(\tilde{\tau})]^2. \end{aligned}$$

Definition 12.4. (i) $\tilde{\tau}$ is an *unbiased* estimator of τ if $\text{E}_\theta(\tilde{\tau}) = \tau(\theta) \forall \theta$, i.e., if $\text{Bias}_\theta(\tilde{\tau}) = 0 \forall \theta$. In this case,

$$(12.43) \quad \text{MSE}_\theta(\tilde{\tau}) = \text{Var}_\theta(\tilde{\tau}).$$

(ii) An unbiased estimator $\hat{\tau}$ of τ is a *uniformly minimum-variance unbiased estimator (UMVUE)* if, for all other unbiased estimators $\tilde{\tau}$,

$$(12.44) \quad \text{Var}_\theta(\hat{\tau}) \leq \text{Var}_\theta(\tilde{\tau}) \quad \forall \theta.$$

If it exists, $\hat{\tau}$ thus has smallest MSE among all unbiased estimators of τ . \square

Note: There may exist *biased* estimators with smaller MSE – see Examples 12.10 and 12.11 and Exercise 12.10.

Lemma 12.2 (*The Improvement Lemma*). Suppose that $T \equiv T(X)$ is a sufficient statistic for θ . If $\tilde{\tau} \equiv \tilde{\tau}(X)$ is an unbiased estimator of $\tau \equiv \tau(\theta)$ then

$$(12.45) \quad \hat{\tau} \equiv \hat{\tau}(T) \equiv E_{\theta}[\tilde{\tau} \mid T] \equiv E[\tilde{\tau} \mid T]$$

does not depend on θ by sufficiency, hence is also a bona fide estimator of $\tau(\theta)$. Furthermore, $\hat{\tau}$ is also unbiased for τ and satisfies (12.44), hence has smaller MSE than $\tilde{\tau}$; in fact, strictly smaller unless $\tilde{\tau}$ is a function of T .

Proof. The estimator $\hat{\tau}(T) \equiv E[\tilde{\tau} \mid T]$ is unbiased for $\tau(\theta)$ because $\tilde{\tau}$ is unbiased:

$$(12.46) \quad E_{\theta}[\hat{\tau}(T)] = E_{\theta}\{E[\tilde{\tau} \mid T]\} = E_{\theta}[\tilde{\tau}(X)] = \tau(\theta) \quad \forall \theta.$$

Then (12.44) follows from (5.11) and (12.45):

$$(12.47) \quad \text{Var}_{\theta}(\tilde{\tau}) = E_{\theta}\{\text{Var}[\tilde{\tau} \mid T]\} + \text{Var}_{\theta}(\hat{\tau}) \geq \text{Var}_{\theta}(\hat{\tau}).$$

Strict inequality holds unless $\text{Var}[\tilde{\tau} \mid T] = 0$, i.e., $\tilde{\tau} = E[\tilde{\tau} \mid T]$, for all T . \square

Lemma 12.3 (*The Uniqueness Lemma*). If $T \equiv T(X)$ is complete, then $\tau(\theta)$ admits at most one unbiased estimator $\hat{\tau}(T)$ depending on T .

Proof. If $\hat{\tau}(T)$ and $\check{\tau}(T)$ are two unbiased estimators of $\tau(\theta)$ then

$$(12.48) \quad E_{\theta}[\hat{\tau}(T) - \check{\tau}(T)] = \tau(\theta) - \tau(\theta) = 0 \quad \forall \theta,$$

so $\hat{\tau}(T) - \check{\tau}(T) = 0$ by completeness, i.e., $\hat{\tau}(T) = \check{\tau}(T)$ as required. \square

Theorem 12.3 (*Rao-Blackwell-Lehmann-Scheffe (RBLS)*). Let $T \equiv T(X)$ be complete and sufficient for θ . If there exists at least one unbiased estimator $\tilde{\tau} \equiv \tilde{\tau}(X)$ for $\tau(\theta)$ then there exists a unique UMVUE $\hat{\tau} \equiv \hat{\tau}(T)$ for $\tau(\theta)$, namely,

$$(12.49) \quad \hat{\tau}(T) \equiv E[\tilde{\tau}(X) \mid T].$$

Proof. Clearly $\hat{\tau}(T)$ is unbiased for $\tau(\theta)$. Let $\check{\tau}(X)$ be any other unbiased estimator for $\tau(\theta)$ and let $\check{\tau}(T) = E[\check{\tau} \mid T]$, so $\check{\tau}(T)$ is also unbiased for θ , hence $\hat{\tau} = \check{\tau}$ by the Uniqueness Lemma. But

$$(12.50) \quad \text{Var}_{\theta}(\hat{\tau}) = \text{Var}_{\theta}(\check{\tau}) \leq \text{Var}_{\theta}(\check{\tau}) \quad \forall \theta$$

by the Improvement Lemma, so $\hat{\tau}$ is the UMVUE for $\tau(\theta)$. \square

Corollary 12.1 (*The UMVUE Supermarket*). *Let $T \equiv T(X)$ be complete and sufficient for θ . Then any function $\phi(T)$ is the UMVUE of its expectation $E_\theta[\phi(T)] \equiv \tau(\theta)$ (provided that this expectation is finite $\forall \theta$).*

Example 12.10. Let X_1, \dots, X_n be i.i.d. $\sim N_1(\mu, \sigma_0^2)$ with σ_0^2 known. This is a 1-parameter exponential family, so $T \equiv \bar{X}_n$ is complete and sufficient for μ . Thus by the UMVUE Supermarket:

(a) $\phi_1(\bar{X}_n) \equiv \bar{X}_n$ is the UMVUE of μ , because $E_\mu(\bar{X}_n) = \mu$.

(b) $\phi_2(\bar{X}_n) \equiv \bar{X}_n^2 - \frac{\sigma_0^2}{n}$ is the UMVUE of μ^2 , because

$$E_\mu[\bar{X}_n^2] = \text{Var}_\mu(\bar{X}_n) + [E_\mu(\bar{X}_n)]^2 = \frac{\sigma_0^2}{n} + \mu^2.$$

Here, however, the UMVUE $\phi_2(\bar{X}_n)$ has the undesirable property that $\phi_2(\bar{X}_n) < 0$ with positive probability, so it is not a valid estimator of $\mu^2 \geq 0$. A more reasonable estimator would be $\phi_2^+ \equiv \max(\phi_2, 0)$, which is no longer unbiased for μ^2 , but has smaller MSE, since obviously

$$(12.51) \quad (\phi_2^+ - \mu^2)^2 \leq (\phi_2 - \mu^2)^2.$$

This shows that unbiasedness is not particularly desirable under quadratic loss when the parameter space is truncated. (Also see Example 12.11.) \square

Example 12.11. Let X_1, \dots, X_n be i.i.d. $\sim N_1(\mu_0, \sigma^2)$ with μ_0 known, so $T \equiv \sum (X_i - \mu_0)^2 \sim \sigma^2 \chi_n^2$ is complete and sufficient for σ^2 . Thus by the UMVUE Supermarket:

(c) $\phi_3(T) \equiv \frac{T}{n} \equiv \frac{1}{n} \sum (X_i - \mu_0)^2$ is the UMVUE of σ^2 .

However, ϕ_3 does *not* have the smallest possible MSE – *there exists a biased estimator whose variance is small enough to reduce the overall MSE (12.42).*

To see this, consider estimators of the form aT , where $a > 0$ is a constant. The UMVUE ϕ_3 is of this form with $a = \frac{1}{n}$. Then from (12.42),

$$\begin{aligned} \text{MSE}_\sigma(aT) &= \text{Var}_\sigma(aT) + [\text{Bias}_\sigma(aT)]^2 \\ &= a^2 \sigma^4 (2n) + (a\sigma^2 n - \sigma^2)^2 \\ &= \sigma^4 [2na^2 + (an - 1)^2]. \end{aligned}$$

The last expression is a quadratic function of a that is minimized when $a = \frac{1}{n+2}$ regardless of σ^2 . Thus the estimator $\frac{T}{n+2}$, although biased, has smaller MSE than the UMVUE $\frac{T}{n}$.

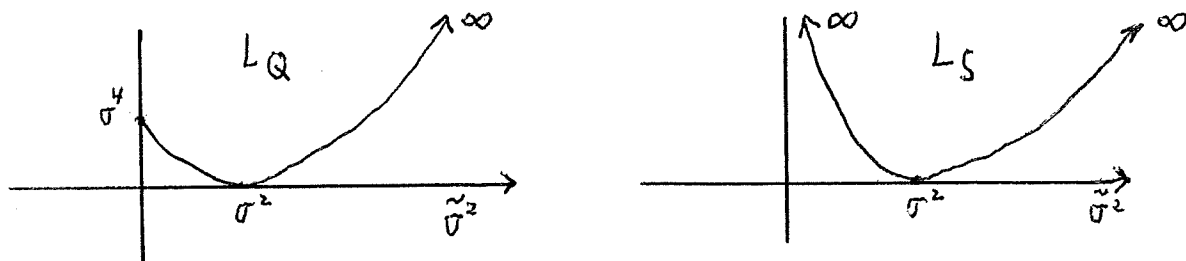
In fact, $\frac{T}{n+2}$ has a *downward* bias:

$$E_{\sigma} \left(\frac{T}{n+2} \right) = \left(\frac{n}{n+2} \right) \sigma^2 < \sigma^2.$$

Why should an estimator with downward bias be preferable to the UMVUE (i.e., have smaller MSE)? This can be explained by the fact that the truncation restriction $\{\sigma^2 > 0\}$ on the parameter space introduces an asymmetry in the quadratic loss function

$$(12.52) \quad L_Q(\tilde{\sigma}^2, \sigma^2) \equiv (\tilde{\sigma}^2 - \sigma^2)^2$$

that MSE uses to measure the accuracy of an estimate $\tilde{\sigma}^2$:



Note that for fixed σ^2 , L_Q is *bounded* for *underestimates* ($L_Q(\tilde{\sigma}^2, \sigma^2) \leq \sigma^4$ as $\tilde{\sigma}^2 \rightarrow 0$) but L_Q is *unbounded* for *overestimates* ($L_Q(\tilde{\sigma}^2, \sigma^2) \rightarrow \infty$ as $\tilde{\sigma}^2 \rightarrow \infty$). Thus underestimates are not penalized as severely as overestimates, which explains why an estimator with downward bias can have smaller MSE than the best unbiased estimator.

In my opinion, however, the solution is not to use a biased estimator but to use a loss function that penalizes underestimates as much as overestimates. One such loss function that is appropriate for the estimation of a positive scale parameter is *Stein's loss function*

$$(12.53) \quad L_S(\tilde{\sigma}^2, \sigma^2) \equiv \left(\frac{\tilde{\sigma}^2}{\sigma^2} \right) - \log \left(\frac{\tilde{\sigma}^2}{\sigma^2} \right) - 1.$$

For fixed σ^2 , L_S is unbounded for both underestimates and overestimates: $L_S(\tilde{\sigma}^2, \sigma^2) \rightarrow \infty$ as $\tilde{\sigma}^2 \rightarrow 0$ or ∞ . When Stein's loss function is used, the UMVUE $\frac{T}{n}$ performs better than the minimum-MSE estimator $\frac{T}{n+2}$.

In fact, $\frac{T}{n}$ is optimal among all estimators of the form aT , $a > 0$:

$$\begin{aligned} E_{\sigma}[L_S(aT, \sigma^2)] &= E_{\sigma} \left[\left(\frac{aT}{\sigma^2} \right) - \log \left(\frac{aT}{\sigma^2} \right) - 1 \right] \\ &= an - \log a - E_{\sigma}(\log \chi_n^2) - 1, \end{aligned}$$

which is minimized by $a = \frac{1}{n}$ regardless of σ^2 .

Note: This is why I prefer the unbiased sample variance $s_n^2 \equiv \frac{\sum (X_i - \bar{X}_n)^2}{n-1}$ to the MLE $\frac{\sum (X_i - \bar{X}_n)^2}{n}$ or the minimum-MSE $\frac{\sum (X_i - \bar{X}_n)^2}{n+1}$ for estimating σ^2 based on $X_1, \dots, X_n \sim N_1(\mu, \sigma^2)$. \square

Example 12.12. Let X_1, \dots, X_n be i.i.d. $\sim N_1(\mu, \sigma^2)$, so $T \equiv (\bar{X}_n, s_n^2)$ is complete and sufficient for (μ, σ^2) . Thus by the UMVUE Supermarket:

(d) $\phi_4(\bar{X}_n, s_n^2) \equiv \bar{X}_n$ is again the UMVUE of μ .

(e) $\phi_5(\bar{X}_n, s_n^2) \equiv s_n^2$ is now the UMVUE of σ^2 .

[But $\left(\frac{n-1}{n+1}\right) s_n^2$ has smaller MSE! - why?]

(f) $\phi_6(\bar{X}_n, s_n^2) \equiv c_n s_n$ is the UMVUE of σ , where $c_n = \sqrt{\frac{n-1}{2}} \frac{\Gamma[(n-1)/2]}{\Gamma[n/2]}$.

(g) $\phi_7(\bar{X}_n, s_n^2) \equiv \bar{X}_n \pm \gamma c_n s_n$ is the UMVUE of $\mu \pm \gamma\sigma$ (tolerance limits).

(h) $\phi_8(\bar{X}_n, s_n^2) \equiv \bar{X}_n^2 - \frac{s_n^2}{n}$ is the UMVUE of μ^2 .

[But ϕ_8^+ is better – see Example 12.10b, p.200.] \square

Example 12.13. Let X_1, \dots, X_n be i.i.d. $\sim \text{Poisson}(\lambda)$. This is a 1-parameter exponential family, so $T \equiv \sum X_i$ is complete and sufficient for λ . Thus by the UMVUE Supermarket:

(a) $\frac{T}{n}$ is the UMVUE for λ .

(b) $\frac{T^2 - T}{n^2}$ is the UMVUE for λ^2 [verify!].

Note: $T = 0 \Rightarrow \frac{T^2 - T}{n^2} = 0$. [This makes sense.]

$T = 1 \Rightarrow \frac{T^2 - T}{n^2} = 0$. [This doesn't make sense: $T = 1 \Rightarrow \lambda > 0$!]

$T \geq 2 \Rightarrow \frac{T^2 - T}{n^2} > 0$. [This makes sense.]

Moral: The UMVUE criterion may not lead to sensible estimators!

(c) Suppose we wish to estimate

$$(12.54) \quad \tau \equiv \tau(\lambda) \equiv e^{-\lambda} = P_\lambda[X_1 = 0] = P[\text{no events occur}].$$

(For example, we may wish to estimate the probability that no accidents occur in a unit time.) To shop at the UMVUE Supermarket, we'd have to guess a function of T that is unbiased for $e^{-\lambda}$. Instead, we can use the *constructive approach* of the RBLT Theorem, as follows.

An obvious unbiased estimator of τ is $\tilde{\tau}(X_1, \dots, X_n) \equiv I_{\{0\}}(X_1)$. Thus, by the RBLT Theorem, the UMVUE $\hat{\tau}$ is found via (12.49): for $t = 0, 1, \dots$,

$$\begin{aligned} \hat{\tau}(t) &= E[\tilde{\tau}(X_1, \dots, X_n) \mid T = t] \\ &= P\left[X_1 = 0 \mid \sum X_i = t\right] \\ &= \frac{P[X_1 = 0, X_2 + \dots + X_n = t]}{P[X_1 + \dots + X_n = t]} \\ &= \frac{P[X_1 = 0] P[X_2 + \dots + X_n = t]}{P[X_1 + \dots + X_n = t]} \\ &= \frac{e^{-\lambda} \cdot e^{-(n-1)\lambda} [(n-1)\lambda]^t / t!}{e^{-n\lambda} (n\lambda)^t / t!} \\ (12.55) \quad &= \left(\frac{n-1}{n}\right)^t. \end{aligned}$$

Thus $\hat{\tau}(t) \equiv \left(\frac{n-1}{n}\right)^T$ is the UMVUE of $e^{-\lambda}$.

Question: How does $\left(\frac{n-1}{n}\right)^T$ compare to the natural estimator $e^{-T/n}$? Because $\frac{T}{n} \rightarrow \lambda$ by the LLN, the ratio

$$\frac{\left(\frac{n-1}{n}\right)^T}{e^{-T/n}} = \left[\frac{\left(1 - \frac{1}{n}\right)^n}{e^{-1}} \right]^{\frac{T}{n}} \rightarrow \left[\frac{e^{-1}}{e^{-1}} \right]^\lambda = 1,$$

so the two estimators are asymptotically equal. \square

Example 12.14. Let X_1, \dots, X_n be i.i.d. $\sim \text{Uniform}(0, \theta]$. This is a 1-parameter truncation family, so $T \equiv X_{(n)} = \max(X_1, \dots, X_n)$ is complete and sufficient for θ . Since $E_\theta(X_{(n)}) = \left(\frac{n}{n+1}\right) \theta$, the UMVUE Supermarket tells us that $\phi(X_{(n)}) \equiv \left(\frac{n+1}{n}\right) X_{(n)}$ is the UMVUE of θ .

Suppose instead that we apply the RBLS constructive approach. One unbiased estimator of θ is $\tilde{\theta}(X_1, \dots, X_n) \equiv 2X_1$, so (12.49) tells us that $E[2X_1 \mid X_{(n)}]$ is the UMVUE of θ . To evaluate this, recall from Remark 11.5 that the conditional distribution of $(X_{(1)}, \dots, X_{(n-1)}) \mid X_{(n)}$ is the same as the distribution of the order statistics $U_{(1)}, \dots, U_{(n-1)}$ for an i.i.d. sample U_1, \dots, U_{n-1} from the Uniform $(0, X_{(n)})$ distribution. Therefore

$$\begin{aligned} E[X_1 \mid X_{(n)}] &= \frac{1}{n} E[X_1 + \dots + X_n \mid X_{(n)}] && \text{[by symmetry]} \\ &= \frac{1}{n} E[X_{(1)} + \dots + X_{(n-1)} + X_{(n)} \mid X_{(n)}] \\ &= \frac{1}{n} E[U_{(1)} + \dots + U_{(n-1)} \mid X_{(n)}] + \frac{1}{n} X_{(n)} \\ &= \frac{1}{n} E[U_1 + \dots + U_{n-1} \mid X_{(n)}] + \frac{1}{n} X_{(n)} \\ &= \left(\frac{n-1}{n}\right) \frac{X_{(n)}}{2} + \frac{X_{(n)}}{n} \\ &= \left(\frac{n+1}{n}\right) \frac{X_{(n)}}{2}, \end{aligned}$$

so the UMVUE $E[2X_1 \mid X_{(n)}]$ is $\left(\frac{n+1}{n}\right) X_{(n)} \equiv \phi(X_{(n)})$, as before. \square

Remark 12.2. In the truncation-parameter Example 12.14, the variance of the UMVUE is

$$\begin{aligned} \text{Var}_\theta \left[\left(\frac{n+1}{n}\right) X_{(n)} \right] &= \left(\frac{n+1}{n}\right)^2 \left[E_\theta(X_{(n)}^2) - (E_\theta(X_{(n)}))^2 \right] \\ &= \theta^2 \left(\frac{n+1}{n}\right)^2 \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2 \right] \\ &= \frac{\theta^2}{n(n+2)}, \end{aligned}$$

which is $O(1/n^2)$. This contrasts with the exponential family Examples 12.10 - 12.13, where the variance of each UMVUE is $O(1/n)$ (also see Remark 13.4). This exhibits a fundamental difference between truncation families, where the support of the distribution P_θ depends on the unknown parameter θ , and “regular families” (including exponential families), where the support of P_θ is the same for every θ . (The asymptotic variance of the MLE in regular families is always $O(1/n)$ – see Theorem 14.9.) \square

As in Example 12.11, however, $\phi(X_{(n)}) \equiv \left(\frac{n+1}{n}\right) X_{(n)}$ does *not* have the smallest possible MSE – there exists a biased estimator whose variance is small enough to reduce the overall MSE (12.42):

Exercise 12.10. (i) In Example 12.14, find that $\tilde{a} > 0$ such that the estimator $\tilde{a}T$ minimizes the MSE $E_\theta[(aT - \theta)^2]$ for all $\theta > 0$. Show that $\tilde{a}T$ is not unbiased, so is not the UMVUE.

(ii) If the Stein loss function (12.53) is used instead of quadratic loss (12.52), show that the UMVUE $\left(\frac{n+1}{n}\right) X_{(n)}$ is the optimal estimator of θ among all estimators of the form $aX_{(n)}$, $a > 0$. \square

12.4. Extension to general convex loss functions.

As in Section 12.3, suppose we wish to estimate $\tau \equiv \tau(\theta)$ based on the observed data $X \sim P_\theta$. Instead of the MSE criterion based on quadratic loss, suppose we evaluate the accuracy of an estimator $\tilde{\tau} \equiv \tilde{\tau}(x)$ by a *convex loss function* $L(\tilde{\tau}, \tau)$, i.e., one that satisfies the following properties:

- (i) $L(\tilde{\tau}, \tau) \geq 0 \ \forall \ \tilde{\tau}, \tau, \ L(\tau, \tau) = 0$.
- (ii) $L(\tilde{\tau}, \tau)$ is a convex function of $\tilde{\tau}$ for each fixed τ .

By (ii), for each fixed τ , $L(\tilde{\tau}, \tau)$ increases as $\tilde{\tau}$ moves away from τ .

Examples of such convex loss functions include [verify!]:

- (a) general power loss: $L(\tilde{\tau}, \tau) = a(\tau) \cdot |\tilde{\tau} - \tau|^\alpha$ for $\alpha \geq 1$ and $a(\tau) > 0$;

(One example is ordinary quadratic loss $(\tilde{\tau} - \tau)^2$. Another is relative quadratic loss: $\left(\frac{\tilde{\tau}}{\tau} - 1\right)^2$ for $\tau > 0$.)

- (b) Stein’s loss function: $L(\tilde{\tau}, \tau) = \left(\frac{\tilde{\tau}}{\tau}\right) - \log\left(\frac{\tilde{\tau}}{\tau}\right) - 1$.

Definition 12.4(ii) is now generalized as follows:

Definition 12.5. Let L be a convex loss function. An unbiased estimator $\hat{\tau}$ of τ is a *uniformly minimum-loss unbiased estimator (UMLUE) w.r.to L* if, for all other unbiased estimators $\tilde{\tau}$,

$$(12.56) \quad E_{\theta}[L(\hat{\tau}, \tau)] \leq E_{\theta}[L(\tilde{\tau}, \tau)] \quad \forall \theta.$$

The Improvement Lemma 12.2 is generalized as follows:

Lemma 12.4 (*The Improvement Lemma for convex loss*). Suppose that $T \equiv T(X)$ is a sufficient statistic for θ . If $\tilde{\tau} \equiv \tilde{\tau}(X)$ is an unbiased estimator of $\tau \equiv \tau(\theta)$ then

$$(12.57) \quad \hat{\tau} \equiv \hat{\tau}(T) \equiv E_{\theta}[\tilde{\tau} | T] \equiv E[\tilde{\tau} | T]$$

does not depend on θ by sufficiency, hence is also a bona fide estimator of $\tau(\theta)$. Furthermore, $\hat{\tau}$ is also unbiased estimator for τ and has smaller expected loss than $\tilde{\tau}$, i.e., satisfies (12.56).

Proof. The proof is essentially the same as that of the original Improvement Lemma, except that the variance inequality (12.47) must be replaced by Jensen's inequality applied conditionally on T :

$$\begin{aligned} E_{\theta}[L(\hat{\tau}, \tau)] &\equiv E_{\theta}[L(E[\tilde{\tau} | T], \tau)] \\ &\leq E_{\theta}\{E_{\theta}[L(\tilde{\tau}, \tau) | T]\} && \text{[Jensen]} \\ &= E_{\theta}[L(\tilde{\tau}, \tau)]. && \square \end{aligned}$$

The Uniqueness Lemma 12.3 remains unchanged, as do the RBLS Theorem 12.3 and the UMLUE Supermarket Corollary 12.1 with UMLUE replaced by UMLUE in both. Thus *if a complete and sufficient statistic T exists, any function $\phi(T)$ is the UMLUE of its (finite) expectation $E_{\theta}[\phi(T)] \equiv \tau(\theta)$ for all convex loss functions.*

Note: As seen in Examples 12.10 and 12.11 and Exercise 12.10 for the case of quadratic loss, there may exist *biased* estimators with smaller expected loss than the UMLUE.

13. The Information Inequality.

We now present the *Cramér-Rao-Frechet (CR) lower bound* for the variance of an unbiased estimator in a regular \equiv smooth statistical model. This bound is called the *Information Inequality*, since it depends on the *Fisher Information Number (FIN)* which measures the intrinsic accuracy of a parametric statistical model.²¹ The Information Inequality provides an alternate approach to determining UMVUEs. Although it is less widely applicable than the RBLT Theorem approach, it is ultimately of greater importance because, as we shall see in Section 14, the FIN determines the asymptotic variance of the MLE in *any* regular statistical model.

13.1. Variance bounds.

Consider a statistical model $(\mathcal{X}, \mathcal{P} \equiv \{P_\theta \mid \theta \in \Omega\})$ where each P_θ is determined by a pdf (continuous case) or pmf (discrete case) $f_\theta(x)$ and where $\Omega \subseteq \mathbf{R}^k$, so $\theta = (\theta_1, \dots, \theta_k)$. In the special case where X_1, \dots, X_n are i.i.d. with $X_i \sim f_\theta(x_i)$, we have $f_\theta(x) = \prod_{i=1}^n f_\theta(x_i)$. We impose the following basic regularity assumption:

Definition 13.1. The family $\{f_\theta(x) \mid \theta \in \Omega\}$ is *regular* if Ω is an open set and $f_\theta(x)$ is a smooth \equiv differentiable²² function of θ for (almost) every x . (Note that truncation families are *not* regular – recall Remark 12.2.) \square

First consider the one-parameter case where θ is real-valued. Let $T(X)$ be any real-valued statistic such that $E_\theta|T(X)| < \infty \forall \theta$. The Information Inequality gives a lower bound for $\text{Var}_\theta[T(X)]$ in terms of $E_\theta[T(X)]$ and the FIN $I_X(\theta)$, an intrinsic characteristic of the model $\{f_\theta(x)\}$: for $k = 1$,

$$(13.1) \quad I_X(\theta) \equiv E_\theta \left\{ \left[\frac{d \log f_\theta(X)}{d\theta} \right]^2 \right\} \geq 0.$$

²¹ Actually it measures the intrinsic accuracy of the particular parametrization chosen to represent the model – see Remark 13.5.

²² We will also assume the existence of higher derivatives as needed – for example, second derivatives are needed in Remark 13.4.

Theorem 13.2 (The Cramér-Rao-Frechet Lower Bound, $k = 1$).
 Assume that $I_X(\theta) > 0$. Then

$$(13.2) \quad \text{Var}_\theta[T(X)] \geq \frac{\left\{ \frac{d}{d\theta} \text{E}_\theta[T(X)] \right\}^2}{I_X(\theta)}.$$

Equality holds in (13.2) iff $\{f_\theta(x)\}$ is a 1-parameter exponential family of the form (13.8).

Proof. Set $Y = Y_\theta(X) \equiv \frac{d \log f_\theta(X)}{d\theta}$. By the Cauchy-Schwartz inequality (5.3),

$$(13.3) \quad \text{Var}_\theta(T) \geq \frac{\{\text{Cov}_\theta(T, Y)\}^2}{\text{Var}_\theta(Y)},$$

which we now show is equivalent to (13.2).

First note that $\int_{\mathcal{X}} f_\theta(x) dx = 1$ (replace \int by \sum in the discrete case), take $\frac{d}{d\theta}$ of both sides, and blithely assume we can exchange $\frac{d}{d\theta}$ and \int :

$$(13.4) \quad \begin{aligned} 0 &= \frac{d}{d\theta} \int f_\theta(x) dx \\ &= \int \frac{d}{d\theta} f_\theta(x) dx \\ &= \int \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\ &= \int \left[\frac{d \log f_\theta(x)}{d\theta} \right] f_\theta(x) dx \\ &\equiv \text{E}_\theta(Y), \end{aligned}$$

so

$$(13.5) \quad \begin{aligned} \text{Cov}_\theta(T, Y) &= \text{E}_\theta(T Y) \\ &= \int T(x) \left[\frac{d \log f_\theta(x)}{d\theta} \right] f_\theta(x) dx \\ &= \int T(x) \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\ &= \frac{d}{d\theta} \int T(x) f_\theta(x) dx \\ &= \frac{d}{d\theta} \text{E}_\theta[T(X)]. \end{aligned}$$

Finally, again apply (13.4) to obtain

$$(13.6) \quad \text{Var}_\theta(Y) = \text{E}_\theta(Y^2) \equiv \text{E}_\theta \left\{ \left[\frac{d \log f_\theta(x)}{d\theta} \right]^2 \right\} \equiv I_X(\theta).$$

Thus (13.2) follows from (13.3) - (13.6).

Next, equality holds in the Cauchy-Schwartz inequality (5.3) iff the variables X, Y are linear related (see §5.1(c)), so equality holds in (13.2) iff $T(X)$ and $\frac{d}{d\theta} \log f_\theta(X)$ are linearly related, i.e., iff

$$(13.7) \quad \frac{d}{d\theta} \log f_\theta(x) = a(\theta) + b(\theta)T(x)$$

for some constants $a(\theta), b(\theta)$ not depending on x . Equivalently,

$$\begin{aligned} \log f_\theta(x) &= \int a(\theta) d\theta + T(x) \int b(\theta) d\theta + c(x) \\ &\equiv A(\theta) + B(\theta)T(x) + c(x), \end{aligned}$$

hence

$$(13.8) \quad f_\theta(x) = e^{A(\theta)} e^{B(\theta)T(x)} e^{c(x)},$$

so $\{f_\theta(x)\}$ is a 1-parameter exponential family as asserted. \square

Corollary 13.3. *Suppose that $T(X)$ is an unbiased estimator of $\tau(\theta)$, a smooth function of θ . Then*

$$(13.9) \quad \text{Var}_\theta[T(X)] \geq \frac{[\tau'(\theta)]^2}{I_X(\theta)},$$

an intrinsic lower bound depending only on the function $\tau(\theta)$ to be estimated and on the model $\{f_\theta(x)\}$. Equality holds in (13.9) iff $\{f_\theta(x)\}$ is a 1-parameter exponential family of the form (13.8).

Proof. Apply Theorem 13.2.

Remark 13.4. *An alternative formula for $I_X(\theta)$:*

$$(13.10) \quad I_X(\theta) = -\text{E}_\theta \left[\frac{d^2 \log f_\theta(X)}{d\theta^2} \right].$$

Proof. First note that

$$(13.11) \quad \frac{d^2 \log f_\theta(X)}{d\theta^2} = \frac{\frac{d^2}{d\theta^2} f_\theta(x)}{f_\theta(x)} - \left[\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} \right]^2. \quad [\text{verify!}]$$

Thus (13.10) follows:

$$\begin{aligned} E_\theta \left[\frac{d^2 \log f_\theta(X)}{d\theta^2} \right] &= \int \left[\frac{d^2 f_\theta(x)}{d\theta^2} \right] \frac{f_\theta(x)}{f_\theta(x)} dx - \int \left[\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} \right]^2 f_\theta(x) dx \\ &= \frac{d^2}{d\theta^2} \underbrace{\int f_\theta(x) dx}_{\equiv 1} - E_\theta \left\{ \left[\frac{d \log f_\theta(X)}{d\theta} \right]^2 \right\} \\ &= 0 - I_X(\theta). \end{aligned} \quad \square$$

Remark 13.5. Let $\theta \equiv \theta(\nu)$ be a smooth function of ν , so $g_\nu(x) \equiv f_{\theta(\nu)}(x)$ is a smooth reparametrization of the model (not necessarily 1-1). Then the information number $I_g(\nu)$ for the model $\{g_\nu(x)\}$ parametrized by ν is related to $I_f(\theta)$ for the model $\{f_\theta(x)\}$ as follows:

$$\begin{aligned} I_g(\nu) &\equiv E_\nu \left\{ \left[\frac{d \log g_\nu(X)}{d\nu} \right]^2 \right\} \\ &= E_\nu \left\{ \left[\frac{d \log f_{\theta(\nu)}(X)}{d\nu} \right]^2 \right\} \\ &= E_\nu \left\{ \left[\frac{d \log f_{\theta(\nu)}(X)}{d\theta} \cdot \frac{d\theta}{d\nu} \right]^2 \right\} \quad [\text{Chain Rule}] \\ (13.12) \quad &= I_f(\theta(\nu)) \cdot \left(\frac{d\theta}{d\nu} \right)^2. \end{aligned}$$

For example, if $\theta = e^\nu$ (so $\nu = \log \theta$), then $I_g(\nu) = I_f(e^\nu) \cdot e^{2\nu}$.

Thus: the information number depends not only on the statistical model (the family of distributions) but also on the particular parametrization chosen to represent the model! \square

Remark 13.6. Suppose that $X = (X_1, \dots, X_n)$, a set of n i.i.d. observations with $X_i \sim \{f_\theta(x_i)\}$, a regular 1-parameter family. Then $f_\theta(x) = \prod_{i=1}^n f_\theta(x_i)$, so the information number for the data X is

$$\begin{aligned}
 I_X(\theta) &= \text{Var}_\theta \left[\frac{d \log f_\theta(X)}{d\theta} \right] && [\text{by (13.6)}] \\
 &= \text{Var}_\theta \left[\sum_{i=1}^n \frac{d \log f_\theta(X_i)}{d\theta} \right] \\
 &= \sum_{i=1}^n \text{Var}_\theta \left[\frac{d \log f_\theta(X_i)}{d\theta} \right] && [\text{by independence}] \\
 (13.13) \quad &= n I_{X_i}(\theta), && [\text{by identical distributions}]
 \end{aligned}$$

where I_{X_i} is the information number for a single observation. Thus the CR lower bound (13.2) or (13.9) is $O(1/n)$ for estimation in a regular family (recall Remark 12.2, p.204). \square

Example 13.7. Let $X \sim \text{Binomial}(n, \theta)$, $0 < \theta < 1$, so for $x = 0, 1, \dots, n$,

$$\begin{aligned}
 f_\theta(x) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \\
 (13.14) \quad \log f_\theta(x) &= \log \binom{n}{x} + x \log \theta + (n - x) \log(1 - \theta),
 \end{aligned}$$

$$(13.15) \quad \frac{d \log f_\theta(x)}{d\theta} = \frac{x}{\theta} - \frac{n - x}{1 - \theta},$$

$$(13.16) \quad \frac{d^2 \log f_\theta(x)}{d\theta^2} = -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2},$$

$$E_\theta \left[\frac{d^2 \log f_\theta(X)}{d\theta^2} \right] = -\frac{n\theta}{\theta^2} - \frac{n(1 - \theta)}{(1 - \theta)^2} = -\frac{n}{\theta(1 - \theta)},$$

so

$$(13.17) \quad I_X(\theta) = \frac{n}{\theta(1 - \theta)} \quad [\text{by (13.10)}].$$

Thus the CR lower bound for the variance of an unbiased estimator of $\tau(\theta) \equiv \theta$ is

$$(13.18) \quad \text{Var}_\theta[T(X)] \geq \frac{[\tau'(\theta)]^2}{I_X(\theta)} = \frac{\theta(1 - \theta)}{n}.$$

Because $\hat{\tau}(X) \equiv \frac{X}{n}$ attains this lower bound, this provides an alternative proof that the unbiased estimator $\frac{X}{n}$ is the UMVUE of θ .

Notice from (13.14) that the Binomial(n, θ) model is an exponential family with $T(x) = x$, so Corollary 13.3 already guarantees that $\hat{\tau}(X) \equiv \frac{X}{n}$ attains the CR bound for the variance of an unbiased estimator of θ .

Now suppose instead that we wish to estimate $\tau(\theta) \equiv \theta(1 - \theta)$. Since $T(X) \equiv X$ is sufficient and complete for θ , the UMVUE Supermarket (Corollary 12.1) provides the UMVUE: if we can find a function $\phi(X)$ that is unbiased for $\tau(\theta)$, then $\phi(X)$ is the UMVUE. An obvious guess leads us to consider

$$\begin{aligned} \mathbb{E}_\theta \left[\frac{X}{n} \left(1 - \frac{X}{n} \right) \right] &= \mathbb{E}_\theta \left(\frac{X}{n} \right) - \mathbb{E}_\theta \left[\left(\frac{X}{n} \right)^2 \right] \\ &= \theta - \left[\text{Var}_\theta \left(\frac{X}{n} \right) + \theta^2 \right] \\ &= \theta - \left[\frac{\theta(1 - \theta)}{n} + \theta^2 \right] \\ &= \left(\frac{n - 1}{n} \right) \theta(1 - \theta), \end{aligned}$$

so $\phi(X) \equiv \left(\frac{n}{n-1} \right) \frac{X}{n} \left(1 - \frac{X}{n} \right)$ is the UMVUE of $\theta(1 - \theta)$. However, $\phi(X)$ is *not* a linear function of $T(X) \equiv X$, so its variance will *not* attain the CR lower bound for the variance of an unbiased estimator of $\theta(1 - \theta)$. \square

Exercise 13.8. In Example 13.7, find the CR lower bound for the variance of an unbiased estimator of $\theta(1 - \theta)$. Find the variance of the UMVUE $\phi(X)$ and show that it strictly exceeds the CR bound. \square

Note: The *Bhattacharya bounds* $B_1 \leq B_2 \leq \dots$ are a sequence of sharper variance bounds, where the CR bound is B_1 and B_r depends on the derivatives $\frac{d^i \log f_\theta(X)}{d\theta^i}$, $i = 1, \dots, r$. Here the variance of the UMVUE $\phi(X)$ for $\theta(1 - \theta)$ attains the second Bhattacharya bound B_2 .

Remark 13.9. As in Remark 13.6, let X_1, \dots, X_n be i.i.d. observations with $X_i \sim \{f_\theta(x_i)\}$, a regular 1-parameter family, and let $\hat{\theta}_n$ be the MLE of θ . If $\tau(\theta)$ is a smooth function, then it follows from Theorem 14.9 that

$$(13.19) \quad \sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)] \xrightarrow{d} N_1 \left(0, \frac{[\tau'(\theta)]^2}{I_{X_i}(\theta)} \right).$$

This shows that estimators based on the MLE *asymptotically attain the CR variance bound*, so are *asymptotically efficient* \equiv *asymptotically optimal*. \square

Now consider the multi-parameter case where $\theta = (\theta_1, \dots, \theta_k)$. For any smooth function $g(\theta)$, let $\nabla_\theta g(\theta) = \left(\frac{\partial g}{\partial \theta_1}, \dots, \frac{\partial g}{\partial \theta_k} \right)' : k \times 1$. Here we let

$$(13.20) \quad I_X(\theta) \equiv \{I_{ij}(\theta) \mid i, j = 1, \dots, k\} : k \times k$$

denote the *Fisher Information Matrix (FIM)*, where

$$(13.21) \quad I_{ij}(\theta) = E_\theta \left\{ \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \right] \cdot \left[\frac{\partial \log f_\theta(X)}{\partial \theta_j} \right] \right\}.$$

Note that $I_X(\theta)$ is positive semidefinite because

$$(13.22) \quad I_X(\theta) = E_\theta \{ [\nabla_\theta \log f_\theta(X)] [\nabla_\theta \log f_\theta(X)]' \}.$$

Theorem 13.10 (The Cramér-Rao-Frechét Lower Bound, $k \geq 2$). Assume that $I_X(\theta)$ is positive definite. For any real-valued statistic $T(X)$ such that $E_\theta |T(X)|^2 < \infty \forall \theta$,

$$(13.23) \quad \text{Var}_\theta[T(X)] \geq \{ \nabla_\theta E_\theta[T(X)] \}' [I_X(\theta)]^{-1} \{ \nabla_\theta E_\theta[T(X)] \}.$$

Proof. Set $Y = Y_\theta(X) \equiv \nabla_\theta \log f_\theta(X) : k \times 1$. Then

$$(13.24) \quad \text{Cov}_\theta \left[\begin{pmatrix} T \\ Y \end{pmatrix} \right] \equiv \begin{pmatrix} \text{Var}_\theta(T) & \text{Cov}_\theta(T, Y') \\ \text{Cov}_\theta(Y, T) & \text{Cov}_\theta(Y) \end{pmatrix}$$

is psd. Furthermore, as in (13.4) on p.208,

$$(13.25) \quad E_\theta [\nabla_\theta \log f_\theta(X)] = 0 \quad [\text{verify}],$$

so from (13.22),

$$(13.26) \quad \text{Cov}_\theta(Y) = I_X(\theta) \quad [\text{compare to (13.6)}].$$

Thus $\text{Cov}_\theta(Y)$ is pd and the Cauchy-Schwartz inequality (5.3) extends as follows (recall (8.30) - (8.34)) :

$$(13.27) \quad \text{Var}_\theta(T) \geq [\text{Cov}_\theta(T, Y')][I_X(\theta)]^{-1}[\text{Cov}_\theta(Y, T)].$$

But

$$(13.28) \quad \text{Cov}_\theta(Y, T) = \nabla_\theta^* \mathbb{E}_\theta[T(X)] \quad [\text{verify; recall (13.5)}],$$

hence (13.27) is equivalent to (13.23). \square

Note: If equality holds in (13.23) then T must be a linear combination of Y (the score vector) [verify!], but this does not imply an exponential family.

Corollary 13.11. *Suppose that $T(X)$ is an unbiased estimator of $\tau(\theta)$, a smooth function of θ . Then*

$$(13.29) \quad \text{Var}_\theta[T(X)] \geq [\nabla_\theta \tau(\theta)]' [I_X(\theta)]^{-1} [\nabla_\theta \tau(\theta)].$$

Proof. Apply Theorem 13.10.

Exercise 13.12. *(An alternative formula for $I_{ij}(\theta)$ in (13.21).) Show that*

$$(13.30) \quad I_{ij}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \log f_\theta(X)}{\partial \theta_i \partial \theta_j} \right] \quad [\text{recall (13.10)}].$$

Remark 13.13. As in Remark 13.6, the FIM $I(\theta)$ for n i.i.d. observations (X_1, \dots, X_n) is just $nI_{X_i}(\theta)$. More generally, *information is additive for independent data*. That is, suppose that

$$X = (U, V), \quad U \perp\!\!\!\perp V, \quad U \sim g_\theta(u), \quad V \sim h_\theta(v).$$

Then

$$f_\theta(x) \equiv f_\theta(u, v) = g_\theta(u)h_\theta(v),$$

so

$$\frac{\partial \log f_\theta(x)}{\partial \theta_i} = \frac{\partial \log g_\theta(u)}{\partial \theta_i} + \frac{\partial \log h_\theta(v)}{\partial \theta_i},$$

hence

$$\nabla_{\theta} \log f_{\theta}(x) = \nabla_{\theta} \log g_{\theta}(u) + \nabla_{\theta} \log h_{\theta}(v).$$

Thus by (13.26) and the independence of U and V , the FIM is additive:

$$\begin{aligned} I_X(\theta) &\equiv I_{U,V}(\theta) = \text{Cov}_{\theta}[\nabla_{\theta} \log f_{\theta}(X)] \\ &= \text{Cov}_{\theta}[\nabla_{\theta} \log g_{\theta}(U)] + \text{Cov}_{\theta}[\nabla_{\theta} \log h_{\theta}(V)] \\ (13.31) \quad &\equiv I_U(\theta) + I_V(\theta). \end{aligned} \quad \square$$

13.2. The role of nuisance parameters.

Suppose that $\theta = (\theta_1, \dots, \theta_k)$ but that the quantity $\tau \equiv \tau(\theta_1)$ to be estimated depends only on θ_1 . In this context, $(\theta_2, \dots, \theta_k)$ are considered to be “nuisance parameters”.

First, if $(\theta_2, \dots, \theta_k)$ are *known* then the 1-parameter CR bound (13.9) is appropriate and here assumes the form

$$(13.32) \quad \text{Var}_{\theta}[T(X)] \geq \frac{(\frac{d\tau}{d\theta_1})^2}{I_{11}(\theta)},$$

for an unbiased estimator T of τ . However, if $(\theta_2, \dots, \theta_k)$ are *unknown* then the k -parameter CR bound (13.29) is appropriate. Here $\nabla_{\theta} \tau = \left(\frac{d\tau}{d\theta_1}, 0, \dots, 0\right)'$, so (13.29) becomes

$$\begin{aligned} \text{Var}_{\theta}[T(X)] &\geq \left(\frac{d\tau}{d\theta_1}, 0, \dots, 0\right) [I(\theta)]^{-1} \left(\frac{d\tau}{d\theta_1}, 0, \dots, 0\right)' \\ (13.33) \quad &= \frac{(\frac{d\tau}{d\theta_1})^2}{I_{11.2}(\theta)} \quad [\text{verify!}], \end{aligned}$$

where the information matrix is now partitioned as

$$I_X(\theta) = \begin{matrix} & \begin{matrix} 1 & k-1 \end{matrix} \\ \begin{matrix} 1 \\ k-1 \end{matrix} & \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix} \end{matrix}$$

and

$$(13.34) \quad I_{11.2}(\theta) \equiv I_{11}(\theta) - I_{12}(\theta)[I_{22}(\theta)]^{-1}I_{21}(\theta).$$

Clearly $I_{11.2}(\theta) \leq I_{11}(\theta)$, so

$$(13.35) \quad \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11}(\theta)} \leq \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11.2}(\theta)}.$$

Thus, for estimating any (smooth) $\tau(\theta_1)$, the presence of the unknown nuisance parameters $(\theta_2, \dots, \theta_k)$ leads to a reduction of (asymptotic) efficiency given by the ratio

$$(13.36) \quad \frac{I_{11.2}(\theta)}{I_{11}(\theta)} \leq 1.$$

No reduction of efficiency is incurred, i.e., this ratio = 1, iff $I_{11.2}(\theta) = I_{11}$, which occurs iff

$$(13.37) \quad I_{12}(\theta) \equiv \text{Cov}_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta_1}, \left(\frac{\partial \log f_\theta(X)}{\partial \theta_2}, \dots, \frac{\partial \log f_\theta(X)}{\partial \theta_k} \right) \right] = 0.$$

If (13.37) holds, we say that the parameter θ_1 is *orthogonal* to the nuisance parameters $(\theta_2, \dots, \theta_k)$. [Also see §14.6.]

Example 13.14. Let $X \equiv (X_1, \dots, X_n)$ be i.i.d. $\sim N_1(\theta_1, \theta_2)$ where $(\theta_1, \theta_2) = (\mu, \sigma^2)$. Calculate the information matrix $I_{X_i}(\theta_1, \theta_2)$ as follows:

$$(13.38) \quad \begin{aligned} f_{\theta_1, \theta_2}(x_i) &= \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x_i - \theta_1)^2}{2\theta_2}}, \\ \log f_{\theta_1, \theta_2}(x_i) &= \log \sqrt{2\pi} - \frac{1}{2} \log \theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2}, \\ \frac{\partial^2 \log f_{\theta_1, \theta_2}(x_i)}{\partial \theta_1^2} &= -\frac{1}{\theta_2}, & \frac{\partial^2 \log f_{\theta_1, \theta_2}(x_i)}{\partial \theta_1 \partial \theta_2} &= -\frac{x_i - \theta_1}{\theta_2^2}, \\ \frac{\partial^2 \log f_{\theta_1, \theta_2}(x_i)}{\partial \theta_2 \partial \theta_1} &= -\frac{x_i - \theta_1}{\theta_2^2}, & \frac{\partial^2 \log f_{\theta_1, \theta_2}(x_i)}{\partial \theta_2^2} &= \frac{1}{2\theta_2^2} - \frac{(x_i - \theta_1)^2}{\theta_2^3}. \end{aligned}$$

Thus by (13.30), the FIM is [verify!]

$$(13.39) \quad I_{X_i}(\theta_1, \theta_2) = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{pmatrix} \equiv \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}.$$

Because $I_{12} = 0$, the parameters $\theta_1 \equiv \mu$ and $\theta_2 \equiv \sigma^2$ are orthogonal, so the CR bound for each is the same whether the other is known or unknown.

For example, the CR lower bound for the variance of an unbiased estimator of $\tau(\theta_1) \equiv \theta_1 \equiv \mu$ is

$$(13.40) \quad \frac{1}{nI_{11}} \equiv \frac{\theta_2}{n} \equiv \frac{\sigma^2}{n},$$

while the CR lower bound for the variance of an unbiased estimator of $\tau(\theta_2) \equiv \theta_2 \equiv \sigma^2$ is

$$(13.41) \quad \frac{1}{nI_{22}} \equiv \frac{2\theta_2^2}{n} \equiv \frac{2\sigma^4}{n}.$$

The former is attained by the UMVUE \bar{X}_n of μ , and when $\mu \equiv \mu_0$ is known the latter is attained by the UMVUE $\frac{1}{n} \sum (X_i - \mu_0)^2 \sim \frac{\sigma^2}{n} \chi_n^2$ of σ^2 (recall Example 12.11(c) on p.200):

$$(13.42) \quad \text{Var}_{\mu_0, \sigma^2} \left[\frac{1}{n} \sum (X_i - \mu_0)^2 \right] = \frac{\sigma^4}{n^2} \cdot (2n) = \frac{2\sigma^4}{n}.$$

When μ is unknown, however, the bound (13.41) is *not* attained by the UMVUE $s_n^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$ of σ^2 (recall Example 12.12(e) on p.202):

$$(13.43) \quad \text{Var}_{\mu, \sigma^2} (s_n^2) = \frac{\sigma^4}{(n-1)^2} \cdot [2(n-1)] = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}.$$

$[(n-1)s_n^2 \equiv \sum X_i^2 - n^{-1}(\sum X_i)^2$ is *not* a linear function of $(\sum X_i, \sum X_i^2)$.]

[In fact, Charles Stein (1964) has shown that when μ is unknown, s_n^2 is *inadmissible* as an estimator of σ^2 with respect to quadratic loss. His idea is to *select and fix an arbitrary* μ_0 and use \bar{X}_n to test $\mu = \mu_0$; if this hypothesis is accepted, use $\frac{1}{n+2} \sum (X_i - \mu_0)^2$ to estimate σ^2 , otherwise use $\frac{1}{n+1} \sum (X_i - \bar{X})^2$. His estimator has smaller MSE for all (μ, σ^2) than any estimator of the form as_n^2 . Because μ_0 is arbitrary, however, I would not seriously consider such an estimator unless I had some prior knowledge that provided a reasonable choice for μ_0 , in which case the Bayesian formulation would be more appropriate anyway.] \square

13.3. Information and sufficiency.

Theorem 13.15. *Let $X \sim \{f_\theta(x)\}$, a regular family of pdfs on \mathcal{X} . Let $T \equiv T(X)$ be a statistic with induced pdf family $\{g_\theta(t)\}$ on the range \mathcal{T} . Then T provides no more information about θ than does X , that is,*

$$(13.44) \quad I_X(\theta) \geq I_T(\theta), \quad \text{i.e., } I_X(\theta) - I_T(\theta) \text{ is psd } \forall \theta,$$

where $I_X(\theta)$ and $I_T(\theta)$ are the information matrices based on X and T , respectively. Equality holds, i.e., $I_X(\theta) = I_T(\theta) \forall \theta$, iff T is sufficient.

