

# Stat 643 Outline

## Spring 2010

Steve Vardeman  
Iowa State University

April 22, 2010

### Abstract

This outline summarizes the main points of the class lectures. Details (including proofs) must be seen in class notes and reference books.

## Contents

<b>1</b>	<b>Characteristic Functions</b>	<b>3</b>
1.1	Some Simple Generalities and Basic Facts About Chf's . . . . .	4
1.2	Chf's and Probabilities . . . . .	5
1.3	Chf's and Convergence in Distribution . . . . .	7
<b>2</b>	<b>Central Limit Theorems</b>	<b>8</b>
2.1	Basic CLTs for Independent Double Arrays . . . . .	8
2.2	Related Limit Results . . . . .	9
2.2.1	CLTs Under Dependence . . . . .	9
2.2.2	CLTs for Sums of Random Numbers of Random Summands	10
2.2.3	Rates for CLTs and Generalizations . . . . .	10
2.2.4	Large Deviation Results . . . . .	11
2.2.5	Results About "Stable Laws" . . . . .	11
2.2.6	Functional Central Limit Theorems and Empirical Processes	11
<b>3</b>	<b>Conditioning</b>	<b>13</b>
3.1	Basic Results About Conditional Expectation . . . . .	14
3.2	Conditional Variance . . . . .	15
3.3	Conditional Analogues of Properties of Ordinary Expectation . .	16
3.4	Conditional Probability . . . . .	16
<b>4</b>	<b>Transition to Statistics</b>	<b>18</b>

<b>5</b>	<b>Sufficiency and Related Concepts</b>	<b>19</b>
5.1	Sufficiency and the Factorization Theorem . . . . .	19
5.2	Minimal Sufficiency . . . . .	21
5.3	Ancillarity and Completeness . . . . .	22
<b>6</b>	<b>Facts About Common Statistical Models</b>	<b>24</b>
6.1	Bayes Models . . . . .	24
6.2	Exponential Families . . . . .	25
6.3	Measures of Statistical Information . . . . .	26
6.3.1	Fisher Information . . . . .	27
6.3.2	Kullback-Leibler Information . . . . .	29
<b>7</b>	<b>Statistical Decision Theory</b>	<b>31</b>
7.1	Basic Framework and Concepts . . . . .	31
7.2	(Finite Dimensional) Geometry of Decision Theory . . . . .	33
7.3	Complete Classes of Decision Rules . . . . .	34
7.4	Sufficiency and Decision Theory . . . . .	35
7.5	Bayes Decision Rules . . . . .	36
7.6	Minimax Decision Rules . . . . .	38
7.7	Invariance/Equivariance Theory . . . . .	38
7.7.1	Location Parameter Estimation and Invariance/Equivariance	39
7.7.2	Generalities of Invariance/Equivariance Theory . . . . .	40
<b>8</b>	<b>Asymptotics of Likelihood Inference</b>	<b>43</b>
8.1	Asymptotics of Likelihood Estimation . . . . .	43
8.1.1	Consistency of Likelihood-Based Estimators . . . . .	43
8.1.2	Asymptotic Normality of Likelihood-Based Estimators . . . . .	46
8.2	Asymptotics of LRT-like Tests and Competitors and Related Confidence Regions . . . . .	47
8.3	Asymptotic Shape of the Loglikelihood and Related Bayesian Asymptotics . . . . .	50
<b>9</b>	<b>Optimality in Finite Sample Point Estimation</b>	<b>52</b>
9.1	Unbiasedness . . . . .	52
9.2	"Information" Inequalities . . . . .	53
<b>10</b>	<b>Optimality in Finite Sample Testing</b>	<b>54</b>
10.1	Simple versus Simple Testing . . . . .	55

# 1 Characteristic Functions

Recall from simple complex analysis that  $i \equiv \sqrt{-1}$  (we'll have to tell from context whether  $i$  is standing for the root of  $-1$  or for something else like an index of summation). For some purposes, complex numbers  $a + bi$  are just a compact representation of points in  $\mathbb{R}^2$ . That is,  $a + bi$  with "real part"  $a$  and "imaginary part"  $b$  is equivalent to the ordered pair  $(a, b) \in \mathbb{R}^2$ . But the complex number rules of operation give the convenience of operating with many of the same rules of computation that we use for real numbers and others that are derivable by simply operating algebraically on  $i$ . For example, a reciprocal of the complex number  $a + bi$  (another complex number that when multiplied times it produces 1) is

$$(a + bi)^{-1} = \frac{a - bi}{a^2 + b^2} = \frac{\overline{(a + bi)}}{(a + bi)\overline{(a + bi)}}$$

If  $g$  is a complex-valued function, we may think of  $g$  in terms of two real-valued functions ( $\operatorname{Re} g(t)$  and  $\operatorname{Im} g(t)$ ) such that

$$g(t) = \operatorname{Re} g(t) + i \operatorname{Im} g(t)$$

So, for example, we may use the shorthand

$$\int g d\mu \equiv \int \operatorname{Re} g d\mu + i \int \operatorname{Im} g d\mu$$