Proof. Let $Y = Y_\theta(X) \equiv \nabla_\theta \log f_\theta(X)$ and $U = U_\theta(T) \equiv \nabla_\theta \log g_\theta(T)$. Then

$$(13.45) \quad \begin{aligned} 0 &\leq E_\theta [(Y - U)(Y - U)'] \\ &= E_\theta (YY') + E_\theta (UU') - E_\theta (YU') - E_\theta (UY') \\ &= I_X(\theta) + I_T(\theta) - E_\theta (YU') - E_\theta (UY'). \quad [\text{by (13.26)}] \end{aligned}$$

The ij -th entry of $E_\theta (YU')$ is

$$e_{ij} \equiv E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \cdot \frac{\partial \log g_\theta(T)}{\partial \theta_j} \right] = E_\theta \left\{ E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \middle| T \right] \cdot \frac{\partial \log g_\theta(T)}{\partial \theta_j} \right\}.$$

However, for all (measurable) $A \subseteq \mathcal{T}$,

$$(13.46) \quad \begin{aligned} \int_A \left[\frac{\partial \log g_\theta(t)}{\partial \theta_i} \right] g_\theta(t) dt &= \int_A \frac{\partial g_\theta(t)}{\partial \theta_i} dt \\ &= \frac{\partial}{\partial \theta_i} \int_A g_\theta(t) dt \\ &\equiv \frac{\partial}{\partial \theta_i} P_\theta [T(X) \in A] \\ &= \frac{\partial}{\partial \theta_i} P_\theta [X \in T^{-1}(A)] \\ &\equiv \frac{\partial}{\partial \theta_i} \int I_{T^{-1}(A)}(x) f_\theta(x) dx \\ &= \int I_{T^{-1}(A)}(x) \frac{\partial f_\theta(x)}{\partial \theta_i} dx \end{aligned}$$

$$\begin{aligned}
&= \int I_{T^{-1}(A)}(x) \left[\frac{\partial \log f_\theta(x)}{\partial \theta_i} \right] f_\theta(x) dx \\
&\equiv E_\theta \left\{ I_{T^{-1}(A)}(X) \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \right] \right\} \\
&= E_\theta \left\{ I_A(T) \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \right] \right\} \\
&= E_\theta \left\{ I_A(T) E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \middle| T \right] \right\} \\
&= \int_A E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \middle| T = t \right] g_\theta(t) dt,
\end{aligned}$$

so from (13.46),

$$(13.47) \quad \frac{\partial \log g_\theta(t)}{\partial \theta_i} = E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta_i} \middle| T = t \right]. \quad [\text{why?}]$$

This implies that

$$e_{ij} = E_\theta \left[\frac{\partial \log g_\theta(T)}{\partial \theta_i} \cdot \frac{\partial \log g_\theta(T)}{\partial \theta_j} \right],$$

which is the ij -th entry of $I_T(\theta)$, hence $E_\theta(YU') = I_T(\theta)$. Similarly $E_\theta(UY') = I_T(\theta)$, so (13.45) becomes $0 \leq I_X(\theta) - I_T(\theta)$, proving (13.44).

Finally, equality holds in (13.45) iff $Y = U$ a.e., that is, iff

$$(13.48) \quad \frac{\partial \log f_\theta(x)}{\partial \theta_i} = \frac{\partial \log g_\theta(t)}{\partial \theta_i} \quad \text{a.e. } (x, t) \quad \text{for } i = 1, \dots, k.$$

By integrating w.r.to θ_i this implies that for a.e. (x, t) ,

$$(13.49) \quad \log f_\theta(x) = \log g_\theta(t) + c_i(x, t, \theta_{-i}), \quad i = 1, \dots, k,$$

where $\theta_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$. Now take $\frac{\partial}{\partial \theta_j}$ for $j \neq i$ to obtain

$$(13.50) \quad \frac{\partial \log f_\theta(x)}{\partial \theta_j} = \frac{\partial \log g_\theta(t)}{\partial \theta_j} + \frac{\partial c_i(x, t, \theta_{-i})}{\partial \theta_j} \quad \text{a.e. } (x, t).$$

By comparing (13.48) with $i = j$ to (13.50) we see that $c_i(x, t, \theta_{-i})$ cannot depend on θ_j for $j \neq i$, hence $c_i(x, t, \theta_{-i}) \equiv c_i(x, t)$ cannot depend on θ at all. Thus from (13.49),

$$(13.51) \quad \log f_\theta(x) = \log g_\theta(t) + c_i(x, t), \quad i = 1, \dots, k,$$

so $c_i(x, t) \equiv c(x, t)$ cannot depend on i either. Therefore

$$f_\theta(x) = g_\theta(t) \cdot e^{c(x, t)} \equiv g_\theta(t) \cdot h(x, t),$$

so T is sufficient for θ by the Factorization Criterion. \square

13.4. A rigorous proof of the variance bound (13.2).

Our proof of the 1-parameter ($k = 1$) Information Inequality (13.2) blithely assumed that $\frac{d}{d\theta}$ and \int could be interchanged in (13.4) and (13.5). Because the latter involves the function $T \equiv T(X)$, our subsequent lower bound (13.9) for the variance of an unbiased estimator T of $\tau(\theta)$ is not applicable for all unbiased T but only for those allowing the interchange of $\frac{d}{d\theta}$ and \int . The following method of proof avoids this restriction.

We continue to assume that $X \sim \{f_\theta(x) \mid \theta \in \Omega\}$, a regular family of pdfs (or pmfs). Also assume that the *support* $S_X(\theta) \equiv \{x \mid f_\theta(x) > 0\}$ of X does not vary with θ . (Again this does not hold for truncation families.) For $\phi \neq \theta \in \Omega$ define

$$(13.52) \quad A(\phi, \theta) = \text{Var}_\theta \left[\frac{f_\phi(X)}{f_\theta(X)} - 1 \right] \leq \infty.$$

Lemma 13.16. *Suppose that $\tau(\theta) \equiv E_\theta[T(X)]$ is finite $\forall \theta$. Then*

$$(13.53) \quad \text{Var}_\theta[T(X)] \geq \sup_\phi \frac{[\tau(\phi) - \tau(\theta)]^2}{A(\phi, \theta)} \geq \limsup_{\phi \rightarrow \theta} \frac{[\tau(\phi) - \tau(\theta)]^2}{A(\phi, \theta)}.$$

Proof. Note that

$$(13.54) \quad E_\theta \left[\frac{f_\phi(X)}{f_\theta(X)} - 1 \right] \equiv \int \left[\frac{f_\phi(x)}{f_\theta(x)} - 1 \right] f_\theta(x) dx = 0,$$

so

$$(13.55) \quad \begin{aligned} \text{Cov}_\theta \left[T(X), \frac{f_\phi(X)}{f_\theta(X)} - 1 \right] &= \int T(x) \left[\frac{f_\phi(x)}{f_\theta(x)} - 1 \right] f_\theta(x) dx \\ &= \int T(x) f_\phi(x) dx - \int T(x) f_\theta(x) dx \\ &= \tau(\phi) - \tau(\theta). \end{aligned}$$

Now (13.53) follows from (13.52), (13.55), and the Cauchy-Schwartz Inequality:

$$\text{Var}_\theta[T(X)] \geq \frac{[\tau(\phi) - \tau(\theta)]^2}{A(\phi, \theta)} \quad \forall \phi. \quad \square$$

Our rigorous version of the Cramér-Rao-Frechet Theorem 13.2 requires the following additional boundedness condition on the model $\{f_\theta(x)\}$ (not on the unbiased estimator $T(X)$):

Condition B: For each $\theta \in \Omega$, \exists an open neighborhood $U(\theta) \subset \Omega$ of θ and a function $G(x; \theta) \geq 0$ such that $E_\theta[G(X; \theta)] < \infty$ and

$$(13.56) \quad \left[\frac{f_\phi(x)}{f_\theta(x)} - 1 \right]^2 \leq (\phi - \theta)^2 \cdot G(x; \theta) \quad \forall \phi \in U(\theta). \quad \square$$

This condition holds if [verify!]

$$(13.57) \quad \sup_{\phi \in U(\theta)} \left[\frac{d \log f_\phi(x)}{d\phi} \right]^2 \leq G(x; \theta) \quad \forall \phi \in U(\theta).$$

In particular, (13.57) holds in an exponential family

$$(13.58) \quad f_\theta(x) = a(\theta) \cdot e^{w(\theta)T(x)} \cdot h(x)$$

provided that $w(\theta)$ is differentiable [verify!].

Theorem 13.17. *Suppose that $\tau(\theta)$ is differentiable and that the model $\{f_\theta(x)\}$ satisfies Condition B. Then*

$$(13.59) \quad \exists \lim_{\phi \rightarrow \theta} \frac{[\tau(\phi) - \tau(\theta)]^2}{A(\phi, \theta)} = \frac{[\tau'(\theta)]^2}{I_X(\theta)}.$$

Thus by (13.53), for any unbiased estimator $T \equiv T(X)$ of $\tau(\theta)$,

$$(13.60) \quad \text{Var}_\theta[T] \geq \frac{[\tau'(\theta)]^2}{I_X(\theta)}. \quad [\text{recall (13.9)}]$$

Proof. Use Condition B to apply the Dominated Convergence Theorem:

$$\begin{aligned}
\lim_{\phi \rightarrow \theta} \frac{A(\phi, \theta)}{(\phi - \theta)^2} &= \lim_{\phi \rightarrow \theta} \int \frac{\left[\frac{f_\phi(x)}{f_\theta(x)} - 1 \right]^2}{(\phi - \theta)^2} f_\theta(x) dx \quad [\text{by (13.54)}] \\
&= \int \lim_{\phi \rightarrow \theta} \frac{\left[\frac{f_\phi(x) - f_\theta(x)}{\phi - \theta} \right]^2}{[f_\theta(x)]^2} f_\theta(x) dx \\
&= \int \left[\frac{d \log f_\theta(x)}{d\theta} \right]^2 f_\theta(x) dx \\
&\equiv I(\theta).
\end{aligned}$$

Thus (13.59) holds:

$$\lim_{\phi \rightarrow \theta} \frac{[\tau(\phi) - \tau(\theta)]^2}{A(\phi, \theta)} = \lim_{\phi \rightarrow \theta} \frac{\left[\frac{\tau(\phi) - \tau(\theta)}{\phi - \theta} \right]^2}{\frac{A(\phi, \theta)}{(\phi - \theta)^2}} = \frac{[\tau'(\theta)]^2}{I_X(\theta)}.$$

14. The Role of the Likelihood Ratio in Statistical Inference; the Method of Maximum Likelihood.

Again consider a statistical model $(\mathcal{X}, \mathcal{P} \equiv \{P_\theta \mid \theta \in \Omega\})$ where each P_θ is determined by a pdf (continuous case) or pmf (discrete case) $f_\theta(x)$ with respect to a dominating measure $d\mu(x)$ (e.g., dx). Recall our two goals of statistical inference: given the observed data $X = x$,

- use x to make inferences about the unknown P_θ that gave rise to x ;
- again using x , assess the accuracy (reliability) of the inferences.

In §11.3 we argued that all relevant information for inference about P_θ is contained in the collection of *likelihood ratios*

$$\left\{ L_{\theta_0, \theta_1}(x) \equiv \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \mid \theta_0, \theta_1 \in \Omega \right\}.$$

Properties of $L_{\theta_0, \theta_1}(x)$:

(a) $L_{\theta_0, \theta_1}(x)$ is invariant under the choice of dominating measure for the family $\{f_\theta(x) \mid \theta \in \Omega\}$ and is the only such invariant quantity: For any measurable $A \subseteq \mathcal{X}$ and any $h(x) > 0$,

$$P_\theta[X \in A] = \int_A f_\theta(x) d\mu(x) = \int_A \left[\frac{f_\theta(x)}{h(x)} \right] [h(x) d\mu(x)] \equiv \int_A \left[\frac{f_\theta(x)}{h(x)} \right] d\nu(x),$$

so the statistical model \mathcal{P} can be represented equally well by the family of pdfs $\left\{ f_\theta^*(x) \equiv \frac{f_\theta(x)}{h(x)} \mid \theta \in \Omega \right\}$ w.r.to the measure $d\nu(x)$. The likelihood ratio $L_{\theta_0, \theta_1}(x)$ is the *only* comparison criterion $\delta(f_{\theta_0}, f_{\theta_1})$ that is invariant under all choices of dominating measure, i.e., under all choices of $h(x)$. That is, if

$$\delta(f_{\theta_0}, f_{\theta_1}) = \delta(f_{\theta_0}^*, f_{\theta_1}^*) \equiv \delta\left(\frac{f_{\theta_0}(x)}{h(x)}, \frac{f_{\theta_1}(x)}{h(x)}\right) \quad \forall h(x) > 0,$$

then δ depends on f_{θ_0} and f_{θ_1} only through L_{θ_0, θ_1} (take $h(x) = f_{\theta_0}(x)$):

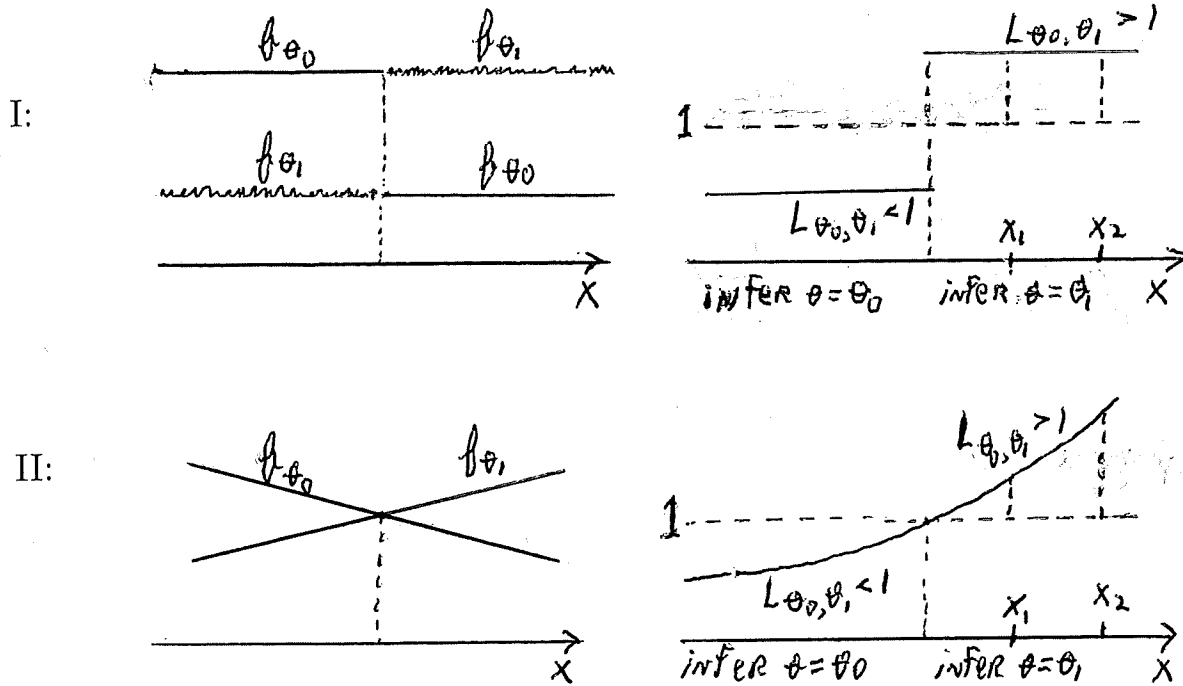
$$\delta(f_{\theta_0}, f_{\theta_1}) = \delta(1, L_{\theta_0, \theta_1}).$$

(For example, $f_{\theta_1} - f_{\theta_0}$ is not a function of L_{θ_0, θ_1} and is not invariant.)

Therefore, since inferences about θ should not depend on the choice of dominating measure, they should be based on the likelihood ratio.

(b) To test $\theta = \theta_0$ vs. $\theta = \theta_1$, the decision should be based on the likelihood ratio L_{θ_0, θ_1} using the method of maximum likelihood: If $X = x$ is observed,

$$\text{infer : } \begin{cases} \theta = \theta_1 & \text{if } L_{\theta_0, \theta_1}(x) \equiv \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > 1; \\ \theta = \theta_0 & \text{if } L_{\theta_0, \theta_1}(x) \equiv \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} < 1; \\ \text{either} & \text{if } L_{\theta_0, \theta_1}(x) \equiv \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = 1. \end{cases}$$



(c) $|L_{\theta_0, \theta_1}(x) - 1|$ indicates the reliability of our choice of θ_0 vs. θ_1 :

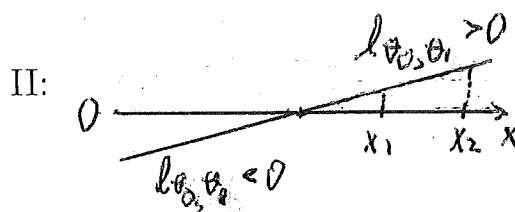
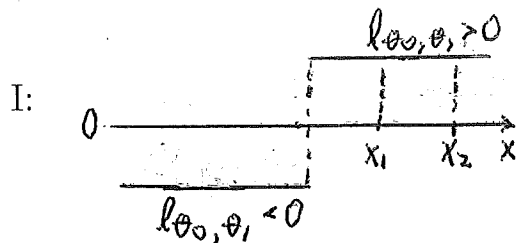
In Case I, $|L_{\theta_0, \theta_1}(x_1) - 1| = |L_{\theta_0, \theta_1}(x_2) - 1| \Rightarrow$ same reliability at x_1 and x_2 for our decision that $\theta = \theta_1$.

In Case II, $|L_{\theta_0, \theta_1}(x_1) - 1| < |L_{\theta_0, \theta_1}(x_2) - 1| \Rightarrow$ greater reliability at x_2 than at x_1 for our decision that $\theta = \theta_1$.

Thus, before observing X , $E_{\theta_0}|L_{\theta_0, \theta_1}(X) - 1|$ or $E_{\theta_0}[(L_{\theta_0, \theta_1}(X) - 1)^2]$ measures the expected accuracy \equiv reliability of inference about θ (when $\theta = \theta_0$). These are *intrinsic* measures of the precision of the model.

(d) To test θ_0 vs. θ_1 , base the decision on the **log** likelihood ratio $l_{\theta_0, \theta_1} \equiv \log L_{\theta_0, \theta_1}$ using the method of maximum likelihood: If $X = x$ is observed,

$$\text{infer : } \begin{cases} \theta = \theta_1 & \text{if } l_{\theta_0, \theta_1}(x) \equiv \log \left[\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \right] > 0; \\ \theta = \theta_0 & \text{if } l_{\theta_0, \theta_1}(x) \equiv \log \left[\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \right] < 0; \\ \text{either} & \text{if } l_{\theta_0, \theta_1}(x) \equiv \log \left[\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \right] = 0. \end{cases}$$

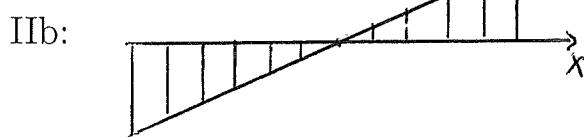
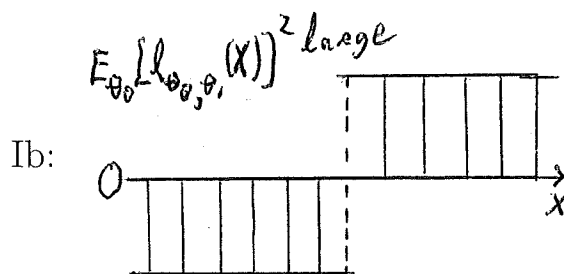
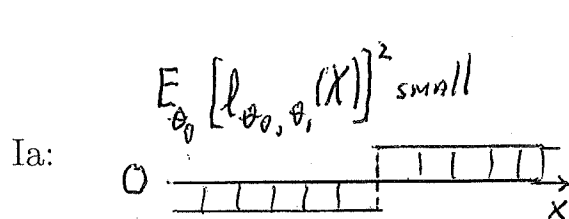


(e) $|l_{\theta_0, \theta_1}(x) - 0| \equiv |l_{\theta_0, \theta_1}(x)|$ indicates the reliability of choosing θ_0 vs. θ_1 :

In Case I, $|l_{\theta_0, \theta_1}(x_1)| = |l_{\theta_0, \theta_1}(x_2)| \Rightarrow$ same reliability at x_1 and x_2 for our decision that $\theta = \theta_1$.

In Case II, $|l_{\theta_0, \theta_1}(x_1)| < |l_{\theta_0, \theta_1}(x_2)| \Rightarrow$ greater reliability at x_2 than at x_1 for our decision that $\theta = \theta_1$.

Thus, before observing X , $E_{\theta_0} |l_{\theta_0, \theta_1}(X)|$ or $E_{\theta_0} \{[l_{\theta_0, \theta_1}(X)]^2\}$ measures the expected accuracy \equiv reliability of inference about θ (when $\theta = \theta_0$). These are *intrinsic* measures of the accuracy attainable in the model. Two different models $\{f_{\theta}\}$ and $\{g_{\theta}\}$ can be compared on the basis of $E_{\theta_0} \{[l_{\theta_0, \theta_1}(X)]^2\}$:



The decision rule for choosing θ_0 vs. θ_1 is the same in Ia and Ib, and the same in IIa and IIb. However, $E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\}$ is greater in Ib (IIb) than in Ia (IIa), reflecting the fact that the expected accuracy \equiv probability of a correct decision is greater in Ib (IIb) than in Ia (IIa).

(f) Let $\{f_\theta\}$ be a regular family with a 1-dimensional parameter θ . Then for $\theta \approx \theta_0$, $E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\}$ is approximately proportional to the Fisher Information Number $I_X(\theta_0)$:

$$(14.1) \quad E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\} \approx (\theta - \theta_0)^2 I_X(\theta_0), \quad \theta \approx \theta_0.$$

Thus $I_X(\theta_0)$, like $E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\}$, is an intrinsic measure of the sensitivity of the model $\{f_\theta(x)\}$ to local changes in θ .

Proof. Apply the first-order Taylor expansion of $\log f_\theta(x)$ about $\theta = \theta_0$:

$$\begin{aligned} E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\} &= \int [\log f_\theta(x) - \log f_{\theta_0}(x)]^2 f_{\theta_0}(x) dx \\ &\approx \int \left[(\theta - \theta_0) \frac{d \log f_\theta(x)}{d\theta} \Big|_{\theta=\theta_0} \right]^2 f_{\theta_0}(x) dx \\ &= (\theta - \theta_0)^2 I_X(\theta_0). \end{aligned}$$

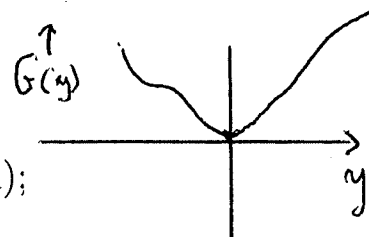
(g) Other measures of the intrinsic precision of a model $\{f_\theta\}$:

Let G be any function on $(-\infty, \infty)$ satisfying

(i) $G(y) > 0$ if $y \neq 0$, $G(0) = 0$;

(ii) $G(y) \nearrow$ as y moves away from 0 (in either direction);

(iii) $G(y)$ is smooth (\equiv differentiable).



Then $E_{\theta_0} \{G[l_{\theta_0, \theta}(X)]\}$, like $E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\}$, is a measure of the intrinsic precision of the model $\{f_\theta\}$. However, the second-order Taylor expansion of $G(y)$ about $y = 0$ is

$$\begin{aligned} G(y) &= G(0) + yG'(0) + y^2G''(0)/2 + O(y^3) \\ (14.2) \quad &= y^2G''(0)/2 + O(y^3) \quad [\text{since } G(0) = G'(0) = 0]. \end{aligned}$$

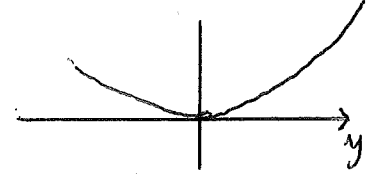
Thus by (14.1) and (14.2),

$$\begin{aligned} E_{\theta_0} \{G[l_{\theta_0, \theta}(X)]\} &\approx E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\} \cdot G''(0)/2 \\ (14.3) \quad &\approx (\theta - \theta_0)^2 I_X(\theta_0) \cdot G''(0)/2, \end{aligned}$$

so the precision measure $E_{\theta_0} \{G[l_{\theta_0, \theta}(X)]\}$ is again approximately proportional to the FIN $I_X(\theta_0)$.

An important example of a precision function G is the *Kullback-Leibler function*

$$(14.4) \quad G_{KL}(y) \equiv e^y - y - 1.$$



This satisfies (i), (ii), (iii) and in fact is convex in y . Here $G''_{KL}(0) = 1$, so by (14.2), $G_{KL}(y) \approx y^2/2$ for $y \approx 0$. Thus by (14.1),

$$\begin{aligned} E_{\theta_0} \{G_{KL}[l_{\theta_0, \theta}(X)]\} &\approx E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\} / 2 \\ (14.5) \quad &\approx (\theta - \theta_0)^2 I_X(\theta_0) / 2. \end{aligned}$$

But

$$\begin{aligned} E_{\theta_0} \{G_{KL}[l_{\theta_0, \theta}(X)]\} &= E_{\theta_0} [e^{l_{\theta_0, \theta}(X)} - l_{\theta_0, \theta}(X) - 1] \\ &= E_{\theta_0} \left[\underbrace{\frac{f_{\theta}(X)}{f_{\theta_0}(X)}}_{=1} \right] - E_{\theta_0} [l_{\theta_0, \theta}(X)] - 1 \\ (14.6) \quad &\equiv -E_{\theta_0} \left\{ \log \left[\frac{f_{\theta}(X)}{f_{\theta_0}(X)} \right] \right\} \end{aligned}$$

$$(14.7) \quad \equiv K(\theta_0, \theta),$$

the *Kullback-Leibler (KL) distance* between f_{θ_0} and f_{θ} . Thus by (14.5),

$$(14.8) \quad K(\theta_0, \theta) \approx (\theta - \theta_0)^2 I_X(\theta_0) / 2 \quad \text{for } \theta \approx \theta_0,$$

so the KL distance is also approximately proportional to the FIN $I_X(\theta_0)$.

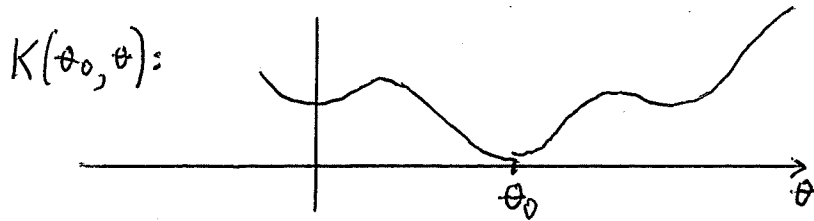
(h) $K(\theta_0, \theta_0) = 0$; $K(\theta_0, \theta) > 0$ if θ is identifiable: $\theta \neq \theta_0 \Rightarrow f_\theta \neq f_{\theta_0}$.

Proof. Because $-\log(\cdot)$ is convex, Jensen's inequality yields

$$\begin{aligned}
 K(\theta_0, \theta) &\equiv -E_{\theta_0} \left\{ \log \left[\frac{f_\theta(X)}{f_{\theta_0}(X)} \right] \right\} \\
 (14.9) \quad &\geq -\log \left\{ E_{\theta_0} \left[\frac{f_\theta(X)}{f_{\theta_0}(X)} \right] \right\} \\
 &= -\log(1) = 0.
 \end{aligned}$$

Furthermore, by identifiability, $\frac{f_\theta(X)}{f_{\theta_0}(X)}$ is a non-degenerate rv [verify!], so since $-\log(\cdot)$ is *strictly* convex, strict inequality must hold in (14.9). \square

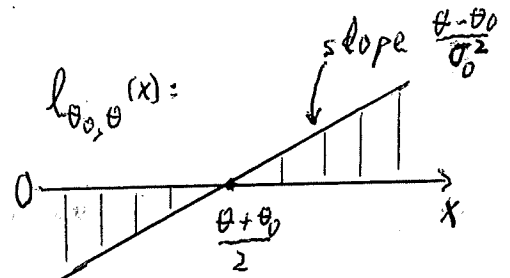
Thus, for a general regular family $\{f_\theta\}$, $K(\theta_0, \theta)$ has a unique minimum at $\theta = \theta_0$:



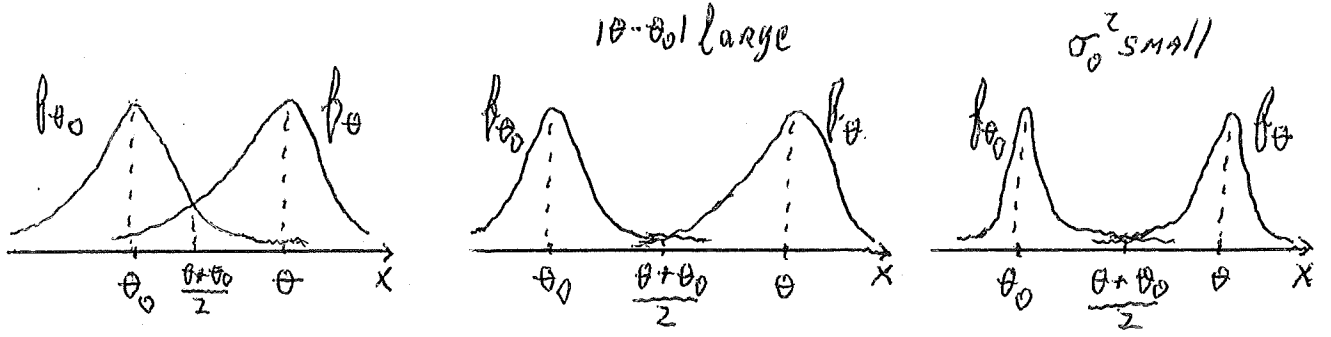
(This also holds if $\theta \equiv (\theta_1, \dots, \theta_k)$ is k -dimensional.)

Example 14.1. $f_\theta = N_1(\theta, \sigma_0^2)$ (σ_0^2 known). Here $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma_0}} e^{-\frac{(x-\theta)^2}{2\sigma_0^2}}$, so for $\theta \neq \theta_0$,

$$\begin{aligned}
 l_{\theta_0, \theta}(x) &\equiv \log \left[\frac{f_\theta(x)}{f_{\theta_0}(x)} \right] \\
 &= \frac{1}{2\sigma_0^2} [(x - \theta_0)^2 - (x - \theta)^2] \\
 (14.10) \quad &= \frac{(\theta - \theta_0)}{\sigma_0^2} \left[x - \left(\frac{\theta + \theta_0}{2} \right) \right]
 \end{aligned}$$



This is a *linear* function of x (as in the examples II above) with slope $\frac{(\theta - \theta_0)}{\sigma_0^2}$. The magnitude of this slope, and hence the “distance” between f_θ and f_{θ_0} , increases if either $|\theta - \theta_0|$ increases or σ_0^2 decreases:



This behavior is reflected in $K(\theta_0, \theta)$ and $E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\}$:

$$\begin{aligned}
 K(\theta_0, \theta) &\equiv -E_{\theta_0} \left\{ \log \left[\frac{f_{\theta}(X)}{f_{\theta_0}(X)} \right] \right\} \\
 &= -\frac{(\theta - \theta_0)}{\sigma_0^2} E_{\theta_0} \left[X - \left(\frac{\theta + \theta_0}{2} \right) \right] \\
 &= \frac{(\theta - \theta_0)^2}{2\sigma_0^2} \\
 (14.11) \quad &\equiv \frac{(\theta - \theta_0)^2}{2} I_X(\theta_0) \quad [\text{by (13.39)}]
 \end{aligned}$$

so (14.8) is exact in this example. Also from (14.10),

$$\begin{aligned}
 E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\} &= \frac{(\theta - \theta_0)^2}{\sigma_0^4} E_{\theta_0} \left\{ \left[X - \left(\frac{\theta + \theta_0}{2} \right) \right]^2 \right\} \\
 &= \frac{(\theta - \theta_0)^2}{\sigma_0^4} \left\{ \text{Var}_{\theta_0}(X) + \left(\frac{\theta - \theta_0}{2} \right)^2 \right\} \\
 &= \frac{(\theta - \theta_0)^2}{\sigma_0^4} \left\{ \sigma_0^2 + \left(\frac{\theta - \theta_0}{2} \right)^2 \right\} \\
 &\approx \frac{(\theta - \theta_0)^2}{\sigma_0^2} \quad [\text{for } \theta \approx \theta_0] \\
 (14.12) \quad &= (\theta - \theta_0)^2 \cdot I_X(\theta_0),
 \end{aligned}$$

in agreement with (14.1). This confirms that both distance criteria $K(\theta_0, \theta)$ and $E_{\theta_0} \{[l_{\theta_0, \theta}(X)]^2\}$ increase if either $|\theta - \theta_0|$ increases or σ_0^2 decreases. \square

14.1. The maximum likelihood estimator.

Definition 14.2. Suppose that $X = x$ is observed. We say that $\hat{\theta} \equiv \hat{\theta}(x)$ is the *maximum likelihood estimator (MLE)* of θ if

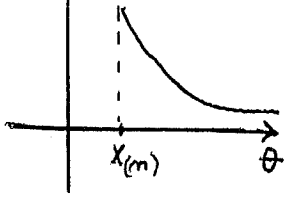
$$(14.13) \quad \exists \max_{\theta \in \Omega} f_{\theta}(x) = f_{\hat{\theta}}(x). \quad \square$$

The MLE may not exist (i.e., the maximum may not exist²³), and if $\hat{\theta}$ does exist it may not be unique. However, if it does exist then it does not depend on the choice of dominating measure for the family of distributions $\{P_{\theta}\}$, for (14.13) can be expressed equivalently in terms of the likelihood ratios:

$$(14.14) \quad \exists \hat{\theta} \text{ s.t. } L_{\theta, \hat{\theta}}(x) \geq 1 \quad \forall \theta \in \Omega.$$

Thus, if the MLE exists, it is a function of the minimal sufficient statistic $T^{**} \equiv$ the set of all likelihood ratios (recall (11.41)).

Example 14.3. $X \equiv (X_1, \dots, X_n) \sim \text{Uniform}(0, \theta]$. Then

$$(14.15) \quad \begin{aligned} f_{\theta}(x_1, \dots, x_n) &= \frac{1}{\theta^n} I_{(0, \theta]}(x_{(n)}) \cdot I_{(0, \infty)}(x_{(1)}) \\ &= \frac{1}{\theta^n} I_{[x_{(n)}, \infty)}(\theta) \cdot I_{(0, \infty)}(x_{(1)}), \end{aligned}$$


so the MLE is $\hat{\theta} = X_{(n)}$. Note that $\hat{\theta}$ differs from the UMVUE $\frac{n+1}{n} X_{(n)}$ (Example 12.14) and the minimum MSE estimator $\frac{n+2}{n+1} X_{(n)}$ (Exercise 12.10).

Example 14.4. $X \equiv (X_1, \dots, X_n) \sim N_1(\mu, \sigma^2)$, $n \geq 2$. Then

$$(14.16) \quad \begin{aligned} \log f_{\mu, \sigma^2}(x) &= -\frac{n \log(2\pi)}{2} - \frac{n \log \sigma^2}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\ &= \text{const.} - \frac{n \log \sigma^2}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 - \frac{n}{2\sigma^2} (\bar{x}_n - \mu)^2. \end{aligned}$$

²³ E.g., let $X \sim N_1(\mu, \sigma^2)$ (a single observation) with μ, σ^2 both unknown [verify!]. However, the MLE does exist for $n \geq 2$ observations – see Example 14.4.

For fixed $\sigma^2 > 0$ this is maximized at $\hat{\mu} = \bar{x}_n$ and the partial maximum is

$$\begin{aligned} \max_{-\infty < \mu < \infty} \log f_{\mu, \sigma^2}(x) &= \text{const.} - \frac{n \log \sigma^2}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \\ (14.17) \quad &\equiv c + \frac{1}{2} \left[n \log \lambda - \lambda \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right] \quad \left[\text{set } \lambda = \frac{1}{\sigma^2} \right]. \end{aligned}$$

Because [verify!] $\gamma(\lambda) \equiv [n \log \lambda - \lambda \sum (x_i - \bar{x}_n)^2]$ is a strictly concave function of $\lambda > 0$ and has its unique maximum at $\hat{\lambda} = \frac{n}{\sum (x_i - \bar{x}_n)^2}$, the partial maximum is itself maximized at $\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x}_n)^2$, so the MLE of (μ, σ^2) is given by

$$(14.18) \quad \hat{\mu} = \bar{X}_n, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Thus the MLE $\hat{\mu}$ coincides with the UMVUE \bar{X}_n of μ , but the MLE $\hat{\sigma}^2$ differs from the UMVUE $s_n^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ of σ^2 (recall Example 12.12). Thus, although MLEs are *asymptotically optimal* for regular models (Theorem 14.9), they need not be optimal²⁴ for finite samples.

Note: If $\mu \equiv \mu_0$ is *known*, then the MLE of σ^2 is $\hat{\sigma}^2 \equiv \frac{1}{n} \sum (X_i - \mu_0)^2$, which is the UMVUE of σ^2 in this case (see Example 12.11). \square

Example 14.5. Let $X \equiv (X_1, \dots, X_n) \sim \text{i.i.d. Cauchy}(\theta)$. Here

$$(14.19) \quad f_{\theta}(x) = \frac{1}{\pi^n} \prod_{i=1}^n \left[\frac{1}{1 + (x_i - \theta)^2} \right]$$

$$(14.20) \quad \log f_{\theta}(x) = -n \log \pi - \sum_{i=1}^n \log[1 + (x_i - \theta)^2]$$

$$(14.21) \quad \frac{1}{2} \frac{d \log f_{\theta}(x)}{d\theta} = \sum_{i=1}^n \frac{x_i - \theta}{1 + (x_i - \theta)^2}.$$

²⁴ This statement is somewhat misleading. By the “MLE” we really mean $(\hat{\mu}, \hat{\sigma}^2)$ considered together, not considered separately. It would be of interest to study the *joint* performance of $(\hat{\mu}, \hat{\sigma}^2)$ for some reasonable loss function $L[(\hat{\mu}, \hat{\sigma}^2), (\mu, \sigma^2)]$.

Thus there is no simple expression for the MLE $\hat{\theta}$ – it is (one of!) the roots of a polynomial equation of degree $2n - 1$. [See Examples 14.17 and 14.34.]

14.2. Strong consistency of the MLE.

Example 14.3 (Uniform) is a truncation family, which is non-regular. Example 14.4 (Normal) is an exponential family, hence regular, while Example 14.5 (Cauchy) is non-exponential but also regular. The first basic property of the MLE is *consistency*, which holds for both regular and non-regular models.

Definition 14.6. Let $X \equiv (X_1, \dots, X_n) \sim \text{i.i.d. } f_\theta(x), \theta \in \Omega$. For each n let $\tilde{\theta}^{(n)} \equiv \tilde{\theta}(X_1, \dots, X_n)$ be an estimator for θ . We say $\tilde{\theta}^{(n)}$ is *consistent in probability* \equiv *weakly consistent* if, for each $\theta_0 \in \Omega$, $\tilde{\theta}^{(n)} \xrightarrow{p} \theta_0$ as $n \rightarrow \infty$ when $\theta = \theta_0$. We say $\tilde{\theta}^{(n)}$ is *strongly consistent* if $\tilde{\theta}^{(n)} \xrightarrow{a.s.} \theta_0$ when $\theta = \theta_0$.

Theorem 14.7. (Abraham Wald, 1949.) Suppose that θ is an identifiable parameter for the family $\{f_\theta \mid \theta \in \Omega\}$. Let $X \equiv (X_1, \dots, X_n)$ consist of i.i.d. observations from $f_\theta(x_i)$, so $f_\theta(x) = \prod_{i=1}^n f_\theta(x_i)$. Suppose that $\theta = \theta_0$.

(i) If $\Omega \equiv \{\theta_1, \dots, \theta_r\}$ is finite, then the MLE $\hat{\theta}^{(n)}$ always exists, is unique for sufficiently large n , and is strongly consistent for θ_0 .

(ii) If Ω is not finite, assume that $f_\theta(x_i)$ is (upper semi-)continuous in θ and that these global dominance and identifiability conditions are satisfied:

$$(14.22) \quad \mathbb{E}_{\theta_0} \left\{ \sup_{\theta \in \Omega} \log^+ \left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \right] \right\} < \infty;$$

$$(14.23) \quad \liminf_{\theta \rightarrow \partial\Omega} K(\theta_0, \theta) > 0.$$

Then with P_{θ_0} -probability 1 the MLE $\hat{\theta}^{(n)}$ exists and is unique for sufficiently large n , and is $\hat{\theta}^{(n)}$ strongly consistent for θ_0 .

Proof. (i) Because Ω is finite, the maximum in (14.13) is always attained, so the MLE always exists. Let $K(\theta_0, \theta)$ denote the KL distance

$$(14.24) \quad K(\theta_0, \theta) = -\mathbb{E}_{\theta_0} \left\{ \log \left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \right] \right\}$$

for a single observation. By identifiability (recall (h)), $K(\theta_0, \theta) > 0$ if $\theta \neq \theta_0$, so in this case the SLLN implies that

$$(14.25) \quad P_{\theta_0} \left[\frac{1}{n} \sum_{i=1}^n \left\{ \log \left[\frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} \right] \right\} \rightarrow -K(\theta_0, \theta) < 0 \text{ as } n \rightarrow \infty \right] = 1,$$

hence

$$(14.26) \quad P_{\theta_0} \left[\prod_{i=1}^n f_{\theta}(X_i) < \prod_{i=1}^n f_{\theta_0}(X_i) \text{ for sufficiently large } n \right] = 1.$$

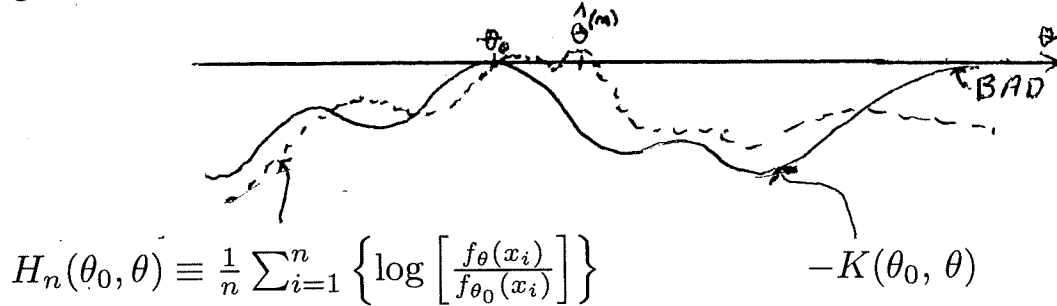
Because Ω is assumed to be finite, this implies that

$$(14.27) \quad P_{\theta_0} \left[\max_{\theta \neq \theta_0} \prod_{i=1}^n f_{\theta}(X_i) < \prod_{i=1}^n f_{\theta_0}(X_i) \text{ for sufficiently large } n \right] = 1,$$

hence, as asserted,

$$(14.28) \quad P_{\theta_0} \left[\hat{\theta}^{(n)} = \theta_0 \text{ for sufficiently large } n \right] = 1.$$

(ii) (sketch) If Ω is not finite, (14.25) and (14.26) remain valid for each individual $\theta \neq \theta_0$, but the maximum in (14.27) must be bounded by means of the global assumptions (14.22) and (14.23), as follows [see figure].



By (14.25), the random function $H_n(\theta_0, \theta) \xrightarrow{a.s.} -K(\theta_0, \theta)$ pointwise for each $\theta \in \Omega$. Assumption (14.22) lets us apply the Dominated Convergence Theorem to show that $-K(\theta_0, \theta)$ is (upper semi-)continuous in θ and that²⁵

$$(14.29) \quad H_n(\theta_0, \theta) \xrightarrow{a.s.} -K(\theta_0, \theta) \text{ uniformly in } \theta \in \Omega.$$

²⁵ Actually, it only shows that $\limsup_{n \rightarrow \infty} H_n(\theta_0, \theta) \leq -K(\theta_0, \theta)$ a.s. uniformly in $\theta \in \Omega$, but this is enough for the remainder of this argument.

Furthermore, by the continuity of $K(\theta_0, \theta)$, (14.23), and a compactness argument, for any open neighborhood $U(\theta_0)$ of θ_0 ,

$$(14.30) \quad \sup_{\theta \notin U(\theta_0)} [-K(\theta_0, \theta)] < 0,$$

so by (14.29) and the definition of $H_n(\theta_0, \theta)$,

$$(14.31) \quad \begin{aligned} 1 &= P_{\theta_0} \left[\sup_{\theta \notin U(\theta_0)} H_n(\theta_0, \theta) < 0 \text{ for sufficiently large } n \right] \\ &= P_{\theta_0} \left[\sup_{\theta \notin U(\theta_0)} \prod_{i=1}^n f_{\theta}(X_i) < \prod_{i=1}^n f_{\theta_0}(X_i) \text{ for sufficiently large } n \right]. \end{aligned}$$

But this is equivalent to

$$(14.32) \quad P_{\theta_0} \left[\hat{\theta}^{(n)} \in U(\theta_0) \text{ for sufficiently large } n \right] = 1,$$

for all open neighborhoods $U(\theta_0)$, which in turn is equivalent to $\hat{\theta}^{(n)} \xrightarrow{a.s.} \theta_0$. Thus the MLE $\hat{\theta}^{(n)}$ is strongly consistent for θ_0 , as asserted. \square

14.3. Asymptotic normality and asymptotic efficiency of the MLE.

The second basic property of the MLE is its *asymptotic normality* as $n \rightarrow \infty$, which holds when $X \equiv (X_1, \dots, X_n)$ consists of n i.i.d. observations²⁶ from a *regular* family of pdfs $\{f_{\theta}(x_i) \mid \theta \in \Omega\}$ with common support, i.e., $S_X(\theta) \equiv \{x \mid f_{\theta}(x) > 0\}$ of X does not vary with θ . Furthermore, its asymptotic variance is essentially the CR lower bound $1/I_X(\theta_0) \equiv 1/nI_{X_i}(\theta_0)$, so the MLE is said to *asymptotically efficient*.

First consider the case where θ is a single real parameter and $\Omega \equiv (a, b) \subseteq \mathbf{R}^1$ is an open interval, possibly infinite. The *likelihood function*

²⁶ The i.i.d. assumption implies that the FIN $I_X(\theta) = nI_{X_i}(\theta) \rightarrow \infty$ as $n \rightarrow \infty$. More generally, asymptotic normality and asymptotic efficiency continue to hold without independence (e.g., the observations could be serially correlated) as long as $I_X(\theta) \rightarrow \infty$.

(LF) $L_n(\theta)$ and *log likelihood function* (LLF) $l_n(\theta)$ are given by

$$(14.33) \quad L_n(\theta) \equiv L_n(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i),$$

$$(14.34) \quad l_n(\theta) \equiv l_n(\theta; x_1, \dots, x_n) = \sum_{i=1}^n \log f_\theta(x_i).$$

The MLE $\hat{\theta}^{(n)}$, if it exists, maximizes $L_n(\theta)$ or, equivalently, maximizes $l_n(\theta)$. Because $l_n(\theta)$ is smooth, we may try to find the MLE by finding the roots of the *likelihood equation* (LEQ)

$$(14.35) \quad \frac{dl_n(\theta)}{d\theta} \equiv \sum_{i=1}^n \frac{d \log f_\theta(x_i)}{d\theta} = 0.$$

The LEQ²⁷ can have no real roots, one real root, or multiple real roots in $\Omega \equiv (a, b)$, and such a root may correspond to a local maximum, local minimum, or inflection point of $l_n(\theta)$. If we solve the LEQ and find a real root $\tilde{\theta}^{(n)}$, we still must determine what kind of root it is. Even if we find it to be a *local* maximum, we cannot yet conclude that it is the MLE $\hat{\theta}^{(n)}$, which is a *global* maximum of the LLF.

Thus, determination of the MLE $\hat{\theta}^{(n)}$ requires global maximization, a harder task than finding a root $\tilde{\theta}^{(n)}$ of the LEQ. Well before Wald's 1949 strong consistency result for $\hat{\theta}^{(n)}$, Fisher and Cramér already established the existence, asymptotic normality, and asymptotic efficiency of *any weakly consistent root $\tilde{\theta}^{(n)}$ of the LEQ*. Of course, if the conditions for Wald's strong consistency theorem hold then the actual MLE $\hat{\theta}^{(n)}$ must itself be a weakly consistent root of the LEQ (since a global maximum over an open set Ω must be a local maximum), hence $\hat{\theta}^{(n)}$ itself *must be asymptotically normal and asymptotically efficient*.

We now present the results of Fisher and Cramér for a single real parameter θ . Throughout the discussion, let θ_0 denote the true value of θ .

²⁷ As in Remark 13.5, let $\theta \equiv \theta(\nu)$ be a smooth monotone function of ν , so $g_\nu(x) \equiv f_{\theta(\nu)}(x)$ is a reparametrization of the model. Since $\frac{d \log g_\nu(x_i)}{d\nu} = \frac{d \log f_\theta(x_i)}{d\theta} \cdot \frac{d\theta}{d\nu}$, if in addition $\frac{d\theta}{d\nu} \neq 0 \forall \nu$ then the roots of the LEQ (14.35) correspond exactly to the roots of the LEQ $\sum \frac{d \log g_\nu(x_i)}{d\nu} = 0$ expressed in terms of ν .