Further, in the obvious way, derivatives of complex functions come from derivatives of two real-valued functions

$$\frac{d}{dt} g(t) = \frac{d}{dt} \operatorname{Re} g(t) + i \frac{d}{dt} \operatorname{Im} g(t)$$

A standard complex analysis convention is (for real  $\theta$ ) to define

$$e^{i\theta} \equiv \cos \theta + i \sin \theta$$

(You should check that this definition is plausible based on the forms of the Maclaurin series for  $\exp t$ ,  $\cos t$ , and  $\sin t$ .) Now it's easy to see that  $|e^{i\theta}| = 1$  so that  $e^{i\theta}$  is on the unit circle in  $\mathbb{C}$  (or  $\mathbb{R}^2$ ). It's also straightforward to verify that for real  $\theta_1$  and  $\theta_2$

$$e^{i(\theta_1 + \theta_2)} = e^{i\theta_1} e^{i\theta_2}$$

Further, the complex-valued function of a single real variable  $g(t) = e^{itb}$  has  $n$ th derivative

$$g^{(n)}(t) = (ib)^n e^{itb}$$

## 1.1 Some Simple Generalities and Basic Facts About Chf's

**Definition 1** If  $\mu$  is a probability distribution on  $\mathbb{R}^k$  (and is the distribution of the random vector  $\mathbf{X}$ ) then the **characteristic function** of  $\mu$  (or of  $\mathbf{X}$ ) is the mapping from  $\mathbb{R}^k$  to  $\mathbb{C}$  defined by

$$\phi(\mathbf{t}) = Ee^{i\mathbf{t}'\mathbf{X}} = \int e^{i\mathbf{t}'\mathbf{x}} d\mu(\mathbf{x}) = \int \cos(\mathbf{t}'\mathbf{x}) d\mu(\mathbf{x}) + i \int \sin(\mathbf{t}'\mathbf{x}) d\mu(\mathbf{x})$$

Characteristic functions are guaranteed to exist for all probability distributions. Both A&L Ch 10 and Chung Ch 6 are full of interesting and useful stuff about chf's. Some simple facts are:

1. If  $\phi$  is a chf, then  $\phi(\mathbf{0}) = 1$ .
2. If  $\phi$  is a chf, then  $|\phi(\mathbf{t})| \leq 1 \forall \mathbf{t}$ .
3. If for  $\phi_1, \dots, \phi_m$  are chf's for  $k$ -dimensional distributions and  $p_1, \dots, p_m$  specify a distribution over  $\{1, 2, \dots, m\}$ , then  $\sum p_l \phi_l$  is the chf of  $\sum p_l \mu_l$ .
4. If  $\mathbf{X}_1, \dots, \mathbf{X}_m$  are independent  $k$ -dimensional random vectors with chf's  $\phi_1, \dots, \phi_m$ , then the sum  $\mathbf{S}_m = \sum_{l=1}^m \mathbf{X}_l$  has chf  $\phi_{\mathbf{S}_m} = \prod_{l=1}^m \phi_l$ .
5. If  $\phi$  is the chf of real-valued  $X$ , then for real numbers  $a$  and  $b$ ,  $Y = a + bX$  has chf  $\phi_Y(t) = e^{iat} \phi(bt)$ .
6. If  $\phi$  is the chf of  $\mathbf{X}$ , then  $\bar{\phi}$  is the chf of  $-\mathbf{X}$ .
7. If  $\mathbf{X}$  and  $\mathbf{Y}$  are iid with chf's  $\phi$ , then (the symmetrically distributed) random vector  $\mathbf{X} - \mathbf{Y}$  has (real-valued) chf  $\phi_{\mathbf{X}-\mathbf{Y}} = |\phi|^2$ .
8. If  $\phi$  is the chf of  $\mathbf{X} \sim \mu$  and  $\nu$  is the distribution of  $-\mathbf{X}$ , then the (symmetric) distribution  $\frac{1}{2}\mu + \frac{1}{2}\nu$  has (real-valued) chf  $\text{Re } \phi$ .
9. If  $\phi_1$  is the chf of the  $k_1$ -dimensional  $\mathbf{X}_1$  and  $\phi_2$  is the chf of the  $k_2$ -dimensional  $\mathbf{X}_2$  and the random vectors  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are independent, then for  $\mathbf{t}_1 \in \mathbb{R}^{k_1}$  and  $\mathbf{t}_2 \in \mathbb{R}^{k_2}$  the random vector  $\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$  has characteristic function  $\phi\left(\begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix}\right) = \phi_1(\mathbf{t}_1) \phi_2(\mathbf{t}_2)$

On pages 155-156 Chung provides a nice list of standard distributions and their chfs. Table 1 is derived from Chung's list.