Proposition 14.8. *There exists at least one strongly consistent sequence $\{\tilde{\theta}^{(n)}\}$ of roots of the LEQ (14.35).*

Proof. For each $k = 1, 2, \dots$, it follows from (14.29)-(14.30) that there exists n_k such that $n \geq n_k \Rightarrow H_n(\theta_0, \theta_0 \pm k^{-1}) < 0$, so $l_n(\cdot)$ has a local maximum $\tilde{\theta}_{n,k}$ in $(\theta_0 \pm k^{-1})$. Because $l_n(\cdot)$ is smooth, $\tilde{\theta}_{n,k}$ must be a root of the LEQ. If we define

$$(14.36) \quad \tilde{\theta}^{(n)} = \tilde{\theta}_{n,k} \quad \text{for } n_k \leq n < n_{k+1},$$

then $\tilde{\theta}^{(n)} \rightarrow \theta_0$ as $n \rightarrow \infty$, so $\{\tilde{\theta}^{(n)}\}$ is strongly consistent. \square

Theorem 14.9 (Fisher-Cramér). *Let $\{\tilde{\theta}^{(n)}\}$ be any weakly consistent sequence of roots of the LEQ (14.35). Assume²⁸ that for $r = 1$ and $r = 2$,*

$$(14.37) \quad E_{\theta_0} \left[\left| \frac{d^r \log f_{\theta}(X_i)}{d\theta^r} \right|_{\theta=\theta_0} \right] < \infty,$$

and that for $r = 3$ \exists an open neighborhood $U(\theta_0)$ of θ_0 such that

$$(14.38) \quad E_{\theta_0} \left[\sup_{\theta \in U(\theta_0)} \left| \frac{d^3 \log f_{\theta}(X_i)}{d\theta^3} \right| \right] \equiv Q(\theta_0) < \infty.$$

If the information number $I(\theta_0) \equiv I_{X_i}(\theta_0) > 0$, then

$$(14.39) \quad \sqrt{n} \left(\tilde{\theta}^{(n)} - \theta_0 \right) \xrightarrow{d} N_1 \left[0, \frac{1}{I(\theta_0)} \right],$$

so $\{\tilde{\theta}^{(n)}\}$ is a (weakly) consistent, asymptotically normal, asymptotically efficient (CANE) sequence of estimators of θ_0 . Thus if the conditions for Wald's Theorem 14.7 holds so that the MLE sequence $\{\hat{\theta}^{(n)}\}$ is a consistent sequence of roots of the LEQ, then $\{\hat{\theta}^{(n)}\}$ is CANE for θ_0 .

²⁸ Actually, (14.39) holds under much weaker conditions: only (14.37) for the first derivative ($r = 1$) is needed (Lecam (1970) *Ann. Math. Statist.*). In fact (14.39) holds for an i.i.d. sample from the double exponential distribution, which is not fully regular. Here the MLE is the sample median, which is asymptotically normal by (10.79).

Proof. Since $\tilde{\theta}^{(n)}$ is a root of the LEQ, $\frac{dl_n(\tilde{\theta}^{(n)})}{d\theta} = 0$. Now expand $\frac{dl_n(\tilde{\theta}^{(n)})}{d\theta}$ in a 2nd-order Taylor expansion about θ_0 to obtain

$$(14.40) \quad 0 = \frac{dl_n(\tilde{\theta}^{(n)})}{d\theta} = \frac{dl_n(\theta_0)}{d\theta} + (\tilde{\theta}^{(n)} - \theta_0) \frac{d^2 l_n(\theta_0)}{d\theta^2} + \frac{(\tilde{\theta}^{(n)} - \theta_0)^2}{2} \frac{d^3 l_n(\theta_n^*)}{d\theta^3}$$

for some $\theta_n^* \in (\theta_0, \tilde{\theta}^{(n)})$. Thus

$$(14.41) \quad (\tilde{\theta}^{(n)} - \theta_0) = \frac{-\frac{dl_n(\theta_0)}{d\theta}}{\frac{d^2 l_n(\theta_0)}{d\theta^2} + \frac{(\tilde{\theta}^{(n)} - \theta_0)}{2} \frac{d^3 l_n(\theta_n^*)}{d\theta^3}},$$

so

$$(14.42) \quad \begin{aligned} \sqrt{n}(\tilde{\theta}^{(n)} - \theta_0) &= \frac{-\sqrt{n} \left[\frac{1}{n} \sum \frac{d \log f_\theta(X_i)}{d\theta} \Big|_{\theta=\theta_0} \right]}{\left[\frac{1}{n} \sum \frac{d^2 \log f_\theta(X_i)}{d\theta^2} \Big|_{\theta=\theta_0} \right] + \frac{(\tilde{\theta}^{(n)} - \theta_0)}{2} \left[\frac{1}{n} \sum \frac{d^3 \log f_\theta(X_i)}{d\theta^3} \Big|_{\theta=\theta_n^*} \right]} \\ &\equiv \frac{-\sqrt{n} \left[\frac{1}{n} \sum U_i \right]}{\left[\frac{1}{n} \sum V_i \right] + \frac{(\tilde{\theta}^{(n)} - \theta_0)}{2} \left[\frac{1}{n} \sum W_i^* \right]}. \end{aligned}$$

Since $E_{\theta_0}(U_i) = 0$ [why?], the Central Limit Theorem yields

$$(14.43) \quad -\sqrt{n} \left[\frac{1}{n} \sum U_i \right] \xrightarrow{d} -N_1[0, \text{Var}_{\theta_0}(U_i)] \equiv N_1[0, I(\theta_0)] \quad [\text{why?}],$$

while the WLLN yields

$$(14.44) \quad \frac{1}{n} \sum V_i \xrightarrow{p} E_{\theta_0}(V_i) = -I(\theta_0) \quad [\text{why?}].$$

Furthermore, since $\tilde{\theta}^{(n)} \xrightarrow{p} \theta_0$, $P_{\theta_0}[\theta_n^* \in U(\theta_0)] \rightarrow 1$ ($U(\theta_0)$ in (14.38)), so

$$P_{\theta_0} \left[\left| \frac{(\tilde{\theta}^{(n)} - \theta_0)}{2n} \sum_{i=1}^n W_i^* \right| \leq \frac{|\tilde{\theta}^{(n)} - \theta_0|}{2n} \sum_{i=1}^n \sup_{\theta \in U(\theta_0)} \left| \frac{d^3 \log f_\theta(X_i)}{d\theta^3} \right| \right] \rightarrow 1.$$

But (14.38), the WLLN, and the weak consistency of $\{\tilde{\theta}^{(n)}\}$ imply that

$$(14.45) \quad \frac{|\tilde{\theta}^{(n)} - \theta_0|}{2n} \sum_{i=1}^n \sup_{\theta \in U(\theta_0)} \left| \frac{d^3 \log f_\theta(X_i)}{d\theta^3} \right| = o_p(1) \cdot [Q(\theta_0) + o_p(1)] = o_p(1),$$

hence also

$$(14.46) \quad \frac{(\tilde{\theta}^{(n)} - \theta_0)}{2n} \sum_{i=1}^n W_i^* = o_p(1).$$

Thus (14.39) follows from (14.42) - (14.46) and Slutsky's Theorem [verify!].

Note: Assumptions (14.37) and (14.38) in the Fisher-Cramér Theorem are called the *Cramér conditions*. These are *local* conditions, since $U(\theta_0)$ is an arbitrarily small neighborhood of θ_0 . By contrast, the Wald condition (14.22) is global, since the supremum is taken over Ω .

Exercise 14.10. The asymptotic normal approximation (14.39) can be used to obtain an approximate $(1 - \alpha)$ -confidence interval $\tilde{\theta}^{(n)} \pm \frac{1}{\sqrt{nI(\theta)}} z_{\frac{\alpha}{2}}$ for θ . In general, however, this cannot be used because θ is unknown. Instead, show that under the Cramér conditions, $I(\theta)$ is continuous in θ and that for any weakly consistent estimating sequence $\bar{\theta}^{(n)}$ for θ ,

$$(14.47) \quad \tilde{\theta}^{(n)} \pm \frac{1}{\sqrt{nI(\bar{\theta}^{(n)})}} z_{\frac{\alpha}{2}} \quad \text{and}$$

is also an approximate $(1 - \alpha)$ -confidence interval for θ .²⁹ [Alternatively, one might try to find a variance-stabilizing transformation.] \square

Exercise 14.11. Let $(N_1, N_2, N_3) \sim \text{Trinomial}(n; \theta, \theta^2, 1 - \theta - \theta^2)$. Specify the allowable range of θ .

- (i) Find a minimal sufficient statistic. Is it complete?
- (ii) Specify the asymptotic distribution of $\hat{\theta}_n$, a CANE root of the LEQ.
- (iii) Now let $(N_1, N_2, N_3) \sim \text{Trinomial}(n; \theta, \theta, 1 - 2\theta)$. Repeat the above analysis for this model, and compare the asymptotic efficiencies of $\hat{\theta}_n$. \square

14.4. The possibility of multiple roots of the LEQ.

Under the Cramér conditions, there is a *unique* weakly consistent root of the LEQ (we only prove this for the case where θ is a single real parameter):

Proposition 14.12. *There exists $\epsilon \equiv \epsilon(\theta_0) > 0$ such that*

$$(14.48) \quad P_{\theta_0} [\exists \geq 2 \text{ roots of the LEQ in } (\theta_0 \pm \epsilon)] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

²⁹ Efron and Hinkley *Biometrika* (1978) recommend replacing $nI(\bar{\theta}^{(n)})$ by the observed information $-\partial^2 l_n(\bar{\theta}^{(n)}) / \partial \theta^2$.

Proof. Choose $\epsilon > 0$ small enough that $\theta_0 \pm \epsilon \in U(\theta_0)$ and $\epsilon < \frac{I(\theta_0)}{2Q(\theta_0)}$ (for definitions, see (14.38)). For all $\theta \in (\theta_0 \pm \epsilon)$, $\exists \theta_n^* \in (\theta_0, \theta)$ such that

$$\begin{aligned} \left| \frac{1}{n} \frac{d^2 l_n(\theta)}{d\theta^2} - \frac{1}{n} \frac{d^2 l_n(\theta_0)}{d\theta^2} \right| &= \frac{|\theta - \theta_0|}{n} \cdot \left| \frac{d^3 l_n(\theta_n^*)}{d\theta^3} \right| \\ &\leq \frac{\epsilon}{n} \cdot \sum_{i=1}^n \sup_{\theta \in U(\theta_0)} \left| \frac{d^3 \log f_\theta(X_i)}{d\theta^3} \right|. \end{aligned}$$

Thus by (14.44), (14.38), and the WLLN,

$$\begin{aligned} \sup_{\theta \in (\theta_0 \pm \epsilon)} \frac{1}{n} \frac{d^2 l_n(\theta)}{d\theta^2} &\leq \frac{1}{n} \frac{d^2 l_n(\theta_0)}{d\theta^2} + \frac{\epsilon}{n} \cdot \sum_{i=1}^n \sup_{\theta \in U(\theta_0)} \left| \frac{d^3 \log f_\theta(X_i)}{d\theta^3} \right| \\ &= [-I(\theta_0) + o_p(1)] + \epsilon \cdot [Q(\theta_0) + o_p(1)] \\ (14.49) \quad &\leq -\frac{I(\theta_0)}{2} + o_p(1). \end{aligned}$$

Thus

$$(14.50) \quad P_{\theta_0}[l_n(\cdot) \text{ is strictly concave on } (\theta_0 \pm \epsilon)] \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

which implies (14.48). □

By combining Propositions 14.8 and 14.12 we conclude that *under the Cramér conditions, there exists a unique weakly consistent root of the LEQ. Therefore, if the LEQ has a unique root, it must be the unique CANE root.*

If the pdf $f_\theta(x)$ is *strictly log concave* in θ (i.e., $\log f_\theta(x)$ is strictly concave), then the LLF $l_n(\theta) \equiv \sum_{i=1}^n \log f_\theta(x_i)$ is strictly concave on Ω , so the uniqueness fact (14.47) can be greatly strengthened: *for all n , the LEQ can have at most one root in the entire parameter space Ω , which therefore must be the unique CANE root.* This occurs, for example, when $\{f_\theta(x_i)\}$ is a 1-parameter exponential family:

Proposition 14.13. (i) *Any exponential family pdf $f_\theta(x) \equiv a(\theta)e^{\theta T(x)}h(x)$ is strictly log concave in its natural parameter θ .*

(ii) *If the equation*

$$(14.51) \quad E_\theta(T) = T(x)$$

has a solution $\hat{\theta} \equiv \hat{\theta}(x)$ then this solution is unique and $\hat{\theta}$ is the MLE of θ .

$$(iii) I_X(\theta) = \text{Var}_\theta[T(X)].$$

Proof. (i) First, $\int f_\theta(x) \equiv a(\theta) \int e^{\theta T(x)} h(x) dx = 1$ implies that

$$(14.52) \quad a(\theta) = \frac{1}{\int e^{\theta T} h(x) dx},$$

$$\begin{aligned} \frac{d \log a(\theta)}{d\theta} &= - \frac{\int T e^{\theta T} h(x)}{\int e^{\theta T} h(x)} \\ &\equiv - \int T \cdot a(\theta) e^{\theta T} h(x) \end{aligned}$$

$$(14.53) \quad \equiv -E_\theta(T),$$

$$\begin{aligned} \frac{d^2 \log a(\theta)}{d\theta^2} &= - \frac{\int e^{\theta T} h(x) \int T^2 e^{\theta T} h(x) - [\int T e^{\theta T} h(x)]^2}{[\int e^{\theta T} h(x)]^2} \\ &= - \int T^2 \cdot a(\theta) e^{\theta T} h(x) + \left[\int T \cdot a(\theta) e^{\theta T} h(x) \right]^2 \end{aligned}$$

$$(14.54) \quad \equiv -\text{Var}_\theta(T).$$

Thus

$$(14.55) \quad \log f_\theta(x) = \log a(\theta) + \theta T(x) + \log h(x),$$

$$(14.56) \quad \frac{d \log f_\theta(x)}{d\theta} = \frac{d \log a(\theta)}{d\theta} + T(x) \equiv T(x) - E_\theta(T),$$

$$(14.57) \quad \frac{d^2 \log f_\theta(x)}{d\theta^2} = \frac{d^2 \log a(\theta)}{d\theta^2} = -\text{Var}_\theta(T) < 0.$$

Thus the LLF is strictly log concave.

Parts (ii) and (iii) now follow from (i), (14.56), and (14.57) [verify!]. \square

Exercise 14.14. Suppose that $X \sim \text{Binomial}(n, p)$ with $0 < p < 1$ unknown. This is an exponential family with $T(X) = X$ and $\theta = \log \frac{p}{1-p}$, so Proposition 14.13(iii) implies that $I_X(\theta) = \text{Var}_\theta(X) \equiv np(1-p)$. This appears to be different than (13.17) (where $\theta \leftrightarrow p$). Explain and resolve this apparent contradiction. [See Remark 13.5.] \square

Exercise 14.15. Show that the Cramér conditions in Theorem 14.9 are satisfied for n i.i.d. observations from any 1-parameter exponential family $\{f_\theta(x_i) \equiv a(\theta)e^{\theta T(x_i)}h(x_i)\}$. [Use (14.52) - (14.57).] \square

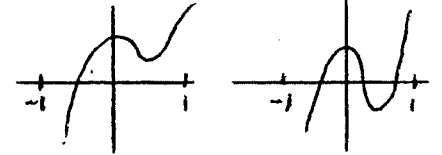
Non-exponential families must be considered on a case-by-case basis. Here are two well-known examples where the LEQ may have multiple roots. Further discussion appears in Kendall & Stuart Vol. II and Perlman (1983).

Example 14.16. Let $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be an i.i.d. sample from the bivariate normal distribution $N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$, where the means and variances are known but the correlation ρ is unknown ($\rho \in \Omega \equiv (-1, 1)$). It can be shown (see Example 14.30 Case B) that:

(a) the LEQ is a cubic equation in ρ , hence has either 1 or 3 real roots;

(b) At least 1 root occurs in $(-1, 1)$;

(c) $P_\rho[\text{exactly 1 root occurs in } (-1, 1)] \rightarrow 1$ as $n \rightarrow \infty$.



Fact (c) is stronger than the asymptotic *local* uniqueness result (14.48) and can be expressed as follows: there is an *asymptotically globally unique* root of the LEQ, which must therefore be the CANE root.³⁰ This is not entirely satisfactory in practice, however, since $P[3 \text{ roots in } (-1, 1)] > 0$ for any *finite* n and we don't know which of the 3 to choose. Several ways to treat this uncertainty are discussed in Example 14.30 Case B and Exercise 14.32. \square

Example 14.17. Let X_1, \dots, X_n be i.i.d. observations from a Cauchy pdf

$$f_\theta(x_i) = \frac{1}{\pi} \frac{1}{[1 + (x_i - \theta)^2]}$$

with $-\infty < \theta < \infty$ unknown. As seen in Example 14.5, the LEQ is a polynomial equation of degree $2n - 1$, hence may have as many as $2n - 1$ real roots. By Proposition 14.12 exactly 1 of these roots is consistent, hence is CANE by Theorem 14.9, but again we don't know which of the multiple roots to choose. Here, unlike Example 14.16, the "extraneous" roots do not cease to exist (with high probability) as $n \rightarrow \infty$. Perlman (1983)

³⁰ which exists by Proposition 14.8 and Theorem 14.9.

proved that the extraneous roots $\xrightarrow{p} \pm\infty$, then Reeds (1980) proved the remarkable result that the *number* of extraneous roots \xrightarrow{d} Poisson($\lambda = 1/\pi$). In particular, as $n \rightarrow \infty$,

$$P[\exists \geq 1 \text{ extraneous roots}] \rightarrow 1 - e^{-1/\pi} \approx 0.273. \quad \square$$

Several methods to deal with the uncertainty caused by multiple roots of the LEQ have been proposed. The first two (I, II) aim to find the CANE root, while the last two (III, IV) produce approximations to this root that are also CANE estimators:

I. Start with some consistent but possibly inefficient estimator $\check{\theta}^{(n)}$, such as the sample median, then select that root $\tilde{\theta}^{(n)}$ of the LEQ that is closest to $\check{\theta}^{(n)}$. By Proposition 14.12 and Theorem 14.9, this $\tilde{\theta}^{(n)}$ must be the unique CANE root of the LEQ.

II. Among all roots of the LEQ, select that root $\tilde{\theta}^{(n)}$ which gives the largest value of the LF. If Wald's condition for strong consistency of the MLE $\hat{\theta}^{(n)}$ is satisfied, then $\tilde{\theta}^{(n)}$ must be the unique CANE root of the LEQ.

III. Use the first Newton-Raphson approximation for the LEQ. Start with some (perhaps inefficient) estimator $\theta_{(0)}^{(n)}$ that is \sqrt{n} -consistent for θ_0 , i.e.,

$$\theta_{(0)}^{(n)} - \theta_0 = O_p(n^{-1/2}).$$

Using $\theta_{(0)}^{(n)}$ as a starting point, begin to solve the LEQ by the iterative Newton-Raphson method. Then the first iterate $\theta_{(1)}^{(n)}$ is a CANE estimator:

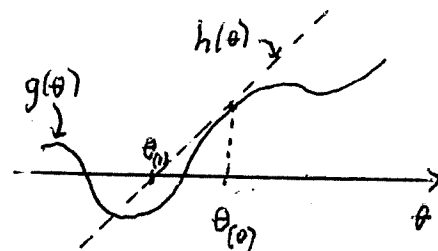
Newton's method for solving $g(\theta) = 0$ for a smooth function g .

Let $\theta_{(0)}$ be an initial guess and consider the first-order (linear) Taylor approximation for g about $\theta = \theta_{(0)}$:

$$(14.58) \quad g(\theta) \approx g(\theta_{(0)}) + (\theta - \theta_{(0)}) g'(\theta_{(0)}) \equiv h(\theta).$$

Now solve the linear approximating equation $h(\theta) = 0$:

$$\theta_{(1)} = \theta_{(0)} - \frac{g(\theta_{(0)})}{g'(\theta_{(0)})}.$$



(If this procedure is iterated, using $\theta_{(1)}$ as a new starting point to obtain $\theta_{(2)}$, etc. then $\theta_{(k)}$ converges to some root of $g(\theta) = 0$ at a geometric rate.)

Theorem 14.18. *Suppose that $f_\theta(x_i)$ satisfies the Cramér conditions and that $\theta_{(0)}^{(n)}$ is \sqrt{n} -consistent for θ_0 . Then the first Newton-Raphson iterate*

$$(14.59) \quad \theta_{(1)}^{(n)} \equiv \theta_{(0)}^{(n)} - \frac{\frac{dl_n(\theta_{(0)}^{(n)})}{d\theta}}{\frac{d^2l_n(\theta_{(0)}^{(n)})}{d\theta^2}}$$

for solving the LEQ $\frac{dl_n(\theta)}{d\theta} = 0$ is a CANE estimator, i.e.

$$(14.60) \quad \sqrt{n} \left(\theta_{(1)}^{(n)} - \theta_0 \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I(\theta_0)} \right).$$

The result (14.60) remains true if $\theta_{(1)}^{(n)}$ is replaced by

$$(14.61) \quad \dot{\theta}_{(1)}^{(n)} \equiv \theta_{(0)}^{(n)} + \frac{\frac{dl_n(\theta_{(0)}^{(n)})}{d\theta}}{nI(\theta_{(0)}^{(n)})} \quad [\text{note the “+” sign}],$$

which may be easier to compute.³¹

Proof. (sketch) In (14.59), expand $\frac{dl_n(\theta_{(0)}^{(n)})}{d\theta}$ twice about θ_0 and expand $\frac{d^2l_n(\theta_{(0)}^{(n)})}{d\theta^2}$ once about θ_0 to obtain

$$(14.62) \quad \theta_{(1)}^{(n)} \equiv \theta_{(0)}^{(n)} - \frac{\frac{dl_n(\theta_0)}{d\theta} + (\theta_{(0)}^{(n)} - \theta_0) \frac{d^2l_n(\theta_0)}{d\theta^2} + \frac{(\theta_{(0)}^{(n)} - \theta_0)^2}{2} \frac{d^3l_n(\theta_n^*)}{d\theta^3}}{\frac{d^2l_n(\theta_0)}{d\theta^2} + (\theta_{(0)}^{(n)} - \theta_0) \frac{d^3l_n(\theta_n^{**})}{d\theta^3}}$$

for some $\theta_n^*, \theta_n^{**} \in (\theta_0, \theta_{(0)}^{(n)})$. Therefore, as in the proof of Theorem 14.9,

$$\sqrt{n} \left(\theta_{(1)}^{(n)} - \theta_0 \right) = \sqrt{n} \left(\theta_{(0)}^{(n)} - \theta_0 \right) \left[1 - \frac{\frac{1}{n} \frac{d^2l_n(\theta_0)}{d\theta^2} + \frac{(\theta_{(0)}^{(n)} - \theta_0)}{2} \frac{1}{n} \frac{d^3l_n(\theta_n^*)}{d\theta^3}}{\frac{1}{n} \frac{d^2l_n(\theta_0)}{d\theta^2} + (\theta_{(0)}^{(n)} - \theta_0) \frac{1}{n} \frac{d^3l_n(\theta_n^{**})}{d\theta^3}} \right]$$

³¹ E.g. in a location family $f(x - \theta)$, $I(\theta)$ does not depend on θ (Example 14.34).

$$\begin{aligned}
& - \frac{\sqrt{n} \frac{1}{n} \frac{dl_n(\theta_0)}{d\theta}}{\frac{1}{n} \frac{d^2 l_n(\theta_0)}{d\theta^2} + (\theta_{(0)}^{(n)} - \theta_0) \frac{1}{n} \frac{d^3 l_n(\theta_{n}^{**})}{d\theta^3}} \\
& \equiv O_p(1) \cdot \left[1 - \frac{-I(\theta_0) + o_p(1)}{-I(\theta_0) + o_p(1)} \right] - \frac{\sqrt{n} \frac{1}{n} \sum U_i}{-I(\theta_0) + o_p(1)} \\
& \xrightarrow{d} N_1 \left(0, \frac{1}{I(\theta_0)} \right)
\end{aligned}$$

by (14.43) and Slutsky's theorem, where $U_i = \left. \frac{d \log f_\theta(X_i)}{d\theta} \right|_{\theta_0}$. \square

Exercise 14.19. Show that (14.60) still holds if $\theta_{(1)}^{(n)}$ is replaced by $\check{\theta}_{(1)}^{(n)}$. \square

IV. Introduce nuisance parameters that actually simplify solving the resulting system of LEQs. Because the values of these nuisance parameters are known, their estimates can be used as covariates to adjust the estimate of θ in order to produce an estimator that is fully efficient, i.e., CANE.

For example, if we begin with a location parameter model with pdf $f(x - \theta)$ (e.g., the Cauchy location family in Example 14.17), we can introduce a scale parameter σ , resulting in a location-scale family $\frac{1}{\sigma} f\left(\frac{x-\theta}{\sigma}\right)$. The resulting set of two LEQs (obtained by differentiating w.r.to both θ and σ) may actually be easier³² to solve, producing a joint CANE $(\tilde{\theta}^{(n)}, \tilde{\sigma}^{(n)})$ with asymptotic covariance matrix $[I(\theta, \sigma)]^{-1}$ given by the inverse of the information matrix – see Theorem 14.21. Because we know that $\sigma = 1$, $\tilde{\theta}^{(n)}$ can be adjusted according to its approximate regression on $\tilde{\sigma}^{(n)} - 1$. The resulting adjusted estimate³³ $\check{\theta}^{(n)}$ will be CANE for θ – see §14.6.

Thus, Method IV requires the extension of the Fisher-Cramér result to the multiparameter case, presented in §14.5. Discussion of Method IV will be continued in §14.6. (Also see Cox and Reid (1990) *Biometrika*, Fosdick and Perlman (2013) *Statist. Prob. Letters*.)

³² In the Cauchy case this system actually has a *unique* solution $(\tilde{\theta}^{(n)}, \tilde{\sigma}^{(n)})$ (cf. Copas *Biometrika* (1975)).

³³ In a location-scale family $\frac{1}{\sigma} f\left(\frac{x-\theta}{\sigma}\right)$ with $f(x) = f(-x)$, the parameters θ and σ are orthogonal, i.e., $I_{12} = 0$ (recall (13.37), so $\check{\theta}^{(n)} = \tilde{\theta}^{(n)}$ (see Example 14.35(ii)).

14.5. The multiparameter case.

Let $\theta \equiv (\theta_1, \dots, \theta_k)$ be k -dimensional with Ω an open subset of \mathbf{R}^k . The *likelihood function (LF)* $L_n(\theta)$ and *log likelihood function (LLF)* $l_n(\theta)$ are given by (14.33) and (14.34). Here we may try to find the MLE $\hat{\theta}^{(n)} \equiv (\hat{\theta}_1^{(n)}, \dots, \hat{\theta}_k^{(n)})$ by finding the roots of the *set of likelihood equations (LEQs)*

$$(14.63) \quad \frac{\partial l_n(\theta)}{\partial \theta_i} \equiv \sum_{i=1}^n \frac{\partial \log f_{\theta}(x_i)}{\partial \theta_i} = 0, \quad i = 1, \dots, k.$$

Again the LEQs can have no real roots, one real root, or multiple real roots in Ω , and such a root may correspond to a local maximum, local minimum, or saddle point of $l_n(\theta)$. If we solve the LEQ and find a real root $\tilde{\theta}^{(n)}$, we still must determine what kind of root it is, and even if we find it to be a *local* maximum, we cannot yet conclude that it is the *global* MLE $\hat{\theta}^{(n)}$.

We again denote the true value of θ by $\theta_0 \equiv (\theta_{10}, \dots, \theta_{k0})$.

Proposition 14.20. *Suppose that \exists an open neighborhood $U(\theta_0)$ of θ_0 s.t.*

$$(14.64) \quad E_{\theta_0} \left\{ \sup_{\theta \in U(\theta_0)} \log^+ \left[\frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} \right] \right\} < \infty.$$

Then there exists at least one strongly consistent sequence $\{\tilde{\theta}^{(n)}\}$ of roots of the system of LEQs (14.63).

Proof. Similar to the proof of Proposition 14.8, except that the Wald-like condition (14.64) is used to show that for all $k = 1, 2, \dots$, $\exists n_k \ni n \geq n_k \Rightarrow \sup_{\|\theta - \theta_0\| = k^{-1}} l_n(\theta) < 0$, hence $l_n(\cdot)$ has a local maximum $\tilde{\theta}_{n,k}$ in the ball $\|\theta - \theta_0\| \leq k^{-1}$. Because $l_n(\cdot)$ is smooth, $\tilde{\theta}_{n,k}$ must be a root of the LEQ. \square

Theorem 14.21 (Fisher-Cramér). *Let $\{\tilde{\theta}^{(n)}\}$ be any weakly consistent sequence of roots of the LEQs (14.63). Assume that for $r = 1, 2, 3$, the r -th order partial derivatives of $\log f_{\theta}(X_i)$ w.r.to $\theta_1, \dots, \theta_k$ satisfy boundedness conditions corresponding to (14.37) – (14.39). If the information matrix $I(\theta_0) \equiv I_{X_i}(\theta_0)$ is positive definite, then*

$$(14.65) \quad \sqrt{n} \left(\tilde{\theta}^{(n)} - \theta_0 \right) \xrightarrow{d} N_k \left(0, [I(\theta_0)]^{-1} \right),$$

so $\{\tilde{\theta}^{(n)}\}$ is a CANE sequence of estimators of θ_0 . Thus if the conditions for Wald's Theorem 14.7 holds so that the MLE sequence $\{\hat{\theta}^{(n)}\}$ is a (weakly) consistent sequence of roots of the LEQ, then $\{\hat{\theta}^{(n)}\}$ is CANE for θ_0 . \square

Under the multiparameter Cramér conditions, there is a *unique* weakly consistent root of the LEQs. Let $B(\theta_0; \epsilon)$ be the ball $\{\theta : \|\theta - \theta_0\| \leq \epsilon\}$.

Proposition 14.22. *There exists $\epsilon \equiv \epsilon(\theta_0) > 0$ such that*

$$(14.66) \quad P_{\theta_0} [\exists \geq 2 \text{ roots of the LEQ in } B(\theta_0; \epsilon)] \rightarrow 0 \text{ as } n \rightarrow \infty. \quad \square$$

Thus, under the conditions of Proposition 14.20 and Theorem 14.21, there exists a unique weakly consistent root of the LEQs. Therefore, if the LEQs have a unique root, it must be the unique CANE root. \square

Again, if the pdf $f_\theta(x)$ is *strictly log concave* in θ then the LLF $l_n(\theta) \equiv \sum_{i=1}^n \log f_\theta(x_i)$ is strictly concave on Ω (now assumed to be an *open and convex* subset of \mathbf{R}^k) and (14.66) can be greatly strengthened: *for all n , the LEQ can have at most one root in the entire parameter space Ω , which therefore must be the unique CANE root.* This occurs when $\{f_\theta(x_i)\}$ is a k -parameter exponential family with natural parameter space Ω :

Proposition 14.23. (i) *An exponential pdf $f_\theta(x) \equiv a(\theta) \exp[\sum \theta_i T_i(x)] h(x)$ is strictly log concave in its natural parameter $\theta \equiv (\theta_1, \dots, \theta_k) \in \Omega$.*

(ii) *The LEQs (14.63) are equivalent to the system of equations*

$$(14.67) \quad E_\theta(T_i) = T_i(x), \quad i = 1, \dots, k.$$

If this system has a solution $\hat{\theta} \equiv \hat{\theta}(x)$ in Ω then this solution is unique and $\hat{\theta}$ is the unique MLE of θ .

(iii) $I_X(\theta) = \text{Cov}_\theta[(T_1(X), \dots, T_k(X))']$. \square

Exercise 14.24. Prove Proposition 14.23 by establishing the following facts: for $i, j = 1, \dots, k$,

$$(14.68) \quad \frac{\partial \log a(\theta)}{\partial \theta_i} = -E_\theta(T_i);$$

$$(14.69) \quad \frac{\partial^2 \log a(\theta)}{\partial \theta_i \partial \theta_j} = -\text{Cov}_\theta(T_i, T_j);$$

$$(14.70) \quad \frac{\partial \log f_\theta(x)}{\partial \theta_i} = \frac{\partial \log a(\theta)}{\partial \theta_i} + T_i(x) \equiv T_i(x) - E_\theta(T_i),$$

$$(14.71) \quad \frac{\partial^2 \log f_\theta(x)}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 \log a(\theta)}{\partial \theta_i \partial \theta_j} = -\text{Cov}_\theta(T_i, T_j). \quad \square$$

Example 14.25. (*Multinomial*) Let $(X_1, \dots, X_k) \sim M_k(n; p_1, \dots, p_k)$, where $0 < p_i < 1$, $p_1 + \dots + p_k = 1$, so Ω is $(k-1)$ -dimensional. From (7.21) the pmf can be written as follows: for $x_1 + \dots + x_k = n$,

$$\begin{aligned} & \frac{n!}{x_1! \dots x_k!} \cdot p_1^{x_1} \dots p_k^{x_k} \\ &= \frac{n!}{x_1! \dots x_k!} \cdot e^{n \log p_k} \cdot e^{x_1 \log \frac{p_1}{p_k} + \dots + x_{k-1} \log \frac{p_{k-1}}{p_k}} \\ (14.72) \quad &= \frac{n!}{x_1! \dots x_k!} \cdot \frac{1}{(e^{\theta_1} + \dots + e^{\theta_{k-1}} + 1)^n} \cdot e^{x_1 \theta_1 + \dots + x_{k-1} \theta_{k-1}}, \end{aligned}$$

where $\theta_i = \log \frac{p_i}{p_k}$, since $p_k = \frac{p_k}{p_1 + \dots + p_k} = \frac{1}{e^{\theta_1} + \dots + e^{\theta_{k-1}} + 1}$. This is a $(k-1)$ -parameter exponential family with natural parameters $\theta \equiv (\theta_1, \dots, \theta_{k-1})$, sufficient statistics $(T_1, \dots, T_{k-1}) = (X_1, \dots, X_{k-1})$, and normalizing constant $a(\theta) = (e^{\theta_1} + \dots + e^{\theta_{k-1}} + 1)^{-n}$. Thus

$$(14.73) \quad -\log a(\theta) = n \log (e^{\theta_1} + \dots + e^{\theta_{k-1}} + 1),$$

$$(14.74) \quad E_\theta(T_i) = -\frac{\partial \log a(\theta)}{\partial \theta_i} = \frac{ne^{\theta_i}}{e^{\theta_1} + \dots + e^{\theta_{k-1}} + 1} \equiv np_i,$$

$$(14.75) \quad \text{Cov}_\theta(T_i, T_j) = -\frac{\partial^2 \log a(\theta)}{\partial \theta_i \partial \theta_j} = \quad [\text{verify}] \quad \equiv n(p_i \delta_{ij} - p_i p_j),$$

for $i, j = 1, \dots, k-1$, where $\delta_{ij} = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{if } i \neq j. \end{cases}$

By (14.67) and (14.74) the MLEs $\hat{p}_1^{(n)}, \dots, \hat{p}_{k-1}^{(n)}$ are given by

$$(14.76) \quad \hat{p}_i^{(n)} = \frac{x_i}{n}, \quad i = 1, \dots, k-1. \quad \left[\text{Also } \hat{p}_k^{(n)} = \frac{x_k}{n}. \right]$$

Thus $\hat{p}^{(n)} \equiv (\hat{p}_1^{(n)}, \dots, \hat{p}_{k-1}^{(n)})'$ is the MLE of $p = (p_1, \dots, p_{k-1})'$.

By Proposition 14.23(iii) and (14.75), the information matrix is

$$(14.77) \quad I_X(\theta) = n(D_p - pp') : (k-1) \times (k-1),$$

where $D_p = \text{diag}(p_1, \dots, p_{k-1})$ (recall (7.33)). Also, recall ((*) p.91) that $D_p - pp'$ is nonsingular since $0 < p_i < 1$ for each i . Thus the multiparameter Fisher-Cramér Theorem 14.21 implies that the MLE $\hat{\theta}^{(n)}$ of θ satisfies

$$(14.78) \quad \sqrt{n}(\hat{\theta}^{(n)} - \theta) \xrightarrow{d} N_{k-1} [0, (D_p - pp')^{-1}].$$

But from (14.74) and (14.75) we have for $i, j = 1, \dots, k-1$,

$$p_i \equiv p_i(\theta) = \frac{e^{\theta_i}}{e^{\theta_1} + \dots + e^{\theta_{k-1}} + 1},$$

$$\frac{\partial p_i}{\partial \theta_j} \equiv \frac{\partial p_i(\theta)}{\partial \theta_j} = p_i \delta_{ij} - p_i p_j,$$

so the $(k-1) \times (k-1)$ matrix of partial derivatives is given by

$$\Delta \equiv \left\{ \frac{\partial p_i}{\partial \theta_j} \mid i, j = 1, \dots, k-1 \right\} = D_p - pp'.$$

Thus, by extending the multivariate propagation-of-error formula (10.31) to the vector-valued function $p(\theta) \equiv (p_1(\theta), \dots, p_{k-1}(\theta))'$, (14.78) yields

$$\sqrt{n} [p(\hat{\theta}^{(n)}) - p(\theta)] \xrightarrow{d} N_{k-1} [0, \Delta(D_p - pp')^{-1}\Delta'],$$

or equivalently,

$$(14.79) \quad \sqrt{n} (\hat{p}^{(n)} - p) \xrightarrow{d} N_{k-1} (0, D_p - pp').$$

(Note that (14.79) is equivalent to (7.34), which we obtained directly from the multivariate Central Limit Theorem.) \square

Exercise 14.26. Apply Proposition 14.23(ii) to find the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$ when $X \equiv (X_1, \dots, X_n)$ consists of n i.i.d. observations from $N_1(\mu, \sigma^2)$ with μ and σ^2 unknown. Specify the natural parameters θ_1, θ_2 and use (14.68) to find $E_\theta(T_i)$. \square

Exercise 14.27. Show that the Cramér conditions in Theorem 14.21 are satisfied for n i.i.d. observations from any k -parameter exponential pdf $f_\theta(x_i) \equiv a(\theta) \exp[\sum_{j=1}^k \theta_j T_j(x_i)] h(x_i)$. [Use (14.68) - (14.71).] \square

Remark 14.28. In Proposition 14.23, the assumption that Ω is an open subset of \mathbf{R}^k is critical. If Ω is a curved surface in \mathbf{R}^k of dimension $d < k$ (e.g., recall the $N_1(\mu, \mu^2)$ Example 11.13(ii), where $d = 1 < 2 = k$) then the model is a curved expo family with $\theta_1, \dots, \theta_k$ expressed as functions of d actual parameters η_1, \dots, η_d . In this case the actual system of LEQs is

$$(14.80) \quad \frac{\partial l_n(\theta(\eta))}{\partial \eta_j} = 0, \quad j = 1, \dots, d,$$

and this system may have multiple roots. Drton and Richardson *Biometrika* (2004) show that this occurs in an apparently simple Gaussian *seemingly unrelated regression model*, contradicting earlier assertions in the econometrics literature. \square

The Wald and Fisher-Cramér theorems show that for sufficiently regular models with i.i.d. observations, the MLE $\hat{\theta}^{(n)} \equiv (\hat{\theta}_1^{(n)}, \dots, \hat{\theta}_k^{(n)})$ is a consistent estimator of $\theta \equiv (\theta_1, \dots, \theta_k)$ as the sample size $n \rightarrow \infty$. Here k , the number of unknown parameters, remains fixed as $n \rightarrow \infty$. However, if $k = k(n) \rightarrow \infty$ as $n \rightarrow \infty$ then $\hat{\theta}^{(n)}$ need *not* be consistent for θ . (Note that $\theta \equiv \theta_n \equiv (\theta_1, \dots, \theta_{k(n)})$ now depends on n .) Here is a classical example:

Exercise 14.28. (*Inconsistency of the MLE in the presence of many nuisance parameters: the Neyman-Scott example, Econometrica (1948).*) For $n = 1, 2, \dots$ let $X(n) = \{X_{ij} \mid i = 1, \dots, n, j = 1, \dots, r\}$ consist of nr independent rvs with $X_{ij} \sim N_1(\mu_i, \sigma^2)$. Suppose that $r \geq 2$ is fixed while $n \rightarrow \infty$. Here the parameter is the $(n+1)$ -vector $(\mu_1, \dots, \mu_n, \sigma^2)$, so $k(n) = n+1 \rightarrow \infty$. Suppose that we wish to estimate σ^2 when μ_1, \dots, μ_n are unknown nuisance parameters.

(i) Show that the (unique) MLE of $(\sigma^2, \mu_1, \dots, \mu_n)$ is $(\hat{\sigma}_n^2, \bar{X}_1, \dots, \bar{X}_n)$, where

$$(14.80) \quad \hat{\sigma}_n^2 = \frac{1}{nr} \sum_{i=1}^n \sum_{j=1}^r (X_{ij} - \bar{X}_{i.})^2, \quad \bar{X}_{i.} = \frac{1}{r} \sum_{j=1}^r X_{ij}.$$

(ii) Show that $\hat{\sigma}_n^2 \xrightarrow{p} \frac{r-1}{r} \sigma^2$ as $n \rightarrow \infty$, so the MLE-based³⁴ estimator $\hat{\sigma}_n^2$ is not consistent for σ^2 . Of course this is easily adjusted: the MLE-based estimator $\frac{r}{r-1} \hat{\sigma}_n^2$ is consistent for σ^2 . \square

14.6. The effect of nuisance parameters on asymptotic efficiency.

As in §14.5, let $\theta \equiv (\theta_1, \dots, \theta_k)'$ be k -dimensional with Ω an open subset of \mathbf{R}^k . Let $\tilde{\theta}^{(n)} \equiv (\tilde{\theta}_1^{(n)}, \dots, \tilde{\theta}_k^{(n)})'$ be any CANE sequence of estimators of θ :

$$(14.81) \quad \sqrt{n} \left(\tilde{\theta}^{(n)} - \theta \right) \xrightarrow{d} N_k \left(0, [I(\theta)]^{-1} \right) \quad [\text{recall (14.65)}],$$

where $I(\theta) \equiv I_{X_i}(\theta)$ is the information matrix for a single observation.

The results of §13.2 readily apply to determine the asymptotic efficiency of $\tilde{\theta}_1^{(n)}$ for estimating θ_1 when $\theta_2, \dots, \theta_k$ are unknown nuisance parameters:

$$(14.82) \quad \sqrt{n} \left(\tilde{\theta}_1^{(n)} - \theta_1 \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I_{11.2}(\theta)} \right),$$

where

$$(14.83) \quad I_{11.2}(\theta) \equiv I_{11}(\theta) - I_{12}(\theta)[I_{22}(\theta)]^{-1}I_{21}(\theta)$$

with $I(\theta)$ partitioned as

$$(14.84) \quad I(\theta) = \begin{matrix} & \begin{matrix} 1 & k-1 \end{matrix} \\ \begin{matrix} 1 \\ k-1 \end{matrix} & \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix} \end{matrix}.$$

³⁴ We say “MLE-based” rather than “MLE” to emphasize that *the MLE refers to an estimate of the entire underlying pdf $f_\theta(x)$ that gave rise to the observed data x* . That is, “the MLE” refers to an estimate of the entire parameter vector $\theta_n \equiv (\theta_1, \dots, \theta_{k(n)})$, not merely to an estimate of one its components.

We know from Theorem 14.9, however, that the optimal asymptotic variance for estimating θ_1 when $\theta_2, \dots, \theta_k$ are known is $1/I_{11}(\theta)$, not $1/I_{11.2}(\theta)$. Thus the presence of the unknown nuisance parameters leads to a reduction of asymptotic efficiency for estimating θ_1 given by the ratio

$$(14.85) \quad \frac{I_{11.2}(\theta)}{I_{11}(\theta)} \leq 1.$$

(No reduction of asymptotic efficiency is incurred if the parameters θ_1 and $(\theta_2, \dots, \theta_k)$ are orthogonal (see (13.37).)

Now suppose that the true values $(\theta_{20}, \dots, \theta_{k0}) \equiv \psi_0$ are *known*. Then as in Method IV in §14.4 (p.244), the estimates $(\tilde{\theta}_2^{(n)}, \dots, \tilde{\theta}_k^{(n)})' \equiv \tilde{\psi}^{(n)}$ can be used as *covariates* to adjust $\tilde{\theta}_1^{(n)}$ in order to produce an estimate that is fully efficient, i.e., CANE, for θ_1 . Here are the details:

From (14.81) and (14.84) we have the approximation

$$(14.86) \quad \begin{aligned} \tilde{\theta}^{(n)} \equiv \begin{pmatrix} \tilde{\theta}_1^{(n)} \\ \tilde{\psi}^{(n)} \end{pmatrix} &\approx N_k \left[\begin{pmatrix} \theta_1 \\ \psi_0 \end{pmatrix}, \frac{1}{n} \begin{pmatrix} I_{11}(\theta_1, \psi_0) & I_{12}(\theta_1, \psi_0) \\ I_{21}(\theta_1, \psi_0) & I_{22}(\theta_1, \psi_0) \end{pmatrix}^{-1} \right] \\ &\equiv N_k \left[\begin{pmatrix} \theta_1 \\ \psi_0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]. \end{aligned}$$

Thus the approximate conditional distribution of $\tilde{\theta}_1^{(n)} \mid \tilde{\psi}^{(n)}$ is given by

$$\tilde{\theta}_1^{(n)} \mid \tilde{\psi}^{(n)} \approx N_1 \left[\theta_1 + \Sigma_{12} \Sigma_{22}^{-1} (\tilde{\psi}^{(n)} - \psi_0), \Sigma_{11.2} \right]$$

(recall (8.72)), so conditionally and therefore unconditionally,

$$\tilde{\theta}_1^{(n)} - \Sigma_{12} \Sigma_{22}^{-1} (\tilde{\psi}^{(n)} - \psi_0) \approx N_1 [\theta_1, \Sigma_{11.2}]$$

But [verify! – (8.33) provides one approach]

$$(14.87) \quad \Sigma_{12} \Sigma_{22}^{-1} = -[I_{11}(\theta_1, \psi_0)]^{-1} I_{12}(\theta_1, \psi_0),$$

$$(14.88) \quad \Sigma_{11.2} = \frac{1}{n I_{11}(\theta_1, \psi_0)},$$

so

$$(14.89) \quad \begin{aligned} \check{\theta}_1^{(n)} &\equiv \tilde{\theta}_1^{(n)} + [I_{11}(\theta_1, \psi_0)]^{-1} I_{12}(\theta_1, \psi_0) (\tilde{\psi}^{(n)} - \psi_0) \\ &\approx N_1 \left(\theta_1, \frac{1}{n I_{11}(\theta_1, \psi_0)} \right), \end{aligned}$$

hence $\sqrt{n}(\check{\theta}_1^{(n)} - \theta_1)$ attains the optimal asymptotic variance $1/I_{11}(\theta_1, \psi_0)$.

However, $\check{\theta}_1^{(n)}$ depends on the unknown value of θ_1 , hence θ_1 must be replaced by some consistent estimate $\bar{\theta}_1^{(n)}$ of θ_1 , (e.g. $\tilde{\theta}_1^{(n)}$) (recall Exercise 14.10). Thus Method IV finally leads to the adjusted estimate

$$(14.90) \quad \check{\theta}_1^{(n)} \equiv \tilde{\theta}_1^{(n)} + [I_{11}(\bar{\theta}_1^{(n)}, \psi_0)]^{-1} I_{12}(\bar{\theta}_1^{(n)}, \psi_0)(\tilde{\psi}^{(n)} - \psi_0).$$

Exercise 14.29. Show that $\check{\theta}_1^{(n)}$ is a CANE estimator for θ_1 , i.e.,

$$(14.91) \quad \sqrt{n}(\check{\theta}_1^{(n)} - \theta_1) \xrightarrow{d} N_1\left(0, \frac{1}{I_{11}(\theta_1, \psi_0)}\right) \quad [\text{recall Exercise 14.10}].$$

Example 14.30. *Case A: Bivariate normal distribution, $\rho, \sigma_x^2, \sigma_y^2$ unknown.* Let $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ be an i.i.d. sample from

$$(14.92) \quad N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right].$$

The joint pdf $f_{\rho, \sigma_x^2, \sigma_y^2}(\dots)$ of the sample has the exponential form [verify!]

$$\begin{aligned} & \frac{1}{(2\pi)^n [\sigma_x^2 \sigma_y^2 (1 - \rho^2)]^{\frac{n}{2}}} \cdot \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[-\frac{2\rho}{\sigma_x \sigma_y} \sum x_i y_i + \frac{1}{\sigma_x^2} \sum x_i^2 + \frac{1}{\sigma_y^2} \sum y_i^2 \right] \right\} \\ &= \frac{1}{(2\pi)^n} (4\theta_2 \theta_3 - \theta_1^2)^{\frac{n}{2}} \cdot \exp \left\{ \theta_1 \sum x_i y_i + \theta_2 \sum x_i^2 + \theta_3 \sum y_i^2 \right\} \\ (14.93) \quad &= a(\theta) \cdot \exp \{ \theta_1 T_1 + \theta_2 T_2 + \theta_3 T_3 \}, \end{aligned}$$

where

$$(14.94) \quad \theta = (\theta_1, \theta_2, \theta_3),$$

$$\theta_1 = \frac{\rho}{(1 - \rho^2)\sigma_x \sigma_y}, \quad \theta_2 = -\frac{1}{2(1 - \rho^2)\sigma_x^2}, \quad \theta_3 = -\frac{1}{2(1 - \rho^2)\sigma_y^2},$$

$$(14.95) \quad a(\theta) = \frac{1}{(2\pi)^n} \cdot (4\theta_2 \theta_3 - \theta_1^2)^{\frac{n}{2}},$$

$$(14.96) \quad T_1 = \sum x_i y_i, \quad T_2 = \sum x_i^2, \quad T_3 = \sum y_i^2.$$

By Proposition 14.23(ii) and (14.67)-(14.68) (cf. pp.246-7), the MLE $\hat{\theta}^{(n)} \equiv (\hat{\theta}_1^{(n)}, \hat{\theta}_2^{(n)}, \hat{\theta}_3^{(n)})$ is the unique solution to the system

$$\begin{aligned} T_1 = E_{\theta}(T_1) &= -\frac{\partial \log a(\theta)}{\partial \theta_1} = \frac{n\theta_1}{4\theta_2\theta_3 - \theta_1^2}, \\ T_2 = E_{\theta}(T_2) &= -\frac{\partial \log a(\theta)}{\partial \theta_2} = -\frac{2n\theta_3}{4\theta_2\theta_3 - \theta_1^2}, \\ T_3 = E_{\theta}(T_3) &= -\frac{\partial \log a(\theta)}{\partial \theta_3} = -\frac{2n\theta_2}{4\theta_2\theta_3 - \theta_1^2}. \end{aligned}$$

More directly, $E_{\theta}(T_1) = n\rho\sigma_x\sigma_y$, $E_{\theta}(T_2) = n\sigma_x^2$, $E_{\theta}(T_3) = n\sigma_y^2$, so the MLE of $(\rho, \sigma_x^2, \sigma_y^2)$ is obtained from the usual moment equations:

$$(14.97) \quad \hat{\rho}^{(n)} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}, \quad \hat{\sigma}_x^{2(n)} = \frac{\sum x_i^2}{n}, \quad \hat{\sigma}_y^{2(n)} = \frac{\sum y_i^2}{n}.$$

By Proposition 14.23(iii) and (14.71), the information matrix $I(\theta)$ for the parameter $\theta \equiv (\theta_1, \theta_2, \theta_3)$ based on a single observation (X_i, Y_i) can be obtained from the partial derivatives $\frac{\partial^2 \log a(\theta)}{\partial \theta_i \partial \theta_j}$ with $n = 1$, then the asymptotic normal distribution of $\hat{\theta}^{(n)}$ is given by (14.65). However, we are more interested in determining the asymptotic distribution of the MLE $(\hat{\rho}^{(n)}, \hat{\sigma}_x^{2(n)}, \hat{\sigma}_y^{2(n)})$. This can be done via the multivariate propagation-of-error method (recall (10.31)), since $\hat{\rho}^{(n)}$, $\hat{\sigma}_x^{2(n)}$, and $\hat{\sigma}_y^{2(n)}$ are smooth functions of $\hat{\theta}^{(n)} \equiv (\hat{\theta}_1^{(n)}, \hat{\theta}_2^{(n)}, \hat{\theta}_3^{(n)})$, found by inverting (14.94). Alternatively, and more easily in this case, we can apply (14.65) directly to $(\hat{\rho}^{(n)}, \hat{\sigma}_x^{2(n)}, \hat{\sigma}_y^{2(n)})$ to obtain

$$(14.98) \quad \sqrt{n} \begin{pmatrix} \hat{\rho}^{(n)} - \rho \\ \hat{\sigma}_x^{2(n)} - \sigma_x^2 \\ \hat{\sigma}_y^{2(n)} - \sigma_y^2 \end{pmatrix} \xrightarrow{d} N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, [I(\rho, \sigma_x^2, \sigma_y^2)]^{-1} \right],$$

where $I(\rho, \sigma_x^2, \sigma_y^2)$ is the information matrix for the parameter $(\rho, \sigma_x^2, \sigma_y^2)$ based on one observation. To find $I(\rho, \sigma_x^2, \sigma_y^2)$, apply (13.30) on p.214 to $\log f_{\rho, \sigma_x^2, \sigma_y^2}(\cdots)$ with $n = 1$, where the partial derivatives are taken w.r.to ρ

and σ_x^2, σ_y^2 (not σ_x, σ_y). After some calculation (Exercise 14.31), we obtain

$$(14.99) \quad I(\rho, \sigma_x^2, \sigma_y^2) = \frac{1}{1 - \rho^2} \begin{pmatrix} \frac{1+\rho^2}{1-\rho^2} & \frac{-\rho}{2\sigma_x^2} & \frac{-\rho}{2\sigma_y^2} \\ \frac{-\rho}{2\sigma_x^2} & \frac{2-\rho^2}{4\sigma_x^4} & \frac{-\rho^2}{4\sigma_x^2\sigma_y^2} \\ \frac{-\rho}{2\sigma_y^2} & \frac{-\rho^2}{4\sigma_x^2\sigma_y^2} & \frac{2-\rho^2}{4\sigma_y^4} \end{pmatrix} \equiv \frac{1}{2} \begin{pmatrix} 1 & 2 \\ I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}.$$

It follows from (14.99) that $I_{12} \neq 0$ if $\rho \neq 0$, so the parameters ρ and (σ_x^2, σ_y^2) are not orthogonal unless $X_i \perp Y_i$. \square

Exercise 14.31. Verify (14.99), then express the efficiency ratio $\frac{I_{11,2}}{I_{11}}$ in terms of $\rho, \sigma_x^2, \sigma_y^2$. What is the minimum value of this ratio? [0.5] \square

Example 14.30. Case B: ρ unknown, $\sigma_x^2 = \sigma_y^2 = 1$ known. This case was introduced in Example 14.16. The joint pdf of $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ is

$$f_{\rho,1,1}(\cdots) = \frac{1}{[2\pi(1 - \rho^2)]^{\frac{n}{2}}} \cdot \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[-2\rho \sum x_i y_i + \left(\sum x_i^2 + \sum y_i^2 \right) \right] \right\}.$$

This is a curved exponential family with 1-dimensional parameter ρ but 2-dimensional minimal sufficient statistic [verify!]

$$(14.100) \quad (S_1, S_2) \equiv \left[\sum x_i y_i, \left(\sum x_i^2 + \sum y_i^2 \right) \right].$$

Therefore

$$(14.101) \quad \log f_{\rho,1,1}(\cdots) = c - \frac{n}{2} \log(1 - \rho^2) - \frac{S_2 - 2\rho S_1}{2(1 - \rho^2)},$$

$$(14.102) \quad \frac{\partial \log f_{\rho,1,1}(\cdots)}{\partial \rho} = \frac{n\rho}{1 - \rho^2} + \frac{S_1}{1 - \rho^2} - \frac{\rho(S_2 - 2\rho S_1)}{(1 - \rho^2)^2},$$

so the LEQ is equivalent to the cubic equation [verify!]

$$(14.103) \quad h(\rho) \equiv n\rho(1 - \rho^2) + S_1(1 + \rho^2) - S_2\rho = 0.$$

Thus the LEQ has either 1 real root or 3 real roots. Note that

$$\begin{aligned} h(-1) &= 2S_1 + S_2 = \sum (x_i + y_i)^2 > 0, \\ h(1) &= 2S_1 - S_2 = -\sum (x_i - y_i)^2 < 0, \end{aligned}$$

so the LEQ has at least 1 real root in $(-1, 1)$. By Theorem 14.9 we know that there exists an asymptotically unique CANE root $\tilde{\rho}^{(n)}$, i.e.,

$$(14.104) \quad \sqrt{n} \left(\tilde{\rho}^{(n)} - \rho \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I_{11}} \right) \equiv N_1 \left(0, \frac{(1 - \rho^2)^2}{1 + \rho^2} \right),$$

where I_{11} is obtained from (14.99). However, $P_\rho[3 \text{ roots in } (-1, 1)] > 0$ for any finite n (see Exercise 14.32), so we now illustrate Methods I-IV for dealing with this uncertainty.

I. The standard estimator $\hat{\rho}^{(n)}$ in (14.97) is consistent for ρ , so select that root $\tilde{\rho}^{(n)}$ of the LEQ that is closest to $\hat{\rho}^{(n)}$. By Proposition 14.12 and Theorem 14.9, this $\tilde{\rho}^{(n)}$ must be the unique CANE root of the LEQ.

II. Select that root $\tilde{\rho}^{(n)}$ which gives the largest value of the LLF (14.101). By Wald's Theorem 14.7, $\tilde{\rho}^{(n)}$ must be the unique CANE root of the LEQ,

III. Use the standard estimator $\hat{\rho}^{(n)}$ in (14.97) as the starting value in the Newton-Raphson algorithm for finding a root of the LEQ $h(\rho) = 0$ in (14.103). Since $\hat{\rho}^{(n)}$ is in fact \sqrt{n} -consistent, Theorem 14.18 implies that the first iterate $\rho_{(1)}^{(n)}$ (or $\dot{\rho}_{(1)}^{(n)}$, cf. (14.61), p.243) is a CANE estimator of ρ .

IV. Lastly, we can apply the asymptotic covariate-adjustment method in (14.81)–(14.91) to obtain a CANE estimator $\check{\rho}^{(n)}$ for ρ . Begin with (14.97) in the role of (14.81) and (14.99) in the role of (14.84) (cf. p.250), then apply (14.90)–(14.91) (p.252) with the correspondences

$$\begin{aligned} \theta &\leftrightarrow (\rho, \sigma_x^2, \sigma_y^2), \quad I(\theta) \equiv I(\rho, \sigma_x^2, \sigma_y^2), \\ \tilde{\theta}_1^{(n)} &\leftrightarrow \hat{\rho}^{(n)}, \quad \tilde{\psi}^{(n)} \leftrightarrow (\hat{\sigma}_x^{2(n)}, \hat{\sigma}_y^{2(n)}), \\ \theta_1 &\leftrightarrow \rho, \quad \psi_0 \leftrightarrow (\sigma_{x0}^2, \sigma_{y0}^2) \equiv (1, 1), \\ I_{11}(\theta_1, \psi_0) &\leftrightarrow \frac{1 + \rho^2}{(1 - \rho^2)^2}, \quad I_{12}(\theta_1, \psi_0) \leftrightarrow \left(\frac{-\rho}{2(1 - \rho^2)}, \frac{-\rho}{2(1 - \rho^2)} \right), \end{aligned}$$

thereby obtaining from (14.90) the adjusted estimator [verify!]

$$(14.105) \quad \check{\rho}^{(n)} = \hat{\rho}^{(n)} \left\{ 1 - \frac{[1 - (\hat{\rho}^{(n)})^2]}{[1 + (\hat{\rho}^{(n)})^2]} \left(\frac{\hat{\sigma}_x^{2(n)} + \hat{\sigma}_y^{2(n)}}{2} - 1 \right) \right\}.$$

From (14.91), $\check{\rho}^{(n)}$ is a CANE estimator for ρ :

$$(14.106) \quad \sqrt{n} \left(\check{\rho}^{(n)} - \rho \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I_{11}} \right) = N_1 \left(0, \frac{(1 - \rho^2)^2}{1 + \rho^2} \right).$$

(Compare to (14.104).) By contrast, from (14.82) the unadjusted estimator $\hat{\rho}^{(n)}$ satisfies [verify!]

$$(14.107) \quad \sqrt{n} \left(\hat{\rho}^{(n)} - \rho \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I_{11.2}} \right) = N_1 \left(0, (1 - \rho^2)^2 \right).$$

(recall Exercise 10.7), so $\hat{\rho}^{(n)}$ is inefficient unless $\rho = 0$. □

Exercise 14.32. (i) Verify that $P_\rho[3 \text{ roots in } (-1, 1)] > 0$ for all finite n .

(ii) Verify that $P_\rho[\text{exactly 1 root occurs in } (-1, 1)] \rightarrow 1$ as $n \rightarrow \infty$.

(iii) Use both (14.59) and (14.61) (cf. p.243) to provide explicit formulas for the first iterates $\rho_{(1)}^{(n)}$ and $\check{\rho}_{(1)}^{(n)}$ of the Newton-Raphson algorithm as approximate solutions of the cubic LEQ (14.103).

(iv) Verify (14.105) and (14.107).

(v) Verify that (S_1, S_2) in (14.100) is minimal sufficient. Show that T_2 and T_3 (given in (14.96) on p.252) are each ancillary, but $S_2 \equiv T_2 + T_3$ is *not* ancillary:

$$(14.108) \quad S_2 \equiv T_2 + T_3 \sim (1 + \rho)\chi_n^2 + (1 - \rho)\chi_n^2,$$

where the two χ^2 variates are independent. [*Explanation:* although $T_2 \sim \chi_n^2$ and $T_3 \sim \chi_n^2$, they are not independent so the joint distribution of (T_2, T_3) is not determined by the marginal distributions of T_2, T_3 .]

(vi) Show that (S_1, S_2) is not complete. [*Hint:* use (14.108); or use (14.110) to show that $S_1 \equiv T_1$ is not independent of the ancillary statistic T_2 .]

(vii) Show that

$$(14.109) \quad \frac{S_2 + 2S_1}{S_2 - 2S_1} \sim \frac{1 + \rho}{1 - \rho} F_{n,n}.$$

Use this to obtain an exact $(1 - \alpha)$ -confidence interval for ρ based on (S_1, S_2) .

(viii)* Let (T_1, T_2, T_3) be the complete sufficient statistic for $(\rho, \sigma_x^2, \sigma_y^2)$ in Case A. Assume that Case B holds. Show that

$$(14.110) \quad \begin{aligned} (T_1, T_2) &\perp\!\!\!\perp T_3 - T_1^2 T_2^{-1}, \\ T_1 \mid T_2 &\sim N_1(\rho T_2, (1 - \rho^2) T_2), \quad T_2 \sim \sigma_y^2 \chi_n^2, \\ T_3 - T_1^2 T_2^{-1} &\sim (1 - \rho^2) \chi_{n-1}^2. \end{aligned}$$

Conclude that

$$(14.111) \quad \frac{\frac{T_1}{\sqrt{T_2}} - \rho \sqrt{T_2}}{\sqrt{\frac{T_3 - T_1^2 T_2^{-1}}{n-1}}} \equiv \frac{\hat{\rho}^{(n)} - \rho \sqrt{\frac{T_2}{T_3}}}{\sqrt{\frac{1 - (\hat{\rho}^{(n)})^2}{n-1}}} \sim t_{n-1},$$

so

$$(14.112) \quad \sqrt{\frac{T_3}{T_2}} \left[\hat{\rho}^{(n)} \pm \sqrt{\frac{1 - (\hat{\rho}^{(n)})^2}{n-1}} \cdot t_{n-1; \alpha/2} \right]$$

is an *exact* $(1 - \alpha)$ -confidence interval for ρ . (Note that it is *not* a function of the minimal sufficient statistic (S_1, S_2) .)

(ix)** From (14.106), an approximate $(1 - \alpha)$ -confidence interval in Case B is given by

$$(14.113) \quad \check{\rho}^{(n)} \pm \frac{1}{\sqrt{n}} \frac{[1 - (\check{\rho}^{(n)})^2]}{\sqrt{1 + (\check{\rho}^{(n)})^2}} z_{\alpha/2}.$$

Compare the accuracies \equiv widths of the confidence intervals for ρ obtained via (14.109), (14.112), and (14.113) (cf. Fosdick and Perlman (2014) *Comm. Statist. Simul. Comp.*). \square

Exercise 14.33. Recall that the MLE of ρ in Case A: σ_x^2, σ_y^2 unknown, is

$$\hat{\rho}^{(n)} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}.$$

In Case B: $\sigma_x^2 = \sigma_y^2 = 1$ known, this suggests the estimator

$$\check{\rho}^{(n)} = \frac{1}{n} \sum x_i y_i.$$

Find the asymptotic distribution of $\sqrt{n}(\check{\rho}^{(n)} - \rho)$. Is $\check{\rho}^{(n)}$ asymptotically efficient? Why might your answer be expected *a priori*? [Consider the range of $\check{\rho}^{(n)}$ and also consider minimal sufficiency.] \square

Example 14.34. (*The general location-scale family*) Let X_1, \dots, X_n be i.i.d. rvs from the pdf $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ on $(-\infty, \infty)$, where $-\infty < \mu < \infty$, $0 < \sigma < \infty$. Assume that f is smooth, strictly positive on $(-\infty, \infty)$, and known.³⁵ Our goal is to estimate μ .

Case A: σ is unknown.

The log likelihood function and two LEQs are obtained as follows:

$$(14.114) \quad l_n(\mu, \sigma) = -n \log \sigma + \sum_{i=1}^n \log f\left(\frac{x_i - \mu}{\sigma}\right);$$

$$(14.115) \quad \frac{\partial l_n(\mu, \sigma)}{\partial \mu} = -\frac{1}{\sigma} \sum_{i=1}^n \frac{f'\left(\frac{x_i - \mu}{\sigma}\right)}{f\left(\frac{x_i - \mu}{\sigma}\right)} = 0,$$

$$(14.116) \quad \frac{\partial l_n(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^n \frac{\left(\frac{x_i - \mu}{\sigma}\right) f'\left(\frac{x_i - \mu}{\sigma}\right)}{f\left(\frac{x_i - \mu}{\sigma}\right)} = 0.$$

If the assumptions of Proposition 14.20 and Theorem 14.21 hold then the unique consistent root $(\hat{\mu}^{(n)}, \hat{\sigma}^{(n)})$ of the two LEQs is CANE and satisfies (14.117)

$$\sqrt{n} \left[\begin{pmatrix} \hat{\mu}^{(n)} \\ \hat{\sigma}^{(n)} \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \right] \xrightarrow{d} N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, [I(\mu, \sigma)]^{-1} \equiv \begin{pmatrix} I_{\mu\mu} & I_{\mu\sigma} \\ I_{\sigma\mu} & I_{\sigma\sigma} \end{pmatrix}^{-1} \right].$$

Here $\begin{pmatrix} I_{\mu\mu} & I_{\mu\sigma} \\ I_{\sigma\mu} & I_{\sigma\sigma} \end{pmatrix}$ is the information matrix for (μ, σ) based on a single observation X_i . It is proportional to σ^{-2} and does not depend on μ ; its

³⁵ This is a “parametric” model with parameters μ, σ . If f is *unknown* also, the model is called “semi-parametric” with μ, σ, f all to be estimated.

entries are given by [verify – Exercise 14.35(i)]

$$(14.118) \quad I_{\mu\mu} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \frac{[f'(y)]^2}{f(y)} dy,$$

$$(14.119) \quad I_{\mu\sigma} = I_{\sigma\mu} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \frac{y[f'(y)]^2}{f(y)} dy,$$

$$(14.120) \quad I_{\sigma\sigma} = \frac{1}{\sigma^2} \left[\int_{-\infty}^{\infty} \frac{y^2[f'(y)]^2}{f(y)} dy - 1 \right].$$

Therefore

$$(14.121) \quad \sqrt{n} \left(\hat{\mu}^{(n)} - \mu \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I_{\mu\mu \cdot \sigma}} \right),$$

where

$$(14.122) \quad I_{\mu\mu \cdot \sigma} \equiv I_{\mu\mu} - \frac{I_{\mu\sigma}^2}{I_{\sigma\sigma}} \leq I_{\mu\mu}.$$

The parameters μ and σ are orthogonal iff $I_{\mu\sigma} = 0$. This holds, for example, if f is symmetric about 0, i.e., $f(x) = f(-x)$ [Exercise 14.35(ii)].

Case B: $\sigma = 1$ is known.

From (14.114) and (14.115) the log likelihood function and single LEQ are

$$(14.123) \quad l_n(\mu) = \sum_{i=1}^n \log f(x_i - \mu);$$

$$(14.124) \quad -\frac{\partial l_n(\mu)}{\partial \mu} = \sum_{i=1}^n \frac{f'(x_i - \mu)}{f(x_i - \mu)} = 0,$$

Under the Cramér conditions of Theorem 14.9, the unique consistent root $\tilde{\mu}^{(n)}$ of the LEQ is CANE and satisfies

$$(14.125) \quad \sqrt{n} \left(\tilde{\mu}^{(n)} - \mu \right) \xrightarrow{d} N_1 \left(0, \frac{1}{I_{\mu\mu}} \right).$$

As in (14.122), $\frac{1}{I_{\mu\mu}} \leq \frac{1}{I_{\mu\mu \cdot \sigma}}$ with strict inequality unless $I_{\mu\sigma} = 0$, so the optimal estimator $\hat{\mu}^{(n)}$ from Case A may be suboptimal in Case B.

However, the optimal estimator $\tilde{\mu}^{(n)}$ may be difficult to determine, whereas $\hat{\mu}^{(n)}$ may be easier to obtain (recall Footnote 32). In this case Method III or IV (using $\hat{\sigma}^{(n)} - 1$ as a covariate) may be used to adjust $\tilde{\mu}^{(n)}$ to produce a CANE estimator for μ in Case B.

However, if f is symmetric about 0 as in the Cauchy case, then μ and σ are orthogonal parameters so no adjustment is needed: $\hat{\mu}^{(n)}$ is also CANE in Case B! (See Exercise 14.36 for another CANE estimator.) \square

Exercise 14.35. (i) Verify (14.118) – (14.120).

(ii) Show that $I_{\mu\sigma} = 0$ if f is symmetric about 0.

Exercise 14.36. Let X_1, \dots, X_n be an i.i.d. sample from the Cauchy location family with scale parameter $\sigma = 1$. (See Examples 14.5 and 14.17.)

(i) Show that $I_{\mu\mu} = \frac{1}{2}$. Thus in Case B ($\sigma = 1$ known), the asymptotic variance of any CANE estimator $\bar{\mu}^{(n)}$ is $\frac{2}{n}$. Recall from (10.83) that the asymptotic variance of the sample median is $\frac{\pi^2}{4n} \approx \frac{2.47}{n}$. What is the asymptotic variance of the sample mean \bar{X}_n ?

(ii) Use both (14.59) and (14.61) to provide explicit formulas for the first iterates $\mu_{(1)}^{(n)}$ and $\dot{\mu}_{(1)}^{(n)}$ of the Newton-Raphson algorithm as approximate solutions of the LEQ (14.124). Use the sample median as the starting value.

(iii*) Show that the LEQs (14.115)–(14.116) have a unique root $(\hat{\mu}^{(n)}, \hat{\sigma}^{(n)})$ [see Copas *Biometrika* (1975).] (Thus $\hat{\mu}^{(n)}$ is a CANE for μ in Case B.)

Exercise 14.37.** Let X_1, \dots, X_n be an i.i.d. sample from the univariate normal location-scale family $N_1(\mu, \sigma^2)$. The MLEs of μ when σ^2 is known and when σ^2 is unknown are both \bar{X}_n , so trivially have the same asymptotic efficiency. This also follows from the fact that the $N_1(0, 1)$ pdf is symmetric about 0, so μ and σ are orthogonal parameters. For a *finite* sample size n , however, different confidence intervals are appropriate for the two cases:

$$\sigma^2 \text{ known: } \bar{X}_n \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \quad \sigma^2 \text{ unknown: } \bar{X}_n \pm \frac{s_n}{\sqrt{n}} t_{n-1; \alpha/2}.$$

Show that the width of the first confidence interval is smaller than the expected width of the second, so that knowing σ^2 actually improves the accuracy of inference about μ (on average).

15. The EM Algorithm for the MLE when Data is Missing.

Data structure: $X_{\text{complete}} = (X_{\text{observed}}, X_{\text{missing}})$

$$X \equiv (Y, Z) \sim f_{\theta}(y, z).$$

The missing data Z may be “observable but missing” or “unobservable” (latent \equiv hidden variables). We seek the MLE based on the data Y actually observed (assuming that θ is identifiable based on Y):

$$\hat{\theta} \equiv \hat{\theta}(y) = \arg \max_{\theta} f_{\theta}(y) \equiv \arg \max_{\theta} \int f_{\theta}(y, z) dz.$$

Basic premise:

- It may be hard to find $f_{\theta}(y)$ – the integral may be complicated.
- It is easier to find $\arg \max_{\theta} f_{\theta}(y, z)$, the MLE based on the complete data, or equivalently, to find

$$\arg \max_{\theta} \log \left[\frac{f_{\theta}(y, z)}{f_{\theta_0}(y, z)} \right] \quad \text{for any fixed } \theta_0.$$

Because z is missing, the EM algorithm, formalized by Dempster, Laird, Rubin (DLR) *JRSSB 1977*, optimistically proposes to replace the target function $h(\theta) \equiv \log \left[\frac{f_{\theta}(y, z)}{f_{\theta_0}(y, z)} \right]$ by its conditional expected value given $Y = y$:

1st E-step: Let $\hat{\theta}_0 \equiv \hat{\theta}_0(y)$ be an initial estimate (or guess). Compute

$$(15.1) \quad E_{\hat{\theta}_0} \left\{ \log \left[\frac{f_{\theta}(Y, Z)}{f_{\hat{\theta}_0}(Y, Z)} \right] \mid Y = y \right\} \equiv J(\theta \mid \hat{\theta}_0(y), y).$$

1st M-step: Find

$$(15.2) \quad \hat{\theta}_1 \equiv \hat{\theta}_1(y) = \arg \max_{\theta} J(\theta \mid \hat{\theta}_0(y), y).$$

(Note that $J(\hat{\theta}_1 \mid \hat{\theta}_0, y) \geq J(\hat{\theta}_0 \mid \hat{\theta}_0, y) = 0$.)

$(k+1)$ -st steps: For $k = 1, 2, \dots$, repeat the E-step and M-step with $\hat{\theta}_0, \hat{\theta}_1$ replaced by $\hat{\theta}_k, \hat{\theta}_{k+1}$. We hope that $\hat{\theta}_{k+1} \rightarrow \hat{\theta}$, the actual MLE.

Cause for hope : The actual likelihood increases at each iteration:

$$(15.3) \quad f_{\hat{\theta}_{k+1}}(y) \geq f_{\hat{\theta}_k}(y).$$

Proof. Trivially,

$$\begin{aligned} \log \left[\frac{f_{\hat{\theta}_{k+1}}(y)}{f_{\hat{\theta}_k}(y)} \right] &= E_{\hat{\theta}_k} \left\{ \log \left[\frac{f_{\hat{\theta}_{k+1}}(Y)}{f_{\hat{\theta}_k}(Y)} \right] \mid Y = y \right\} \\ &= E_{\hat{\theta}_k} \left\{ \log \left[\frac{f_{\hat{\theta}_{k+1}}(Y, Z)}{f_{\hat{\theta}_k}(Y, Z)} \right] \mid Y = y \right\} - E_{\hat{\theta}_k} \left\{ \log \left[\frac{f_{\hat{\theta}_{k+1}}(Z|Y)}{f_{\hat{\theta}_k}(Z|Y)} \right] \mid Y = y \right\} \\ &\equiv \underbrace{J(\hat{\theta}_{k+1} \mid \hat{\theta}_k, y)}_{\geq 0} + \underbrace{K_{Z|Y}(\hat{\theta}_k, \hat{\theta}_{k+1})}_{\geq 0}. \end{aligned}$$

(Here $K_{Z|Y}$ is the conditional KL distance.) □

However: The EM algorithm is *not guaranteed to always converge to a limit, and if it does converge, the limit may not be the actual MLE but instead may be another stationary point of the actual likelihood function.*

Furthermore, convergence may be slow and sensitive to the choice of starting point $\hat{\theta}_0$. See the DLR paper and accompanying discussion (e.g., by Murray), also Wu (1983) *Ann. Statist.* Various improvements to EM have been proposed to speed up convergence, e.g., ECM, EMCM, ECMC, etc. (cf. X.-L. Meng, D. Van Dyk, Balakrishnan/Wainwright/Yu (2017), etc.)

Relation between EM and “imputation” of missing data in an exponential family. The EM algorithm assumes a relatively simple form in an exponential family, where it can be interpreted as “imputing” the value of the missing data Z based on the observed data $Y = y$:

Suppose that the complete data $X \equiv (Y, Z)$ has pdf of the canonical k -parameter exponential form

$$(15.4) \quad f_{\theta}(y, z) = a(\theta) e^{\theta' T(y, z)} \cdot h(y, z) \equiv f_{\theta}[T(y, z)] \cdot h(y, z),$$

where θ and $T(y, z)$ are $k \times 1$. Then

$$(15.5) \quad \log \left[\frac{f_{\theta}(y, z)}{f_{\hat{\theta}_0}(y, z)} \right] = \log \left[\frac{a(\theta)}{a(\hat{\theta}_0)} \right] + (\theta - \hat{\theta}_0)' T(y, z),$$

so

$$\begin{aligned}
 J(\theta \mid \hat{\theta}_0, y) &= \log \left[\frac{a(\theta)}{a(\hat{\theta}_0)} \right] + (\theta - \hat{\theta}_0)' \underbrace{E_{\hat{\theta}_0}[T(Y, Z) \mid Y = y]}_{\equiv \hat{T}_1(y)} \\
 (15.6) \qquad &\equiv \log \left[\frac{f_{\theta}[\hat{T}_1(y)]}{f_{\hat{\theta}_0}[\hat{T}_1(y)]} \right],
 \end{aligned}$$

which is just the complete-data LLR using the *imputed value* $\hat{T}_1 \equiv \hat{T}_1(y)$. Thus the $(k+1)$ -st E-step consists simply of imputing $T(y, z)$ as

$$(15.7) \qquad \hat{T}_{k+1} \equiv \hat{T}_{k+1}(y) = E_{\hat{\theta}_k}[T(Y, Z) \mid Y = y],$$

then the $(k+1)$ -st M-step chooses $\hat{\theta}_{k+1}$ to maximize the complete-data LLR based on the imputed value \hat{T}_{k+1} .

This shows that the EM algorithm can be expressed very easily for multivariate normal models or multinomial models with missing data, where the regression functions are simple, in fact linear. (In these cases, the EM approach was known long before 1977.)

Example 15.1. (*Multivariate normal (MVN) model with missing data*) Suppose that the complete data set is

$$X \equiv \begin{matrix} p_1 : \\ p_2 : \end{matrix} \begin{pmatrix} R_1 \\ S_1 \end{pmatrix}, \dots, \begin{pmatrix} R_l \\ S_l \end{pmatrix}, \begin{pmatrix} T_1 \\ U_1 \end{pmatrix}, \dots, \begin{pmatrix} T_m \\ U_m \end{pmatrix}, \begin{pmatrix} V_1 \\ W_1 \end{pmatrix}, \dots, \begin{pmatrix} V_n \\ W_n \end{pmatrix},$$

an i.i.d. sample of size $l+m+n$ from the (p_1+p_2) -variate normal distribution

$$(15.8) \qquad N_{p_1+p_2} \left[\mu \equiv \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma \equiv \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right], \quad \Sigma \text{ known.}$$

Suppose that the U 's and V 's are missing, so the observed data is

$$Y \equiv \begin{pmatrix} R_1 \\ S_1 \end{pmatrix}, \dots, \begin{pmatrix} R_l \\ S_l \end{pmatrix}, \begin{pmatrix} T_1 \end{pmatrix}, \dots, \begin{pmatrix} T_m \end{pmatrix}, \begin{pmatrix} W_1 \end{pmatrix}, \dots, \begin{pmatrix} W_n \end{pmatrix}$$

and the missing data is

$$Z \equiv \begin{pmatrix} U_1 \end{pmatrix}, \dots, \begin{pmatrix} U_m \end{pmatrix}, \begin{pmatrix} V_1 \end{pmatrix}, \dots, \begin{pmatrix} V_n \end{pmatrix}.$$

Set $\theta \equiv \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \Sigma^{-1}\mu$, so the complete-data pdf is given by [verify!]
(15.9)

$$f_{\theta}(x) = a(\theta)e^{\theta'_1(\sum r_i + \sum t_i + \sum v_i) + \theta'_2(\sum s_i + \sum u_i + \sum w_i)} \cdot h(x),$$

which is of the exponential form (15.4) with

$$(15.10) \quad T \equiv T(y, z) = \begin{pmatrix} \sum r_i + \sum t_i + \sum v_i \\ \sum s_i + \sum u_i + \sum w_i \end{pmatrix}.$$

Thus from (15.7) the $(k+1)$ -st E-step simply imputes $T(y, z)$ by

$$(15.11) \quad \hat{T}_{k+1}(y) = E_{\hat{\mu}_k}[T(y, Z) \mid Y = y] = \begin{pmatrix} \sum r_i + \sum t_i + \sum \hat{v}_{i,k+1} \\ \sum s_i + \sum \hat{u}_{i,k+1} + \sum w_i \end{pmatrix},$$

where, from the usual MVN regression formulas,

$$(15.12) \quad \hat{u}_{i,k+1} = E_{\hat{\mu}_k}[U_i \mid T_i = t_i] = \hat{\mu}_{2,k} + \Sigma_{21}\Sigma_{11}^{-1}(t_i - \hat{\mu}_{1,k}),$$

$$(15.13) \quad \hat{v}_{i,k+1} = E_{\hat{\mu}_k}[V_i \mid W_i = w_i] = \hat{\mu}_{1,k} + \Sigma_{12}\Sigma_{22}^{-1}(w_i - \hat{\mu}_{2,k}),$$

and where $\hat{\mu}_k \equiv \hat{\mu}_k(y) \equiv \begin{pmatrix} \hat{\mu}_{1,k}(y) \\ \hat{\mu}_{2,k}(y) \end{pmatrix}$ is the estimate from the k -th M-step.³⁶

Now the $(k+1)$ -st M-step chooses $\hat{\mu}_{k+1}$ to maximize the complete-data LF based on the updated statistic \hat{T}_{k+1} . For this MVN case,

$$(15.14) \quad \hat{\mu}_{k+1} \equiv \begin{pmatrix} \hat{\mu}_{1,k+1} \\ \hat{\mu}_{2,k+1} \end{pmatrix} = \frac{1}{l+m+n} \begin{pmatrix} \sum r_i + \sum t_i + \sum \hat{v}_{i,k+1} \\ \sum s_i + \sum \hat{u}_{i,k+1} + \sum w_i \end{pmatrix}.$$

The algorithm is very easy to program in this case. □

³⁶ The initial estimate $\hat{\mu}_0$ may be based on the observed data: use the sample mean vectors $\hat{\mu}_{1,0} = \frac{1}{l+m}(\sum r_i + \sum t_i)$, $\hat{\mu}_{2,0} = \frac{1}{l+n}(\sum s_i + \sum w_i)$.

Exercise 15.2*. If (15.14) is used with the initial estimates in Footnote 36, find $\lim_{k \rightarrow \infty} \hat{\mu}_{1,k}$ and $\lim_{k \rightarrow \infty} \hat{\mu}_{2,k}$ *without electronic assistance*. \square

Remark 15.3. Suppose the observed data pattern is *monotone* \equiv *nested*:

$$(15.15) \quad Y \equiv \begin{pmatrix} R_1 \\ S_1 \end{pmatrix}, \dots, \begin{pmatrix} R_l \\ S_l \end{pmatrix}, \begin{pmatrix} T_1 \end{pmatrix}, \dots, \begin{pmatrix} T_m \end{pmatrix}.$$

Then the actual MLE $(\hat{\mu}, \hat{\Sigma})$ is easy to obtain explicitly in one step, even when Σ is unknown, as follows. Since

$$(15.16) \quad S_i | R_i = r_i \sim N_{p_2} \left[\underbrace{\mu_2 - \Sigma_{21} \Sigma_{11}^{-1} \mu_1}_{\alpha} + \underbrace{\Sigma_{21} \Sigma_{11}^{-1} r_i}_{\beta r_i}, \underbrace{\Sigma_{22 \cdot 1}}_{\Lambda} \right]$$

the joint pdf of the observed data can be factored as

$$(15.17) \quad \prod_{i=1}^l f_{\mu, \Sigma}(r_i, s_i) \cdot \prod_{i=1}^m f_{\mu_1, \Sigma_{11}}(t_i) = \prod_{i=1}^l f_{\alpha, \beta, \Lambda}(s_i | r_i) \cdot \prod_{i=1}^l f_{\mu_1, \Sigma_{11}}(r_i) \prod_{i=1}^m f_{\mu_1, \Sigma_{11}}(t_i).$$

The first factor on the right is the joint pdf of a linear regression model, for which the MLEs $\hat{\alpha}, \hat{\beta}, \hat{\Lambda}$ coincide with the least squares estimators. (See the solution to CB Exercise 7.18). The second and third factors on the right together constitute the joint pdf for a sample of size $l+m$ from $N_{p_1}(\mu_1, \Sigma_{11})$, for which the MLEs μ_1, Σ_{11} are simply the p_1 -dimensional sample mean vector and sample covariance matrix.

This simple factorization method for finding the MLEs obviously extends to any MVN monotone missing data model, where the observed data has the form

$$(15.18) \quad Y = \begin{pmatrix} * \\ * \\ * \end{pmatrix} \dots \begin{pmatrix} * \\ * \\ * \end{pmatrix} \dots \begin{pmatrix} * \\ * \\ * \end{pmatrix}.$$

This factorization method also applies to *non-monotone* MVN missing data models, provided that one is willing to impose certain conditional independence constraints on Σ determined by the observed data pattern. (Andersson and Perlman, *Stat. Prob. Letters* 1990; Perlman and Wu, *JSPI* 1999.)

This is closely related to the theory of Gaussian *graphical Markov models* for acyclic directed graphs (Andersson and Perlman, *JMVA* 1998). \square

Standard errors for the MLE with missing data. The observed data Y usually does *not* consist of n i.i.d. observations, so the MLE $\hat{\theta}(y)$ based on Y may *not* satisfy the usual asymptotic relation $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [I(\theta)]^{-1})$, where $I(\theta)$ is the Fisher information number or matrix for a single complete observation. Instead it is usually the case that

$$(15.19) \quad \hat{\theta} - \theta \approx N(0, [I_Y(\theta)]^{-1})$$

if $I_Y(\theta) \rightarrow \infty$, where I_Y denotes the information number or matrix for Y .

Calculation of I_Y can often be simplified as follows:

$$(15.20) \quad \begin{aligned} f_{\theta}(y) &= \frac{f_{\theta}(y, z)}{f_{\theta}(z | y)}, \\ \log f_{\theta}(y) &= \log f_{\theta}(y, z) - \log f_{\theta}(z | y), \\ \frac{\partial^2 \log f_{\theta}(y)}{\partial \theta_i \partial \theta_j} &= \frac{\partial^2 \log f_{\theta}(y, z)}{\partial \theta_i \partial \theta_j} - \frac{\partial^2 \log f_{\theta}(z | y)}{\partial \theta_i \partial \theta_j}, \end{aligned}$$

so

$$(15.21) \quad I_Y(\theta) = I_{Y,Z}(\theta) - E_{\theta} [I_{Z|Y}(\theta)].$$

Here $I_{Y,Z}(\theta) \equiv I_X(\theta)$ is the complete-data information, while

$$(15.22) \quad I_{Z|Y}(\theta) \equiv -E_{\theta} \left[\frac{\partial^2 \log f_{\theta}(Z | Y)}{\partial \theta_i \partial \theta_j} \mid Y \right]$$

denotes the conditional information in the missing data Z given the observed data Y .

Example 15.4. We illustrate the use of (15.21) in Example 15.1 (Σ is known). The information matrix $I(\mu)$ for μ in a single complete observation $X_i \sim N_{p_1+p_2}(\mu, \Sigma)$ is Σ^{-1} [Exercise 15.5(i)], so by the additivity of information,

$$(15.23) \quad I_X(\mu) = (l + m + n)\Sigma^{-1} \equiv (l + m + n) \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}.$$

Next, as in (8.72),

$$U_i | T_i \sim N_{p_2}(\alpha + \Sigma_{21}\Sigma_{11}^{-1}T_i, \Sigma_{22.1}),$$

where $\alpha \equiv \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1$ is unknown, so

$$I_{U_i|T_i}(\alpha) = \Sigma_{22.1}^{-1}.$$

Thus, by extending Remark 13.5 (p.210) to the multiparameter case,

$$\begin{aligned} I_{U_i|T_i}(\mu) \equiv I_{U_i|T_i}(\mu_1, \mu_2) &= \begin{pmatrix} \left(\frac{\partial \alpha}{\partial \mu_1} \right)' \\ \left(\frac{\partial \alpha}{\partial \mu_2} \right)' \end{pmatrix} I_{U_i|T_i}(\alpha) \begin{pmatrix} \frac{\partial \alpha}{\partial \mu_1} & \frac{\partial \alpha}{\partial \mu_2} \end{pmatrix} \\ (15.24) \qquad \qquad \qquad &= \begin{pmatrix} -\Sigma_{11}^{-1}\Sigma_{12} \\ I_{p_2} \end{pmatrix} \Sigma_{22.1}^{-1} \begin{pmatrix} -\Sigma_{21}\Sigma_{11}^{-1} & I_{p_2} \end{pmatrix}, \end{aligned}$$

and similarly

$$(15.25) \quad I_{V_i|W_i}(\mu) \equiv I_{V_i|W_i}(\mu_1, \mu_2) = \begin{pmatrix} I_{p_1} \\ -\Sigma_{22}^{-1}\Sigma_{21} \end{pmatrix} \Sigma_{11.2}^{-1} \begin{pmatrix} I_{p_1} & -\Sigma_{12}\Sigma_{22}^{-1} \end{pmatrix}.$$

Furthermore, by the (conditional) independence of $U_1, \dots, U_m, V_1, \dots, V_n$ and the additivity of (conditional) information,

$$(15.26) \qquad I_{Z|Y}(\theta) = mI_{U_i|T_i}(\mu) + nI_{V_i|W_i}(\mu).$$

By (15.24) and (15.25) the right side of (15.26) does not depend on Y , so

$$E_\theta [I_{Z|Y}(\theta)] = mI_{U_i|T_i}(\mu) + nI_{V_i|W_i}(\mu)$$

as well. Thus we conclude from (15.21) and (15.24)-(15.25) that the information in the observed data Y is [verify! Exercise 15.5]

$$\begin{aligned} I_Y(\mu) &= I_X(\mu) - mI_{U_i|T_i}(\mu) - nI_{V_i|W_i}(\mu) \\ &= (l + m + n)\Sigma^{-1} - m \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} \\ -\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22.1}^{-1} \end{pmatrix} \\ &\quad - n \begin{pmatrix} \Sigma_{11.2}^{-1} & -\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1} & \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
&= (l + m + n) \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} - m \begin{pmatrix} \Sigma^{12}(\Sigma^{22})^{-1}\Sigma^{21} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \\
&\quad - n \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{21}(\Sigma^{11})^{-1}\Sigma^{12} \end{pmatrix} \\
&= \begin{pmatrix} l\Sigma^{11} + m\Sigma^{11.2} & l\Sigma^{12} \\ l\Sigma^{21} & l\Sigma^{22} + n\Sigma^{22.1} \end{pmatrix} \\
(15.27) \quad &= l\Sigma^{-1} + \begin{pmatrix} m\Sigma_{11}^{-1} & 0 \\ 0 & n\Sigma_{22}^{-1} \end{pmatrix}.
\end{aligned}$$

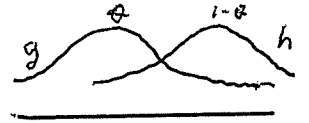
This shows the contribution of the partially observed data $T_1, \dots, T_m, W_1, \dots, W_n$ to the overall information about $\mu \equiv (\mu_1, \mu_2)$. \square

Exercise 15.5. (i) If $X \sim N_p(\mu, \Sigma)$, show that $I_X(\mu) = \Sigma^{-1}$.

(ii) Verify the algebra leading to (15.27). Give a direct (shorter) derivation of (15.27), using the facts that $T_i \sim N_{p_1}(\mu_1, \Sigma_{11})$ and $W_i \sim N_{p_2}(\mu_2, \Sigma_{22})$.

Example 15.6. (*A mixture model*) Suppose that $Y \equiv (Y_1, \dots, Y_n)$ is an i.i.d. sample from the *mixture pdf*

$$(15.28) \quad f_\theta(y_i) = \theta g(y_i) + (1 - \theta)h(y_i), \quad 0 < \theta < 1,$$



where g, h are *known* pdfs (or pmfs) and θ is unknown. The pdf of Y is

$$(15.29) \quad f_\theta(y) = \prod_{i=1}^n [\theta g(y_i) + (1 - \theta)h(y_i)].$$

This is log concave in θ , so the LF has at most one mode (a maximum) and the LEQ

$$\sum_{i=1}^n \frac{g(y_i) - h(y_i)}{h(y_i) + \theta[g(y_i) - h(y_i)]} = 0,$$

has at most one solution, which must be the MLE if it exists.³⁷

³⁷ No solution need exist, for example if all $g(y_i) > h(y_i)$ in which case the LF approaches its maximum as $\theta \rightarrow 1$, or all $g(y_i) < h(y_i)$ in which case the LF approaches its maximum as $\theta \rightarrow 0$.

The LEQ is equivalent to a polynomial equation of degree $n-1$. Rather than solving this equation we can apply the EM algorithm by introducing the i.i.d “missing data” $Z \equiv (Z_1, \dots, Z_n)$, where $Z_i \sim \text{Bernoulli}(\theta)$ is the indicator variable that determines whether Y_i is drawn from g or from h :

$$Y_i \sim \begin{cases} g(y_i) & \text{if } Z_i = 1; \\ h(y_i) & \text{if } Z_i = 0. \end{cases}$$

Thus

$$(15.30) \quad f_\theta(y_i | z_i) = [g(y_i)]^{z_i} [h(y_i)]^{1-z_i},$$

$$(15.31) \quad f_\theta(z_i) = \theta^{z_i} (1 - \theta)^{1-z_i},$$

so the joint (mixed) pdf of (Y, Z) has the exponential form

$$(15.32) \quad f_\theta(y, z) = \prod_{i=1}^n [\theta g(y_i)]^{z_i} [(1 - \theta)h(y_i)]^{1-z_i}.$$

This has the exponential form (15.4) with $T(y, z) = \sum_{i=1}^n z_i$. Therefore by (15.7), p.263, the $(k+1)$ -st E-step requires the calculation

$$(15.33) \quad \begin{aligned} E_{\hat{\theta}_k} \left[Z_i \mid Y_1 = y_1, \dots, Y_n = y_n \right] &= P_{\hat{\theta}_k} \left[Z_i = 1 \mid Y_i = y_i \right] \\ &= \frac{\hat{\theta}_k g(y_i)}{\hat{\theta}_k g(y_i) + (1 - \hat{\theta}_k) h(y_i)}, \end{aligned}$$

which follows from (15.30)-(15.31) and Bayes formula (4.14) [Exercise 15.7]. Finally [Exer. 15.7], the $(k+1)$ -st M-step yields the updated estimate

$$(15.34) \quad \hat{\theta}_{k+1} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{\theta}_k g(y_i)}{\hat{\theta}_k g(y_i) + (1 - \hat{\theta}_k) h(y_i)}.$$

By the unimodality of the LF, this will converge to the MLE $\hat{\theta}$ (or to one of the boundary points $\theta = 0$ or $\theta = 1$ – see Footnote 37, p.268). \square

Exercise 15.7. (i) Verify (15.33) and (15.34) of Example 15.6.

(ii) Suggest a reasonable starting value $\hat{\theta}_0$. [Hint: consider $\{y | g(y) > h(y)\}$.]

(iii) Find an integral expression for the information number $I_Y(\theta)$. \square

Example 15.8. In most mixture problems arising in applications, the pdfs g and h are assumed to be *unknown* members of a parametric family $\{f_\lambda\}$, so (15.28) now appears as

$$(15.35) \quad f_{\theta, \lambda, \mu}(y_i) = \theta f_\lambda(y_i) + (1 - \theta) f_\mu(y_i), \quad 0 < \theta < 1,$$

with θ, λ, μ unknown. The EM algorithm in Example 15.6 readily extend to this case – details appear in many textbooks, e.g. K. Knight *Mathematical Statistics* Chapman & Hall, 2000. Here is an example.

Suppose that Y_i is a θ -mixture of Poisson(λ) and Poisson(μ) rvs, with θ, λ, μ all unknown. Thus (15.29) and (15.32) become, respectively,

$$(15.36) \quad f_{\theta, \lambda, \mu}(y) = \prod_{i=1}^n \left[\theta \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} + (1 - \theta) \frac{e^{-\mu} \mu^{y_i}}{y_i!} \right],$$

$$(15.37) \quad f_{\theta, \lambda, \mu}(y, z) = \prod_{i=1}^n \left[\theta \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right]^{z_i} \left[(1 - \theta) \frac{e^{-\mu} \mu^{y_i}}{y_i!} \right]^{1-z_i}$$

$$(15.38) \quad = \left[\frac{1 - \theta}{e^\mu} \right]^n \left[\frac{\theta e^{\mu - \lambda}}{1 - \theta} \right]^{\sum z_i} \mu^{\sum y_i} \left(\frac{\lambda}{\mu} \right)^{\sum y_i z_i} \frac{1}{\prod y_i!}.$$

Thus the complete-data likelihood $f_\theta(y, z)$ has the exponential family form given in (15.4) with $T(y, z) = (\sum z_i, \sum y_i, \sum y_i z_i)$, so by (15.7) and Bayes' formula, the $(k + 1)$ -st E-step simply imputes z_i by

$$(15.39) \quad \begin{aligned} \hat{z}_{i, k+1} &\equiv E_{\hat{\theta}_k, \hat{\lambda}_k, \hat{\mu}_k} [Z_i \mid Y_i = y_i] \\ &= P_{\hat{\theta}_k, \hat{\lambda}_k, \hat{\mu}_k} [Z_i = 1 \mid Y_i = y_i] \\ &= \frac{\hat{\theta}_k e^{-\hat{\lambda}_k} \hat{\lambda}_k^{y_i}}{\hat{\theta}_k e^{-\hat{\lambda}_k} \hat{\lambda}_k^{y_i} + (1 - \hat{\theta}_k) e^{-\hat{\mu}_k} \hat{\mu}_k^{y_i}}. \end{aligned}$$

Then, because the complete-data MLEs are [Exercise 15.11]

$$(15.40) \quad \bar{\theta} = \frac{1}{n} \sum z_i, \quad \bar{\lambda} = \frac{\sum y_i z_i}{\sum z_i}, \quad \bar{\mu} = \frac{\sum y_i (1 - z_i)}{\sum (1 - z_i)},$$

the $(k + 1)$ -st M-step yields the updated estimates

$$\hat{\theta}_{k+1} = \frac{1}{n} \sum \hat{z}_{i, k+1}, \quad \hat{\lambda}_{k+1} = \frac{\sum y_i \hat{z}_{i, k+1}}{\sum \hat{z}_{i, k+1}}, \quad \hat{\mu}_{k+1} = \frac{\sum y_i (1 - \hat{z}_{i, k+1})}{\sum (1 - \hat{z}_{i, k+1})}. \quad \square$$

Remark 15.9. Whereas it is possible to find a reasonable “all-purpose” starting value $\hat{\theta}_0$ when g and h are known (see Exercise 15.7), this is not the case when they are unknown. Instead, as in Example 15.8, one may begin with a histogram of y_1, \dots, y_n and attempt to discern the two mixture component pdfs f_λ and f_μ and their relative weights θ and $1 - \theta$ by eye. As a default option, one may simply take $\hat{\theta}_0 = .5$ and $\hat{\lambda}_0 = \hat{\mu}_0 = \bar{y}_n$, the mean of the pooled data set. This option is problematic, however, in view of the notorious sensitivity of the EM algorithm to the choice of the starting value. (Also, the histogram may reveal more than two modes, suggesting a mixture of more than two pdfs – this is called “bump-hunting”.) \square

Remark 15.10. When the mixture component pdfs f_λ and f_μ are unknown as in Example 15.8, it is apparent from (15.35) that *the parameters θ, λ, μ are actually not identifiable on the basis of the observed data Y alone, since*

$$(15.41) \quad f_{\theta, \lambda, \mu}(y_i) = f_{1-\theta, \mu, \lambda}(y_i).$$

To attain identifiability a constraint should be imposed, such as $\lambda \leq \mu$. Therefore, if $\hat{\lambda}_k > \hat{\mu}_k$, replace each by their average $\frac{1}{2}(\hat{\lambda}_k + \hat{\mu}_k)$. [why?] \square

Exercise 15.11. Verify (15.40) in Example 15.8. What happens if $\sum z_i = 0$ or $\sum z_i = n$? \square

Exercise 15.12. Repeat Example 15.8 when the two mixture component distributions are Binomial(m, λ) and Binomial(m, μ) (rather than Poisson). What goes wrong when $m = 1$? What about $m = 2$? \square

Example 15.13. (*Multinomial (categorical data) model with missing data*)

Example 15.14. (*Censored data model*) [Also see Supplement 1, p.342.]