Per Prof. Nordman's chf notes and 10.1.3 of A&L, there is the promise that chf's are continuous.

**Theorem 2** If  $\phi$  is a chf, then  $\phi$  is uniformly continuous on  $\mathbb{R}^k$ .

Per Nordman's Theorem 7.1, 10.1.4 of A&L, or page 67 of Lamperti, there is the fact that when they exist, moments of random variables can be deduced from derivatives of chf's.

Table 1: Chung's List of Chf's

Distribution	pdf or pmf	Chf
Unit Point Mass at $a$		$e^{iat}$
Mass $\frac{1}{2}$ at each of $\pm 1$		$\cos t$
Ber( $p$ )	$p^x (1-p)^{1-x} I[x \text{ is } 0 \text{ or } 1]$	$(1-p) + pe^{it}$
Bi( $n, p$ )	$\binom{n}{x} p^x (1-p)^{n-x} I[x \in \{0, 1, \dots, n\}]$	$((1-p) + pe^{it})^n$
Geo( $p$ )	$p(1-p)^x I[x \in \{0, 1, \dots\}]$	$p(1 - (1-p)e^{it})^{-1}$
Poisson( $\lambda$ )	$\frac{e^{-\lambda} \lambda^x}{x!} I[x \in \{0, 1, \dots\}]$	$e^{\lambda(e^{it}-1)}$
Exp( $\lambda$ ) (mean $\lambda^{-1}$ )	$\lambda e^{-\lambda x} I[x > 0]$	$(1 - \lambda^{-1}it)^{-1}$
U( $-a, a$ )	$\frac{1}{2a} I[-a < x < a]$	$\frac{\sin at}{at}$
Triangular on $[-a, a]$	$\frac{a -  x }{a^2} I[-a < x < a]$	$\left(\frac{\sin \frac{at}{2}}{\frac{at}{2}}\right)^2$
N( $\mu, \sigma^2$ )	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right)$
N(0, 1)		$e^{-t^2/2}$
Cauchy with scale $\beta > 0$	$\frac{\beta}{\pi(\beta^2 + x^2)}$	$e^{-\beta t }$

**Theorem 3** *If  $\phi$  is the chf of the random variable  $X$ , and for some positive integer  $r$ ,  $E|X|^r < \infty$ , then  $\phi$  has a continuous  $r$ th derivative  $\phi^{(r)}$  on  $\mathbb{R}$  and*

$$\phi^{(r)}(0) = i^r EX^r$$

As a bit of an aside, an interesting and important additional property of characteristic functions (provided by Bochner's Theorem) is that they are *non-negative definite* complex-valued functions. That means that any *real-valued* characteristic function can serve as a lag-correlation function for a second-order stationary stochastic process. That fact gives large classes of models for time series and spatial statistics (that are defined in terms of their lag-correlation functions). Related to this use of characteristic functions are theorems that provide sufficient conditions for a real-valued function to be a characteristic function. One such theorem is on page 191 of Chung and is 10.1.8 of A&L.

**Theorem 4** *If  $f$  on  $\mathbb{R}$  is decreasing and convex on  $\mathbb{R}_+ = [0, \infty)$  with  $f(0) = 1$ ,  $f(t) \geq 0$ , and  $f(t) = f(-t)$ , then  $f$  is a chf.*

## 1.2 Chf's and Probabilities

Theorem 3 shows that when they exist, chf's encode the moments of a distribution. In fact, a chf encodes all there is to know about a distribution. One

can get all probabilities from a distribution's chf. 10.2.6 of A&L and Theorem 11.1.1 of Rosenthal is next.

**Theorem 5** (*Fourier Inversion Theorem*) *If  $\phi$  is the chf of a probability distribution  $\mu$  on  $\mathbb{R}$ , then for any real  $a < b$*

$$\lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt = \mu((a, b)) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\})$$

Two simple consequences of Theorem 5 are next.

**Corollary 6** *Under the hypotheses of Theorem 5, if  $a$  and  $b$  are continuity points of the cdf  $F$  corresponding to  $\mu$ , then*

$$\lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt = \mu((a, b))$$

**Corollary 7** (*Fourier Uniqueness Theorem*) *Chf's uniquely determine probability distributions.*

Corollary 7 is 10.2.5 of A&L.

We will say that  $\mathbf{X}$  is symmetrically distributed iff  $\mathbf{X}$  and  $-\mathbf{X}$  have the same distribution.

**Corollary 8**  *$\mathbf{X}$  is symmetrically distributed iff its characteristic function  $\phi$  is real-valued.*

There are all sorts of properties of distributions that are reflected in particular properties of characteristic functions. Per Nordman's Theorem 7.2 and Corollary 7.4 (that is 10.2.4. of A&L) there are the following two results, the second of which follows from Corollary 6.

**Theorem 9** (*Riemann-Lebesgue Lemma*) *If the distribution of a random variable  $X$  has a density  $f$  wrt Lebesgue measure on  $\mathbb{R}$ , then the chf of  $X$ ,  $\phi$ , satisfies*

$$\phi(t) \rightarrow 0 \text{ as } |t| \rightarrow \infty$$

**Corollary 10** *If the probability distribution  $\mu$  on  $\mathbb{R}$  has a chf,  $\phi$ , for which  $\int |\phi(t)| dt < \infty$ , then  $\mu$  has density wrt Lebesgue measure (pdf) on  $\mathbb{R}$*

$$f(x) = \frac{1}{2\pi} \int e^{-itx} \phi(t) dt$$

A&L's 10.2.6b is also interesting. It is next.

**Theorem 11** *For a distribution on  $\mathbb{R}$  with chf  $\phi$  and real number  $a$ ,*

$$\mu(\{a\}) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T e^{-ita} \phi(t) dt$$

### 1.3 Chf's and Convergence in Distribution

The major remaining major concerning characteristic functions is that they are useful for establishing convergence in distribution. That is, for distributions  $\mu_n$  with chf's  $\phi_n$  and a distribution  $\mu$  with chf  $\phi$ ,  $\mu_n \xrightarrow{d} \mu$  iff  $\phi_n(\mathbf{t}) \rightarrow \phi(\mathbf{t}) \forall \mathbf{t} \in \mathbb{R}^k$ . Notice that such a result is a kind of continuity theorem for the mappings that 1) produce chf's from distributions and 2) produce distributions from chf's. (Continuity of a function means that proximity of arguments implies proximity of function values.)

To get to this kind of continuity result, a technical lemma is required. The following is 10.3.3 of A&L and 11.1.13 of Rosenthal.

**Lemma 12** *Let  $\{\mu_n\}$  be a sequence of distributions on  $\mathbb{R}$  with corresponding sequence of characteristic functions  $\{\phi_n\}$ . If there exists  $g : \mathbb{R} \rightarrow \mathbb{C}$  that is continuous at 0 and  $t_0 > 0$  such that*

$$\phi_n(t) \rightarrow g(t) \quad \forall t \in (-t_0, t_0)$$

*then  $\{\mu_n\}$  is tight.*

This then enables proof of the important continuity theorem (that is 10.3.4 of A&L).

**Theorem 13** *(The  $\mathbb{R}^1$  version of the Continuity Theorem) If distributions  $\{\mu_n\}$  on  $\mathbb{R}$  have characteristic functions  $\{\phi_n\}$  and distribution  $\mu$  has chf  $\phi$ ,*

$$\mu_n \xrightarrow{d} \mu \quad \text{iff} \quad \phi_n(t) \rightarrow \phi(t) \quad \forall t \in \mathbb{R}^1$$

(See A&L 10.4.4 for an  $\mathbb{R}^k$  version of this.)

A technical lemma provides a simple Central Limit Theorem as a direct consequence of Theorem 13. This concerns the complex number version of a limit from freshman calculus.

**Lemma 14** *If complex numbers  $c_n$  converge to  $c$ , then*

$$\lim_{n \rightarrow \infty} \left(1 + \frac{c_n}{n}\right)^n = e^c$$

**Theorem 15** *Suppose the random variables  $\{X_n\}$  are iid with  $EX_1^2 < \infty$ . Let  $EX_1 = \mu$  and  $\text{Var}X_1 = \sigma^2 < \infty$ . Then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

A second very useful consequence of Theorem 13 is the so-called "Cramér-Wold device." This appears as 10.4.5 of A&L.

**Theorem 16** *(Cramér-Wold Device) Suppose that  $\{\mathbf{X}_n\}$  is a sequence of  $k$ -dimensional random vectors. Then*

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \quad \text{iff} \quad \mathbf{a}'\mathbf{X}_n \xrightarrow{d} \mathbf{a}'\mathbf{X} \quad \forall \mathbf{a} \in \mathbb{R}^k$$

## 2 Central Limit Theorems

The basic idea of Central Limit Theory is that sums of small, not too highly dependent, random variables should be approximately normal. The simplest such theorems are for sequences of summands  $X_1, X_2, \dots$ , where for  $S_n \equiv \sum_{i=1}^n X_i$  there are sequences of constants  $a_n$  and  $b_n$  such that

$$\frac{S_n - a_n}{b_n} \xrightarrow{d} N(0, 1)$$

A more general and useful framework for CLTs is that of double arrays.