Example 15.15. (*MVN model with patterned covariance matrix*)

Example 15.16. (*Combining cells in a multinomial distribution*)

16. Bayes Estimators.

It may be appropriate to assume that the unknown parameter θ , is itself *random*. That is, θ is the realized value of a random variable Θ taking values in the parameter space Ω . If we know the *prior* pdf $\pi(\theta)$ of Θ (continuous or discrete), then we should incorporate this prior information into our inferences concerning θ . In fact, even with no data, a reasonable prior guess for θ is

$$E(\Theta) \equiv \int \theta \pi(\theta) d\theta \quad \left(\text{or } \sum \theta_i \pi(\theta_i) \right).$$

If data is available, the best estimator of θ is often a *weighted average* of this prior guess and the best unbiased estimator of θ

For example, if the data consists of a random sample X_1, \dots, X_n from a univariate normal distribution $N_1(\theta, \sigma^2)$ with σ^2 *known*, and if Θ has prior distribution $N_1(\eta, \tau^2)$ with η and τ^2 *known*, then $E(\Theta) \equiv \eta$ is our best prior guess of θ , while the sample average (\equiv MLE) \bar{X}_n is the best unbiased estimator of θ . We will see that the Bayes estimator is a weighted average of η and \bar{X}_n , with the weights depending on the ratio $\sigma^2/n\tau^2$.

In the general case, let $X \equiv (X_1, \dots, X_n)$ denote the observed random data vector with pdf $f_\theta(x)$ (continuous or discrete) where, as above, θ is the realized value of Θ . It is now appropriate to write $f_\theta(x)$ in the form of a conditional pdf $f(x|\theta)$ and to interpret the prior pdf $\pi(\theta)$ as a marginal pdf. Then the joint pdf of (X, Θ) is

$$f(x, \theta) = f(x|\theta) \pi(\theta),$$

so the conditional \equiv *posterior* pdf of $\Theta|X$ is, by Bayes' formula (4.14),

$$(16.1) \quad f(\theta|x) = \frac{f(x|\theta) \pi(\theta)}{f(x)} = \frac{f(x|\theta) \pi(\theta)}{\int f(x|\theta) \pi(\theta) d\theta}.$$

Given $f(x|\theta)$, $\pi(\theta)$, and the observed value $X = x$, it is easy to find the optimal \equiv *Bayes* estimator $\hat{\theta} \equiv \hat{\theta}(x)$ of θ w.r.to the loss criterion given by the MSE $E(\hat{\theta}(X) - \Theta)^2$. (Here we are treating the case of a one-dimensional parameter θ but the generalization to the multiparameter case is straightforward.) That is, $\hat{\theta}(x)$ is the value that minimizes the mean-squared error

$$(16.2) \quad E[(\Theta - \hat{\theta}(X))^2] = E\{E[(\Theta - \hat{\theta}(x))^2 | X = x]\}.$$

This can be minimized by minimizing the *expected posterior loss*

$$E[(\Theta - \hat{\theta})^2 \mid X]$$

for each x , which is accomplished by setting

$$(16.3) \quad \hat{\theta} \equiv \hat{\theta}(x) = E[\Theta \mid X = x],$$

the *posterior mean* of Θ , which is often called the *Bayes estimator* of θ .

Remark 16.1. If we adopted the loss criterion $|\hat{\theta} - \theta|$, then the appropriate Bayes estimator would be the posterior *median* of $\Theta \mid X = x$. Alternatively, the posterior *mode*

$$\check{\theta} \equiv \operatorname{argmax}_{\theta} f(\theta|x) = \operatorname{argmax}_{\theta} f(x|\theta)\pi(\theta)$$

is sometimes used. For the *uniform* “prior” pdf $\pi(\theta) \propto 1$ on Ω , this coincides with the MLE (though this is an *improper* prior if $\int_{\Omega} d\theta = \infty$; see §16.1.)

Example 16.2. As above, X_1, \dots, X_n is a random sample from $N_1(\theta, \sigma^2)$ with σ^2 known and Θ has prior distribution $N_1(\eta, \tau^2)$ with η, τ^2 known. In terms of the sufficient statistic \bar{X}_n for θ we summarize this as

$$(16.4) \quad \bar{X}_n \mid \Theta \sim N_1 \left(\Theta, \frac{\sigma^2}{n} \right),$$

$$(16.5) \quad \Theta \sim N_1(\eta, \tau^2) \quad (\equiv \pi).$$

To find the Bayes estimator $E[\Theta \mid X]$ we must find the posterior distribution of $\Theta \mid \bar{X}_n$. As in Example 5.1 and Exercise 8.7, (16.4) and (16.5) imply that the joint distribution of (\bar{X}_n, Θ) is bivariate normal, so we need only find its mean vector and covariance matrix:

$$\begin{aligned} E(\bar{X}_n) &= E[E(\bar{X}_n \mid \Theta)] = E[\Theta] = \eta; \\ \operatorname{Var}(\bar{X}_n) &= E[\operatorname{Var}(\bar{X}_n \mid \Theta)] + \operatorname{Var}[E(\bar{X}_n \mid \Theta)] \\ &= E[\sigma^2/n] + \operatorname{Var}(\Theta) \\ &= (\sigma^2/n) + \tau^2; \\ E(\Theta) &= \eta; \quad \operatorname{Var}(\Theta) = \tau^2; \\ \operatorname{Cov}(\bar{X}_n, \Theta) &= \operatorname{Cov}[E(\bar{X}_n \mid \Theta), \Theta] = \operatorname{Cov}(\Theta, \Theta) = \operatorname{Var}(\Theta) = \tau^2. \end{aligned}$$

Thus

$$(16.6) \quad \begin{pmatrix} \bar{X}_n \\ \Theta \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \eta \\ \eta \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix} \right],$$

so [verify!]

$$(16.7) \quad \Theta | \bar{X}_n \sim N_1 \left[\frac{\frac{n\bar{X}_n}{\sigma^2} + \frac{\eta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right],$$

hence the Bayes estimator of θ is

$$(16.8) \quad \hat{\theta}(\bar{X}_n) \equiv E[\Theta | \bar{X}_n] = \frac{\frac{n\bar{X}_n}{\sigma^2} + \frac{\eta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},$$

a weighted average of \bar{X}_n and η as predicted above. The weights are proportional to $\frac{n}{\sigma^2}$ and $\frac{1}{\tau^2}$, the *precisions* of the normal distributions given by the model distribution (16.4) and the prior distribution (16.5), respectively. The ratio of these weights is $n\tau^2/\sigma^2$, which determines the relative weight assigned to the MLE \bar{X}_n . Thus the weight assigned to \bar{X}_n :

- increases as n increases (because \bar{X}_n becomes more precise)
- increases as σ^2 decreases (because \bar{X}_n becomes more precise)
- increases as τ^2 increases (because the prior dist'n becomes more diffuse)

Note that the Bayes estimator $\hat{\theta}(\bar{X}_n)$ in (16.8) is *biased*:

$$E[\hat{\theta}(\bar{X}_n) | \Theta = \theta] = \frac{\frac{n\theta}{\sigma^2} + \frac{\eta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \neq \theta \quad \text{unless } \theta = \eta,$$

and its MSE for each fixed value of θ is given by

$$\begin{aligned} E[(\hat{\theta}(\bar{X}_n) - \theta)^2 | \Theta = \theta] &= \text{Var}[\hat{\theta}(\bar{X}_n) | \Theta = \theta] + \{E[\hat{\theta}(\bar{X}_n) | \Theta = \theta] - \theta\}^2 \\ &= \left(\frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)^2 \text{Var}[\bar{X}_n | \Theta = \theta] + \left(\frac{\frac{n\theta}{\sigma^2} + \frac{\eta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \theta \right)^2 \\ &= \frac{\frac{n}{\sigma^2}}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^2} + \left(\frac{\frac{\eta}{\tau^2} - \frac{\theta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)^2 \\ (16.9) \quad &= \frac{\frac{n}{\sigma^2} + \frac{(\theta - \eta)^2}{\tau^4}}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^2}, \end{aligned}$$

which is smallest when $\theta = \eta \equiv E(\Theta)$, as should be expected. Finally, the optimal (minimum) Bayes risk is attained by $\hat{\theta}(\bar{X}_n)$ and is given by

$$(16.10) \quad r_{\hat{\theta}}(\pi) \equiv E_{\pi} \left[\frac{\frac{n}{\sigma^2} + \frac{(\Theta - \eta)^2}{\tau^4}}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^2} \right] = \frac{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^2} = \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)}. \quad \square$$

Example 16.3. As in Example 4.3, suppose that

$$(16.11) \quad X \mid \Theta \sim \text{Binomial}(n, \Theta),$$

$$(16.12) \quad \Theta \sim \text{Uniform}(0, 1).$$

By Bayes formula (16.1),

$$\begin{aligned} f(\theta \mid x) &= \frac{\binom{n}{x} \theta^x (1 - \theta)^{n-x}}{\int_0^1 \binom{n}{x} \theta^x (1 - \theta)^{n-x} d\theta} \\ &= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^x (1 - \theta)^{n-x}, \end{aligned}$$

hence the posterior distribution is given by

$$(16.13) \quad \Theta \mid X \sim \text{Beta}(X+1, n-X+1).$$

Thus the Bayes estimator of θ is

$$(16.14) \quad E[\Theta \mid X] = \frac{X+1}{n+2} = \left(\frac{n}{n+2}\right) \left(\frac{X}{n}\right) + \left(\frac{2}{n+2}\right) \left(\frac{1}{2}\right),$$

again a weighted average of the unbiased MLE $\frac{X}{n}$ and the prior mean $\frac{1}{2}$. The ratio of the weights is $\frac{n}{2}$ so, as in Example 16.1, the weight assigned to $\frac{X}{n}$ increases to 1 as $n \rightarrow \infty$. Finally, the Bayes estimator is again *biased*:

$$E\{E[\Theta \mid X] \mid \Theta = \theta\} = \frac{n\theta + 1}{n+2} \neq \theta \quad \text{unless } \theta = \frac{1}{2}. \quad \square$$

Example 16.4. Suppose that $X \mid \Theta = \pm\Theta$ with probability $\frac{1}{2}$ each, where $\Theta > 0$. Then $|X| = \Theta$, so the Bayes estimator $E[\Theta \mid X] = \Theta$, that is, it is a

perfect estimator of Θ (regardless of the choice of prior distribution of Θ). In this case the Bayes estimator is unbiased:

$$E\{E[\Theta | X] | \Theta = \theta\} = E\{\Theta | \Theta = \theta\} = \theta \quad \forall \theta. \quad \square$$

Exercise 16.5. Show that the Bayes estimator $\hat{\theta} \equiv E[\Theta | X]$ is unbiased for θ iff it is perfect, i.e., iff $\hat{\theta} \equiv \Theta$ (which does not occur in practice). \square

Exercise 16.6. In Example 16.3, generalize the prior distribution (16.12) to $\Theta \sim \text{Beta}(\alpha, \beta)$ for $\alpha, \beta > 0$. Show that (16.13) and (16.14) become

$$(16.15) \quad \Theta | X \sim \text{Beta}(X + \alpha, n - X + \beta),$$

$$(16.16) \quad E[\Theta | X] = \frac{X + \alpha}{n + \alpha + \beta} = \left(\frac{n}{n + \alpha + \beta} \right) \left(\frac{X}{n} \right) + \left(\frac{\alpha + \beta}{n + \alpha + \beta} \right) \left(\frac{\alpha}{\alpha + \beta} \right).$$

Again $E[\Theta | X]$ is a weighted average of the MLE $\frac{X}{n}$ and the prior mean $E(\Theta) = \frac{\alpha}{\alpha + \beta}$. \square

Exercise 16.7. Suppose that $X | \Lambda \sim \text{Poisson}(\Lambda)$ and Λ has the prior pdf $\pi(\lambda) = \frac{1}{\nu} e^{-\frac{\lambda}{\nu}}$, an exponential distribution with $\nu > 0$. Find the posterior distribution of $\Lambda | X$ and the Bayes estimator $E[\Lambda | X]$, and express the latter as a weighted average of the MLE X and the prior mean $E(\Lambda) = \nu$. \square

Remark 16.8. The priors in Exercises 16.2, 16.6, and 16.7 are *conjugate priors*. These occur for parametric models $\{f(x|\theta) | \theta \in \Omega\}$ such that for each fixed x , $f(x|\theta) \propto \pi_{\tau(x)}(\theta)$ with $\{\pi_{\tau}(\theta) | \tau \in \Xi\}$ a second parametric family of prior pdfs such that for each pair (x, τ_0) , the posterior pdf $f_{\tau_0}(\theta|x)$ obtained from the prior $\pi_{\tau_0}(\theta)$ remains a member of the second family; i.e.,

$$f_{\tau_0}(\theta|x) \propto \pi_{\tau(x)}(\theta) \pi_{\tau_0}(\theta) \propto \pi_{\tau(x, \tau_0)}(\theta)$$

for some $\tau(x, \tau_0) \in \Xi$. In this case $\{\pi_{\tau}(\theta) | \tau \in \Xi\}$ is called a *conjugate family of prior pdfs* for the model $\{f(x|\theta) | \theta \in \Omega\}$. The conjugate families in Exercises 16.2, 16.6, and 16.7 are the $N(\eta, \tau^2)$, $\text{Beta}(\alpha, \beta)$, and $\text{Exponential}(\nu)$ families, respectively.

A conjugate family of priors can greatly simplify the calculation of the posterior distribution. However, priors should always be chosen on the basis of prior knowledge rather than mathematical convenience. \square

16.1. Prior distributions: proper vs. improper, informative vs. uninformative.

Many so-called Bayesian analyses use “improper” and/or “uninformative” prior distributions. An improper distribution is one with infinite mass, e.g., Lebesgue measure on $(-\infty, \infty)$. An uninformative prior is one supposedly used to represent prior “ignorance”, e.g., using the Uniform(0, 1) distribution for a binomial probability p . In my opinion, however, neither of these belong in a valid Bayesian analysis.

(a) *There is no such thing as an improper prior distribution.*

On one level, this is tautological: a “distribution” refers to a probability distribution, and no measure with infinite mass can be normalized to become a probability measure, period. Attempts that formally invoke Bayes formula to convert improper prior “distributions” into proper posterior distributions thus rest upon a nonexistent foundation. Attempts to justify improper prior distributions as limits of proper prior distributions (e.g. J. Berger, *Statistical Decision Theory and Bayesian Analysis*) are invalid in general (see Example 16.8 below.) Equally troubling, severe difficulties are encountered if one attempts to interpret an improper measure as a *uninformative* prior “distribution” – see (c).

(b) *An “uninformative” proper prior distribution may be informative.*³⁸

As a simple example, consider the model given by an observation $X \sim \text{Binomial}(n, \theta)$, $0 < \theta < 1$, with θ unknown. If we have no prior knowledge about θ we might be tempted to “represent this prior ignorance” by a “uninformative” prior distribution, and an obvious choice is the uniform prior distribution Uniform(0, 1). But this prior is far from “uninformative”: for example, it tells us that $E(\Theta) = 1/2$ *a priori*. In fact, the use of an “uninformative” prior violates the rationale of the Bayesian model, whereby prior *knowledge* is to be combined efficiently with sample data to yield optimal *a posteriori* inferences.

³⁸ An invariant proper prior distribution may be uninformative. For example, in a directional data model, if the parameter $\theta \in [0, 2\pi)$ represents a direction in R^2 then the uniform distribution on the unit circle is invariant under rotations and uninformative.

Furthermore, in the Binomial example this choice of a uniform prior to represent “complete ignorance” about θ is somewhat arbitrary because it is dependent on the choice of parametrization. For example, the Binomial model also can be parametrized by the logit $\gamma \equiv \log \frac{\theta}{1-\theta} \in (-\infty, \infty)$; an “uninformative” prior for γ would appear to be the uniform (improper) distribution on $(-\infty, \infty)$. However, the uniform prior for θ induces a standard logistic prior for γ , which is non-uniform on $(-\infty, \infty)$. [See “Jeffreys prior”.]

(c) *There is no such thing as an uninformative improper prior distribution.*

The difficulties encountered in (a) and (b) are compounded by attempts to construct “uninformative” improper prior “distributions” – examples appear in Lehmann TSH Ch.5 §9. As a simple example, consider the model given by an observation $X \sim N_1(\theta, 1)$ with $-\infty < \theta < \infty$ unknown. Lehmann (TSH Example 12 p. 226) notes the possibility of using Lebesgue measure $d\theta$ over the entire parameter space $(-\infty, \infty)$ as an improper prior “distribution” to represent “indifference”. (This prior is often called a “flat \equiv uniform” prior.) Note, however, that this improper prior “distribution” *assigns infinite mass to the complement of every bounded interval $[-M, M]$* no matter how large M is, and thereby (at first glance, at least) will bias any inference away from 0 and favor large values of θ . (Furthermore, this improper prior suffers from the same arbitrariness (non-invariance under reparameterization) that we encountered in the Binomial example.)

At second glance, however, this discussion makes no sense because we may fix *any* value η and apply the same “reasoning” to conclude that the improper Lebesgue prior “distribution” biases any inference away from η , no matter how large $|\eta|$ may be. The correct conclusion is that this discussion is vacuous because Bayes Theorem applies only to proper probability distributions – improper priors have no place in the Bayesian paradigm.

Lehmann (TSH Example 12 p.226) suggests that the flat Lebesgue prior might be viewed as an approximation to a *proper* normal prior distribution $\theta \sim N_1(\eta, \tau^2)$ with τ^2 very large, so that inferences made on the basis of the improper prior may be viewed as approximations to inferences based on a proper but very diffuse prior. (Again, however, η is arbitrary.) The following example due to L. J. Savage (cf. Perlman and Rasmussen (1975) *Comm. Statist.* 4 455-468) shows that this suggestion is invalid. [This illustrates the “Marginalization Paradox”, c.f. Dawid, Stone, Zidek (1973).]

Example 16.8. In this example, a multivariate generalization of Example 16.2, we observe $X \sim N_p(\theta, I_p)$ with $\theta \equiv (\theta_1, \dots, \theta_p) \in \mathbf{R}^p$ unknown. We are interested in the case where p is *large*. Suppose that $\Theta \sim N_p(\eta, \tau^2 I_p)$ with η and τ^2 both known, and take $\eta = 0$ for simplicity of notation. A standard Bayesian calculation (see Example 16.2 for the univariate case $p = 1$) shows that the posterior distribution of $\Theta|X$ is [verify]

$$(16.17) \quad \Theta|X \sim N_p\left(\frac{\tau^2}{1+\tau^2}X, \frac{\tau^2}{1+\tau^2}I_p\right).$$

As the prior variance $\tau^2 \rightarrow \infty$, the prior $N_p(0, \tau^2 I_p)$ becomes increasingly diffuse and the posterior distribution (16.17) approaches

$$(16.18) \quad \Theta|X \sim N_p(X, I_p),$$

which is also the “posterior distribution” obtained by using Lebesgue measure as an improper prior “distribution” and formally applying Bayes formula [verify]. The “Bayes estimator” of θ derived from (16.18) is X itself. Because X is complete and sufficient for θ , X is the UMVUE of θ as well.

Suppose, however, that we wish to estimate

$$(16.19) \quad \delta_p \equiv \frac{1}{p}\|\theta\|^2 \equiv \frac{1}{p}(\theta_1^2 + \dots + \theta_p^2),$$

the “average noncentrality per coordinate”. Because $\|X\|^2 \mid \Theta \sim \chi_p^2(\|\Theta\|^2)$, a noncentral χ^2 variate with p degrees of freedom and noncentrality parameter $\|\Theta\|^2$, $E(\|X\|^2 \mid \Theta) = p + \|\Theta\|^2$, so the UMVUE of δ_p is

$$(16.20) \quad \tilde{\delta}_p \equiv \frac{1}{p}\|X\|^2 - 1.$$

However, the posterior “distribution” (16.18) derived from the improper Lebesgue prior “distribution” yields the posterior “distribution”

$$(16.21) \quad \|\Theta\|^2 \mid X \sim \chi_p^2(\|X\|^2),$$

so the corresponding “improper Bayes estimator” of δ_p is

$$(16.22) \quad \hat{\delta}_p \equiv E\left[\frac{1}{p}\|\Theta\|^2 \mid X\right] = \frac{1}{p}\|X\|^2 + 1.$$

Next, under the *proper* prior $N_p(0, \tau^2 I_p)$, it follows from (16.17) that

$$(16.23) \quad \|\Theta\|^2 | X \sim \frac{\tau^2}{1 + \tau^2} \cdot \chi_p^2 \left(\frac{\tau^2 \|X\|^2}{1 + \tau^2} \right),$$

and the corresponding *proper* Bayes estimator of δ_p is

$$(16.24) \quad (\hat{\delta}_p)_\tau \equiv E_\tau \left[\frac{1}{p} \|\Theta\|^2 \mid X \right] = \frac{\tau^2}{1 + \tau^2} \left(\frac{\tau^2 \|X\|^2}{p(1 + \tau^2)} + 1 \right).$$

We now assert that *no matter how large τ^2 may be, the difference*

$$(16.25) \quad |(\hat{\delta}_p)_\tau - (\frac{1}{p} \|X\|^2 - 1)| = O(p^{-1/2})$$

uniformly in τ w.r.to the marginal \equiv unconditional distribution of X . By (16.22), this will imply that

$$(16.26) \quad |(\hat{\delta}_p)_\tau - \hat{\delta}_p| = 2 + O(p^{-1/2})$$

uniformly in τ^2 , so that *the proper Bayes estimator $(\hat{\delta}_p)_\tau$ does not approximate the “improper Bayes estimator” $\hat{\delta}_p$ no matter how large τ^2 may be, i.e., no matter how diffuse the proper prior $N_p(0, \tau^2 I_p)$. Instead, $(\hat{\delta}_p)_\tau$ approximates the UMVUE $\tilde{\delta}_p \equiv \frac{1}{p} \|X\|^2 - 1$.*

To establish (16.25), note that

$$X | \Theta \sim N_p(\Theta, I_p) \text{ and } \Theta \sim N_p(0, \tau^2 I_p) \implies X \sim N_p(0, (1 + \tau^2) I_p),$$

so $\|X\|^2 \sim (1 + \tau^2) \chi_p^2$. Therefore

$$\begin{aligned} & \left| \frac{\tau^2}{1 + \tau^2} \left(\frac{\tau^2 \|X\|^2}{p(1 + \tau^2)} + 1 \right) - \left(\frac{1}{p} \|X\|^2 - 1 \right) \right| \\ &= \left| \frac{\|X\|^2}{p} \left[\frac{\tau^4}{(1 + \tau^2)^2} - 1 \right] + \left[\frac{\tau^2}{1 + \tau^2} + 1 \right] \right| \\ &= \left| \left(\frac{1 + 2\tau^2}{1 + \tau^2} \right) \left(1 - \frac{\|X\|^2}{p(1 + \tau^2)} \right) \right| \\ &= \left| \left(\frac{1 + 2\tau^2}{1 + \tau^2} \right) \left(1 - \frac{\chi_p^2}{p} \right) \right| \\ (16.27) \quad & \leq 2 \cdot O(p^{-1/2}) \end{aligned}$$

uniformly in τ^2 , as asserted. □

17. The Elements of Statistical Decision Theory (non-sequential).

- *Statistical model* $\mathcal{P} \equiv \{P_\theta \mid \theta \in \Omega\}$, *sample space* \mathcal{X} , *data* $X \sim P_\theta$.
- *Action space* (\equiv *decision space*) $\mathcal{A} \equiv \{a\}$.

Examples:

(a) $\mathcal{A} = \Omega$: here $a \in \Omega$, so a is an *estimate* of θ .

(b) Partition $\Omega = \Omega_0 \dot{\cup} \Omega_1$. Let H_i be the hypothesis that $\theta \in \Omega_i$, $i = 0, 1$:

$$\mathcal{A} = \{ \{ \text{accept } H_0 \}, \{ \text{reject } H_0 \} \} \leftrightarrow \text{test } H_0 \text{ vs. } H_1.$$

(c) $\mathcal{A} = \{ \text{all intervals } (c, d) \}$: *confidence interval* for θ .

- *Loss function* $L(a, \theta) = \text{loss incurred by action } a \text{ when } \theta \text{ is true.}$

Examples:

(a) *Estimation*: $L(a, \theta) = (a - \theta)^2$ or $|a - \theta|$, etc.

$$(b) \text{ Testing: } L_{c_{01}, c_{10}}(a, \theta) = \begin{matrix} & \begin{matrix} H_0 & H_1 \end{matrix} \\ \begin{matrix} a = \text{"accept } H_0\text{"} \\ a = \text{"reject } H_0\text{"} \end{matrix} & \begin{pmatrix} 0 & c_{01} \\ c_{10} & 0 \end{pmatrix} \end{matrix},$$

where $c_{01}, c_{10} > 0$. $L_{1,1}$ is called *the 0-1 loss function*.

- *Decision rule* $d(x) : \mathcal{X} \mapsto \mathcal{A}$. If $X = x$ is observed, $d(x) = \text{action taken}$

Examples:

(a) *Estimator*: $d(x_1, \dots, x_n) = \bar{x}_n$ or x_{median} or s_n^2 , etc.

(b) *Test*: $d(x) = \begin{cases} \text{"accept } H_0\text{"} & \text{if } x \in \text{acceptance region } A \subset \mathcal{X}, \\ \text{"reject } H_0\text{"} & \text{if } x \in \text{rejection region } R \equiv \mathcal{X} \setminus A. \end{cases}$

Thus a (non-randomized) test \leftrightarrow a partitioning $\mathcal{X} = A \dot{\cup} R$.

- *Randomized decision rule* $d(x) : \mathcal{X} \mapsto \mathcal{P}(\mathcal{A})$, the set of all probability distributions on \mathcal{A} . If $X = x$ is observed, the action taken is determined by randomizing over \mathcal{A} according to the distribution $d(x)$.

For example, a randomized test has the following form:

$$d(x) \equiv d^\phi(x) = \begin{cases} \text{accept } H_0 & \text{w. probability } 1 - \phi(x), \\ \text{reject } H_0 & \text{w. probability } \phi(x), \end{cases}$$

where $\phi(x) : \mathcal{X} \mapsto [0, 1]$ is a *test function*, interpreted as follows:

$$(17.1) \quad \phi(x) = P[d^\phi \text{ rejects } H_0 \mid X = x].$$

Thus, d^ϕ is a non-randomized test with rejection region R iff $\phi(x) = I_R(x)$. Because any test $d \equiv d^\phi$, randomized or non-randomized, is completely determined by its test function $\phi(x)$, we drop “ d ” and simply refer to “the test ϕ ”.

Proposition 17.1. *The set Φ of all (randomized and non-randomized) test functions ϕ is convex: $\phi_1, \phi_2 \in \Phi \Rightarrow \eta_1 \phi_1 + \eta_2 \phi_2 \in \Phi$ for all $\eta_1 > 0, \eta_2 > 0, \eta_1 + \eta_2 = 1$.*

Proof. This is immediate from (17.1), which says that Φ consists of *all* (measurable) functions ϕ on \mathcal{X} that satisfy $0 \leq \phi(x) \leq 1$. \square

• *Risk function of a decision rule d : $R_d(\theta) = E_\theta[L(d(X), \theta)]$, the average loss incurred by decision rule d when θ is the true parameter value.*

Examples:

(a) *Estimation:* $R_d(\theta) = E_\theta[(d(X) - \theta)^2]$ (MSE) or $E_\theta[|d(X) - \theta|]$.

(b) *Testing:* $R_\phi(\theta) = \begin{cases} c_{10} P_\theta[\phi \text{ rejects } H_0] & \text{if } \theta \in \Omega_0, \\ c_{01} P_\theta[\phi \text{ accepts } H_0] & \text{if } \theta \in \Omega_1, \end{cases}$

$$(17.2) \quad \equiv \begin{cases} c_{10} \pi_\phi(\theta) & \text{if } \theta \in \Omega_0, \\ c_{01} [1 - \pi_\phi(\theta)] & \text{if } \theta \in \Omega_1, \end{cases}$$

where

$$(17.3) \quad \pi_\phi(\theta) \equiv P_\theta[\phi \text{ rejects } H_0] \equiv E_\theta[\phi(X)]$$

is the *power function*. Obviously the *ideal power function* is

$$\pi_{\text{ideal}}(\theta) \equiv \begin{cases} 0 & \text{if } \theta \in \Omega_0, \\ 1 & \text{if } \theta \in \Omega_1. \end{cases}$$



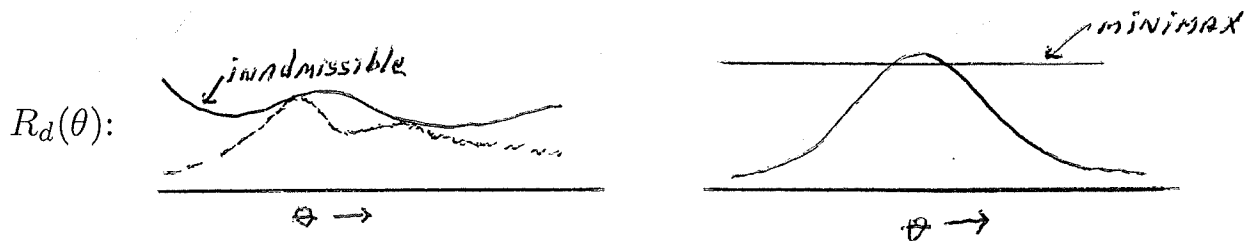
Typically this is unattainable for a finite sample size n but approached as $n \rightarrow \infty$.

- Decision rules are compared on the basis of their risk functions (\equiv expected loss). This is the standard formulation of decision theory (called *game theory* in economics). Thus:

d_1 dominates d_2 if $R_{d_1}(\theta) \leq R_{d_2}(\theta) \forall \theta \in \Omega$, with $<$ for at least one $\theta \in \Omega$.

d is *inadmissible* if it is dominated by some d' ; otherwise it is *admissible*.

d is *minimax* if $\max_{\theta \in \Omega} R_d(\theta) = \min_{d'} \max_{\theta \in \Omega} R_{d'}(\theta)$.



17.1. Bayes decision rules.

Suppose that the value of θ is the realization of a random variable Θ with range Ω . The *prior probability distribution* of Θ is specified by a pdf (or pmf) $\psi(\theta)$.

- The *Bayes risk* $r_d(\psi)$ is again the expected loss, but now averaged w.r.to both x and θ :

$$\begin{aligned} r_d(\psi) &= E\{L(d(X), \Theta)\} \\ (17.4) \quad &= E_\psi\{E[L(d(X), \Theta) \mid \Theta]\} \end{aligned}$$

$$(17.5) \quad \equiv E_\psi\{R_d(\Theta)\}.$$

- A *Bayes decision rule* for the prior distribution ψ is any decision rule d_ψ that minimizes the Bayes risk w.r.to ψ :

$$r_{d_\psi}(\psi) = \min_d r_d(\psi).$$

A Bayes rule need not exist and/or need not be unique. If it does exist, however, it is easy to specify. Reverse the iteration in (17.4) to obtain

$$\begin{aligned} r_d(\psi) &= E\{E[L(d(X), \Theta) \mid X]\} \\ (17.6) \quad &\equiv E[\text{expected posterior loss for } d(X)]. \end{aligned}$$

Thus if $X = x$, a Bayes rule $d_\psi(x)$ takes the (not necessarily unique) action $a(x)$ that minimizes the expected posterior loss. For this we use Bayes' formula to calculate the posterior (conditional) pdf $\psi(\theta | x)$ of $\Theta | X = x$.

Note: We can always find a non-randomized Bayes rule, but sometimes a minimax rule *must* be randomized (combine Exercises 18.8 and 18.14(i)).

Examples:

(a) *Estimation:* if $L(a, \theta) = (a - \theta)^2$, the expected posterior loss is $E[(a - \Theta)^2 | X]$, which is minimized³⁹ when $a \equiv a(x)$ is the mean of the posterior distribution of $\Theta | X = x$. Thus the Bayes estimator is the *posterior mean*:

$$(17.7) \quad d_\psi(x) = E[\Theta | X = x].$$

If $L(a, \theta) = |a - \theta|$, then the Bayes estimator is the *posterior median*:

$$(17.8) \quad d_\psi(x) = \text{median}[\Theta | X = x].$$

(b) *Hypothesis testing with loss $L_{c_{01}, c_{10}}$:* the expected posterior loss is

$$E[L_{c_{01}, c_{10}}(a, \Theta) | X = x] = \begin{cases} c_{01}P[\Theta \in \Omega_1 | X = x] & \text{if } a = \text{"accept } H_0\text{"}, \\ c_{10}P[\Theta \in \Omega_0 | X = x] & \text{if } a = \text{"reject } H_0\text{"}. \end{cases}$$

Thus a Bayes test ϕ_ψ must have the form

$$(17.9) \quad \phi_\psi(x) = \begin{cases} \text{"accept } H_0\text{"} & \text{if } c_{01}P[\Theta \in \Omega_1 | X = x] < c_{10}P[\Theta \in \Omega_0 | X = x], \\ \text{"reject } H_0\text{"} & \text{if } c_{01}P[\Theta \in \Omega_1 | X = x] > c_{10}P[\Theta \in \Omega_0 | X = x], \\ \text{either} & \text{if } c_{01}P[\Theta \in \Omega_1 | X = x] = c_{10}P[\Theta \in \Omega_0 | X = x]. \end{cases}$$

Equivalently, the Bayes test compares the *posterior odds ratio* and *cost ratio*:

$$(17.10) \quad \phi_\psi(x) = \begin{cases} \text{"accept } H_0\text{"} & \text{if } \frac{P[\Theta \in \Omega_1 | X = x]}{P[\Theta \in \Omega_0 | X = x]} < \frac{c_{10}}{c_{01}}, \quad \text{etc.} \end{cases}$$

Special case: $\Omega_0 \equiv \{\theta_0\}$ and $\Omega_1 \equiv \{\theta_1\}$ are both *simple*, i.e., singletons. Here Bayes formula (4.14) gives

$$P[\Theta = \theta_i | X = x] = f(x | \theta_i) \psi_i / f(x), \quad i = 0, 1,$$

³⁹ Note that this minimizing value of a is unique [verify].

where $f(x | \theta_i)$ is the pdf or pmf specified by the statistical model for X and $\psi_i \equiv P[\Theta = \theta_i]$, $i = 0, 1$, are the prior probabilities. Thus (17.10) becomes

$$(17.11) \quad \phi_\psi(x) = \begin{cases} \text{"accept } H_0" & \text{if } \frac{f(x|\theta_1)}{f(x|\theta_0)} < \frac{\psi_0 c_{10}}{\psi_1 c_{01}}, \quad \text{etc. ,} \end{cases}$$

which depends solely on the likelihood ratio $\frac{f(x|\theta_1)}{f(x|\theta_0)}$ (also see Prop. 18.4).

Example 17.2. (*Discriminating between two multivariate normal distributions with common covariance matrix*) Suppose we observe $X \sim N_p(\mu, \Sigma)$ and wish to test $\mu = \mu_0$ vs. $\mu = \mu_1$. For simplicity suppose that Σ is known (and positive definite). Assume prior probabilities ψ_0, ψ_1 for μ_0, μ_1 , respectively. Then

$$f(x | \mu_i) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)}$$

so the log likelihood ratio (LLR) is given by [verify]

$$(17.12) \quad \log \left[\frac{f(x | \mu_1)}{f(x | \mu_0)} \right] = (\mu_1 - \mu_0)'\Sigma^{-1}x - \frac{1}{2} (\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0).$$

Thus the Bayes test $\phi_\psi(x)$ in (17.11) assumes the following form:

$$(17.13) \quad \phi_\psi(x) = \begin{cases} \text{"choose } \mu_0" & \text{if } (\mu_1 - \mu_0)'\Sigma^{-1}x < c^*, \\ \text{"choose } \mu_1" & \text{if } (\mu_1 - \mu_0)'\Sigma^{-1}x > c^*, \end{cases}$$

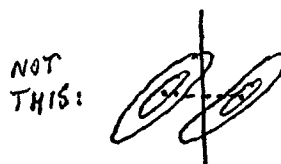
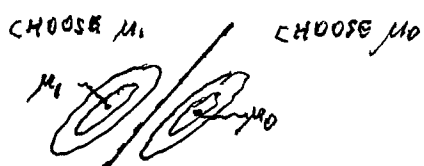
where

$$(17.14) \quad c^* = \log \left(\frac{\psi_0 c_{10}}{\psi_1 c_{01}} \right) + \frac{1}{2} (\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0).$$

The linear function $(\mu_1 - \mu_0)'\Sigma^{-1}x \equiv d'x$ is called *Fisher's discriminant function*. The Bayes test ϕ_ψ partitions \mathbf{R}^p into two halfspaces $\{x | d'x < c^*\}$ and $\{x | d'x > c^*\}$, then chooses μ_0 or μ_1 accordingly [see figure below]. \square

Exercise 17.3. In Example 17.2, find the two error probabilities

$$(17.15) \quad P[\phi_\psi \text{ chooses } \mu_1 | \mu_0 \text{ true}] \quad \text{and} \quad P[\phi_\psi \text{ chooses } \mu_0 | \mu_1 \text{ true}].$$



17.2. Admissible Bayes estimators.

Proposition 17.4. A unique Bayes rule w.r.to a general loss function L is admissible w.r.to L .

Proof. Obvious [verify].

Example 17.5. As in Exercise 16.6, suppose that

$$(17.16) \quad X \mid \Theta \sim \text{Binomial}(n, \theta),$$

$$(17.17) \quad \Theta \sim \text{Beta}(\alpha, \beta) \quad \text{for } \alpha, \beta > 0.$$

For the quadratic loss function $L(a, \theta) = (a - \theta)^2$, the Bayes estimator is

$$(17.18) \quad d_{\alpha, \beta}(X) = E[\Theta \mid X] = \frac{X + \alpha}{n + \alpha + \beta}$$

by (16.16). Since $d_{\alpha, \beta}(X)$ is the unique Bayes estimator for the prior (17.17) (see Footnote 39), it is admissible by Proposition 17.4. \square

Exercise 17.6. What about the admissibility of the unbiased MLE $\frac{X}{n}$?
By (17.18),

$$(17.19) \quad \frac{X}{n} = \lim_{\alpha, \beta \rightarrow 0} d_{\alpha, \beta}(X),$$

a limit of Bayes estimators, but Proposition 17.4 is not directly applicable. Suppose, however, that we change the loss function to the scaled quadratic loss function

$$(17.20) \quad \tilde{L}(a, \theta) = \frac{(a - \theta)^2}{\theta(1 - \theta)}.$$

Show that if $\alpha, \beta \geq 1$ and we replace quadratic loss L by \tilde{L} , then from (17.6) the unique Bayes estimator becomes

$$(17.21) \quad \tilde{d}_{\alpha, \beta}(X) = \frac{X + \alpha - 1}{n + \alpha + \beta - 2}.$$

Now set $\alpha = \beta = 1$ to obtain $\tilde{d}_{1,1}(X) = \frac{X}{n}$, which shows that $\frac{X}{n}$ is admissible w.r.to the loss function \tilde{L} . Finally, show that admissibility w.r.to L is equivalent to admissibility w.r.to \tilde{L} , hence $\frac{X}{n}$ is admissible w.r.to the ordinary MSE criterion. \square

Example 17.7. Let X_1, \dots, X_n be a random sample from $N_1(\theta, \sigma^2)$ with σ^2 known and let Θ have prior distribution $\psi_\tau \equiv N_1(\eta, \tau^2)$ with η and τ known. From Example 16.2, the Bayes estimator of θ (w.r.to MSE) is

$$(17.22) \quad d_\tau(\bar{X}_n) = E[\Theta \mid \bar{X}_n] = \frac{\frac{n\bar{X}_n}{\sigma^2} + \frac{\eta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},$$

a weighted average of \bar{X}_n and η . Note that, similar to (17.19),

$$(17.23) \quad \lim_{\tau \rightarrow \infty} d_\tau(X) = \bar{X}_n,$$

i.e., the Bayes estimator converges to the unbiased MLE \bar{X}_n as $\tau \rightarrow \infty$, that is, as the prior distribution ψ_τ becomes increasingly diffuse. Unlike (17.20) in Exercise 17.6, however, there is no obvious modification \tilde{L} of the quadratic loss function $L(a, \theta) = (a - \theta)^2$ such that \bar{X}_n is the Bayes estimator w.r.to \tilde{L} . Thus a different argument is needed to establish the admissibility of \bar{X}_n w.r.to L . One approach is the following:

If \bar{X}_n is inadmissible w.r.to L , it is dominated w.r.to MSE by some estimator d , that is,

$$(17.24) \quad R_d(\theta) \leq R_{\bar{X}_n}(\theta) \equiv \frac{\sigma^2}{n}$$

with strict inequality at some θ^* . Because $\{N_1(\theta, \sigma^2) \mid -\infty < \theta < \infty\}$ is an exponential family and the quadratic loss function $L(a, \theta)$ is continuous in θ , both risk functions are continuous in θ , so $\exists \epsilon > 0$ s.t.

$$(17.25) \quad R_d(\theta) < \frac{\sigma^2}{n} - \epsilon \quad \text{if} \quad |\theta - \theta^*| < \epsilon.$$

We shall show that the Bayes risk $r_d(\psi_\tau)$ of d satisfies

$$(17.26) \quad r_d(\psi_\tau) < \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2} \equiv r_{\bar{X}_n}(\psi_\tau) \quad \text{for sufficiently large } \tau,$$

(recall (16.10) where d_τ was denoted as $\hat{\theta}$). Since (17.26) contradicts the Bayesian optimality of d_τ w.r.to ψ_τ , we can conclude that \bar{X}_n is admissible.

First note that (17.26) is equivalent to

$$(17.27) \quad \frac{\sigma^2}{n} - r_d(\psi_\tau) > \frac{\sigma^4}{n(\sigma^2 + n\tau^2)} \text{ for sufficiently large } \tau.$$

Now let $I_{(\theta^* \pm \epsilon)}$ denote the indicator function of $(\theta^* - \epsilon, \theta^* + \epsilon)$. Then

$$\begin{aligned} & \frac{\sigma^2}{n} - r_d(\psi_\tau) \\ & \equiv \frac{\sigma^2}{n} - E_{\psi_\tau} [R_d(\Theta)] \\ & = E_{\psi_\tau} \left[\left(\frac{\sigma^2}{n} - R_d(\Theta) \right) I_{(\theta^* \pm \epsilon)}(\Theta) + \left(\frac{\sigma^2}{n} - R_d(\Theta) \right) I_{(\theta^* \pm \epsilon)^c}(\Theta) \right] \\ & \geq \epsilon P[|\Theta - \theta^*| < \epsilon] \quad \text{[by (17.24) and (17.25)]} \\ & = \frac{\epsilon}{\sqrt{2\pi}\tau} \int_{\theta^* - \epsilon}^{\theta^* + \epsilon} e^{-\frac{(\theta - \eta)^2}{2\tau^2}} d\theta \\ (17.28) \quad & \geq \frac{2\epsilon^2}{\sqrt{2\pi}\tau} \min \left(e^{-\frac{(\theta^* + \epsilon - \eta)^2}{2\tau^2}}, e^{-\frac{(\theta^* - \eta - \epsilon)^2}{2\tau^2}} \right) \\ & > \frac{\sigma^4}{n(\sigma^2 + n\tau^2)} \end{aligned}$$

for sufficiently large τ , which implies (17.27). \square

Remark 17.8. Extend Example 17.7 to \mathbf{R}^p as follows. Let X_1, \dots, X_n be a random sample from $N_p(\theta, \sigma^2 I_p)$ with σ^2 known and let Θ have prior distribution $N_p(\eta, \tau^2 I_p)$ with η and τ^2 known ($\eta \in \mathbf{R}^p$). The above proof of the admissibility of \bar{X}_n does *not* extend to \mathbf{R}^p , however, because (17.28) is replaced by $O(\tau^{-p})$ [verify], whereas the right-hand side of (17.27) remains $O(\tau^{-2})$ [verify]. In fact, Charles Stein (1956, 1962) showed that \bar{X}_n is *inadmissible* for $p \geq 3$, where it is dominated by the renowned *James-Stein* estimator – see §22. (He also showed that \bar{X}_n is admissible for $p = 2$ by a different argument, based on the Information Inequality.) \square

(recall (16.10) where d_τ was denoted as $\hat{\theta}$). Since (17.26) contradicts the Bayesian optimality of d_τ w.r.to ψ_τ , we can conclude that \bar{X}_n is admissible.

First note that (17.26) is equivalent to

$$(17.27) \quad \frac{\sigma^2}{n} - r_d(\psi_\tau) > \frac{\sigma^4}{n(\sigma^2 + n\tau^2)} \text{ for sufficiently large } \tau.$$

Now let $I_{(\theta_1 \pm \epsilon)}$ denote the indicator function of $(\theta_1 - \epsilon, \theta_1 + \epsilon)$. Then

$$\begin{aligned} & \frac{\sigma^2}{n} - r_d(\psi_\tau) \\ & \equiv \frac{\sigma^2}{n} - E_{\psi_\tau} [R_d(\Theta)] \\ & = E_{\psi_\tau} \left[\left(\frac{\sigma^2}{n} - R_d(\Theta) \right) I_{(\theta_1 \pm \epsilon)}(\Theta) + \left(\frac{\sigma^2}{n} - R_d(\Theta) \right) I_{(\theta_1 \pm \epsilon)^c}(\Theta) \right] \\ & \geq \epsilon P[|\Theta - \theta_1| < \epsilon] \quad [\text{by (17.24) and (17.25)}] \\ & = \frac{\epsilon}{\sqrt{2\pi} \tau} \int_{\theta_1 - \epsilon}^{\theta_1 + \epsilon} e^{-\frac{(\theta - \theta_0)^2}{2\tau^2}} d\theta \\ (17.28) \quad & \approx \frac{2\epsilon^2}{\sqrt{2\pi} \tau} \\ & > \frac{\sigma^4}{n(\sigma^2 + n\tau^2)} \end{aligned}$$

for sufficiently large τ , which implies (17.27). \square

Remark 17.8. Extend Example 17.7 to \mathbf{R}^p as follows. Let X_1, \dots, X_n be a random sample from $N_p(\theta, \sigma^2 I_p)$ with σ^2 known and let Θ have prior distribution $N_p(\theta_0, \tau^2 I_p)$ with θ_0 and τ^2 known ($\theta_0 \in \mathbf{R}^p$). The above proof of the admissibility of \bar{X}_n does *not* extend to \mathbf{R}^p , however, because (17.28) is replaced by $O(\tau^{-p})$ [verify], whereas the right-hand side of (17.27) remains $O(\tau^{-2})$ [verify]. In fact, Charles Stein (1956, 1962) showed that \bar{X}_n is inadmissible for $p \geq 3$, where it is dominated by the renowned *James-Stein* estimator – see §22. (He also showed that \bar{X}_n is admissible for $p = 2$ by a different argument, based on the Information Inequality.) \square

18. Testing Statistical Hypotheses.

18.1 Testing a simple hypothesis vs. a simple alternative.

In this section we consider the simplest nontrivial decision problem, testing a *simple* hypothesis H_0 vs. a *simple* alternative H_1 , i.e., H_i consists of a single distribution with pdf f_i , $i = 0, 1$. Thus we wish to test

$$(18.1) \quad H_0 : X \sim f_0 \quad \text{vs.} \quad H_1 : X \sim f_1.$$

Here both the action space and parameter space have exactly two members and we can completely characterize the set of all admissible decision rules. Unless otherwise specified we shall assume the 0-1 loss function.

The risk function $R \equiv R_\phi$ is now a risk *vector* (recall (17.2)):

$$(18.2) \quad R_\phi \equiv (R_0, R_1) = (P_0[\phi \text{ rejects } H_0], P_1[\phi \text{ accepts } H_0])$$

$$= (E_0[\phi(X)], 1 - E_1[\phi(X)])$$

$$(18.3) \quad \equiv (\pi_\phi(0), 1 - \pi_\phi(1)).$$

where $P_0 \equiv P_{f_0}$, $P_1 \equiv P_{f_1}$, $E_0 \equiv E_{f_0}$, $E_1 \equiv E_{f_1}$. The *ideal risk vector* is $(0,0)$. For this simple testing problem we can visualize the set of all attainable risk vectors:

$$(18.4) \quad \mathcal{R} \equiv \{R_\phi \mid \phi \in \Phi\}$$

as a subset of the unit square in \mathbf{R}^2 [see figure].

The following result is fundamental:

Proposition 18.1. (i) \mathcal{R} is a convex subset of \mathbf{R}^2 .

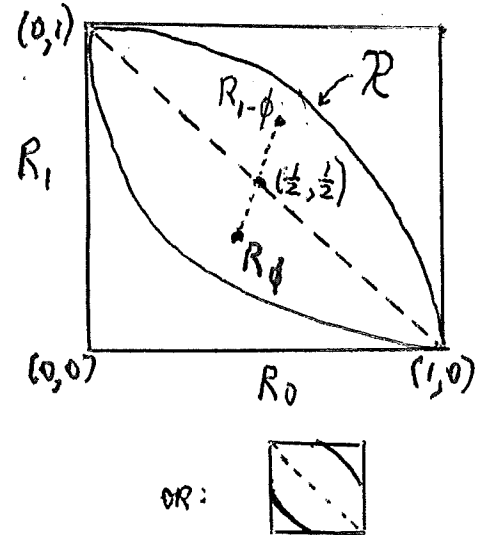
(ii) \mathcal{R} contains the diagonal $\{(\eta, 1 - \eta) \mid 0 \leq \eta \leq 1\}$.

(iii) \mathcal{R} is symmetric about $(\frac{1}{2}, \frac{1}{2})$.

(iv) \mathcal{R} is a closed subset of \mathbf{R}^2 .

Proof. (i) It follows from the convexity of Φ (see Proposition 17.1), from (18.2), and the linearity of expectation that if $R_{\phi_1}, R_{\phi_2} \in \mathcal{R}$ then [verify]

$$\eta_1 R_{\phi_1} + \eta_2 R_{\phi_2} = R_{\eta_1 \phi_1 + \eta_2 \phi_2} \in \mathcal{R} \quad \text{if } \eta_1 + \eta_2 = 1.$$



(ii) Let $\phi_0(x) \equiv 0$ (resp., $\phi_1(x) \equiv 1$) be the trivial test that ignores the data x and *always* accepts H_0 (resp., H_1). Then from (18.2), $R_{\phi_0} = (0, 1)$ and $R_{\phi_1} = (1, 0)$, so (ii) follows from the convexity of \mathcal{R} .