**Definition 17**  $\{X_{ij}\}$  for  $i = 1, 2, \dots$  and  $j = 1, 2, \dots, n_i$  where  $n_i \xrightarrow{i \rightarrow \infty} \infty$  will be called a **double array** of random variables, and in the case where  $n_i = i$  we will call the collection a **triangular array**. Standard notation will be to let

$$S_i = \sum_{j=1}^{n_i} X_{ij}$$

It is easy to see that even if one assumes that  $\{X_{ij}\}_{j=1,2,\dots,n_i}$  is a set of independent random variables, there is no guarantee that there are constants  $a_i$  and  $b_i$  such that

$$\frac{S_i - a_i}{b_i} \xrightarrow{d} N(0, 1)$$

### 2.1 Basic CLTs for Independent Double Arrays

The most famous condition under which Central Limit Theorems are proved is the famous "Lindeberg condition."

**Definition 18** A double array of independent random variables  $\{X_{ij}\}$  with  $EX_{ij} = 0 \forall i, j$  for which each  $EX_{ij}^2 < \infty$  is said to satisfy **the Lindeberg condition** provided for  $B_i^2 \equiv \text{Var}S_i = \sum_{j=1}^{n_i} EX_{ij}^2$ ,

$$\forall \epsilon > 0 \quad \frac{1}{B_i^2} \sum_{j=1}^{n_i} EX_{ij}^2 I[|X_{ij}| > \epsilon B_i] \xrightarrow{i \rightarrow \infty} 0$$

This condition says that the tails of the  $X_{ij}$  contribute less and less to the variance of  $S_i$  with increasing  $i$ . Next is 11.1.1 of A&L

**Theorem 19** (Lindeberg Central Limit Theorem) If  $\{X_{ij}\}$  is a double array of independent mean 0 random variables with finite variances that satisfy the Lindeberg condition, then

$$\frac{S_i}{B_i} \xrightarrow{d} N(0, 1)$$

A simpler-looking condition under which the Lindeberg condition holds is Liapounov's condition.

**Definition 20** A double array of independent random variables  $\{X_{ij}\}$  with  $EX_{ij} = 0 \forall i, j$  for which each  $E|X_{ij}|^{2+\delta} < \infty$  for some fixed  $\delta > 0$  is said to satisfy **the Liapounov condition** provided for  $B_i^2 \equiv \text{Var}S_i = \sum_{j=1}^{n_i} EX_{ij}^2$

$$\frac{1}{B_i^{2+\delta}} \sum_{j=1}^{n_i} E|X_{ij}|^{2+\delta} \xrightarrow{i \rightarrow \infty} 0$$

The Liapounov condition says that the ratio of the total absolute  $2 + \delta$  moment to the  $(1 + \frac{\delta}{2})$  power of the total absolute 2nd moment goes to 0.

Next is a result on page 348 of A&L.

**Theorem 21** If a double array of independent random variables  $\{X_{ij}\}$  with  $EX_{ij} = 0 \forall i, j$  for which each  $E|X_{ij}|^{2+\delta} < \infty$  for some fixed  $\delta > 0$  satisfies the Liapounov condition, it also satisfies the Lindeberg condition.

## 2.2 Related Limit Results

The basic central limit theorems can be generalized in a number of directions. A few such directions are mentioned here.

### 2.2.1 CLTs Under Dependence

One direction is to allow some (not too severe) dependencies in the variables being summed.

**Definition 22** A sequence of random variables  $X_1, X_2, \dots$  is said to be *m-dependent* if

$$(X_1, \dots, X_j) \text{ is independent of } (X_{j+m}, X_{j+m+1}, \dots) \quad \forall j$$

On his page 224, Chung gives an  $m$ -dependent central limit theorem.

**Theorem 23** If an  $m$ -dependent sequence of uniformly bounded random variables  $X_1, X_2, \dots$  is such that

$$\frac{\sigma_n}{n^{1/3}} \equiv \frac{\sqrt{\text{Var}S_n}}{n^{1/3}} \rightarrow \infty$$

then

$$\frac{S_n - ES_n}{\sigma_n} \xrightarrow{d} N(0, 1)$$

The condition on the variance of the partial sum in Theorem 23 says that the variance of  $S_n$  can not grow too slowly. Note that an iid model of course has  $\sigma_n = \sqrt{n}\sigma$ .

### 2.2.2 CLTs for Sums of Random Numbers of Random Summands

There are limit theorems for random sums of random variables like that of A&L problem 11.11 and Chung page 226.

**Theorem 24** *Suppose that  $X_1, X_2, \dots$  is a sequence of iid mean 0 variance 1 random variables and  $\nu_1, \nu_2, \dots$  is a sequence of random variables taking only positive integer values such that for some  $c > 0$*

$$\frac{\nu_n}{n} \xrightarrow{P} c$$

then

$$\frac{S_{\nu_n}}{\sqrt{\nu_n}} \xrightarrow{d} N(0, 1)$$

This kind of theorem is fairly easy to prove in the case that  $\{X_j\}$  and  $\{\nu_j\}$  are independent (using characteristic function and conditioning). But this result is harder to prove and much more general.

### 2.2.3 Rates for CLTs and Generalizations

There are theorems about the rate of convergence of distributions to normal. That is, Theorem 15 says that in the iid finite variance case

$$P \left[ \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq t \right] \rightarrow \Phi(t) \quad \forall t$$

This together with Polya's Theorem (9.14 of A&L) says that

$$\Delta_n \equiv \sup_t \left| P \left[ \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq t \right] - \Phi(t) \right| \rightarrow 0$$

and there are rate theorems about  $\Delta_n$  like 11.4.1 of A&L.

**Theorem 25** *(The Berry-Esseen Theorem) Suppose the random variables  $\{X_j\}$  are iid with  $E|X_1|^3 < \infty$ . Let  $EX_1 = \mu$  and  $\text{Var}X_1 = \sigma^2$ . Then there is a  $C \in (0, \infty)$  such that*

$$\Delta_n \leq C \frac{E|X_1 - \mu|^3}{\sigma^3 \sqrt{n}}$$

In fact, the  $C$  in Theorem 25 is no more than 5.05.

There are also rate theorems for "higher order" approximations called "Edgeworth Expansions." See for example, Theorem 11.4 of A&L, that says that under conditions like those of Theorem 25, with

$$\Delta'_n \equiv \sup_t \left| P \left[ \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq t \right] - \left[ \Phi(t) - \frac{E|X_1 - \mu|^3}{6\sigma^3 \sqrt{n}} (t^2 - 1) \phi(t) \right] \right|$$

$$\Delta'_n = o\left(n^{-\frac{1}{2}}\right) \text{ or even } \Delta'_n = o\left(n^{-1}\right).$$

### 2.2.4 Large Deviation Results

There are theorems about approximating other probabilities related to  $S_n$ . For  $X_1, X_2, \dots$  a sequence of iid variables with  $EX_1 = \mu$  and  $t > \mu$  so called "large deviation" results concern approximation of  $P\left[\frac{S_n}{n} > t\right]$ . Note that typical Central Limit Theorems say that

$$P\left[\frac{S_n - n\mu}{\sqrt{n}\sigma} > t\right] = P\left[\frac{S_n}{n} > \mu + t\frac{\sigma}{\sqrt{n}}\right] \rightarrow 1 - \Phi(t)$$

and so CLTs provide approximate cumulative probabilities for values of  $\frac{S_n}{n}$  differing from  $\mu$  by amounts that shrink to 0 at rate  $n^{-1/2}$ . Large deviation results instead concern probabilities of exceeding a fixed value  $t > \mu$ . Theorem 11.4.6 of A&L says

$$\left(P\left[\frac{S_n}{n} > t\right]\right)^n \rightarrow C(t)$$

for  $C(t)$  some function depending upon the distribution of  $X_1$ .

### 2.2.5 Results About "Stable Laws"

CLTs say that there are sequences of constants  $a_n$  and  $b_n$  such that

$$\frac{S_n - a_n}{b_n} \tag{1}$$

has a standard normal limit. A sensible question is whether quantities (1) can have any other types of limit distributions.

**Definition 26** A non-degenerate distribution  $G$  on  $\mathbb{R}$  is **stable** iff for  $X_1, X_2, \dots$  iid  $G$ ,  $\exists$  sequences of constants  $a_n$  and  $b_n$  such that

$$\frac{S_n - a_n}{b_n} \stackrel{d}{=} X_1$$

**Theorem 27** A non-degenerate distribution  $G$  on  $\mathbb{R}$  is stable iff  $\exists$  some distribution  $F$  and sequences of constants  $a_n$  and  $b_n$  such that with  $X_1, X_2, \dots$  iid  $F$

$$\frac{S_n - a_n}{b_n} \xrightarrow{d} G$$

Section 11.2 of A&L is full interesting facts about stable laws (including representations of their characteristic functions).

### 2.2.6 Functional Central Limit Theorems and Empirical Processes

For  $C[0, 1]$  the space of continuous functions on the unit interval with metric

$$d(f, g) = \sup_t |f(t) - g(t)|$$

one may think of putting distributions on the Borel sets in the complete separable metric space  $C[0, 1]$ . Thinking of the argument of an  $f \in C[0, 1]$  as "time," such a distribution specifies a stochastic process model.