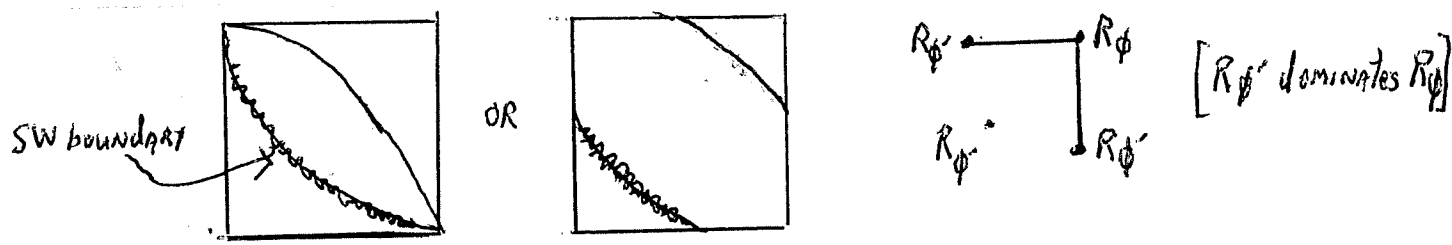
(iii) $\phi \in \Phi \Rightarrow (1 - \phi) \in \Phi$, so $R_\phi \in \mathcal{R} \Rightarrow R_{1-\phi} \in \mathcal{R}$. But

$$\frac{1}{2}R_\phi + \frac{1}{2}R_{1-\phi} = R_{\frac{\phi}{2} + \frac{1-\phi}{2}} = R_{\frac{1}{2}} = \left(\frac{1}{2}, \frac{1}{2}\right),$$

so (iii) holds.

(iv) This follows from the Weak Compactness Theorem of functional analysis (see Lehmann, *TSH* Appendix 4). \square

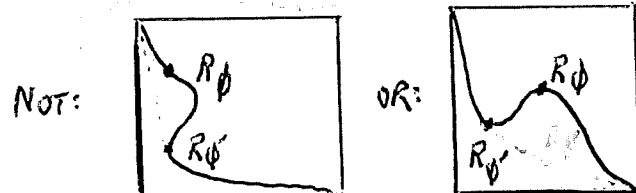
Since \mathcal{R} is closed it contains its boundary, in particular \mathcal{R} contains its *southwest (SW) boundary* [see figure], which corresponds to the risk vectors of all admissible (non-dominated) tests. The compactness of \mathcal{R} guarantees that its SW boundary is nonempty [Exercise 18.2(i)], so *admissible tests exist!*



Exercise 18.2. (i) Show that the SW boundary of \mathcal{R} is nonempty.

Hint: Consider that $R \in \mathcal{R}$ closest to $(0, 0)$.

(ii) Show that the SW boundary of \mathcal{R} is the graph of a strictly decreasing function:

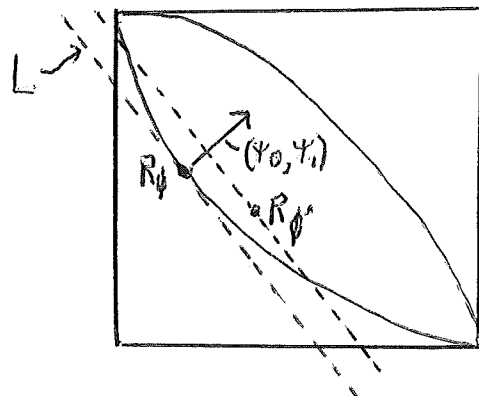


(iii) Show that there exists a *unique* admissible test ϕ^* iff ϕ^* is a perfect test, i.e., $R_{\phi^*} = (0, 0)$. Show that this occurs iff f_0 and f_1 have disjoint supports, so it is possible to distinguish between them *without error*. \square

Since H_0 and H_1 are simple, a prior distribution $\psi \equiv (\psi_0, \psi_1)$ is specified by the two prior probabilities $\psi_0 = P[H_0]$ and $\psi_1 = P[H_1]$ ($\psi_0 + \psi_1 = 1$).

Proposition 18.3. ϕ is admissible $\Rightarrow \phi$ is a Bayes test, i.e., $\phi = \phi_\psi$ for some prior $\psi \equiv (\psi_0, \psi_1)$. (The converse is not true in general.)

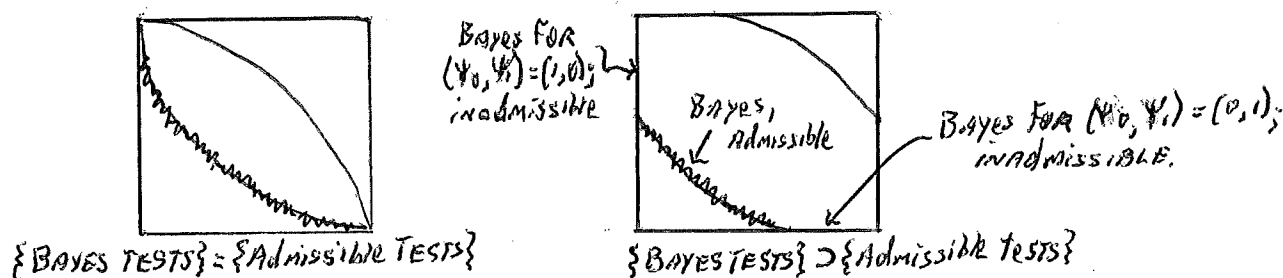
Proof. Suppose that ϕ is admissible, i.e., $R_\phi \in \text{SW boundary of } \mathcal{R}$. Since \mathcal{R} is convex, there exists a supporting line (\equiv tangent line) L for \mathcal{R} that passes through R_ϕ . Since the SW boundary is the graph of a decreasing function, this line has negative slope, so $L = \{(x_0, x_1) \mid \psi_0 x_0 + \psi_1 x_1 = c\}$ for some $\psi_0, \psi_1 \geq 0$, $\psi_0 + \psi_1 > 0$, $c \geq 0$. Dividing by $\psi_0 + \psi_1$ if necessary, we can ensure that $\psi_0 + \psi_1 = 1$. We shall show that ϕ is Bayes for the prior distribution $\psi \equiv (\psi_0, \psi_1)$:



Because L is a tangent line to \mathcal{R} that passes through its SW boundary point $R_\phi \equiv (R_0, R_1)$, the Bayes risk of ϕ w.r.to ψ satisfies

$$r_\phi(\psi) \equiv \psi_0 R_0 + \psi_1 R_1 \leq \psi_0 R'_0 + \psi_1 R'_1 \equiv r_{\phi'}(\psi)$$

for any other $R_{\phi'} \equiv (R'_0, R'_1) \in \mathcal{R}$, i.e., for any other test $\phi' \in \Phi$. This implies that $\phi = \phi_\psi$, i.e., ϕ is a Bayes test for ψ . \square



Proposition 18.4. $\phi \equiv \phi_\psi$ is a Bayes test $\Rightarrow \phi$ is a likelihood ratio (LR) test, i.e., has the form (recall (17.11))

$$(18.5) \quad \phi(x) \equiv \phi_c(x) = \begin{cases} 0 & (\equiv \text{accept } H_0) & \text{if } \lambda(x) \equiv \frac{f_1(x)}{f_0(x)} < c, \\ 1 & (\equiv \text{reject } H_0) & \text{if } \lambda(x) \equiv \frac{f_1(x)}{f_0(x)} > c, \\ \gamma(x) & (\equiv \text{randomize}) & \text{if } \lambda(x) \equiv \frac{f_1(x)}{f_0(x)} = c \end{cases}$$

for some $c \in [0, \infty]$ and some (measurable) $0 \leq \gamma \leq 1$.

Proof. From (17.9), a Bayes test $\phi_\psi(x)$ must have the form

$$(18.6) \quad \phi_\psi(x) = \begin{cases} 0 & \text{if } P[H_1 | X = x] < P[H_0 | X = x], \\ 1 & \text{if } P[H_1 | X = x] > P[H_0 | X = x], \\ \gamma(x) & \text{if } P[H_1 | X = x] = P[H_0 | X = x]. \end{cases}$$

By Bayes' formula, however [verify!],

$$(18.7) \quad P[H_1 | X = x] = \frac{f_1(x)\psi_1}{f_0(x)\psi_0 + f_1(x)\psi_1},$$

$$(18.8) \quad P[H_0 | X = x] = \frac{f_0(x)\psi_0}{f_0(x)\psi_0 + f_1(x)\psi_1},$$

so (18.5) holds with $c = \frac{\psi_0}{\psi_1}$. □

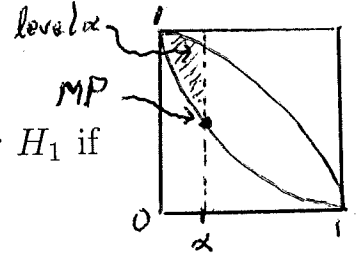
Definition 18.5. *The Neyman-Pearson (NP) criterion.* Fix $0 < \alpha < 1$. A test ϕ is *level α* (resp., *size α*) for H_0 if [see figure]

$$\pi_\phi(0) \equiv E_0[\phi(X)] \leq \alpha \quad (\text{resp., } \pi_\phi(0) = \alpha).$$

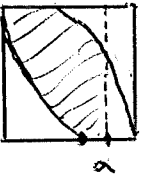
A level α test for $H_0 : \theta \in \Omega_0$ is *most powerful (MP) level α* for H_1 if

$$\pi_\phi(1) \equiv E_1[\phi(X)] = \sup_{\phi'} \pi_{\phi'}(1),$$

where ϕ' ranges over all level α tests for H_0 . □



Propositions 18.3 and 18.4 together imply that *any admissible test, i.e., any test with risk vector on the SW boundary of \mathcal{R} , must be a LR test of the form (18.5)*. Furthermore, if $0 < \alpha < 1$ then the risk vector of any MP level α test ϕ must lie on the SW boundary (provided that its power is < 1 , i.e., $E_1(\phi) < 1$ [see Figure]). Thus *any such MP level α test must be a LR test*. The *Neyman-Pearson Lemma* provides the converse.



Theorem 18.6. The Neyman-Pearson Lemma. *Let ϕ be a LR test of the form (18.5) with $c < \infty$ and set $\alpha = E_0[\phi(X)]$. Then ϕ is a MP level α test for testing H_0 vs H_1 . That is, if ϕ' is any other test such that*

$$(18.9) \quad E_0[\phi'(X)] \leq \alpha,$$

then

$$(18.10) \quad E_1[\phi'(X)] \leq E_1[\phi(X)].$$

Proof. By assumption we have

$$(18.11) \quad \int \phi(x) f_0(x) dx = \alpha,$$

$$(18.12) \quad \int \phi'(x) f_0(x) dx \leq \alpha.$$

Thus

$$\begin{aligned} & E_1[\phi(X)] - E_1[\phi'(X)] \\ &= \int \phi(x) f_1(x) dx - \int \phi'(x) f_1(x) dx \\ &= \int_{\{\frac{f_1(x)}{f_0(x)} > c\}} \underbrace{(\phi - \phi')}_{\geq 0} f_1 + \int_{\{\frac{f_1(x)}{f_0(x)} < c\}} \underbrace{(\phi - \phi')}_{\leq 0} f_1 + \int_{\{\frac{f_1(x)}{f_0(x)} = c\}} (\phi - \phi') f_1 \\ &\geq \int_{\{\frac{f_1(x)}{f_0(x)} > c\}} (\phi - \phi') c f_0 + \int_{\{\frac{f_1(x)}{f_0(x)} < c\}} (\phi - \phi') c f_0 + \int_{\{\frac{f_1(x)}{f_0(x)} = c\}} (\phi - \phi') c f_0 \\ &= c \int (\phi - \phi') f_0 \geq c(\alpha - \alpha) = 0 \end{aligned}$$

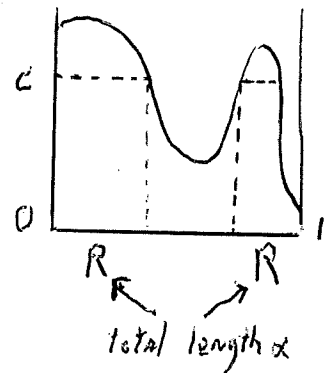
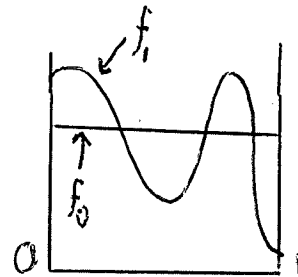
by (18.11) and (18.12), so (18.10) holds. \square

Explanation why LR tests are MP:

Suppose that $f_0 = \text{Uniform}[0, 1]$. Restrict attention to non-randomized tests $\phi(x) = I_R(x)$. The size constraint (18.9) becomes $\int_R dx \leq \alpha$, i.e., $\text{length}(R) \leq \alpha$. Subject to this constraint we wish to find R to maximize $\int_R f_1(x) dx$. Clearly [see figure] we should choose R to be that interval (or union of intervals) of total length α on which $f_1(x)$ achieves its largest values. Thus we should choose R to be of the form

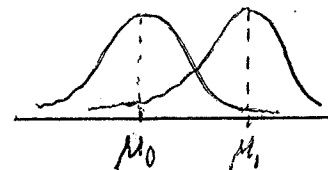
$$R_c \equiv \{x \mid f_1(x) \geq c\} \equiv \left\{x \mid \frac{f_1(x)}{f_0(x)} \geq c\right\},$$

where c is chosen so that $\text{length}(R_c) = \alpha$, which is a LR test.



Example 18.7. Let $X \equiv (X_1, \dots, X_n)$ be an i.i.d. sample from $N_1(\mu, \sigma^2)$ with σ^2 known. We wish to test

$$(18.13) \quad H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1, \quad \mu_0 < \mu_1.$$



The pdf of X is

$$\begin{aligned} f_\mu(x) &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \cdot e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2} \\ &= \text{const} \cdot e^{-\frac{1}{2\sigma^2} \sum x_i^2} \cdot e^{\frac{\mu}{\sigma^2} \sum x_i} \cdot e^{-\frac{\mu^2}{2\sigma^2}}, \end{aligned}$$

so the LR is

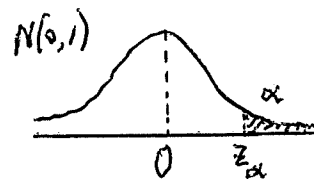
$$(18.14) \quad \frac{f_{\mu_1}(x)}{f_{\mu_0}(x)} = e^{\frac{(\mu_1 - \mu_0)n\bar{x}_n}{\sigma^2}} \cdot e^{\frac{\mu_0^2 - \mu_1^2}{2\sigma^2}},$$

a strictly increasing function of \bar{x}_n . Thus a LR test (18.5) has the form

$$(18.15) \quad \phi(x) = \begin{cases} 0 & (\equiv \text{accept } H_0) & \text{if } \bar{x}_n < c, \\ 1 & (\equiv \text{reject } H_0) & \text{if } \bar{x}_n > c, \\ \gamma(x) & (\equiv \text{randomize}) & \text{if } \bar{x}_n = c \end{cases}$$

Note that this is essentially a non-randomized test since $P[\bar{X}_n = c] = 0$. Finally, the MP level α test is given by the LR test (18.15) with $c \equiv c_\alpha$ chosen to satisfy

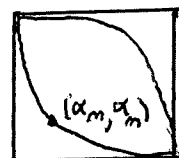
$$(18.16) \quad P_{\mu_0}[\bar{X}_n > c_\alpha] = \alpha,$$



hence $c_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$.

Note: Because $\bar{X}_n \sim N_1\left(\theta, \frac{\sigma^2}{n}\right)$ is sufficient for θ , we could work directly with the pdf $f_\mu(\bar{x}_n)$: the LR $\frac{f_{\mu_1}(\bar{x}_n)}{f_{\mu_0}(\bar{x}_n)}$ yields the same test as (18.15). \square

Exercise 18.8. In Example 18.7, show that the risk set $\mathcal{R} \equiv \mathcal{R}_n$ [see figure] fills out the unit square at an exponential rate as $n \rightarrow \infty$. That is, show that the minimax risk vector $(\alpha_n, \alpha_n) \rightarrow (0, 0)$ exponentially fast.



Example 18.9. In Example 18.7, suppose instead that we wish to test

$$(18.17) \quad H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1 : \sigma^2 = \sigma_1^2, \quad \sigma_0^2 < \sigma_1^2,$$

with μ known. Now the LR is

$$(18.18) \quad \frac{f_{\sigma_1^2}(x)}{f_{\sigma_0^2}(x)} = \left(\frac{\sigma_1^2}{\sigma_0^2}\right)^{\frac{n}{2}} \cdot e^{\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) \sum (x_i - \mu)^2},$$

a strictly increasing function of $\sum (x_i - \mu)^2$. Thus the size α LR test rejects H_0 in favor of H_1 if [verify]

$$(18.19) \quad \sum (X_i - \mu)^2 > \sigma_0^2 \chi_{n,\alpha}^2.$$

Example 18.10. The results of Examples 18.7 and 18.9 extend immediately to any 1-parameter exponential family, including Binomial, Poisson, Exponential, etc. That is, if $X \equiv (X_1, \dots, X_n)$ has joint pdf

$$(18.20) \quad f_\theta(x) = [a(\theta)]^n \exp \left[\theta \sum_{i=1}^n T(x_i) \right] \cdot \prod_{i=1}^n h(x_i),$$

where $\theta \in \Omega$ is a real parameter, then the MP level α test for testing $H_0 : \theta_0$ vs. $H_1 : \theta_1$ ($\theta_0 < \theta_1$) is a LR test given by

$$(18.21) \quad \phi(x) = \begin{cases} 0 & (\equiv \text{accept } H_0) & \text{if } T < c_\alpha, \\ 1 & (\equiv \text{reject } H_0) & \text{if } T > c_\alpha, \\ \gamma_\alpha(x) & (\equiv \text{randomize}) & \text{if } T = c_\alpha, \end{cases}$$

where $T = \sum T(X_i)$. If T is continuous then c_α can be chosen as in (18.16), but if T is discrete then we can select c_α and $\gamma_\alpha(x) \equiv \gamma_\alpha$ to satisfy

$$(18.22) \quad P_{\theta_0} [T > c_\alpha] + \gamma_\alpha P_{\theta_0} [T = c_\alpha] = \alpha.$$

Remark 18.11. Note that the MP test ϕ specified by (18.21) and (18.22) does not depend on the alternative $\theta_1 > \theta_0$, hence is a uniformly most powerful (UMP) level α for testing $H_0 : \theta = \theta_0$ vs. the extended one-sided alternative $H_0^> : \theta > \theta_0$ (also see Theorem 18.20.)

Remark 18.12. In most scientific applications, the attainment of an exact Type I error probability α is not relevant, so “randomization on the boundary” as in (18.22) would not be used. Instead one would report the *p-value* \equiv *attained significance level*

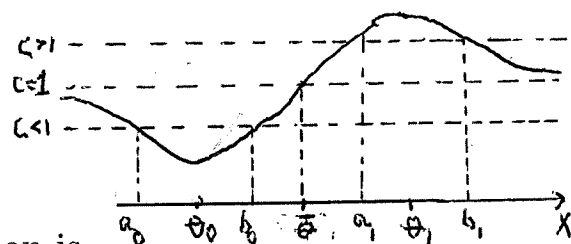
$$(18.23) \quad p \equiv p(T_{\text{observed}}) = P_{\theta_0}[T \geq T_{\text{observed}}].$$

A *small* *p-value* is evidence that favors H_1 over H_0 . [Also see §18.11, p.321.]

Note: Both H_0 and H_1 must be specified to form the LR $\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}$, and thence to obtain a *p-value*. It is not enough to specify H_0 alone. \square

Example 18.13. Let X be a single observation from a Cauchy(θ) distribution and suppose we wish to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ ($\theta_0 < \theta_1$). Then a LR test rejects H_0 iff [see figure]

$$(18.24) \quad \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \frac{1 + (x - \theta_0)^2}{1 + (x - \theta_1)^2} > c.$$



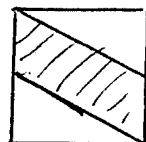
Thus, setting $\bar{\theta} = (\theta_0 + \theta_1)/2$, the rejection region is

$$(18.25) \quad \begin{cases} \text{an interval } (a_1, b_1) \subset (\bar{\theta}, \infty) & \text{if } c > 1, \\ \text{a complement } (a_0, b_0)^c \text{ with } (a_0, b_0) \subset (-\infty, \bar{\theta}) & \text{if } c < 1, \\ \text{the one-sided interval } (\bar{\theta}, \infty) & \text{if } c = 1. \end{cases}$$

This irregular behavior is due to the fact that the LR (18.24) is *not* a monotone function of x . By contrast, the LR is a strictly monotone function of T in all 1-parameter exponential families. (See the discussion of monotone likelihood ratio in §18.2.) \square

The following exercise presents an example where the LR is monotone but not strictly monotone. Compare the shapes of the risk sets to that depicted in Exercise 18.8.

Exercise 18.14. (i) Let $X = (X_1, \dots, X_n)$ be an i.i.d sample from the Uniform $[0, \theta]$ distribution. Consider the problem of testing $H_0 : \theta = 1$ vs. $H_1 : \theta = 2$ with $n = 1$. Specify the SW boundary of the risk set \mathcal{R} . Specify the tests corresponding to the risk vectors on this SW boundary.



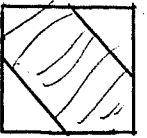
(ii) Repeat (i) for $n \geq 2$. Show that as $n \rightarrow \infty$, the risk set $\mathcal{R} \equiv \mathcal{R}_n$ fills out the unit square at an exponential rate.

(iii) Repeat (i) and (ii) for testing $H_0 : \theta = 2$ vs. $H_1 : \theta = 1$. Show that for $2^{-n} \leq \alpha < 1$, the MP level α test has power 1. Explain this behavior.



(iv) Repeat (i) and (ii) for testing $H_0 : X_i \sim U[0, 2]$ vs. $H_1 : X_i \sim U[1, 3]$.

Theorem 18.15. (*Consistency of MP and LR tests for i.i.d. samples.*) Consider the problem of testing



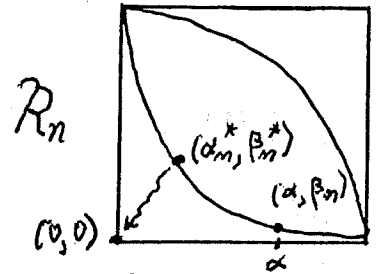
$$(18.26) \quad H_0 : X_i \sim f_0 \quad \text{vs.} \quad H_1 : X_i \sim f_1$$

based on an i.i.d. sample X_1, \dots, X_n . For each $0 < \alpha < 1$ the MP level α test ϕ_n is consistent, i.e. its power at H_1 approaches 1 as $n \rightarrow \infty$:

$$P_1[\phi_n(X_1, \dots, X_n) \text{ rejects } H_0] \rightarrow 1.$$

Proof. Note that this consistency is equivalent to

$$\beta_n \equiv P_1[\phi_n(X_1, \dots, X_n) \text{ accepts } H_0] \rightarrow 0.$$



By the geometry of the risk set \mathcal{R}_n [see figure], this will follow if we can exhibit a sequence of tests $\{\phi_n^* \equiv \phi_n^*(X_1, \dots, X_n)\}$ whose risk vectors

$$(18.27) \quad R_{\phi_n^*} \equiv (\alpha_n^*, \beta_n^*) \rightarrow (0, 0) \quad \text{as } n \rightarrow \infty.$$

Let ϕ_n^* be a LR test with $c = 1$:

$$(18.28) \quad \phi_n^*(x_1, \dots, x_n) = \begin{cases} 0 & (\equiv \text{accept } H_0) & \text{if } \lambda_n \equiv \frac{\prod f_1(x_i)}{\prod f_0(x_i)} \leq 1, \\ 1 & (\equiv \text{reject } H_0) & \text{if } \lambda_n \equiv \frac{\prod f_1(x_i)}{\prod f_0(x_i)} > 1. \end{cases}$$

Now apply⁴⁰ the WLLN and the inequality

$$E_0 \left\{ \log \left[\frac{f_1(X_i)}{f_0(X_i)} \right] \right\} \equiv -K(f_0, f_1) < 0$$

⁴⁰ Recall (14.26) in the proof of Wald's Theorem 14.7(i) for the consistency of the MLE when Ω is finite.

to see that

$$(18.29) \quad \alpha_n^* \equiv P_0 \left[\frac{\prod f_1(X_i)}{\prod f_0(X_i)} > 1 \right] = P_0 \left\{ \frac{1}{n} \sum \log \left[\frac{f_1(X_i)}{f_0(X_i)} \right] > 0 \right\} \rightarrow 0.$$

Similarly $\beta_n^* \rightarrow 0$, so (18.27) is established.

In fact, α_n^* and $\beta_n^* \rightarrow 0$ exponentially fast: by Markov's inequality,

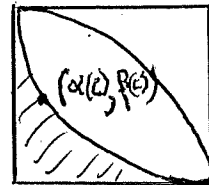
$$(18.30) \quad \begin{aligned} \alpha_n^* &= \Pr_0[\lambda_n > 1] = \Pr_0[\lambda_n^{1/2} > 1] \leq E_0(\lambda_n^{1/2}) \\ &= \int \cdots \int \left[\prod f_0(x_i) \right]^{1/2} \left[\prod f_1(x_i) \right]^{1/2} \Pi dx_i \equiv \rho^n \rightarrow 0, \end{aligned}$$

where $\rho = \int [f_0(x)]^{1/2} [f_1(x)]^{1/2} dx = 1 - \frac{1}{2} \int (f_0^{1/2} - f_1^{1/2})^2 < 1. \quad \square$

Exercise 18.16. Let $\lambda(X) \equiv \frac{f_1(X)}{f_0(X)}$ denote the LR statistic for testing $H_0 : X \sim f_0$ vs. $H_1 : X \sim f_1$. Assume that the range of λ is an interval (a, b) (necessarily $a < 1 < b$ [why?]) Let G_0 and G_1 be the cdfs of λ under H_0 and H_1 , respectively; i.e., for $a < c < b$,

$$G_0(c) = P_{f_0}[\lambda(X) \leq c] \equiv 1 - \alpha(c),$$

$$G_1(c) = P_{f_1}[\lambda(X) \leq c] \equiv \beta(c).$$



Assume that G_0 and G_1 are continuous and strictly increasing for $a < c < b$. Here $R_\phi \equiv (\alpha(c), \beta(c))$ is the risk vector of the LRT $\phi \equiv \phi_c$ in (18.5).

(i) Let $\bar{G}_i(c) = 1 - G_i(c)$, $i = 0, 1$. Show that

$$\begin{cases} G_1(c) < c G_0(c) & \text{if } a < c < 1, \\ \bar{G}_1(c) > c \bar{G}_0(c) & \text{if } 1 < c < b. \end{cases}$$

Conclude that $\alpha(c) + \beta(c) < 1$ and that $G_1(c) < G_0(c)$ for all $a < c < b$. The latter inequality shows that λ is stochastically larger under H_1 than under H_0 . In particular, $E_1(\lambda) > E_0(\lambda) = 1$.

(ii) Let $g_i(c) = G'_i(c)$, $i = 0, 1$. Show that $\frac{g_1(c)}{g_0(c)} = c$, $a < c < b$. [Since g_i is the pdf of λ under H_i , this result says that “the LR of the LR is the LR”.]

Hint: One method is to consider the moment generating functions of $l \equiv \log \lambda$ under H_0 and H_1 , but there is a short direct method.

(iii) Let “AUB” denote the area of the shaded region under the SW boundary of the risk set \mathcal{R} . Show that $\text{AUB} = \int_a^b G_1(t) dG_0(t) \equiv \int_a^b G_1(t) g'_0(t) dt$. Combine this with (i) to show that $\text{AUB} < \frac{1}{2}$.

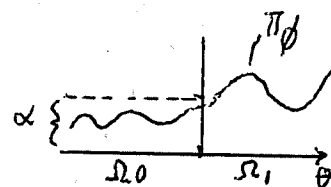
(iv) If X is an i.i.d. sample as in Theorem 18.15, show that $AUB_n \rightarrow 0$ as $n \rightarrow \infty$. [Hint: just apply (18.27).]

(v) Let X_1, \dots, X_n be i.i.d. $\text{Exponential}(\lambda)$ rvs with pdf $\lambda e^{-\lambda x}$, $0 < x < \infty$. Consider the problem of testing $\lambda = 1$ vs. $\lambda = \lambda_1$ ($\neq 1$). Find AUB_n for $n = 1$ and $n \geq 2$, and show that $AUB_n \rightarrow \infty$ at an exponential rate. \square

18.2. Testing composite hypotheses and/or alternatives.

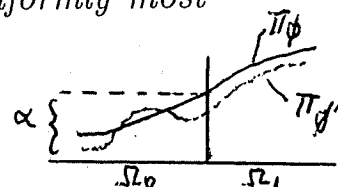
Definition 18.17. *The Neyman-Pearson (NP) criterion.* Fix $0 < \alpha < 1$. A test ϕ is *level α* (resp., *size α*) for testing $H_0 : \theta \in \Omega_0$ if

$$\sup_{\theta \in \Omega_0} \pi_\phi(\theta) \leq \alpha \quad (\text{resp., } \sup_{\theta \in \Omega_0} \pi_\phi(\theta) = \alpha).$$



A level α test ϕ for testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_1$ is *uniformly most powerful (UMP) level α* if

$$\pi_\phi(\theta) = \sup_{\phi' \text{ level } \alpha} \pi_{\phi'}(\theta) \quad \forall \theta \in \Omega_1.$$



The NP criterion treats H_0 and H_1 asymmetrically: we control the probability of a “Type I error” (reject H_0 when it is true) at level α , then minimize the probability of a “Type II error” (accept H_0 when false) subject to this constraint. This treats a Type I error as more serious than a Type II error.

Hypothesis-testing terminology reflects this asymmetry: H_0 is often called the “null hypothesis”, for example representing the effect of a standard treatment, while H_1 is called the “alternative hypothesis”, representing the effect of a new treatment. Strong evidence (indicated by a small p -value) is required to reject the null hypothesis in favor of the alternative.

Remark 18.18. The NP criterion requires only that the *supremum* of the Type I error probabilities be controlled for $\theta \in \Omega_0$ - it does not consider the detailed behavior of the power function on Ω_0 . For certain non-standard multiparameter testing problems where Ω_0 is *not* a linear subspace (or smooth manifold) - e.g., if Ω_0 is the complement of the positive orthant - this can lead to serious anomalies. See “The Emperor’s New Tests”, Perlman and Wu (1999), *Statistical Science*. \square

18.3. One-parameter testing problems with one-sided alternatives.

Let X have pdf or pmf $f_\theta(x)$ where θ is a real parameter whose parameter space Ω is an interval (possibly infinite). Consider the problems of testing

$$(18.31) \quad H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1^> : \theta > \theta_0,$$

$$(18.32) \quad H_0^< : \theta \leq \theta_0 \quad \text{vs.} \quad H_1^> : \theta > \theta_0.$$

In each case, $H_1^>$ is a *one-sided alternative*. We will show that UMP level α tests exist for these problems when $f_\theta(x)$ has monotone likelihood ratio.

Definition 18.19. $\{f_\theta(x)\}$ has (*strict*) *monotone likelihood ratio (MLR)* if there exists a real-valued statistic $T \equiv T(X)$ such that for each pair $\theta_1 < \theta_2 \in \Omega$, the LR $\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$ is a (strictly) increasing function of $T(x)$, i.e.,

$$(18.33) \quad \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = g_{\theta_1, \theta_2}(T(x))$$

where $g_{\theta_1, \theta_2}(t)$ is (strictly) increasing in t . □

Note: It follows from (18.33) and the Factorization Criterion that $T(X)$ is a real-valued sufficient statistic for θ . (Set $\theta_2 = \theta$ in (18.33).)

In Example 18.10 we saw that if $X \equiv (X_1, \dots, X_n)$ is an i.i.d. sample from any 1-parameter exponential family, then $f_\theta(x)$ has strictly MLR. By contrast, for $n \geq 2$ the Cauchy, double exponential, and logistic families do not have MLR. This can be verified by direct examination of their LRs (e.g. see Example 18.12), or it follows since we know that for each family the order statistics are minimal sufficient, hence no real-valued sufficient statistic can exist. Similarly, the Uniform $[\theta, \theta + 1]$ family cannot have MLR because $(X_{(1)}, X_{(n)})$ is a 2-dimensional minimal sufficient statistic.

Other strict MLR families include the noncentral $\chi_n^2(\theta)$, $F_{m,n}(\theta)$, and $t_n(\theta)$. The Uniform $[0, \theta]$ family has MLR but not strict MLR [verify – see Exercise 18.14].

Theorem 18.20. Let $f_\theta(x)$ have MLR in T and let $\phi \equiv \phi(T)$ be the test

$$(18.34) \quad \phi(t) = \begin{cases} 0 & \text{if } t < c_\alpha, \\ 1 & \text{if } t > c_\alpha, \\ \gamma_\alpha & \text{if } t = c_\alpha, \end{cases}$$

where c_α and γ_α are chosen to satisfy

$$(18.35) \quad P_{\theta_0} [T > c_\alpha] + \gamma_\alpha P_{\theta_0} [T = c_\alpha] = \alpha.$$

Then ϕ is UMP level α for testing (18.31) and (18.32).

Proof. By the MLR property, the test ϕ is a size α LR test for testing θ_0 vs. any fixed $\theta_1 > \theta_0$, hence is MP level α for θ_0 vs. θ_1 by the NP Lemma. Since ϕ does not depend on θ_1 , it must be UMP level α for (18.31), as in Remark 18.11.

Next consider (18.32). Because

$$\{\text{all level } \alpha \text{ tests for } H_0^\leq\} \subseteq \{\text{all level } \alpha \text{ tests for } H_0\},$$

and ϕ is the UMP level α test for H_0 vs. $H_1^>$, it suffices to verify that ϕ remains level α for H_0^\leq . That is, we need show that for all $\theta' < \theta_0$,

$$(18.36) \quad \alpha' \equiv E_{\theta'}[\phi(T)] \leq E_{\theta_0}[\phi(T)] \equiv \alpha.$$

Method 1. By invoking the MLR property and the NP Lemma as above, we see that ϕ is MP level α' for testing θ' vs. θ_0 . (Here θ' is the null hypothesis and θ_0 is the alternative.) However, the trivial purely randomized test $\phi' \equiv \alpha'$ satisfies

$$E_{\theta'}[\phi'] = E_{\theta_0}[\phi'] = \alpha',$$

hence is also level α' for θ' and has power α' at θ_0 . Since ϕ is MP level α' for θ' vs. θ_0 , this implies (18.36).

Method 2. From its definition (18.34), $\phi(t)$ is non-decreasing in t . Thus (18.36) is a consequence of the following lemma. \square

Lemma 18.21. Suppose that $f_\theta(x)$ has (strict) MLR in $T \equiv T(X)$ and that $g(t)$ is (strictly) increasing in t . Then $E_\theta[g(T)]$ is (strictly) increasing in θ . Therefore T is “stochastically increasing” in θ .

Proof. For simplicity, suppose $f_\theta(x) > 0$ for all $x \in \mathcal{X}$. For $\theta_2 > \theta_1$,

$$E_{\theta_2}[g(T)] = \int g(T(x))f_{\theta_2}(x)dx$$

$$\begin{aligned}
&= \int g[(T(x))] \left[\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} \right] f_{\theta_1}(x) dx \\
&= E_{\theta_1} \left\{ \underbrace{g(T)}_{\nearrow \text{ in } T} \underbrace{\left[\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} \right]}_{\nearrow \text{ in } T} \right\} \\
&\stackrel{*}{\geq} E_{\theta_1}[g(T)] \cdot E_{\theta_1} \left[\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} \right] \\
&= E_{\theta_1}[g(T)],
\end{aligned}$$

where (*) follows from Chebyshev's Other Inequality (Lemma 5.1). \square

Exercise 18.22. Prove Lemma 18.21 without assuming that $f_{\theta}(x) > 0$. \square

Note 1: Theorem 18.20 is also valid if (18.31) and (18.32) are replaced by

$$(18.37) \quad H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1^< : \theta < \theta_0,$$

$$(18.38) \quad H_0^{\geq} : \theta \geq \theta_0 \quad \text{vs.} \quad H_1^< : \theta < \theta_0,$$

and “<” and “>” are interchanged in (18.34) and (18.35). \square

Note 2: The Cauchy location family does not have MLR; UMP tests do not exist for these one-sided testing problems (recall Example 18.13). However, *locally most powerful* tests can be found via the NP Lemma. \square

18.4. One-parameter testing problems with two-sided alternatives.

Let X have pdf or pmf $f_{\theta}(x)$ where θ is a real parameter whose parameter space Ω is an interval (possibly infinite). Consider the problems of testing

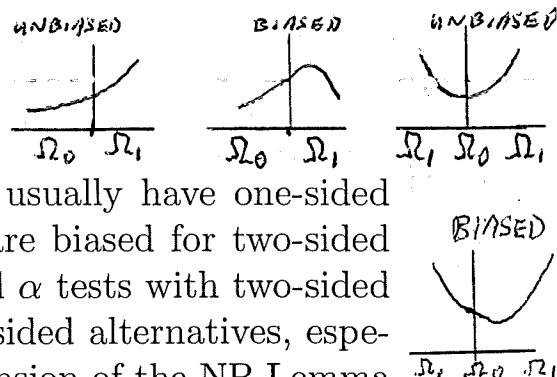
$$(18.39) \quad H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1^{\neq} : \theta \neq \theta_0,$$

$$(18.40) \quad H_{a,b} : a \leq \theta \leq b \quad \text{vs.} \quad H_{a,b}^c : \theta < a \text{ or } b < \theta.$$

Both H_1^{\neq} and $H_{a,b}^c$ are *two-sided alternatives*. In view of Theorem 18.20 and Note 1, UMP level α tests *do not exist* for these problems when $f_{\theta}(x)$ has monotone likelihood ratio [explain]. Instead, often we can find UMP *unbiased* level α tests for (18.39) and (18.40).

Definition 18.23. A test ϕ is *unbiased* for $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_1$ if it is more likely to reject H_0 when H_0 is false than when it is true. That is, its power function π_ϕ satisfies [see figures]

$$(18.41). \quad \sup_{\theta \in \Omega_0} \pi_\phi(\theta) \leq \inf_{\theta \in \Omega_1} \pi_\phi(\theta).$$



UMP level α tests for one-sided alternatives usually have one-sided rejection regions and are unbiased [why?], hence are biased for two-sided alternatives. Instead, UMP *unbiased* (UMPU) level α tests with two-sided rejection regions often can be constructed for two-sided alternatives, especially in 1-parameter exponential families. An extension of the NP Lemma is needed that incorporates not only the level constraint but also the unbiasedness constraint. This is straightforward but will not be included here.

Example 18.24. (a) As in Example 18.7, let X_1, \dots, X_n be an i.i.d. sample from $N_1(\mu, \sigma^2)$. Suppose we wish to test

$$(18.42) \quad H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1^\neq : \mu \neq \mu_0$$

with σ^2 known. The one-sided tests with rejection regions

$$\left\{ \bar{X}_n > \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \right\} \quad \text{and} \quad \left\{ \bar{X}_n < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha \right\}$$

are, respectively, UMP for H_0 vs. $H_1^>$ and for H_0 vs. $H_1^<$, but neither is unbiased for (18.42). Instead, the UMP unbiased level α test for (18.42) has the symmetric two-sided rejection region

$$|\bar{X}_n - \mu_0| > \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}.$$

(b) As in Example 18.9, suppose instead that we wish to test

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1^\neq : \sigma^2 \neq \sigma_0^2$$

with μ known. Here the UMP unbiased level α test has an (asymmetric!) two-sided rejection region

$$\left\{ \frac{\sum (X_i - \mu)^2}{\sigma_0^2} < c_1 \right\} \cup \left\{ \frac{\sum (X_i - \mu)^2}{\sigma_0^2} > c_2 \right\},$$

where $c_1 \equiv \chi_{n,1-\alpha_1}^2$ and $c_2 \equiv \chi_{n,\alpha_2}^2$ are chosen so that $c_1^n e^{-c_1} = c_2^n e^{-c_2}$ and $\alpha_1 + \alpha_2 = \alpha$. (See Lehmann *Testing Statistical Hypotheses* §4.2.)

18.5. One-parameter testing problems with nuisance parameters.

[Similar tests, Neyman structure, UMPU tests.]

18.6. Testing composite hypotheses; the general LRT.

The general problem is this: based on data $X \sim f_\theta$, test

$$(18.43) \quad H_0 : \theta \in \Omega_0 \quad \text{vs.} \quad H_1 : \theta \in \Omega_1 \equiv \Omega \setminus \Omega_0.$$

Case Ia: Ω_0 and Ω_1 are uniformly KL-separated:

$$(18.44a) \quad \inf_{\theta_0 \in \Omega_0, \theta_1 \in \Omega_1} K(\theta_0, \theta_1) > 0, \quad \inf_{\theta_0 \in \Omega_0, \theta_1 \in \Omega_1} K(\theta_1, \theta_0) > 0.$$

Case Ib: Ω_0 and Ω_1 are pointwise KL-separated:

$$(18.44b) \quad \inf_{\theta \in \Omega_1} K(\theta_0, \theta) > 0 \quad \forall \theta_0 \in \Omega_0, \quad \inf_{\theta \in \Omega_0} K(\theta_1, \theta) > 0 \quad \forall \theta_1 \in \Omega_1.$$

Clearly uniform separation \Rightarrow pointwise separation. Here the LRT is

$$(18.45) \quad \phi_n^*(x_1, \dots, x_n) = \begin{cases} 0 & (\equiv \text{accept } H_0) \quad \text{if } \frac{\prod f_{\hat{\theta}_1}(x_i)}{\prod f_{\hat{\theta}_0}(x_i)} \leq 1, \\ 1 & (\equiv \text{reject } H_0) \quad \text{if } \frac{\prod f_{\hat{\theta}_1}(x_i)}{\prod f_{\hat{\theta}_0}(x_i)} > 1, \end{cases}$$

where $\hat{\theta}_i$ is the MLE of θ under H_i . This LRT behaves similarly to the LRT for a simple hypothesis vs. a simple alternative. If $X \equiv (X_1, \dots, X_n)$ are i.i.d. observations $X_i \sim f_\theta(x_i)$, $f_\theta(x_i)$ is (upper semi-)continuous in θ , and

$$(18.46) \quad E_{\theta_0} \left\{ \sup_{\theta \in \Omega_1} \log \left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \right] \right\} < \infty, \quad E_{\theta_1} \left\{ \sup_{\theta \in \Omega_0} \log \left[\frac{f_\theta(X_i)}{f_{\theta_1}(X_i)} \right] \right\} < \infty$$

for each $\theta_0 \in \Omega_0$ and $\theta_1 \in \Omega_1$, respectively (recall (14.22)), then

$$\begin{aligned} (18.47) \quad & P_{\theta_0} \left[\frac{\prod f_{\hat{\theta}_1}(X_i)}{\prod f_{\hat{\theta}_0}(X_i)} < 1 \right] \geq P_{\theta_0} \left[\frac{\prod f_{\hat{\theta}_1}(X_i)}{\prod f_{\theta_0}(X_i)} < 1 \right] \\ & = P_{\theta_0} \left\{ \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_{\hat{\theta}_1}(X_i)}{f_{\theta_0}(X_i)} \right] < 0 \right\} = P_{\theta_0} \left\{ \sup_{\theta \in \Omega_1} \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \right] < 0 \right\} \\ & \equiv P_{\theta_0} \left\{ \sup_{\theta \in \Omega_1} H_n(\theta_0, \theta) < 0 \right\} \rightarrow 1 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

by (14.29) (for Ω_1) and (18.44a). Similarly $P_{\theta_1} \left[\frac{\prod f_{\hat{\theta}_1}(X_i)}{\prod f_{\hat{\theta}_0}(X_i)} > 1 \right] \rightarrow 1$ as $n \rightarrow \infty$. Thus as in the proof of Theorem 18.15, the LRT ϕ_n^* is consistent under both H_0 and H_1 [usually at an exponential rate].

Case II: $\Omega_0 \subset \Omega$ are subspaces (or submanifolds) of \mathbf{R}^k . Let $\theta = (\theta_1, \dots, \theta_k)$. Let $\Omega_0 \subset \Omega$ be smooth, relatively open manifolds in \mathbf{R}^k , determined by nested sets of differentiable constraints (with nonzero gradients):

$$(18.48) \quad \Omega = \{\theta \mid g_1(\theta) = \dots = g_r(\theta) = 0\},$$

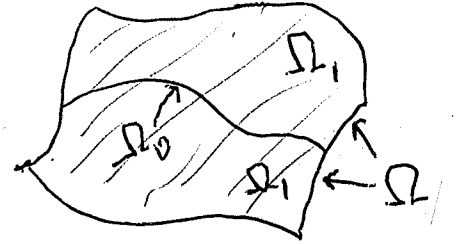
$$(18.49) \quad \Omega_0 = \{\theta \mid g_1(\theta) = \dots = g_r(\theta) = g_{r+1}(\theta) = \dots = g_{r+s}(\theta) = 0\},$$

so that

$$d \equiv \text{dimension}(\Omega) = k - r,$$

$$d_0 \equiv \text{dimension}(\Omega_0) = k - r - s,$$

$$(18.50) \quad d - d_0 \equiv \dim(\Omega) - \dim(\Omega_0) = s.$$



For such Ω_0 and Ω , (18.43) allows two-sided alternatives but *not* one-sided alternatives.)

For Case II testing problems the usual form⁴¹ of the LR statistic is

$$(18.51) \quad \lambda \equiv \lambda(x) = \frac{\sup_{\theta \in \Omega_0} f_{\theta}(x)}{\sup_{\theta \in \Omega} f_{\theta}(x)} \equiv \frac{f_{\hat{\theta}_0}(x)}{f_{\hat{\theta}}(x)},$$

where $\hat{\theta}_0$ and $\hat{\theta}$ are the MLEs of θ under Ω_0 and Ω respectively. Note that $0 \leq \lambda \leq 1$. The LRT takes the form

$$(18.52) \quad \phi(x) = \begin{cases} 0 & \text{if } \lambda > c, \quad [\text{Note: not } \lambda < c!] \\ 1 & \text{if } \lambda < c, \quad [\text{Note: not } \lambda > c!] \\ \gamma & \text{if } \lambda = c, \end{cases}$$

and the associated p -value is (cf. (18.112), p.321)

$$(18.53) \quad p \equiv p(\lambda_{\text{observed}}) = \sup_{\theta_0 \in \Omega_0} P_{\theta_0}[\lambda \leq \lambda_{\text{observed}}].$$

⁴¹ The main reason that Wilks used λ rather than $1/\lambda$ is that λ reduces to a beta statistic when testing normal linear models, as in Example 18.27.

Example 18.25. (*The 1-sample normal model.*) Let $X = (X_1, \dots, X_n)$ be i.i.d. with $X_i \sim N_1(\mu, \sigma^2)$, $i = 1, \dots, n$. Consider the problem of testing

$$(18.54) \quad H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0, \quad \text{with } \sigma^2 \text{ unknown.}$$

The pdf of X is

$$f_{\mu, \sigma^2}(x) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}.$$

The unrestricted parameter space Ω is $\{(\mu, \sigma^2) \mid -\infty < \mu < \infty, \sigma^2 > 0\}$ and from (14.18) the unrestricted MLE of (μ, σ^2) is given by

$$\hat{\mu} = \bar{x}_n, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2,$$

so [verify]

$$(18.55) \quad \sup_{-\infty < \mu < \infty, \sigma^2 > 0} f_{\mu, \sigma^2}(x) = \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{n}{2}}} e^{-\frac{n}{2}}.$$

The restricted parameter space Ω_0 is $\{(\mu_0, \sigma^2) \mid \sigma^2 > 0\}$ and the MLE of σ^2 is

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum (x_i - \mu_0)^2,$$

so [verify]

$$(18.56) \quad \sup_{\sigma^2 > 0} f_{\mu_0, \sigma^2}(x) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{\frac{n}{2}}} e^{-\frac{n}{2}}.$$

Thus the LRT for (18.54) rejects $H_0 : \mu = \mu_0$ for *small* values of

$$(18.57) \quad \lambda^{\frac{2}{n}} \equiv \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum (x_i - \bar{x}_n)^2}{\sum (x_i - \mu_0)^2} = \frac{\sum (x_i - \bar{x}_n)^2}{\sum (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu_0)^2},$$

or, equivalently, for *large* values of

$$(18.58) \quad t^2 \equiv \frac{(n-1)n(\bar{x}_n - \mu_0)^2}{\sum (x_i - \bar{x}_n)^2} \equiv \frac{n(\bar{x}_n - \mu_0)^2}{s_n^2}.$$

Under H_0 , $t^2 \sim t_{n-1}^2 (\equiv F_{1,n-1})$ regardless of the value of σ^2 , so the p -value can be easily obtained:

$$(18.59) \quad p \equiv p(t_{\text{observed}}^2) = P[t_{n-1}^2 \geq t_{\text{observed}}^2]. \quad \square$$

Exercise 18.26. (*The 2-sample normal model*) Let $X_1, \dots, X_m, Y_1, \dots, Y_n$ be independent with

$$(18.60) \quad X_i \sim N_1(\mu, \sigma^2), \quad Y_j \sim N_1(\nu, \tau^2), \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

Consider the problem of testing $\mu = \nu$ vs. $\mu \neq \nu$ under the assumption that the variances are unknown but equal: $\sigma^2 = \tau^2$. The unrestricted and restricted models are written in the forms (18.48) and (18.49) as follows:

$$(18.61) \quad \begin{aligned} \Omega : \sigma^2 &= \tau^2 \\ \Omega_0 : \sigma^2 &= \tau^2, \quad \mu = \nu. \end{aligned}$$

Here $k = 4$, $r = 1$, $s = 1 = d - d_0$. Show that the unrestricted and restricted MLEs of μ, ν, σ^2 are, respectively,

$$\begin{aligned} \hat{\mu} &= \bar{x}_m, \quad \hat{\nu} = \bar{y}_n, \quad \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x}_m)^2 + \sum (y_j - \bar{y}_n)^2}{m+n}, \\ \hat{\mu}_0 &= \hat{\nu}_0 = \frac{m\bar{x}_m + n\bar{y}_n}{m+n}, \quad \hat{\sigma}_0^2 = \frac{\sum (x_i - \bar{x}_m)^2 + \sum (y_j - \bar{y}_n)^2 + \frac{mn}{m+n}(\bar{x}_m - \bar{y}_n)^2}{m+n}. \end{aligned}$$

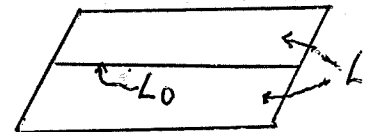
Show that the LRT is equivalent to the 2-sided t -test that rejects H_0 for large values of

$$(18.62) \quad t^2 \equiv \frac{\frac{mn}{m+n}(\bar{x}_m - \bar{y}_n)^2}{\frac{\sum (x_i - \bar{x}_m)^2 + \sum (y_j - \bar{y}_n)^2}{m+n-2}}.$$

Show that under H_0 , $t^2 \sim t_{m+n-2}^2$ regardless of the values of μ and σ^2 . \square

Exercise 18.27. (*Testing linear models*) Based on a single observation $X \sim N_p(\xi, \sigma^2 I_p)$ with σ^2 unknown, consider the problem of testing

$$(18.63) \quad H_0 : \xi \in L_0 \quad \text{vs.} \quad H_1 : \xi \in L \setminus L_0,$$



where $L_0 \subset L \subseteq \mathbf{R}^p$ are nested linear subspaces of dimensions d_0 and d respectively. Because an l -dimensional linear subspace is determined by a set of $p - l$ linear constraints, Ω and Ω_0 are of the forms (18.48) and (18.49) respectively. Find the LRT statistic λ for (18.63) and show that it is equivalent to an F -statistic with $d - d_0$ and $p - d$ degrees of freedom (recall Remark 8.4, where $L_0 = \{0\}$). \square

Exercise 18.28. An example of the testing problem (18.63) is that of testing a simple linear regression model (8.131) vs. a quadratic regression model (8.137):

$$(18.64) \quad EX_i = a + bt_i \quad \text{vs.} \quad EX_i = a + bt_i + ct_i^2, \quad i = 1, \dots, n.$$

that is, test $c = 0$ vs. $c \neq 0$. Here $d_0 = 2$, $d = 3$ [why?]. Show that the LRT statistic for (18.64) is equivalent to a t^2 statistic with $n - 3$ d.f. \square

Example 18.29. (*Exact case of Wilks' Theorem.*) Based on an observation

$$(18.65) \quad X \equiv \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{p_1+p_2} \left[\mu \equiv \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma \equiv \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right],$$

where $\mu_1 : p_1 \times 1$, $\mu_2 : p_2 \times 1$, $\Sigma_{11} : p_1 \times p_1$, $\Sigma_{12} : p_1 \times p_2$, etc. and Σ is a known pd matrix, consider the problem of testing

$$(18.66) \quad H_0 : \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_2 \neq 0 \quad \text{with } \mu_1 \text{ unknown.}$$

Note that (18.66) is a special case of the linear hypothesis (18.63) in Exercise 18.27, with $L_0 \leftrightarrow \{\mu \mid \mu_2 = 0\}$ and $d_0 \leftrightarrow p_1$, $L \leftrightarrow \mathbf{R}^k$ and $d \leftrightarrow p_1 + p_2$, so $d - d_0 \leftrightarrow p_2$. (However, Σ is assumed to be completely known.)