Suppose that  $\{X_j\}$  are iid with  $EX_1 = 0$  and  $\text{Var}X_1 = 1$  and define  $W_n(t)$  on  $[0, 1]$  by

$$W_n\left(\frac{j}{n}\right) = \frac{1}{\sqrt{n}}S_j \quad \text{for } j = 0, 1, \dots, n$$

and by linear interpolation for  $t$  not of the form  $j/n$  for any  $j = 0, 1, \dots, n$ .  $W_n(t)$  is a random element of  $C[0, 1]$  and there is a probability measure  $\mu_n$  such that  $W_n \sim \mu_n$ . The multivariate CLT can be used to show that  $\forall 0 \leq t_1 \leq t_2 \leq \dots \leq t_k \leq 1$

$$\begin{pmatrix} W_n(t_1) \\ W_n(t_2) \\ \vdots \\ W_n(t_k) \end{pmatrix} \xrightarrow{d} \text{MVN}_k(\mathbf{0}, \Sigma)$$

where

$$\Sigma = (\sigma_{ij}) \quad \text{with } \sigma_{ij} = t_i \wedge t_j \tag{2}$$

It further turns out that there is a distribution  $\mu$  on  $C[0, 1]$  called the Standard Brownian Motion such that  $\forall 0 \leq t_1 \leq t_2 \leq \dots \leq t_k \leq 1$ ,  $W \sim \mu$  implies that

$$\begin{pmatrix} W(t_1) \\ W(t_2) \\ \vdots \\ W(t_k) \end{pmatrix} \sim \text{MVN}_k(\mathbf{0}, \Sigma)$$

where  $\Sigma$  has form (2). ( $\mu$  is a Gaussian process, meaning all finite dimensional marginals are Gaussian.) In addition, under these assumptions there is a "functional CLT" (11.4.9 of A&L).

**Theorem 28** For  $\{X_j\}$  iid with  $EX_1 = 0$  and  $\text{Var}X_1 = 1$ , and  $W_n(t)$ ,  $\mu_n$ , and  $\mu$  as defined above

$$\mu_n \xrightarrow{d} \mu$$

The distribution of  $W_n$  converges to a standard Brownian motion. (Convergence in distribution in this context is most naturally convergence of probabilities of sets of functions with boundaries that are null sets under the limiting distribution. See A&L Section 9.3 for convergence in distribution in abstract spaces.)

A related line of argument concerns "empirical processes" and the "Brownian Bridge" distribution on  $C[0, 1]$  and is discussed in Section 11.4.5 of A&L. That is, suppose that  $W \sim \mu$  for  $\mu$  on  $C[0, 1]$  the standard Brownian motion distribution. Let

$$B(t) = W(t) - tW(1)$$

Then  $B(t)$  has a Gaussian process distribution called the "Brownian Bridge" with

$$B(0) = B(1) = 0 \text{ and } EB(t) = 0 \forall t$$

and

$$\text{Cov}(B(t_1), B(t_2)) = t_1 \wedge t_2 - t_1 t_2$$

This distribution on  $C[0, 1]$  can be thought of as arising as a limit distribution in the following way. For  $\{U_j\}$  iid  $U(0, 1)$  and (the empirical cumulative distribution function for the  $U$ 's)

$$F_n(t) \equiv \frac{1}{n} \sum_{j=1}^n I[U_j \leq t] \text{ for } t \in [0, 1]$$

let

$Y_n(t)$  = the linear interpolation of  $\sqrt{n}(F_n(t) - t)$  made from values at the jump points of  $F_n$  (i.e. where  $t$  is some  $U_j$ )

Then there is 11.4.1. of A&L.

**Theorem 29** *With  $B(t)$  and  $Y_n(t)$  as above*

$$Y_n(t) \xrightarrow{d} B(t)$$

$Y_n(t)$  is the "empirical process" for the  $U_j$  and Theorem 29 says that approximations for probabilities for the process (and thence for  $F_n(t)$ ) can be had from use of the Brownian bridge distribution.

### 3 Conditioning

One needs properly general notions of "conditioning" in order to deal coherently with general statistical models. Some thought about even fairly simple not-completely-discrete problems leaves one convinced that ideas of conditional expectation must focus on conditioning  $\sigma$ -algebras, not conditioning variables (a conditioning variable generates a corresponding  $\sigma$ -algebra, but it is the latter that is intrinsic, not the former, as many slightly different variables can generate the same  $\sigma$ -algebra). Here is the standard definition.

**Definition 30** *Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $\mathcal{G} \subset \mathcal{F}$  be a sub  $\sigma$ -algebra. Let  $X$  be a real-valued measurable function (a random variable) on  $\Omega$  with  $E|X| = \int |X(\omega)| dP(\omega) < \infty$ . Then **the conditional expectation of  $X$  given  $\mathcal{G}$** , written  $E(X|\mathcal{G})$ , is a random variable  $\Omega \rightarrow \mathbb{R}$  with the properties that*

1.  $E(X|\mathcal{G})$  is  $\mathcal{G}$ -measurable, and

2. for any  $A \in \mathcal{G}$

$$\int_A E(X|\mathcal{G}) dP = \int_A X dP$$

In other notation, the basic requirement 2. of Definition 30 is that  $E I_A E(X|\mathcal{G}) = E I_A X$  for all  $A \in \mathcal{G}$ . The existence of  $E(X|\mathcal{G})$  satisfying Definition 30 follows from the Radon-Nikodym Theorem.

The case  $X = I_B$  for  $B \in \mathcal{F}$  gets its own notation.