To find the LRT λ (18.51) (here $\theta \leftrightarrow \mu$), factor the pdf of (X_1, X_2) as

$$(18.67) \quad \begin{aligned} f_{\mu_1, \mu_2}(x_1, x_2) &= f_{\mu_1, \mu_2}(x_1 | x_2) \cdot f_{\mu_2}(x_2) \\ &\equiv N_{p_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11.2}) \cdot N_{p_2}(\mu_2, \Sigma_{22}). \end{aligned}$$

Thus

$$\begin{aligned} \lambda &= \frac{\sup_{\mu_1} [f_{\mu_1, 0}(x_1 | x_2) f_0(x_2)]}{\sup_{\mu_1, \mu_2} [f_{\mu_1, \mu_2}(x_1 | x_2) f_{\mu_2}(x_2)]} \\ &= \frac{[\sup_{\mu_1} f_{\mu_1, 0}(x_1 | x_2)] \cdot f_0(x_2)}{\sup_{\mu_2} \{ [\sup_{\mu_1} f_{\mu_1, \mu_2}(x_1 | x_2)] \cdot f_{\mu_2}(x_2) \}}. \end{aligned}$$

But $\sup_{\mu_1} f_{\mu_1,0}(x_1|x_2) = \sup_{\mu_1} f_{\mu_1,\mu_2}(x_1|x_2)$ for each fixed μ_2 [verify], so

$$\lambda = \frac{f_0(x_2)}{\sup_{\mu_2} f_{\mu_2}(x_2)} = \frac{(\text{const}) \cdot e^{-\frac{1}{2}x_2'\Sigma_{22}^{-1}x_2}}{(\text{const}) \cdot e^{-\frac{1}{2}(0)}},$$

the LR statistic for testing $\mu_2 = 0$ vs. $\mu_2 \neq 0$ based on $X_2 \sim N_{p_2}(\mu_2, \Sigma_{22})$ alone. Thus the LRT rejects H_0 for *large* values of the quadratic form

$$(18.68) \quad -2 \log \lambda \equiv X_2' \Sigma_{22}^{-1} X_2 \sim \chi_{p_2}^2(\mu_2' \Sigma_{22}^{-1} \mu_2),$$

a noncentral χ^2 distribution (recall (8.110)). Under $H_0 : \mu_2 = 0$,

$$-2 \log \lambda \sim \chi_{p_2}^2 \equiv \chi_{d-d_0}^2 \quad (\text{a central } \chi^2 \text{ distribution}),$$

so Wilks' approximation (18.69) below is *exact* in this testing problem. \square

In Case II the null distribution of the LRT statistic λ is particularly easy to approximate for i.i.d. samples from a regular family of pdfs.

Theorem 18.30. (S. S. Wilks, 1938) *Let X_1, \dots, X_n be an i.i.d. sample from a regular family $\{f_\theta\}$ that satisfies the Fisher-Cramér and Wald conditions, so the MLE $\hat{\theta}^{(n)}$ is CANE for θ . Let $\lambda_n \equiv \lambda_n(X_1, \dots, X_n)$ be the LRT statistic (18.51) for testing (18.49) vs. (18.48). Then under H_0 ,*

$$(18.69) \quad -2 \log \lambda_n \xrightarrow{d} \chi_{d-d_0}^2 \quad \text{as } n \rightarrow \infty.$$

Proof. See §18.8. \square

Example 18.31. (Multinomial) Let $(X_1, \dots, X_k) \sim M_k(n; p_1, \dots, p_k)$. Consider the problem of testing a simple hypothesis

$$(18.70) \quad H_0 : \mathbf{p} = \mathbf{p}^0 \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \mathbf{p}^0$$

based on $X \equiv (X_1, \dots, X_k) \sim M_k(n; \mathbf{p})$, where

$$\mathbf{p} = (p_1, \dots, p_k), \quad \mathbf{p}^0 = (p_1^0, \dots, p_k^0) \quad (\text{all } p_i^0 > 0).$$

The unrestricted and restricted parameter spaces can be expressed in the forms (18.48) and (18.49) as follows:

$$(18.71) \quad \begin{aligned} \Omega &: p_1 + \cdots + p_k = 1, \\ \Omega_0 &: p_1 + \cdots + p_k = 1, \quad p_1 - p_1^0 = 0, \dots, p_{k-1} - p_{k-1}^0 = 0. \end{aligned}$$

Here $k = k$, $r = 1$, $s = k - 1 = d - d_0$. The multinomial pmf of X is

$$f_{\mathbf{p}}(x) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \quad \text{if } \sum x_i = n.$$

The unrestricted MLE of \mathbf{p} is given by (recall (14.76))

$$\hat{p}_i = \frac{x_i}{n}, \quad i = 1, \dots, k,$$

so the LRT for testing H_0 vs. H_1 rejects H_0 for *small* values of

$$(18.72) \quad \lambda_n \equiv \prod_{i=1}^k \left(\frac{np_i^0}{x_i} \right)^{x_i}.$$

By Wilks' Theorem 18.30, under H_0 we have

$$(18.73) \quad -2 \log \lambda_n \xrightarrow{d} \chi_{k-1}^2 \quad \text{as } n \rightarrow \infty. \quad \square$$

Remark 18.32. In (7.32) of §7 we saw that Pearson's χ^2 -statistic

$$(18.74) \quad \chi^2 \equiv \sum_{i=1}^k \frac{(X_i - np_i^0)^2}{np_i^0} \equiv \sum_{i=1}^k \frac{(\text{Observed}_i - \text{Expected}_i^0)^2}{\text{Expected}_i^0}$$

also has an asymptotic χ_{k-1}^2 distribution under $H_0 : \mathbf{p} = \mathbf{p}^0$. This is no coincidence: it will be shown in §18.7 that the LRT statistic $-2 \log \lambda_n$ and the Pearson χ^2 statistic are asymptotically equal as $n \rightarrow \infty$. \square

Example 18.33. (*Testing independence in a 2-way contingency table.*) Let

$$\{X_{ij}\} \sim M_{\bar{r}\bar{c}}(n; \{p_{ij}\})$$

be the entries in a 2-way contingency table with \bar{r} rows and \bar{c} columns:

X_{11}			$X_{1\bar{c}}$	$X_{1\cdot}$
				\vdots
$X_{\bar{r}1}$			$X_{\bar{r}\bar{c}}$	$X_{\bar{r}\cdot}$
$X_{\cdot 1}$	---		$X_{\cdot \bar{c}}$	n

P_{11}			$P_{1\bar{c}}$	$P_{1\cdot}$
				\vdots
$P_{\bar{r}1}$			$P_{\bar{r}\bar{c}}$	$P_{\bar{r}\cdot}$
$P_{\cdot 1}$	---		$P_{\cdot \bar{c}}$	1

Consider the problem of testing the hypothesis of independence

$$(18.75) \quad H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, \dots, \bar{r}, \quad j = 1, \dots, \bar{c},$$

against the unrestricted alternative, where $p_{i\cdot} = \sum_{j=1}^{\bar{c}} p_{ij}$, $p_{\cdot j} = \sum_{i=1}^{\bar{r}} p_{ij}$. The null and alternative hypotheses can be expressed as follows:

$$(18.76) \quad \Omega : \sum_{i=1}^{\bar{r}} \sum_{j=1}^{\bar{c}} p_{ij} - 1 = 0,$$

$$\Omega_0 = \Omega \cap \{p_{ij} - p_{i\cdot} p_{\cdot j} = 0, \quad i = 1, \dots, \bar{r} - 1, \quad j = 1, \dots, \bar{c} - 1\}.$$

[Why $\bar{r} - 1$, $\bar{c} - 1$ rather than \bar{r} , \bar{c} ?] [See figure for $\bar{r} = \bar{c} = 2$.]

Here $k = \bar{r}\bar{c}$, $r = 1$, $s = (\bar{r} - 1)(\bar{c} - 1) = d - d_0$. Note that

$$d = \bar{r}\bar{c} - 1,$$

$$d_0 = (\bar{r} - 1) + (\bar{c} - 1) \quad [\text{interpret}].$$

The multinomial pmf of X is

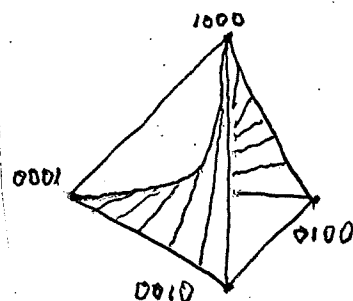
$$(18.77) \quad \frac{n!}{\prod_{i=1}^{\bar{r}} \prod_{j=1}^{\bar{c}} x_{ij}!} \prod_{i=1}^{\bar{r}} \prod_{j=1}^{\bar{c}} p_{ij}^{x_{ij}} \quad \text{if} \quad \sum_{i=1}^{\bar{r}} \sum_{j=1}^{\bar{c}} x_{ij} = n,$$

so the unrestricted MLE of $\{p_{ij}\}$ is given by (recall (14.76))

$$\hat{p}_{ij} = \frac{x_{ij}}{n}, \quad i = 1, \dots, \bar{r}, \quad j = 1, \dots, \bar{c}.$$

Under H_0 the pmf of X is

$$(18.78) \quad \frac{n!}{\prod_{i=1}^{\bar{r}} \prod_{j=1}^{\bar{c}} x_{ij}!} \prod_{i=1}^{\bar{r}} \prod_{j=1}^{\bar{c}} (p_{i\cdot} p_{\cdot j})^{x_{ij}} = \frac{n!}{\dots} \left(\prod_{i=1}^{\bar{r}} p_{i\cdot}^{x_{i\cdot}} \right) \left(\prod_{j=1}^{\bar{c}} p_{\cdot j}^{x_{\cdot j}} \right),$$



so the MLEs of $(\{p_{i\cdot}\} \text{ and } \{p_{\cdot j}\})$ under H_0 are given by

$$(18.79) \quad \hat{p}_{i\cdot}^0 = \frac{x_{i\cdot}}{n} \quad \text{and} \quad \hat{p}_{\cdot j}^0 = \frac{x_{\cdot j}}{n}.$$

Thus the LRT rejects H_0 for *small* values of

$$(18.80) \quad \lambda_n \equiv \prod_{i=1}^{\bar{r}} \prod_{j=1}^{\bar{c}} \frac{\left(\frac{x_{i\cdot}}{n} \frac{x_{\cdot j}}{n}\right)^{x_{ij}}}{\left(\frac{x_{ij}}{n}\right)^{x_{ij}}} = \frac{(\prod_{i=1}^{\bar{r}} x_{i\cdot}^{x_{i\cdot}})(\prod_{j=1}^{\bar{c}} x_{\cdot j}^{x_{\cdot j}})}{n^n \prod_{i=1}^{\bar{r}} \prod_{j=1}^{\bar{c}} x_{ij}^{x_{ij}}}.$$

By Wilks' Theorem 18.30, under H_0 we have

$$(18.81) \quad -2 \log \lambda_n \xrightarrow{d} \chi_{(\bar{r}-1)(\bar{c}-1)}^2 \quad \text{as } n \rightarrow \infty. \quad \square$$

18.7. Relation between $-2 \log \lambda_n$ and Pearson's χ^2 for multinomial hypotheses.

Examples 18.31 and 18.33 are special cases of a more general Case II testing problem in a multinomial distribution: based on

$$X \equiv (X_1, \dots, X_k) \sim M_k(n; \mathbf{p} \equiv (p_1, \dots, p_k)),$$

test

$$(18.82) \quad H_0 : \mathbf{p} \in \Omega_0 \quad \text{vs.} \quad H_1 : \mathbf{p} \in \mathcal{P}_k \setminus \Omega_0,$$

where Ω_0 is a submanifold of the probability simplex $\mathcal{P}_k = \{\mathbf{p} \mid \sum p_i = 1\}$. As above, the LRT for (18.82) rejects H_0 for *small* values of

$$(18.83) \quad \lambda_n \equiv \prod_{i=1}^k \left(\frac{n \hat{p}_i^0}{X_i} \right)^{X_i},$$

where $(\hat{p}_1^0, \dots, \hat{p}_k^0)$ is the MLE of (p_1, \dots, p_k) under H_0 . By Wilks' Theorem,

$$(18.84) \quad -2 \log \lambda_n \xrightarrow{d} \chi_{k-1-d_0}^2 \quad \text{as } n \rightarrow \infty$$

when H_0 is true, where $d_0 = \dim(\Omega_0)$.

A second reasonable test statistic for the testing problem (18.82) is Pearson's goodness-of-fit statistic (recall (18.74))

$$(18.85) \quad \chi^2 \equiv \sum_{i=1}^k \frac{(X_i - n\hat{p}_i^0)^2}{n\hat{p}_i^0} \equiv \sum_{i=1}^k \frac{(O_i - \hat{E}_i^0)^2}{\hat{E}_i^0}.$$

It was shown in (7.32) that Pearson's χ^2 statistic also has an asymptotic $\chi_{k-1-d_0}^2$ distribution under H_0 when $d_0 = 0$, i.e., for the case of a simple hypothesis $\Omega_0 = \{\mathbf{p}_0\}$. In fact this holds for all values of d_0 , i.e., for the general testing problem (18.82). We now establish this by showing that under H_0 , $-2\log \lambda_n$ and Pearson χ^2 are asymptotically equal as $n \rightarrow \infty$.

From the Taylor expansion $\log(1-x) = -x - \frac{1}{2}x^2 + O(x^3)$ we have

$$\begin{aligned} -2\log \lambda_n &= -2 \sum_{i=1}^k X_i \log \left[1 - \left(1 - \frac{n\hat{p}_i^0}{X_i} \right) \right] \\ &= 2 \sum_{i=1}^k (X_i - n\hat{p}_i^0) + \sum_{i=1}^k \frac{(X_i - n\hat{p}_i^0)^2}{X_i} + O \left[\sum_{i=1}^k \frac{(X_i - n\hat{p}_i^0)^3}{X_i^2} \right] \\ &= 0 + \sum_{i=1}^k \frac{(X_i - n\hat{p}_i^0)^2}{n\hat{p}_i^0} + O \left[\sum_{i=1}^k (X_i - n\hat{p}_i^0)^3 \left(\frac{1}{n\hat{p}_i^0 X_i} + \frac{1}{X_i^2} \right) \right] \\ &= 0 + \chi^2 + O_p \left(n^{-\frac{1}{2}} \right) \end{aligned}$$

under H_0 , since $X_i - n\hat{p}_i^0 = (X_i - np_i^0) + n(p_i^0 - \hat{p}_i^0) = O_p(n^{\frac{1}{2}}) + O_p(n^{\frac{1}{2}})$ under H_0 [verify]. Thus $(-2\log \lambda_n) - \chi^2 = O_p(n^{-\frac{1}{2}})$ as $n \rightarrow \infty$, which establishes their asymptotic equivalence under H_0 .

Thus in Example 18.33, $\chi^2 \xrightarrow{d} \chi_{(\bar{r}-1)(\bar{c}-1)}^2$ under H_0 (cf. (18.81)).

Exercise 18.34. In Example 18.33, show that Pearson's chi-square statistic for testing independence in a 2-way table can be expressed as

$$(18.86) \quad \chi^2 = \sum_{i=1}^{\bar{r}} \sum_{j=1}^{\bar{c}} \frac{(nX_{ij} - X_{i.}X_{.j})^2}{nX_{i.}X_{.j}}. \quad \square$$

approximately equivalent to the requirement $np > 4$. This suggests that the condition $\hat{E}_i^0 \equiv n\hat{p}_i^0 > 4$, $i = 1, \dots, k$, is necessary for the accuracy of the approximation $\chi^2 \approx \chi_{k-1-d_0}^2$ to the null distribution of Pearson's statistic, and thus also of the LRT statistic $-2\log \lambda_n$. (Unfortunately, variance-stabilizing transformations do not exist for the multinomial distribution or other multivariate distributions – see P. Holland (1973) *Ann. Statist.*) \square

Remark 18.36. The MLEs \hat{p}_i^0 that appear in the Pearson χ^2 statistic (18.85) must be based on the multinomial counts X_i , not on any underlying data from which the X_i 's may have been derived. For example, suppose we observe continuous i.i.d. rvs Y_1, \dots, Y_n from an unknown pdf f on $-\infty, \infty$ and wish to test $H_0 : f = N_1(\mu, \sigma^2)$ with (μ, σ^2) unspecified. One approach is to group the data into k cells (i.e., form a sample histogram) and use the observed cell counts X_1, \dots, X_k to test H_0 . Then the cell probabilities $p_i^0 \equiv p_i^0(\mu, \sigma^2)$ depend on the unknown (μ, σ^2) , which must be estimated.

However, we may *not* use the usual MLEs $\hat{\mu} = \bar{Y}_n$, $\hat{\sigma}^2 = \frac{n-1}{n} s_n^2$ based on the Y_i 's but rather must use the MLEs $\tilde{\mu}, \tilde{\sigma}^2$ based only on the X_i 's – otherwise the approximation $\chi^2 \approx \chi_{k-3}^2$ (here $d = k - 1$, $d_0 = 2$) is invalid – see Chernoff and Lehmann (1954) *Ann. Math. Statist.* \square

18.8. Proof of Wilks' Theorem; consistency of the Case II LRT.

Theorem 18.30 (repeated). Let X_1, \dots, X_n be an i.i.d. sample from a regular family $\{f_\theta\}$ that satisfies the Fisher-Cramér and Wald conditions, so that the MLE $\hat{\theta}^{(n)}$ is CANE for θ . Let $\lambda_n \equiv \lambda_n(X_1, \dots, X_n)$ be the LRT statistic (18.51) for testing (18.49) vs. (18.48). Then under H_0 ,

$$(18.69) \quad -2\log \lambda_n \xrightarrow{d} \chi_{d-d_0}^2 \quad \text{as } n \rightarrow \infty.$$

Proof (sketch). We will indicate the proof for the simple case where Ω_0 is a single point $\{\theta_0\}$ and Ω is an open subset of \mathbf{R}^k , so $d_0 = 0$, $d = k$, and $d - d_0 = k$. Because $\hat{\theta} \equiv \hat{\theta}^{(n)}$ is a CANE estimator (recall (14.65)),

$$(18.87) \quad \sqrt{n} \left(\hat{\theta}^{(n)} - \theta_0 \right) \xrightarrow{d} N_k \left(0, [I(\theta_0)]^{-1} \right)$$

when $H_0 : \theta = \theta_0$ holds, so by Slutsky's Theorem the quadratic form

$$(18.88) \quad Q_n \equiv n \left(\hat{\theta}^{(n)} - \theta_0 \right)' I(\theta_0) \left(\hat{\theta}^{(n)} - \theta_0 \right) \xrightarrow{d} \chi_k^2 \quad \text{as } n \rightarrow \infty.$$

Thus (18.69) will follow if we can show that

$$(18.89) \quad -2 \log \lambda_n - Q_n = o_p(1) \quad \text{as } n \rightarrow \infty.$$

First take $k = 1$ so θ is 1-dimensional. Let $l_n(\theta) \equiv \sum_{i=1}^n \log f_\theta(x_i)$ denote the log likelihood function. From (18.51) (p.305) and the 2nd-order Taylor expansion of l_n about $\hat{\theta}^{(n)}$ [not θ_0],

$$\begin{aligned} -\log \lambda_n &\equiv -\left[l_n(\theta_0) - l_n(\hat{\theta}^{(n)})\right] \\ &= (\hat{\theta}^{(n)} - \theta_0) \frac{dl_n(\hat{\theta}^{(n)})}{d\theta} - \frac{(\hat{\theta}^{(n)} - \theta_0)^2}{2} \frac{d^2 l_n(\hat{\theta}^{(n)})}{d\theta^2} + \frac{(\hat{\theta}^{(n)} - \theta_0)^3}{6} \frac{d^3 l_n(\theta_n^*)}{d\theta^3} \end{aligned}$$

for some $\theta_n^* \in (\theta_0, \hat{\theta}^{(n)})$. But $\hat{\theta}^{(n)}$ satisfies the LEQ $\frac{dl_n(\hat{\theta}^{(n)})}{d\theta} = 0$, so by arguments similar to those for Theorems 14.9, 14.18, and Exercise 14.10,

$$\begin{aligned} -2 \log \lambda_n &= -(\hat{\theta}^{(n)} - \theta_0)^2 \frac{d^2 l_n(\hat{\theta}^{(n)})}{d\theta^2} + o_p(1) \\ &= n(\hat{\theta}^{(n)} - \theta_0)^2 I(\theta_0) + o_p(1). \end{aligned}$$

By (18.88), this establishes (18.89).

If $k \geq 2$, use the multivariate 2nd-order Taylor expansion instead.

The proof for the case of a general composite Ω_0 may be found in the original paper by S. S. Wilks (1938): "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *Annals of Mathematical Statistics* **9** pp. 60-62; also see the book *Approximation Theorems of Mathematical Statistics* by R. Serfling (1980). Again the idea is to use the asymptotic normality (18.87) of the MLE $\hat{\theta}^{(n)}$ and the smoothness of Ω_0 and Ω to reduce the problem to one of testing a linear hypothesis about the mean of a multivariate normal distribution, as in Example 18.29. \square

In §18.6 we established the consistency of the LRT (18.45) based on the statistic ϕ_n^* for Cases Ia,b where Ω_0 and Ω_1 are separated. This remains true for Case II where the appropriate LRT statistic is given by (18.51), but a different proof is required.

Theorem 18.37. (*Consistency of the LRT in Case II.*) Under the conditions of Wilks' Theorem 18.30, for each $\theta_1 \in \Omega_1 \equiv \Omega \setminus \Omega_0$,

$$(18.90) \quad P_{\theta_1}[-2 \log \lambda_n > \chi_{d-d_0, \alpha}^2] \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Proof. For fixed θ_1 , select $\theta_0^* \equiv \theta_0^*(\theta_1) \in \Omega_0$ such that

$$(18.91) \quad K(\theta_1, \theta_0^*) = \min_{\theta_0 \in \Omega_0} K(\theta_1, \theta_0) > 0.$$

Let $\hat{\theta}_0 \equiv \hat{\theta}_0^{(n)}$ and $\hat{\theta} \equiv \hat{\theta}^{(n)}$ be the MLEs of θ under Ω_0 and Ω , respectively. Then

$$(18.92) \quad \begin{aligned} -\log \lambda_n &= -[l_n(\hat{\theta}_0) - l_n(\hat{\theta})] \\ &= -[l_n(\hat{\theta}_0) - l_n(\theta_0^*)] - [l_n(\theta_0^*) - l_n(\theta_1)] - [l_n(\theta_1) - l_n(\hat{\theta})]. \end{aligned}$$

By Wilks' Theorem with $\Omega_0 = \{\theta_1\}$ (so $d_0 = 0$),

$$(18.93) \quad -[l_n(\theta_1) - l_n(\hat{\theta})] \xrightarrow{d} \frac{1}{2} \chi_d^2 = O_p(1) \quad \text{as } n \rightarrow \infty.$$

By the WLLN,

$$(18.94) \quad \begin{aligned} -[l_n(\theta_0^*) - l_n(\theta_1)] &= -n \cdot \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_{\theta_0^*}(X_i)}{f_{\theta_1}(X_i)} \right] \\ &\approx n \cdot K(\theta_1, \theta_0^*) \rightarrow \infty \quad \text{as } n \rightarrow \infty. \end{aligned}$$

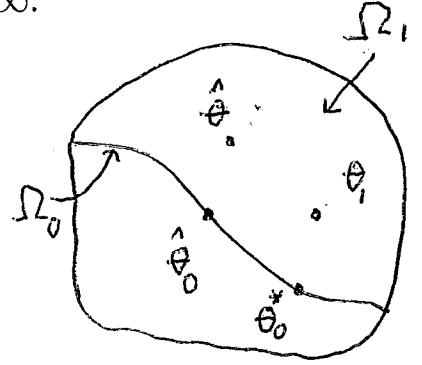
Finally, note that when θ_1 is the actual value of θ , $\hat{\theta}_0$ is the MLE of θ under the “wrong” model Ω_0 . Just as for the MLE under the correct model, it can be shown (see (18.102) in §18.9) that from the choice of θ_0^* in (18.91),

$$(18.95) \quad \hat{\theta}_0 - \theta_0^* = O_p\left(n^{-\frac{1}{2}}\right).$$

If we now⁴² let $\nabla_{\theta_0}^2 l_n(\theta_0)$ denote the $d_0 \times d_0$ Hessian matrix $\left(\frac{\partial^2 l_n(\theta_0)}{\partial \theta_{0i} \partial \theta_{0j}}\right)$, the multivariate 2nd-order Taylor approximation of $l_n(\theta_0^*)$ about $\hat{\theta}_0$ gives

$$l_n(\theta_0^*) - l_n(\hat{\theta}_0) \approx (\theta_0^* - \hat{\theta}_0)' \nabla_{\theta_0} l_n(\hat{\theta}_0) + \frac{1}{2} (\theta_0^* - \hat{\theta}_0)' \nabla_{\theta_0}^2 l_n(\hat{\theta}_0) (\theta_0^* - \hat{\theta}_0)$$

⁴² Note that under the null model determined by Ω_0 , both θ_0 and $\nabla_{\theta_0} l_n(\theta_0)$ are actually d_0 -dimensional not k -dimensional, $\nabla_{\theta_0}^2 l_n(\theta_0)$ is $d_0 \times d_0$ not $k \times k$, etc. Perhaps our notation should be changed to reflect this.



$$\begin{aligned}
&= +\frac{1}{2}(\theta_0^* - \hat{\theta}_0)' \nabla_{\theta_0}^2 l_n(\hat{\theta}_0)(\theta_0^* - \hat{\theta}_0) \quad [\text{since } \nabla_{\theta_0} l_n(\hat{\theta}_0) = 0] \\
&\approx \frac{1}{2}(\theta_0^* - \hat{\theta}_0)' \nabla_{\theta_0}^2 l_n(\theta_0^*)(\theta_0^* - \hat{\theta}_0) \quad [\text{verify}] \\
&= \frac{1}{2} \left[\sqrt{n}(\theta_0^* - \hat{\theta}_0) \right]' \left[\frac{1}{n} \nabla_{\theta_0}^2 l_n(\theta_0^*) \right] \left[\sqrt{n}(\theta_0^* - \hat{\theta}_0) \right] \\
(18.96) \quad &= O_p(1)
\end{aligned}$$

by (18.95) and the fact that under θ_1 ,

$$(18.97) \quad \frac{1}{n} \nabla_{\theta_0}^2 l_n(\theta_0^*) \rightarrow E_{\theta_1} \left\{ \left[\nabla_{\theta_0}^2 \log f_{\theta_0}(X) \right]_{\theta_0^*(\theta_1)} \right\} \equiv -J(\theta_1)$$

by the LLN. Thus by (18.92), (18.93), (18.94), and (18.96),

$$-2 \log \lambda_n = O_p(1) + n K(\theta_1, \theta_0^*) + O_p(1) \rightarrow \infty \quad \text{as } n \rightarrow \infty,$$

which establishes the consistency result (18.90). \square

18.9. Properties of the MLE when the model is incorrect.

As in §14.3 and §14.5, let $\{f_\theta \mid \theta \in \Omega\}$ be a regular family of pdfs, where Ω is an open subset of \mathbf{R}^k . It was shown there, cf. Propositions 14.8, 14.20, and Theorems 14.9, 14.21, that under the Fisher-Cramér-Wald conditions, the MLE $\hat{\theta} \equiv \hat{\theta}^{(n)}$ based on i.i.d. $X_1, \dots, X_n \sim f_\theta$ is a CANE estimator of θ . The key results were the following. When $\theta = \theta_0$, i.e., $X_i \sim f_{\theta_0}$,

(a) $E_{\theta_0} \left\{ \log \left[\frac{f_\theta(X)}{f_{\theta_0}(X)} \right] \right\} \equiv -K(\theta_0, \theta) \leq 0$ has a maximum value 0 attained uniquely at $\theta = \theta_0$, so $E_{\theta_0} \{ [\nabla_\theta \log f_\theta(X)]_{\theta_0} \} = [\nabla_\theta K(\theta_0, \theta)]_{\theta_0} = 0$.

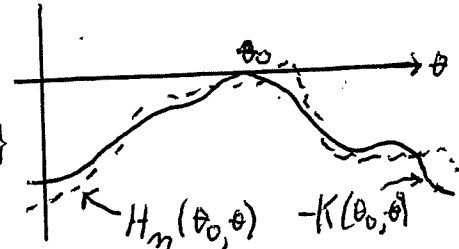
(b) $H_n(\theta_0, \theta) \equiv \frac{1}{n} \sum_{i=1}^n \left\{ \log \left[\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \right] \right\} \rightarrow -K(\theta_0, \theta)$ uniformly in θ .

(c) $\sqrt{n} (\hat{\theta}^{(n)} - \theta_0) \xrightarrow{d} N_k(0, [I(\theta_0)]^{-1})$, where

$$(18.98) \quad I(\theta) = E_\theta \{ [\nabla_\theta \log f_\theta(X)] [\nabla_\theta \log f_\theta(X)]' \}$$

$$(18.99) \quad = -E_\theta [\nabla_\theta^2 \log f_\theta(X)]$$

$$(18.100) \quad = [\nabla_{\theta'} K(\theta, \theta')]_{\theta'=\theta}$$



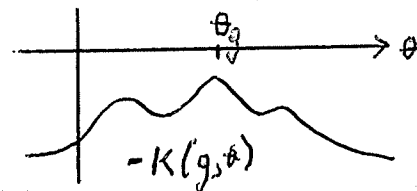
The first of these relations shows that $I(\theta_0)$ is positive semidefinite.

Now suppose that the model $\{f_\theta \mid \theta \in \Omega\}$ is *incorrect*, so that $X_i \sim g$ rather than $X_i \sim f_{\theta_0}$, where $g \neq f_\theta$ for any $\theta \in \Omega$. If the Fisher-Cramér-Wald conditions remain valid when $X_i \sim g$ then (a), (b), and (c) can be directly extended as follows:

(a') $E_g \left\{ \log \left[\frac{f_\theta(X)}{g(X)} \right] \right\} \equiv -K(g, \theta) < 0 \forall \theta \in \Omega$ (by Jensen's inequality).

Assume \exists a *unique* $\theta_g \in \Omega$ such that [see figure]

$$(18.101) \quad K(g, \theta_g) = \min_{\theta \in \Omega} K(g, \theta) > 0.$$



(b') $H_n(g, \theta) \equiv \frac{1}{n} \sum_{i=1}^n \left\{ \log \left[\frac{f_\theta(X_i)}{g(X_i)} \right] \right\} \rightarrow -K(g, \theta)$ uniformly in θ .

(c') Let $\hat{\theta}^{(n)}$ be the MLE of θ under the incorrect model $\{f_\theta \mid \theta \in \Omega\}$. Then

$$(18.102) \quad \sqrt{n} \left(\hat{\theta}^{(n)} - \theta_g \right) \xrightarrow{d} N_k \left(0, [J(g)]^{-1} I(g) [J(g)]^{-1} \right)$$

(recall (14.42)-(14.44), p.237), where $I(g)$ is the $k \times k$ psd matrix given by

$$(18.103) \quad I(g) = E_g \left\{ [\nabla_\theta \log f_\theta(X)]_{\theta_g} [\nabla_\theta \log f_\theta(X)]'_{\theta_g} \right\} \quad [\text{compare (18.98)}]$$

and $J(g)$ is the $k \times k$ symmetric matrix given by

$$(18.104) \quad J(g) = -E_g \left\{ [\nabla_\theta^2 \log f_\theta(X)]_{\theta_g} \right\} \quad [\text{compare to (18.99)}]$$

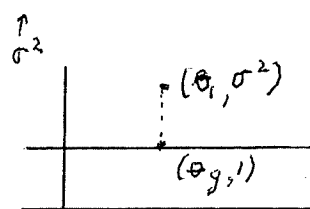
$$(18.105) \quad = [\nabla_\theta^2 K(g, \theta)]_{\theta_g} \quad [\text{compare to (18.100)}].$$

Note that $I(g)$ and $J(g)$ are psd [verify] by (18.103) and (18.101,105); so both must be *assumed* to be positive definite.

Exercise 18.38. Derive (18.102) for $k = 1$ [cf. (14.42)-(14.46), p.237]. \square

Example 18.39. Take $f_\theta = N_1(\theta, 1)$, $g = N_1(\theta_1, \sigma^2)$ with $\sigma^2 \neq 1$. The MLE of θ under the "wrong" model $\{f_\theta\}$ remains $\hat{\theta}^{(n)} = \bar{X}_n$. From (a'),

$$(18.106) \quad \begin{aligned} K(g, \theta) &= -E_{\theta_1, \sigma^2} \left[\log \sigma + \frac{(X - \theta_1)^2}{2\sigma^2} - \frac{(X - \theta)^2}{2} \right] \\ &= -\log \sigma - \frac{1}{2} + \frac{\sigma^2 + (\theta_1 - \theta)^2}{2}, \end{aligned}$$



which is minimized when $\theta = \theta_1 \equiv \theta_g$. Thus from (18.103), (18.105), and (18.106),

$$\begin{aligned} I(g) &= E_{\theta_1, \sigma^2} (X - \theta_1)^2 = \sigma^2, \\ J(g) &= 1, \end{aligned}$$

so from (18.102),

$$\sqrt{n} (\bar{X}_n - \theta_1) \xrightarrow{d} N_1(0, \sigma^2),$$

the “wrong” variance. (This also follows by the Central Limit Theorem.) \square

Exercise 18.40. By its definition, the MLE $\hat{\theta}^{(n)}$ under the model $\{f_\theta\}$ satisfies the following system of k *estimating equations* (i.e., the LEQs):

$$(18.107) \quad \frac{1}{n} \sum_{i=1}^n [\nabla_\theta \log f_\theta(x_i)]_{\hat{\theta}^{(n)}} = \left[\nabla_\theta \left(\frac{1}{n} \sum_{i=1}^n \log f_\theta(x_i) \right) \right]_{\hat{\theta}^{(n)}} = 0,$$

while it follows from (18.101) that when $X_i \sim g$ (compare to (a), p.317):

$$(18.108) \quad E_g \left\{ [\nabla_\theta \log f_\theta(X)]_{\theta_g} \right\} = [\nabla_\theta K(g, \theta)]_{\theta_g} = 0.$$

More generally, we can replace the *score function* $\nabla_\theta \log f_\theta(x_i)$ by a general *estimating function* $\eta_\theta(x_i)$ and define $\tilde{\theta}^{(n)}$ and θ_g to be the solutions (assumed to exist and be unique) of the systems

$$(18.109) \quad \frac{1}{n} \sum_{i=1}^n \eta_{\tilde{\theta}^{(n)}}(x_i) = 0,$$

$$(18.110) \quad E_g [\eta_{\theta_g}(X)] = 0.$$

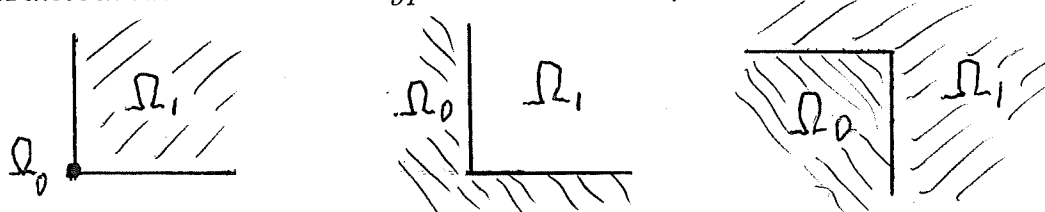
Under appropriate Fisher-Cramér-type regularity conditions, show that when $X_i \sim g$, $\tilde{\theta}^{(n)}$ is a CAN estimator of θ_g , that is, satisfies (18.102) with $I(g)$ and $J(g)$ suitably redefined (how?). [Also see CB pp.485-8.] \square

Remark 18.41. Wilks’ Theorem is based on the asymptotic normality of the MLE as obtained in the Fisher-Cramér Theorems 14.9 and 14.21. We have stated each of these results for an i.i.d. sample X_1, \dots, X_n from a

regular family of pdfs $\{f_\theta(x_i)\}$, but they hold more generally, for example, in the two-sample normal model of Exercise 18.26 and in models with dependent observations such as auto-regressive time series models. For such models the key requirement for asymptotic normality of the MLE is that the total Fisher information concerning the parameters to be estimated should approach infinity. \square

18.10. Case III: Non-separated but non-smooth hypotheses.

Multivariate one-sided hypotheses. Examples include:

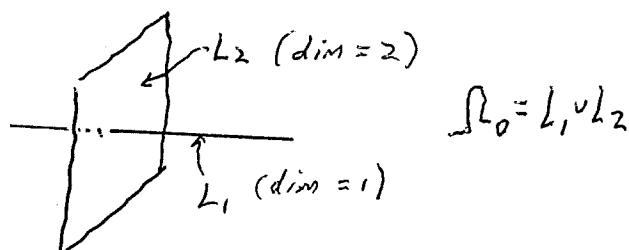


The appropriate form of the LR statistic λ is again (18.51) as in Case II, but now for i.i.d. sampling, Wilks' asymptotic χ^2 approximation (18.69) of the null distribution of λ_n is often replaced by

$$(18.111) \quad -2 \log \lambda_n \xrightarrow{d} \omega_1 \chi_1^2 + \cdots + \omega_k \chi_k^2 \quad \text{as } n \rightarrow \infty,$$

where the χ^2 variates are mutually independent and $\omega_1, \dots, \omega_k$ are weights depending on the geometry of Ω_0 and Ω_1 . Furthermore, if Ω_0 is composite, these weights may depend on the value of $\theta_0 \in \Omega_0$. See books by Barlow, Bartholomew, Bremner, & Brunk (1972), Robertson, Dykstra, & Wright (1988), and Silvapulle & Sen (2005). For a cautionary note see MDP & Chaudhuri "The role of reversals in order-restricted inference", *Can. J. Stat.*, 2004; MDP & Wu "The Emperor's New Tests", *Stat. Sci.* 1999.

Null hypotheses with varying dimensionalities. If $\Omega_0 = \cup_{i=1}^r \Omega_{0i}$, a union of non-nested sets of varying dimensionalities, then no form of the LRT statistic is appropriate – see Perlman and Wu "On the validity of the likelihood ratio and maximum likelihood methods", *JSPI* 2003. Instead, the *union-intersection* or *intersection-union* methods may be used – see CB Ch. 8. (Perhaps a better approach is to restate the problem as a multiple decision problem, rather than use the simpler hypothesis-testing form.)



18.11. Properties of p -values.

It is often stated that “under the null hypothesis H_0 , the p -value (cf. (18.112)) is uniformly distributed on $[0,1]$ ”. This may be true for test statistics T with continuous distributions (recall Example 2.2a, p.20), but must be modified for discrete or general distributions, as follows.

For any random variable T (continuous, discrete, or a mixture thereof), its cdf $F(t) \equiv P[T \leq t]$ is nondecreasing, right continuous, and satisfies $F(t-) = P[T < t]$. Similarly, $G(t) \equiv P[T \geq t]$ is nonincreasing, left continuous, and satisfies $G(t+) = P[T > t]$. Let $U \sim \text{Uniform}[0,1]$.

Lemma 18.42. (i) $F(T) \succeq_{\text{stoch}} U$ and $G(T) \succeq_{\text{stoch}} U$. That is,

$$P[F(T) \leq u] \leq u \text{ and } P[G(T) \leq u] \leq u \text{ for all } 0 \leq u \leq 1.$$

(ii) If T is a continuous rv, $F(T) \sim U$ and $G(T) \sim U$.

Exercise 18.43. Prove (i) and (ii). □

Now let T be a test statistic for testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_1$, where H_0 is rejected for large values of T . Set $G_\theta(t) = P_\theta[T \geq t]$. The p -value \equiv attained significance level associated with T is

$$(18.112) \quad p \equiv p(T_{\text{obs}}) \equiv \sup_{\theta \in \Omega_0} G_\theta(T_{\text{obs}}),$$

where $T_{\text{obs}} \equiv T_{\text{observed}} \sim T$. Clearly $p \equiv p(T_{\text{obs}})$ is nonincreasing in T_{obs} . It follows from Lemma 18.42(i) that when $\theta \in \Omega_0$,

$$P_\theta[p \leq u] \leq P_\theta[G_\theta(T_{\text{obs}}) \leq u] \leq u.$$

Thus under H_0 , the p -value is stochastically larger than $U \equiv \text{Uniform}[0,1]$.

Finally, for a real-valued parameter θ , consider the problem of testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ on a test statistic $T \sim f_\theta$, where $f_\theta(t)$ has monotone likelihood ratio in θ (cf. §18.3). Then by Lemma 18.21 (p.301), $G_\theta(t)$ is increasing in θ , so $p \equiv p(T_{\text{obs}}) = G_{\theta_0}(T_{\text{obs}})$ and p is stochastically decreasing in θ . Thus by Lemma 18.42(ii), if T is continuous then $p \sim \text{Uniform}[0,1]$ when $\theta = \theta_0$.

19. Sequential Tests and Estimators.

19.1. The sequential probability ratio test.

During WWII, the quest for increased industrial efficiency led Abraham Wald in the U.S. and George Barnard in the U.K. independently to study statistical tests and estimates based on *sequential sampling procedures*. In hypothesis testing, for example, traditionally the sample size is predetermined so that prespecified (small) Type I and Type II error probabilities will be obtained. However, it may occur that the evidence for or against the null hypothesis accumulates so rapidly that the entire sample need not be examined, thereby reducing the time and cost of the study. We shall illustrate this by the *sequential probability ratio test (SPRT)* for testing a simple hypothesis H_0 vs. a simple alternative H_1 .

Let X_1, X_2, \dots be an infinite series of successive possible observations, *not necessarily i.i.d.* Let $\mathbf{X}_n \equiv (X_1, \dots, X_n)$ have pdf $f_j(\mathbf{x}_n)$ under H_j , $j = 0, 1$, and let λ_n denote the likelihood ratio (LR) for \mathbf{X}_n :

$$(19.1) \quad \lambda_n \equiv \lambda_n(\mathbf{x}_n) = \frac{f_1(\mathbf{x}_n)}{f_0(\mathbf{x}_n)}.$$

Consider the nonrandomized LRTs based on \mathbf{X}_n (recall (18.5):

$$(19.2) \quad \phi(\mathbf{x}_n) \equiv \phi^k(\mathbf{x}_n) = \begin{cases} 0 & (\equiv \text{choose } H_0) \text{ if } \lambda(\mathbf{x}_n) \leq k, \\ 1 & (\equiv \text{choose } H_1) \text{ if } \lambda(\mathbf{x}_n) > k, \end{cases}$$

with error probabilities

$$(19.3) \quad \alpha_n(k) \equiv P_0[\lambda_n > k],$$

$$(19.4) \quad \beta_n(k) \equiv P_1[\lambda_n \leq k].$$

For *fixed* n we can control *one* of the error probabilities via the choice of k :

$$(19.5) \quad \begin{aligned} \alpha_n(k) &= \int_{\{\lambda_n(\mathbf{x}_n) > k\}} f_0(\mathbf{x}_n) < \frac{1}{k} \int_{\{\lambda_n(\mathbf{x}_n) \geq k\}} f_1(\mathbf{x}_n) \\ &= \frac{1}{k} [1 - \beta_n(k)] \leq \frac{1}{k}, \end{aligned}$$

so $\alpha_n(k) \rightarrow 0$ as $k \rightarrow \infty$, but in this case $\beta_n(k) \rightarrow 1$ by (19.4).

In the case where X_1, X_2, \dots are i.i.d., we saw in Theorem 18.15 (cf. (18.28)-(18.30)) that we can control *both* $\alpha_n(k)$ and $\beta_n(k)$ by fixing k (e.g., $k = 1$) and letting $n \rightarrow \infty$. Wald and Barnard realized, however, that by sampling sequentially, that is, by letting the sample size depend on the data as it accrues, both α and β can be controlled with *smaller expected sample size*. Sometimes the initial observations may clearly reveal which of H_0 or H_1 is true, hence sampling may be stopped very early. Other times extended sampling may be needed, but on average the sample size is reduced – see Theorem 19.8.

The SPRT(B, A): fix $0 < B < 1 < A < \infty$. For $n = 1, 2, \dots$,

$$(19.6) \quad \begin{cases} \text{stop sampling and choose } H_0 & \text{if } \lambda_n \leq B, \\ \text{stop sampling and choose } H_1 & \text{if } \lambda_n \geq A, \\ \text{continue sampling} & \text{if } B < \lambda_n < A. \end{cases}$$

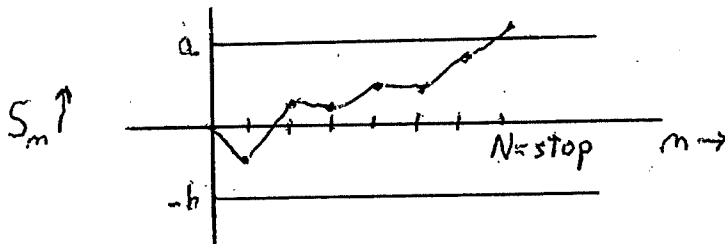
Note that SPRT(B, A) is defined even if X_1, X_2, \dots are not i.i.d. The sample size $N_{B,A} \equiv N_{B,A}(X_1, X_2, \dots)$ is itself a random *stopping time*:

Definition 19.1. A positive-integer-valued rv $N \equiv N(X_1, X_2, \dots)$ is a *stopping time* (\equiv *Markov time* \equiv *optional time*) if for each $n = 1, 2, \dots$, the event $\{N = n\}$ depends only on (X_1, \dots, X_n) . That is, the decision to stop at time n depends only on the past and present, not on the future. \square

Proposition 19.2. If X_1, X_2, \dots are i.i.d. then $P_j[N_{B,A} < \infty] = 1$ and $E_j(N_{B,A}) < \infty$, $j = 0, 1$.

Proof. Suppose that $X_i \sim f_j$ under H_j , $j = 0, 1$, so

$$(19.7) \quad \log \lambda_n = \sum_{i=1}^n \log \left[\frac{f_1(X_i)}{f_0(X_i)} \right] \equiv \sum_{i=1}^n Z_i \equiv S_n.$$



Here $\{S_n \mid n = 1, 2, \dots\}$ is a *random walk*, i.e., a discrete-time stochastic process consisting of consecutive sums of the i.i.d. rvs Z_1, Z_2, \dots . Then

$$\{N_{B,A} < \infty\} = \{S_n \leq -b \text{ or } S_n \geq a \text{ for some } n = 1, 2, \dots\},$$

where

$$(19.8) \quad -b \equiv \log B < 0 < \log A \equiv a.$$

But $P_j[Z_i \neq 0] > 0$ for $j = 0, 1$ since $f_0 \neq f_1$, so $P_j[Z_i > 0] > 0$ or $P_j[Z_i < 0] > 0$ or both. WLOG suppose that $P_j[Z_i > 0] > 0$. (Abbreviate $N_{B,A}$ by N .)

(a) Suppose in fact that $P_j[Z_i \geq b + a] \equiv \delta > 0$. Then

$$\{N \leq n\} \supseteq \{Z_1 \geq b + a\} \cup \cdots \cup \{Z_n \geq b + a\},$$

so

$$\begin{aligned} P_j[N > n] &\leq P_j[Z_1 < b + a, \dots, Z_n < b + a] \\ &= P_j[Z_1 < b + a]^n \\ &= (1 - \delta)^n. \end{aligned}$$

Thus

$$P_j[N < \infty] = \lim_{n \rightarrow \infty} P_j[N \leq n] \geq \lim_{n \rightarrow \infty} [1 - (1 - \delta)^n] = 1.$$

Also,

$$E_j(N) = \sum_{n=0}^{\infty} P_j[N > n] \leq \sum_{n=0}^{\infty} (1 - \delta)^n < \infty.$$

[In fact, all positive moments $E_j(N^r)$ are finite.]

(b) If $P_j[Z_i \geq b + a] = 0$, $\exists \epsilon > 0$ s.t. $P_j[Z_i \geq \epsilon] \equiv \gamma > 0$. Choose $r \geq \frac{b+a}{\epsilon}$, so $r\epsilon \geq b + a$. Then

$$\begin{aligned} P_j[Z_1 + \cdots + Z_r \geq b + a] &\geq P_j[Z_1 \geq \epsilon, \dots, Z_r \geq \epsilon] \\ &= P_j[Z_1 \geq \epsilon] \cdots P_j[Z_r \geq \epsilon] \\ &= \gamma^r > 0. \end{aligned}$$

Now apply the argument in (a) with $\delta = \gamma^r$ and Z_i replaced by $Z_{r(i-1)+1} + \cdots + Z_{ri}$. \square

Proposition 19.3. *Suppose that $P_{H_j}[N_{B,A} < \infty] = 1$, $j = 0, 1$. Define*

$$(19.9) \quad \alpha(B, A) \equiv P_0[\text{SPRT}(B, A) \text{ chooses } H_1] \equiv P_0[\{\lambda_n\} \text{ hits } A \text{ before } B],$$

$$(19.10) \quad \beta(B, A) \equiv P_1[\text{SPRT}(B, A) \text{ chooses } H_0] \equiv P_1[\{\lambda_n\} \text{ hits } B \text{ before } A].$$

Then whether X_1, X_2, \dots are i.i.d. or not,

$$(19.11) \quad \alpha(B, A) \leq \frac{1 - \beta(B, A)}{A} \leq \frac{1}{A},$$

$$(19.12) \quad \beta(B, A) \leq (1 - \alpha(B, A))B \leq B.$$

[Compare (19.11) to (19.5) with $k = A$. Also, note that these bounds do not depend on the actual sequence of pdfs $\{f_j(\mathbf{x}_n)\}$ under H_j , $j = 0, 1$.]

Proof. Abbreviate $N_{B,A}$ by N . Then

$$\begin{aligned} \alpha(B, A) &= P_0[\lambda_N \geq A] = \sum_{n=1}^{\infty} P_0[\lambda_N \geq A, N = n] \\ &= \sum_{n=1}^{\infty} \int_{\{\lambda_n(\mathbf{x}_n) \geq A, N=n\}} f_0(\mathbf{x}_n) d\mathbf{x}_n \\ (19.13) \quad &\leq \frac{1}{A} \sum_{n=1}^{\infty} \int_{\{\lambda_n(\mathbf{x}_n) \geq A, N=n\}} f_1(\mathbf{x}_n) d\mathbf{x}_n \\ &= \frac{1}{A} \sum_{n=1}^{\infty} P_1[\lambda_N \geq A, N = n] \\ &= \frac{1}{A} P_1[\lambda_N \geq A] \\ &= \frac{1}{A} \{1 - P_1[\lambda_N \leq B]\} \quad [\text{since } P_1[N < \infty] = 1] \\ &\equiv \frac{1}{A} (1 - \beta(B, A)), \end{aligned}$$

so (19.11) holds. A similar argument establishes (19.20). \square

Remark 19.4. (*Wald's approximations.*) By (19.6), either $\lambda_{N_{B,A}} \leq B$ or $\lambda_{N_{B,A}} \geq A$. If the successive ratios λ_n/λ_{n-1} are nearly one then approximate equality holds in (19.13) hence holds for the first inequality in (19.11), and similarly in (19.12):

$$(19.14) \quad B \approx \frac{\beta(B, A)}{1 - \alpha(B, A)}, \quad A \approx \frac{1 - \beta(B, A)}{\alpha(B, A)},$$

which in turn yield the following approximations for $\alpha(B, A)$ and $\beta(B, A)$:

$$(19.15) \quad \alpha(B, A) \approx \frac{1 - B}{A - B}, \quad \beta(B, A) \approx \frac{B(A - 1)}{A - B}.$$

Furthermore, the approximations in (19.14) can be used to select B and A so that the SPRT(B, A) approximately attains prespecified error probabilities α and β . Note that the B and A so obtained will not depend on f_0 or f_1 . (Usually this approximation will be conservative, i.e., the actual error probabilities will be smaller than the prespecified values.) \square

Proposition 19.5. (Wald's Lemma for a randomly stopped sum.)
Suppose that U_1, U_2, \dots is a sequence of i.i.d. rvs with $E|U_i| < \infty$ and let N be a general stopping time for $\{U_n\}$. If $E(N) < \infty$ then

$$(19.16) \quad E(S_N) \equiv E(U_1 + \dots + U_N) = E(N) E(U_1).$$

Proof. Since N is a stopping time, the indicator function

$$I_n \equiv I_{\{N \geq n\}} = 1 - I_{\{N \leq n-1\}}$$

depends only on U_1, \dots, U_{n-1} , so $I_n \perp\!\!\!\perp U_n$. Therefore

$$\begin{aligned} E(S_N) &= E\left(\sum_{n=1}^N U_n\right) = E\left(\sum_{n=1}^{\infty} I_n U_n\right) \\ &\stackrel{*}{=} \sum_{n=1}^{\infty} E(I_n U_n) = \sum_{n=1}^{\infty} E(I_n) E(U_n) \\ &= \left(\sum_{n=1}^{\infty} \Pr[N \geq n]\right) E(U_1) \\ &= E(N) E(U_1). \end{aligned}$$

This completes the proof, except for justification of the interchange of E and \sum at (*). This relies on the *Dominated Convergence Theorem*, which implies that $E(\sum_{n=1}^{\infty} Y_n) = \sum_{n=1}^{\infty} EY_n$ provided that either all $Y_n \geq 0$ or $E(\sum_{n=1}^{\infty} |Y_n|) < \infty$. Set $Y_n = I_n U_n$. Since $|I_n U_n| \geq 0$ we have

$$\begin{aligned} E\left(\sum_{n=1}^{\infty} |I_n U_n|\right) &= \sum_{n=1}^{\infty} E|I_n U_n| = \sum_{n=1}^{\infty} E(I_n) E|U_n| \\ &= \left(\sum_{n=1}^{\infty} \Pr[N \geq n]\right) E|U_1| \\ &= E(N) E|U_1| < \infty. \end{aligned}$$

Thus the interchange of E and \sum at (*) is justified. \square

Remark 19.6. (*Approximating $E_0(N_{B,A})$ and $E_1(N_{B,A})$ for the SPRT(B, A) in the i.i.d. case:*) Apply Wald's Lemma (19.16) with $U_i = Z_i \equiv \log \left[\frac{f_1(X_i)}{f_0(X_i)} \right]$ to obtain

$$(19.17) \quad E_j(S_{N_{B,A}}) = E_j(N_{B,A}) E_j(Z_1), \quad j = 0, 1.$$

Here $E_j(N_{B,A}) < \infty$ by Proposition 19.2, while we must assume that $E_j|Z_1| < \infty$, $j = 0, 1$. If the conditions for Wald's approximations hold, then under H_0 ,

$$S_{N_{B,A}} \approx \begin{cases} -b & \text{with probability } 1 - \alpha, \\ a & \text{with probability } \alpha, \end{cases}$$

where $\alpha = \alpha(B, A)$, so

$$(19.18) \quad E_0(S_{N_{B,A}}) \approx -(1 - \alpha)b + \alpha a.$$

Since $E_0(Z_i) = -K(f_0, f_1)$, it follows by (19.17), (19.18), and (19.14) that

$$(19.19) \quad \begin{aligned} E_0(N_{B,A}) &\approx \frac{-(1 - \alpha)b + \alpha a}{-K(f_0, f_1)} \\ &\approx \frac{(1 - \alpha) \log \frac{1 - \alpha}{\beta} + \alpha \log \frac{\alpha}{1 - \beta}}{K(f_0, f_1)}, \end{aligned}$$

and similarly

$$(19.20) \quad E_1(N_{B,A}) \approx \frac{(1 - \beta) \log \frac{1 - \beta}{\alpha} + \beta \log \frac{\beta}{1 - \alpha}}{K(f_1, f_0)}. \quad \square$$

Remark 19.7. For the symmetric SPRT(A^{-1}, A) (where $b = a$), Wald's approximations (19.15) become

$$(19.21) \quad \alpha(A^{-1}, A) \approx \frac{1}{A + 1}, \quad \beta(A^{-1}, A) \approx \frac{1}{A + 1},$$

while (19.19), (19.20), and (19.21) combine to yield

$$\begin{aligned}
 (19.22) \quad E_0(N_{A^{-1}, A}) &\approx \frac{\frac{A-1}{A+1} \log A}{K(f_0, f_1)} \approx \frac{\log A}{K(f_0, f_1)} \quad \text{if } A \text{ is large,} \\
 E_1(N_{A^{-1}, A}) &\approx \frac{\frac{A-1}{A+1} \log A}{K(f_1, f_0)} \approx \frac{\log A}{K(f_1, f_0)} \quad \text{if } A \text{ is large.} \quad \square
 \end{aligned}$$

Theorem 19.8 (Optimality of the SPRT in the i.i.d. case). *Let X_1, X_2, \dots be i.i.d. Consider the problem of testing the simple hypotheses*

$$(19.23) \quad H_0 : X_i \sim f_0 \quad \text{vs.} \quad H_1 : X_i \sim f_1.$$

Let $N_{B,A}$ denote the stopping time of the SPRT(B, A) $\equiv T$, whose error probabilities α and β are given by (19.9) and (19.10). Let $T' \equiv (N', \{\phi'_n\})$ be any other test (fixed or sequential sample size N') with the same error probabilities as T and such that $E_0(N') < \infty$, $E_1(N') < \infty$. Then

$$(19.24) \quad E_0(N_{B,A}) \leq E_0(N'),$$

$$(19.25) \quad E_1(N_{B,A}) \leq E_1(N').$$

Proof. We give the non-rigorous proof due to Wald in his 1947 book *Sequential Analysis*. (Also Wolfowitz, Kiefer, Wijsman.) Define the events

$$F = \{T' \text{ chooses } H_0\}, \quad F^c = \{T' \text{ chooses } H_1\}.$$

We will show below that

$$(19.26) \quad E_0[\lambda_{N'} \mid F] = \frac{\beta}{1 - \alpha},$$

$$(19.27) \quad E_0[\lambda_{N'} \mid F^c] = \frac{1 - \beta}{\alpha},$$

$$(19.28) \quad E_1[\lambda_{N'} \mid F] = \frac{1 - \alpha}{\beta},$$

$$(19.29) \quad E_1[\lambda_{N'} \mid F^c] = \frac{\alpha}{1 - \beta}.$$

Now apply Wald's Lemma to $S_{N'} \equiv \sum_{i=1}^{N'} Z_i \equiv \log \lambda_{N'}$ (cf. (19.7)): Then

$$\begin{aligned}
-K(f_0, f_1) E_0(N') &\equiv E_0(N') E_0(Z_i) \\
&= E_0(S_{N'}) && [\text{by (19.16)}] \\
&= E_0[\log \lambda_{N'} \mid F](1 - \alpha) + E_0[\log \lambda_{N'} \mid F^c]\alpha \\
&\leq \log E_0[\lambda_{N'} \mid F](1 - \alpha) + \log E_0[\lambda_{N'} \mid F^c]\alpha \\
&= (1 - \alpha) \log \frac{\beta}{1 - \alpha} + \alpha \log \frac{1 - \beta}{\alpha}
\end{aligned}$$

by (19.26) and (19.27). Thus (19.24) holds *approximately*:

$$(19.30) \quad E_0(N') \geq \frac{(1 - \alpha) \log \frac{1 - \alpha}{\beta} + \alpha \log \frac{\alpha}{1 - \beta}}{K(f_0, f_1)} \approx E_0(N_{B,A})$$

by (19.19), and (19.25) is similarly approximated.

It remains to establish (19.26) – (19.29). For (19.26):

$$\begin{aligned}
E_0[\lambda_{N'} \mid F] &= \frac{E_0[\lambda_{N'} I_F]}{P_0[F]} \\
&= \frac{1}{1 - \alpha} E_0 \left[\sum_{n=1}^{\infty} \lambda_n I_{F \cap \{N'=n\}} \right] && [P_0(N' < \infty) = 1] \\
&= \frac{1}{1 - \alpha} \sum_{n=1}^{\infty} E_0[\lambda_n I_{F \cap \{N'=n\}}] \\
&\stackrel{*}{=} \frac{1}{1 - \alpha} \sum_{n=1}^{\infty} \int_{F \cap \{N'=n\}} \frac{f_1(\mathbf{x}_n)}{f_0(\mathbf{x}_n)} f_0(\mathbf{x}_n) d\mathbf{x}_n \\
&= \frac{1}{1 - \alpha} \sum_{n=1}^{\infty} \int_{F \cap \{N'=n\}} f_1(\mathbf{x}_n) d\mathbf{x}_n \\
&\stackrel{*}{=} \frac{1}{1 - \alpha} \sum_{n=1}^{\infty} P_1[F \cap \{N' = n\}] \\
&= \frac{1}{1 - \alpha} P_1(F) && [P_1(N' < \infty) = 1] \\
&\equiv \frac{\beta}{1 - \alpha}.
\end{aligned}$$

(*): These equalities hold because $(N', \{\phi'_n\})$ is a sequential test, which implies that the event $F \cap \{N' = n\}$ depends only on $\mathbf{x}_n \equiv (x_1, \dots, x_n)$.

The relations (19.27) – (19.29) are established similarly. (Note that the proof of (19.26) – (19.29) does not require the i.i.d. assumption.) \square

Example 19.9. In Theorem 19.8, suppose that $f_0 = N_1(\mu_0, \sigma^2)$ and $f_1 = N_1(\mu_1, \sigma^2)$. Let T' be the *fixed sample size* test that achieves specified error probabilities $\alpha, \beta > 0$. It is well known that the required fixed sample size is

$$N' = \frac{\sigma^2(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2}.$$

From (14.11) (p.229) the K-L distance $K(f_0, f_1)$ is given by

$$K(f_0, f_1) = \frac{(\mu_1 - \mu_0)^2}{2\sigma^2},$$

so by (19.19),

$$\begin{aligned} \frac{E_0(N_{B,A})}{N'} &\approx \frac{(1 - \alpha) \log \frac{1-\alpha}{\beta} + \alpha \log \frac{\alpha}{1-\beta}}{K(f_0, f_1) [\sigma^2(z_\alpha + z_\beta)^2 / (\mu_1 - \mu_0)^2]} \\ &= \frac{(1 - \alpha) \log \frac{1-\alpha}{\beta} + \alpha \log \frac{\alpha}{1-\beta}}{[(z_\alpha + z_\beta)^2 / 2]}. \end{aligned}$$

Thus in the symmetric case where $B = A^{-1}$ (so $\alpha = \beta$), we have

$$\frac{E_j(N_{A^{-1},A})}{N'} = \frac{(1 - 2\alpha) \log \frac{1-\alpha}{\alpha}}{2z_\alpha^2} = \begin{cases} .49 & \text{if } \alpha = .05, \\ .42 & \text{if } \alpha = .01, \end{cases}$$

for $j = 0, 1$. Here the SPRT requires, on average, fewer than half as many observations as the fixed sample size test. \square

19.2. Uniform consistency of tests; need for sequential sampling.

For testing a simple hypothesis vs. a simple alternative, the SPRT reduces the expected sample size needed to control both error probabilities, but sequential sampling is not necessary to achieve this control. In this section we examine composite hypotheses to determine when sequential sampling is indeed necessary.

Let X_1, X_2, \dots be an (infinite) sequence of i.i.d. random variables or random vectors with common unknown distribution $P \in \mathcal{P}$ (the statistical model). Suppose that $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$, where $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. We wish to test

$$(19.31) \quad H_0 : P \in \mathcal{P}_0 \quad \text{vs.} \quad H_1 : P \in \mathcal{P}_1.$$

Let $\phi_n \equiv \phi_n(X_1, \dots, X_n)$ denote a test function based on the finite set of observations X_1, \dots, X_n only. Such ϕ_n is called a *finite sample test*.

Definition 19.10. A sequence of finite sample tests $\{\phi_n\}$ is *consistent* if

$$(19.32) \quad \lim_{n \rightarrow \infty} E_P[\phi_n] = \begin{cases} 0, & \text{if } P \in \mathcal{P}_0; \\ 1, & \text{if } P \in \mathcal{P}_1; \end{cases}$$

that is, if the power approaches 0 on the null hypothesis and 1 on the alternative. The sequence is *uniformly consistent* if this convergence occurs uniformly on \mathcal{P}_0 and \mathcal{P}_1 , that is, if

$$(19.33) \quad \begin{cases} \lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0} E_P[\phi_n] = 0; \\ \lim_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_1} E_P[\phi_n] = 1. \end{cases}$$

If there exists a uniformly consistent sequence of *finite sample tests* $\{\phi_n\}$, then \mathcal{P}_0 and \mathcal{P}_1 are called *finitely distinguishable (f.d.)*. \square

Proposition 19.11. [A. Berger (1951) *Ann. Math. Statist.*] A sufficient condition for finite distinguishability of \mathcal{P}_0 and \mathcal{P}_1 is the following:

$$(19.34) \quad d(\mathcal{P}_0, \mathcal{P}_1) := \sup_A \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} |P_0(A) - P_1(A)| > 0.$$

Proof. By (19.34) we can choose A , $\delta > 0$ such that

$$(19.35) \quad |P_0(A) - P_1(A)| > \delta \quad \forall P_0, P_1.$$

Define $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n I_A(X_i)$, so $\hat{p}_n \rightarrow p \equiv P(A)$ when P obtains. Define $B_0 = \{p \mid P \in \mathcal{P}_0\}$ and define the finite sample test ϕ_n by

$$(19.36) \quad \phi_n = \begin{cases} 1, & \text{if } |\hat{p}_n - B_0| > \frac{\delta}{2}; \\ 0, & \text{if } |\hat{p}_n - B_0| \leq \frac{\delta}{2}. \end{cases}$$

Fix $\epsilon > 0$ and choose $n \geq \frac{1}{\delta^2 \epsilon}$. Then by Chebyshev's inequality,

$$(19.37) \quad \Pr_P \left[|\hat{p}_n - p| > \frac{\delta}{2} \right] \leq \frac{4p(1-p)}{\delta^2 n} \leq \epsilon \quad \forall P.$$

If $P \in \mathcal{P}_0$ then $|\hat{p}_n - B_0| > \frac{\delta}{2} \Rightarrow |\hat{p}_n - p| > \frac{\delta}{2}$, so

$$(19.38) \quad \sup_{P \in \mathcal{P}_0} \Pr_P \left[|\hat{p}_n - B_0| > \frac{\delta}{2} \right] \leq \sup_{P \in \mathcal{P}_0} \Pr_P \left[|\hat{p}_n - p| > \frac{\delta}{2} \right] \leq \epsilon.$$

If $P \in \mathcal{P}_1$ then $|\hat{p}_n - p| < \frac{\delta}{2} \Rightarrow |\hat{p}_n - B_0| > \frac{\delta}{2}$, so

$$(19.39) \quad \inf_{P \in \mathcal{P}_1} \Pr_P \left[|\hat{p}_n - B_0| > \frac{\delta}{2} \right] \geq \inf_{P \in \mathcal{P}_1} \Pr_P \left[|\hat{p}_n - p| < \frac{\delta}{2} \right] \geq 1 - \epsilon.$$

Thus, $\{\phi_n\}$ is uniformly consistent, as required. \square

Remark 19.12. Berger's condition (19.34) is sufficient for finite distinguishability (f.d.), but not necessary. To see this, consider $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1 = \{P_1, P_2, P_3\}$, four discrete distributions on $\{a, b, c\}$:

$$P_0 = \left(\frac{2}{6}, \frac{2}{6}, \frac{2}{6} \right), \quad P_1 = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6} \right), \quad P_2 = \left(\frac{2}{6}, \frac{3}{6}, \frac{1}{6} \right), \quad P_3 = \left(\frac{3}{6}, \frac{1}{6}, \frac{2}{6} \right).$$

Then there are 2^3 possibilities for the sets A in (19.6): $\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}$. But for each such A , $P_0(A) = P_i(A)$ for some $i = 1, 2, 3$, so $d(\mathcal{P}_0, \mathcal{P}_1) = 0$. Nonetheless, \mathcal{P}_0 and \mathcal{P}_1 are f.d, by Prop. 19.14 below. Thus, (19.34) is not necessary for f.d. \square

Exercise 19.13. (i) Show, however, that $\mathcal{P}_0^2 \equiv \{P_0 \times P_0\}$ and $\mathcal{P}_1^2 \equiv \{P_1 \times P_1, P_2 \times P_2, P_3 \times P_3\}$ do satisfy Berger's condition (19.6), where $P \times P$ represents the distribution of two i.i.d. repetitions X_1, X_2 from P .

(ii)*** [class research project]. Study the following *Conjecture*:

A necessary and sufficient condition that \mathcal{P}_0 and \mathcal{P}_1 are finitely distinguishable is that $d(\mathcal{P}_0^n, \mathcal{P}_1^n) > 0$ for some $n = 1, 2, 3, \dots$ \square

Proposition 19.14. *Finite families \mathcal{P}_0 and \mathcal{P}_1 are finitely distinguishable.*

Proof. Consider the sequence $\{\phi_n^*\}$ of finite sample size tests where ϕ_n^* is the LRT based on $\frac{\prod f_{\hat{\theta}_1}(x_i)}{\prod f_{\hat{\theta}_0}(x_i)} \leq (>) 1$ as defined in (18.45). Then (14.24)-(14.28) in the proof of Theorem 14.7(i) shows that $\{\phi_n^*\}$ is consistent for testing \mathcal{P}_0 vs. \mathcal{P}_1 . Because \mathcal{P}_0 and \mathcal{P}_1 are finite this is equivalent to uniform consistency, hence these families are finitely distinguishable. \square

A necessary and sufficient condition for finite distinguishability has been given by Hoeffding and Wolfowitz. (Also see related work by Kraft, LeCam, D. Freedman....)

Definition 19.15. For two probability measures P_0 and P_1 on $(\mathcal{X}, \mathcal{S})$, define the *total variation (TV) distance*

$$(19.40) \quad D(P_0, P_1) = \sup_{A \in \mathcal{S}} |P_0(A) - P_1(A)|.$$

Clearly $0 \leq D(P_0, P_1) \leq 1$, $D(P_0, P_1) = 0$ iff $P_0 = P_1$, and $D(P_0, P_1) = 1$ iff the supports of P_0 and P_1 are disjoint. \square

For two disjoint sets of probability measures \mathcal{P}_0 and \mathcal{P}_1 , define

$$(19.41) \quad D(\mathcal{P}_0, \mathcal{P}_1) = \inf_{P_0 \in \mathcal{P}_0, P_1 \in \mathcal{P}_1} D(P_0, P_1).$$

Then \mathcal{P}_0 and \mathcal{P}_1 are *uniformly TV-separated* if

$$(19.42) \quad D(\mathcal{P}_0, \mathcal{P}_1) > 0,$$

while \mathcal{P}_0 and \mathcal{P}_1 are *pointwise TV-separated* if

$$(19.43) \quad D(P_0, P_1) > 0 \quad \forall P_0 \in \mathcal{P}_0, \quad D(P_0, P_1) > 0 \quad \forall P_1 \in \mathcal{P}_1.$$

(Compare to (18.44a) and (18.44b), and see Exercise 19.26(ii).)

Hoeffding and Wolfowitz (1958) *Ann. Math. Statist.* showed that *uniform TV-separation is necessary and sufficient for finite distinguishability*:

$$(19.44) \quad \mathcal{P}_0 \text{ and } \mathcal{P}_1 \text{ are f.d.} \iff D(\mathcal{P}_0, \mathcal{P}_1) > 0.$$

Note that Berger's sufficient condition (19.34) implies H-W's necessary and sufficient condition (19.44) because $D(\mathcal{P}_0, \mathcal{P}_1) \geq d(\mathcal{P}_0, \mathcal{P}_1)$. We shall prove (\Rightarrow) in (19.44) by means of the following well-known result.

Lemma 19.16. *Let P, Q be probability measures on $(\mathcal{X}, \mathcal{S})$ and let p, q be their corresponding pdfs w.r.to some measure ν . Then*

$$(19.45) \quad 1 - \int_{\mathcal{X}} (p \wedge q) d\nu = \frac{1}{2} \int_{\mathcal{X}} |p - q| d\nu$$

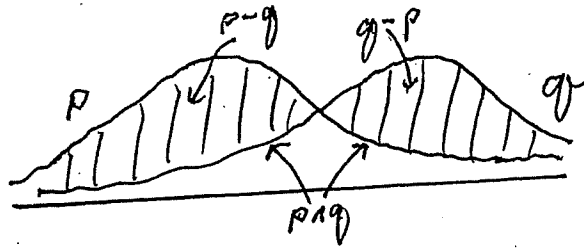
$$(19.46) \quad = \int_{\{p > q\}} (p - q) d\nu$$

$$(19.47) \quad = D(P, Q).$$

Proof. First, (19.45) follows from [see figure]

$$(19.48) \quad \int_{\mathcal{X}} |p - q| = \int_{\{p > q\}} (p - q) + \int_{\{q > p\}} (q - p),$$

$$(19.49) \quad 2 = \int_{\mathcal{X}} p + \int_{\mathcal{X}} q = \int_{\{p > q\}} (p - q) + \int_{\{q > p\}} (q - p) + 2 \int_{\mathcal{X}} (p \wedge q).$$



Next,

$$(19.50) \quad 1 = \int_{\{p > q\}} (p - q) + \int_{\mathcal{X}} (p \wedge q) = \int_{\{q > p\}} (q - p) + \int_{\mathcal{X}} (p \wedge q)$$

$$(19.51) \quad \Rightarrow \int_{\{p > q\}} (p - q) = \int_{\{q > p\}} (q - p)$$

$\Rightarrow (19.46) \text{ holds} \quad [\text{by (19.48)}].$

Last, clearly

$$(19.52) \quad D(P, Q) \geq \int_{\{p > q\}} (p - q) \quad [\text{take } A = \{p > q\}].$$

Also, for any A ,

$$(19.53) \quad \begin{aligned} \int_{\{p > q\}} (p - q) &\geq \int_{A \cap \{p > q\}} (p - q) \geq \int_{A \cap \{p > q\}} (p - q) + \int_{A \cap \{q > p\}} (p - q) \\ &\geq \int_{A \cap \{q > p\}} (p - q) \geq - \int_{\{q > p\}} (q - p) = - \int_{\{p > q\}} (p - q), \end{aligned}$$

where the equality follows from (19.51). But

$$\int_{A \cap \{p > q\}} (p - q) + \int_{A \cap \{q > p\}} (p - q) = \int_A (p - q) = P(A) - Q(A),$$

so from (19.53),

$$|P(A) - Q(A)| \leq \int_{\{p > q\}} (p - q) \quad \forall A.$$

Therefore $D(P, Q) \leq \int_{\{p > q\}} (p - q)$, which together with (19.52) implies (19.47). \square

Proposition 19.17. (\Rightarrow) holds in (19.44).

Proof. Since \mathcal{P}_0 and \mathcal{P}_1 are f.d., $\exists \epsilon > 0$, $\exists n$, and \exists a finite sample test ϕ_n such that for all $P_0 \in \mathcal{P}_0$, $P_1 \in \mathcal{P}_1$,

$$\begin{aligned} \epsilon &< E_{P_1}[\phi_n] - E_{P_0}[\phi_n] = \int \phi_n \left[\prod_{i=1}^n f_1(x_i) - \prod_{i=1}^n f_0(x_i) \right] \\ &\leq \int_{\{\prod f_1(x_i) > \prod f_0(x_i)\}} [\prod f_1(x_i) - \prod f_0(x_i)] \\ &= \frac{1}{2} \int |\prod f_1(x_i) - \prod f_0(x_i)| \quad [\text{by (19.46)}] \end{aligned}$$

$$\begin{aligned}
&= 1 - \int (\prod f_0(x_i)) \wedge (\prod f_1(x_i)) && \text{[by (19.45)]} \\
&\leq 1 - \int \prod (f_0(x_i) \wedge f_1(x_i)) \\
&= 1 - \left(\int f_0(x_i) \wedge f_1(x_i) \right)^n \\
&= 1 - \left(1 - \frac{1}{2} \int |f_1(x) - f_0(x)| \right)^n, && \text{[by (19.45)]}
\end{aligned}$$

where f_0, f_1 are pdfs for P_0, P_1 , resp. Thus

$$\begin{aligned}
(1 - \epsilon) &\geq \left[1 - \frac{1}{2} \int |f_1(x) - f_0(x)| \right]^n \\
&= [1 - D(P_0, P_1)]^n \quad \forall P_0, P_1,
\end{aligned}$$

so

$$D(P_0, P_1) \geq 1 - (1 - \epsilon)^{1/n} > 0. \quad \square$$

Example 19.18. Let $\mathcal{P}_i = \{N(\mu_i, \sigma^2) \mid 0 < \sigma^2 < \infty\}$, $i = 0, 1$, where $\mu_0 < \mu_1$. Note that the power of the size α t -test for testing \mathcal{P}_0 vs. \mathcal{P}_1 approaches α on \mathcal{P}_1 as $\sigma \rightarrow \infty$. [Verify; see Lehmann, TSH, Ch.5, Problem 2.] We apply Proposition 19.17 to show that in fact, \mathcal{P}_0 and \mathcal{P}_1 are *not distinguishable by any sequence of finite sample tests*. For this, it suffices to show that $D(\sigma) \equiv D(N(\mu_0, \sigma^2), N(\mu_1, \sigma^2)) \rightarrow 0$ as $\sigma \rightarrow \infty$:

$$\begin{aligned}
D(\sigma) &= \int_{\{f_{\mu_0, \sigma}(x) > f_{\mu_1, \sigma}(x)\}} [f_{\mu_0, \sigma}(x) - f_{\mu_1, \sigma}(x)] && \text{[by (19.46)]} \\
&= \int_{\{x < (\mu_0 + \mu_1)/2\}} [f_{\mu_0, \sigma}(x) - f_{\mu_1, \sigma}(x)] \\
&= \Phi\left(\frac{\mu_1 - \mu_0}{2\sigma}\right) - \Phi\left(\frac{\mu_0 - \mu_1}{2\sigma}\right) \\
&\rightarrow \frac{1}{2} - \frac{1}{2} = 0 \quad \text{as } \sigma \rightarrow \infty. && \square
\end{aligned}$$

Example 19.18 demonstrates the need for *tests based on sequential sampling rules*, or, simply, *sequential tests*. Let X_1, X_2, \dots be a sequence of random variates (not necessarily independent or real-valued). A *sequential*

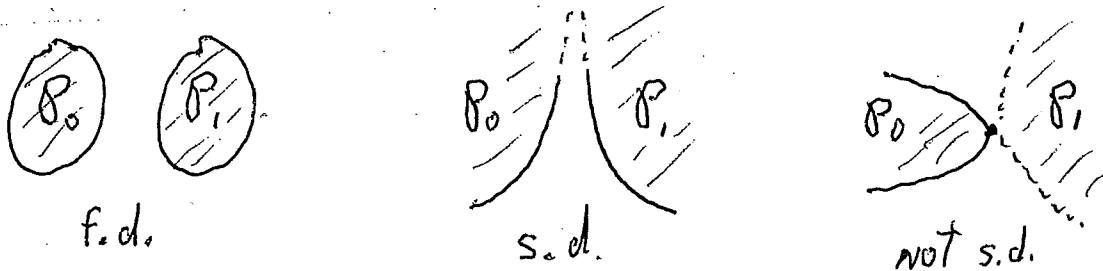
sampling rule \equiv *stopping time* is an integer-valued random variable N such that for each $n = 1, 2, \dots$, the event $\{N = n\}$ depends only on X_1, \dots, X_n . That is, the decision to stop at time n depends only on X_1, \dots, X_n (not on subsequent X_i 's). Unless otherwise noted, we shall require that the procedure stops with probability 1, that is, $\Pr[N < \infty] = 1$.

A *sequential test* $(N, \{\phi_n\})$ consists of a stopping rule N and a sequence $\{\phi_n\}$ of finite sample tests applied as follows: when $N = n$, the sampling is stopped and the test $\phi_n \equiv \phi_n(X_1, \dots, X_n)$ is applied.

Definition 19.19. \mathcal{P}_0 and \mathcal{P}_1 are *sequentially distinguishable (s.d.)* if, for each $\epsilon > 0$, \exists a sequential test $(N^\epsilon, \{\phi_n^\epsilon\})$ such that

$$(19.54) \quad \begin{aligned} \sup_{P \in \mathcal{P}_0} E_P[\phi_{N^\epsilon}^\epsilon] &\leq \epsilon, \\ \inf_{P \in \mathcal{P}_1} E_P[\phi_{N^\epsilon}^\epsilon] &\geq 1 - \epsilon. \end{aligned}$$

Hoeffding and Wolfowitz (1958) *Ann. Math. Statist.* showed that *pointwise TV-separation* (19.43) is necessary and sufficient for sequential distinguishability [see figure].



We will not prove this but will illustrate the necessity by an example:

Example 19.20. (*Stein's (second) two-stage testing procedure*) [See TSH Ch.5, Problems 26(i), (iv) and part of 28(iii).] Let X_1, X_2, \dots be a sequence of i.i.d. $N(\mu, \sigma^2)$ r.v.s. Let

$$(19.55) \quad \begin{aligned} \mathcal{P}_0 &= \{N(\mu, \sigma^2) \mid \mu \leq \mu_0, 0 < \sigma^2 < \infty\}, \\ \mathcal{P}_1 &= \{N(\mu, \sigma^2) \mid \mu \geq \mu_1, 0 < \sigma^2 < \infty\}, \end{aligned}$$

where $\mu_0 < \mu_1$ are fixed. Then it is readily shown that \mathcal{P}_0 and \mathcal{P}_1 are pointwise TV-separated (19.43). [Verify: consider $A = [\mu_1, \infty)$ and $(-\infty, \mu_0]$.] We now show directly that \mathcal{P}_0 and \mathcal{P}_1 are s.d. Fix $0 < \alpha < \frac{1}{2}$ and $0 < \beta < \frac{1}{2}$.

Stage 1: Fix an initial sample size $m \geq 2$ and observe X_1, \dots, X_m . Set

$$\bar{X}_m = \frac{1}{m} \sum_{i=1}^m X_i, \quad s_m^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X}_m)^2.$$

Stage 2: Let $N = \text{smallest integer} \geq \max \left\{ m, \left(\frac{t_{m-1,\alpha} + t_{m-1,\beta}}{\mu_1 - \mu_0} \right)^2 s_m^2 \right\}$,

Take $N - m \geq 0$ additional observations X_{m+1}, \dots, X_N . Then N ($\geq m$) trivially is an (unbounded) random stopping time, since N depends on $\{X_i\}$ only through s_m^2 , hence only through X_1, \dots, X_m with m fixed. Set

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i, \quad \phi_N(X_1, \dots, X_N) = \begin{cases} 0, & \text{if } \bar{X}_N \leq c, \\ 1, & \text{if } \bar{X}_N > c \end{cases}.$$

where $c = \left(\frac{t_{m-1,\alpha}\mu_1 + t_{m-1,\beta}\mu_0}{t_{m-1,\alpha} + t_{m-1,\beta}} \right)$. Then:

$$(i) \quad \forall \mu, \sigma^2, \quad \Pr_{\mu, \sigma^2} [N < \infty] = 1.$$

Proof of (i): Note that N depends on s_m^2 only, so its distribution depends only on σ^2 , not μ . Then

$$N \leq \max \left\{ m, \left(\frac{t_{m-1,\alpha} + t_{m-1,\beta}}{\mu_1 - \mu_0} \right)^2 s_m^2 \right\} + 1 < \infty \quad \text{w.pr.1.}$$

$$(ii) \quad \sup_{\mathcal{P}_0} \Pr_{\mu, \sigma^2} [\bar{X}_N > c] \leq \alpha, \quad \sup_{\mathcal{P}_1} \Pr_{\mu, \sigma^2} [\bar{X}_N \leq c] \leq \beta.$$

Proof of (ii): Let the true parameter values be μ, σ^2 and set $T = \frac{\sqrt{N}(\bar{X}_N - \mu)}{s_m}$. Clearly T does not depend on μ . We claim that $T \sim t_{m-1}$ (not depending on σ). To see this, consider the conditional distribution

$$\begin{aligned} T \mid s_m &\sim T \mid s_m, N & [N = N(s_m)] \\ &= \frac{\sqrt{N}(\bar{X}_N - \mu)}{s_m} \mid s_m, N \\ &\sim N\left(0, \frac{\sigma^2}{s_m^2}\right). & [\sqrt{N}(\bar{X}_N - \mu) \perp\!\!\!\perp s_m \text{ (why?)}] \end{aligned}$$

Therefore $\frac{s_m}{\sigma}T \mid s_m \sim N(0, 1)$, so $\frac{s_m}{\sigma}T \perp\!\!\!\perp s_m$. Thus

$$(19.56) \quad T \equiv \frac{(s_m/\sigma)T}{s_m/\sigma} \sim \frac{N(0, 1)}{\sqrt{\frac{\chi_{m-1}^2}{m-1}}} \sim t_{m-1}.$$

Therefore, for any $\mu \leq \mu_0$,

$$\begin{aligned} \Pr_{\mu, \sigma}[\bar{X}_N > c] &= \Pr_{\mu, \sigma}[\bar{X}_N - \mu > c - \mu] \\ &= \Pr_{\sigma}\left[T > \frac{\sqrt{N}(c - \mu)}{s_m}\right] \\ &\leq \Pr_{\sigma}\left[T > \frac{\sqrt{N}(c - \mu_0)}{s_m}\right] \\ &= \Pr_{\sigma}\left[T > \frac{\sqrt{N}t_{m-1, \alpha}(\mu_1 - \mu_0)}{s_m(t_{m-1, \alpha} + t_{m-1, \beta})}\right] \\ &\leq \Pr[T > t_{m-1, \alpha}] \equiv \alpha. \end{aligned} \quad [\text{verify}]$$

Similarly, for any $\mu \geq \mu_1$, $\Pr_{\mu, \sigma}[\bar{X}_N \leq c] \leq \beta$. This establishes (ii). \square

Example 19.21. (*Stein's two-stage fixed-width confidence interval*) [See TSH Ch.5, Problem 27(ii).] The preceding ideas are related to the notion of *fixed-width confidence intervals*. Suppose we want a confidence interval for μ of specified width L and specified confidence (\geq) γ , e.g., $\gamma = .95$. Choose $c > 0$ such that

$$\Pr\left[|t_{m-1}| \leq \frac{1}{2\sqrt{c}}\right] = \gamma.$$

Fix m and observe X_1, \dots, X_m . Let $N = \max\left\{m, \left\lceil \frac{s_m^2}{c} \right\rceil + 1\right\}$. Then

$$\begin{aligned} \Pr_{\mu, \sigma}\left[|\bar{X}_N - \mu| \leq \frac{L}{2}\right] &= \Pr_{\mu, \sigma}\left[\frac{\sqrt{N}|\bar{X}_N - \mu|}{s_m} \leq \frac{\sqrt{N}L}{2s_m}\right] \\ &\geq \Pr_{\sigma}\left[|T| \leq \frac{L}{2\sqrt{c}}\right] \equiv \gamma. \quad [\text{since } T \sim t_{m-1}] \end{aligned}$$

Thus $\bar{X}_N \pm \frac{L}{2}$ is a confidence interval for μ of width L and confidence $\geq \gamma$.

Example 19.22. (No fixed-sample-size, fixed-length confidence interval for μ exists) [See TSH Ch.5, Problem 25.] Let $\mathbf{X}_n = (X_1, \dots, X_n)$. Any confidence interval for μ of fixed length L must have the form $\delta(\mathbf{X}_n) \pm \frac{L}{2}$. We shall show that this interval has confidence 0, i.e.,

$$(19.57) \quad \inf_{-\infty < \mu < \infty, 0 < \sigma < \infty} \Pr_{\mu, \sigma} \left[|\delta(\mathbf{X}_n) - \mu| \leq \frac{L}{2} \right] = 0,$$

(This also shows that any confidence interval with random but bounded width also has confidence 0.)

Define $S_i = \{\mathbf{X}_n \mid |\delta(\mathbf{X}_n) - \mu_i| \leq \frac{L}{2}\}$, where $\mu_i = 2iL$, $i = 1, 2, \dots$. Thus S_1, S_2, \dots are mutually disjoint. We shall show that for any integer $K = 1, 2, \dots$, $\exists \sigma_0(K) > 0$ such that

$$(19.58) \quad \sigma > \sigma_0(K) \implies |\Pr_{\mu_i, \sigma}(S_i) - \Pr_{\mu_1, \sigma}(S_i)| \leq \frac{1}{K}, \quad i = 1, \dots, K.$$

But $\min_{i=1, \dots, K} \Pr_{\mu_1, \sigma}(S_i) \leq \frac{1}{K}$ since S_1, \dots, S_K are disjoint, so

$$(19.59) \quad \sigma > \sigma_0(K) \implies \min_{i=1, \dots, K} \Pr_{\mu_i, \sigma}(S_i) \leq \frac{2}{K}, \quad i = 1, \dots, K.$$

Now let $K \rightarrow \infty$ in (19.59) to obtain (19.57).

To establish (19.58), first note that $\mathbf{X}_n \sim N_n(\mu \mathbf{e}_n, \sigma^2 I_n)$, where $\mathbf{e}_n = (1, \dots, 1)$. Then

$$(19.60) \quad \begin{aligned} & D(N_n(\mu_i \mathbf{e}_n, \sigma^2 I_n), N_n(\mu_1 \mathbf{e}_n, \sigma^2 I_n)) \\ &= D\left(N_n\left(\frac{\mu_i}{\sigma} \mathbf{e}_n, I_n\right), N_n\left(\frac{\mu_1}{\sigma} \mathbf{e}_n, I_n\right)\right) \\ \text{[verify]} & \rightarrow 0 \quad \text{as } \sigma \rightarrow \infty. \end{aligned}$$

[Both normal pdfs (p_σ, q_σ , say) in (19.60) converge to the $N_n(0, I_n)$ pdf (p_0 , say) as $\sigma \rightarrow \infty$, so Scheffe's convergence theorem applies to $|p_\sigma - q_\sigma| \leq |p_\sigma - p_0| + |q_\sigma - p_0|$ - see TSH, Appendix, Lemma 4. Now apply (19.45)-(19.47).] Thus we can choose $\sigma_0(K)$ large enough so that $\sigma > \sigma_0(K) \implies$

$$D\left(N_n\left(\frac{\mu_i}{\sigma} \mathbf{e}_n, I_n\right), N_n\left(\frac{\mu_1}{\sigma} \mathbf{e}_n, I_n\right)\right) < \frac{1}{K}, \quad i = 1, \dots, K.$$

Remark 19.23. Let $\mathcal{P}_0 = \{N(\mu, 1) \mid \mu \leq \mu_0, \}$, $\mathcal{P}_1 = \{N(\mu, 1) \mid \mu > \mu_0, \}$. Here \mathcal{P}_0 and \mathcal{P}_1 are not pointwise D -separated so they are not s.d. Robbins and Siegmund (ca. 1970) devised sequential tests for \mathcal{P}_0 vs. \mathcal{P}_1 whose power can be made uniformly close to 1 on \mathcal{P}_1 , at the cost, however, that sampling may never terminate if \mathcal{P}_0 holds. They argue that this may not be a drawback in practical terms, since if the null hypothesis \mathcal{P}_0 is true, it is ok if we never stop to reject it. \square

Exercise 19.24. Problems 26, 27, 28 in TSH Ch.5. Some parts of these problems already have been treated in the above Examples – you need not repeat their derivations, but you may refer to them. \square

Definition 19.25. Let P, Q be probability measures on a measure space $(\mathcal{X}, \mathcal{S})$ and let p, q be their corresponding pdfs w.r.to some measure ν . The *Hellinger distance* $H(P, Q)$ is defined via

$$(19.61) \quad H^2(P, Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\nu = 1 - \int \sqrt{pq} d\nu.$$

Clearly $0 \leq H(P, Q) \leq 1$, $H(P, Q) = 0$ iff $P = Q$, and $H(P, Q) = 1$ iff the supports of P and Q are disjoint. Note that $0 \leq \int \sqrt{pq} d\nu \leq 1$. \square

Exercise 19.26. *Relations among Hellinger, total variation, and Kullback-Leibler distances.* Show that

- (i) $H^2(P, Q) \leq D(P, Q) \leq 2H(P, Q),$
- (ii) $H^2(P, Q) \leq \frac{1}{2}K(P, Q).$

These imply that TV-separation is equivalent to H-separation, and that both imply KL-separation.

(iii)** Does KL-separation imply TV \equiv H-separation? \square

Exercise 19.27. Finite families \mathcal{P}_0 and \mathcal{P}_1 are finitely distinguishable with an exponential error rate. (This strengthens Proposition 19.14).

Hint: Use Hellinger distance – recall (18.30). \square

Supplement 1: Censored Data exercise with solution. Let T_1, \dots, T_n be i.i.d. survival times with $T_i \sim \text{Exponential}(\theta)$, that is, T_i has pdf

$$f_\theta(t) = \theta e^{-\theta t}, \quad 0 < t < \infty, \quad 0 < \theta < \infty.$$

Let $\tau > 0$ be a known, fixed time point and define

$$U_i = \begin{cases} 1 & \text{if } T_i < \tau, \\ 0 & \text{if } T_i \geq \tau, \end{cases} \quad V_i = \min(T_i, \tau) = \begin{cases} T_i & \text{if } T_i < \tau, \\ \tau & \text{if } T_i \geq \tau. \end{cases}$$

(i) If only U_1, \dots, U_n are observed, what is the MLE $\tilde{\theta}$ of θ ? Does it always exist?

(ii) If only V_1, \dots, V_n are observed, what is the MLE $\hat{\theta}$ of θ ? Does it always exist?

Hint: express the “pdf” of V_i as a mixture of the $\text{Exponential}(\theta)$ pdf on $(0, \tau)$ and the discrete pmf that puts mass $e^{-\theta\tau}$ at τ .

(iii) Find the asymptotic normal distributions of $\tilde{\theta}$ and $\hat{\theta}$ (suitably standardized). Which is asymptotically more efficient? Why is your answer not surprising?

Solution:

(i) $U_i \sim \text{Bernoulli}(p)$, where $p = 1 - e^{-\theta\tau}$, so $0 < p < 1$. Thus

$$S \equiv \sum U_i \sim \text{Binomial}(n, p),$$

so this is a standard problem. The MLE of p is

$$\tilde{p} = \begin{cases} \frac{S}{n} & \text{if } 1 \leq S \leq n-1, \\ \text{does not exist} & \text{if } S = 0, n. \end{cases}$$

Because $\theta = -[\log(1-p)]/\tau$, the MLE of θ is

$$\tilde{\theta} = \begin{cases} -[\log(1 - \frac{S}{n})]/\tau & \text{if } 1 \leq S \leq n-1, \\ \text{does not exist} & \text{if } S = 0, n. \end{cases}$$

(ii) The distribution of V_i is a mixture of a continuous component on $(0, \tau)$ and a discrete component at τ . Thus the likelihood function is $\prod_{i=1}^n f_\theta(v_i)$,

where $f_\theta(v_i)$ is the "pdf" of this mixture w.r.to the dominating measure defined as normalized Lebesgue measure on $(0, \tau)$ and point mass 1 at τ . Thus

$$(1) \quad f_\theta(v_i) = \theta e^{-\theta v_i} \frac{1}{\tau} I_{(0, \tau)}(v_i) + e^{-\theta \tau} I_{\{\tau\}}(v_i),$$

$$\prod_{i=1}^n f_\theta(v_i) = \tau^{-S} \cdot \theta^S e^{-\theta T} e^{-\theta \tau(n-S)} = \tau^{-S} \cdot \theta^S e^{-\theta[T + \tau(n-S)]},$$

where $T = \sum U_i T_i$. Thus the MLE of θ is

$$\hat{\theta} = \begin{cases} \frac{S}{T + \tau(n-S)} & \text{if } 1 \leq S \leq n, \\ \text{does not exist} & \text{if } S = 0. \end{cases}$$

(iii) In both cases, $P[\text{MLE does not exist}] \rightarrow 0$ so we can ignore the non-existence for the asymptotic distributions. For $\tilde{\theta}$, $\sqrt{n}(\tilde{p} - p) \xrightarrow{d} N(0, p(1-p))$ and $d\theta/dp = 1/(1-p)\tau$, so apply propagation of error to obtain

$$(2) \quad \sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{p}{(1-p)\tau^2}\right) = N\left(0, \frac{e^{\theta\tau} - 1}{\tau^2}\right).$$

For $\hat{\theta}$, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, 1/I(\theta))$, where the Fisher information is obtained from (1):

$$nI(\theta) = -E\left[\frac{\partial^2 \log \prod_{i=1}^n f_\theta(V_i)}{\partial \theta^2}\right] = E\left(\frac{S}{\theta^2}\right) = \frac{np}{\theta^2} = \frac{n(1-e^{-\theta\tau})}{\theta^2},$$

so $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{\theta^2}{1-e^{-\theta\tau}}\right)$. It is straightforward to show that

$$\frac{\theta^2}{1-e^{-\theta\tau}} < \frac{e^{\theta\tau} - 1}{\tau^2} \quad \forall \theta,$$

so $\hat{\theta}$ is more efficient than $\tilde{\theta}$. This is not surprising since U_i is a function of V_i .

- 20. Estimation and Hypothesis Testing with Normal Data.**
[supplementary notes will be handed out]
- 21. Invariant Tests and Equivariant Estimators.**
[supplementary notes will be handed out]
- 22. The James-Stein Estimator.**
[supplementary notes will be handed out]
- 23. How Likely is Simpson's Paradox?**
[supplementary notes will be handed out]
- 24. Sharpening Buffon's Needle.**
[supplementary notes will be handed out]
- 25. Estimating the Face Probabilities of Shaved Dice.**
[supplementary notes will be handed out]
- 26. Circular and Spherical Copulas.**
[supplementary notes will be handed out]
- 27. The Emperor's New Tests.**
[supplementary notes will be handed out]
- 28. The Role of Reversals in Order-restricted Inference.**
[supplementary notes will be handed out]
- 29. Predicting Extinction or Explosion in a Galton-Watson
Branching Process.** [supplementary notes will be handed out]
- 30. Variance-stabilizing Transformations for a Normal
Correlation Coefficient.** [supplementary notes will be handed out]