**Definition 31** For  $B \in \mathcal{F}$  and  $\mathcal{G} \subset \mathcal{F}$  a sub  $\sigma$ -algebra, the **conditional probability of  $B$  given  $\mathcal{G}$**  is

$$P(B|\mathcal{G}) \equiv E(I_B|\mathcal{G})$$

Several comments about the notion of conditional expectation are in order. First,  $E(X|\mathcal{G})$  is clearly unique only up to sets of  $P$  measure 0. Second, the case where  $\mathcal{G} = \sigma(W)$  for some random variable  $W$  is

$$E(X|W) \equiv E(X|\sigma(W))$$

And third, A&L have what is a very clever and possibly more natural (at least for statisticians) definition of conditioning that applies to  $X$  with second moments. Namely,  $E(X|\mathcal{G})$  is the projection of  $X$  onto the linear subspace of  $L^2(\Omega, \mathcal{F}, P)$  consisting of  $\mathcal{G}$ -measurable functions. (More on this later.)

If  $\mathcal{G}$  is the trivial sub  $\sigma$ -algebra, i.e.  $\mathcal{G} = \{\emptyset, \Omega\}$ , then  $E(X|\mathcal{G})$  must be constant and thus  $E(X|\mathcal{G}) = EX$ . On the other hand, if  $\mathcal{G} = \mathcal{F}$ , then it's easy to argue that  $E(X|\mathcal{G}) = EX$ . Intuitively, the smaller is  $\mathcal{G}$ , the "less random" is  $E(X|\mathcal{G})$ , i.e. the "more averaging has gone on in its creation," i.e. the fewer values it takes on.

### 3.1 Basic Results About Conditional Expectation

Some basic facts about conditional expectation are collected in the next theorem.

**Theorem 32** Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $\mathcal{G} \subset \mathcal{F}$  be a sub  $\sigma$ -algebra.

1. If  $c \in \mathbb{R}$ ,  $E(c|\mathcal{G}) = c$  a.s.
2. If  $X$  has a first moment and is  $\mathcal{G}$ -measurable, then  $E(X|\mathcal{G}) = X$  a.s.
3. If  $X$  has a first moment and  $X \geq 0$  a.s., then  $E(X|\mathcal{G}) \geq 0$  a.s.
4. If  $X_1, \dots, X_k$  have first moments and  $c_1, \dots, c_k$  are real, then

$$E\left(\sum_{i=1}^k c_i X_i | \mathcal{G}\right) = \sum_{i=1}^k c_i E(X_i | \mathcal{G}) \quad \text{a.s.}$$

5. If  $X$  has a first moment then  $|E(X|\mathcal{G})| \leq E(|X| | \mathcal{G})$  a.s.

Next is 12.1.5ii of A&L and 13.2.6 of Rosenthal.

**Theorem 33** *Suppose that  $X$  and  $Y$  are random variables and  $\mathcal{G} \subset \mathcal{F}$  is a sub  $\sigma$ -algebra. If  $Y$  and  $XY$  have first moments and  $X$  is  $\mathcal{G}$ -measurable, then*

$$E(XY|\mathcal{G}) = XE(Y|\mathcal{G}) \text{ a.s.}$$

(A  $\mathcal{G}$ -measurable  $X$  can be "factored out" of a conditional expectation.)

Then there is a standard result from 12.1.5i of A&L or 13.2.7 of Rosenthal.

**Theorem 34** *Let  $X$  be a random variable with first moment and  $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$  be two sub  $\sigma$ -algebras. Then*

$$E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(X|\mathcal{G}_1) \text{ a.s.}$$

And there is a natural result about independence.

**Theorem 35** *Suppose that  $X$  has a first moment and  $\mathcal{G} \subset \mathcal{F}$  is a sub  $\sigma$ -algebra. If  $\mathcal{G}$  and  $\sigma(X)$  are independent (in the sense that  $A \in \mathcal{G}$  and  $B \in \sigma(X)$  implies that  $P(A \cap B) = P(A)P(B)$ ) then*

$$E(X|\mathcal{G}) = EX \text{ a.s.}$$

### 3.2 Conditional Variance

Definition 31 is one important special case of conditional expectation. Another is conditional variance.

**Definition 36** *If  $EX^2 < \infty$  and  $\mathcal{G} \subset \mathcal{F}$  is a sub  $\sigma$ -algebra, define the **conditional variance***

$$\text{Var}(X|\mathcal{G}) = E(X^2|\mathcal{G}) - (E(X|\mathcal{G}))^2$$

Theorem 12.2.6 of A&L is next.

**Theorem 37** *If  $EX^2 < \infty$  and  $\mathcal{G} \subset \mathcal{F}$  is a sub  $\sigma$ -algebra, then*

$$\text{Var}X = \text{Var}(E(X|\mathcal{G})) + E(\text{Var}(X|\mathcal{G}))$$

The next result shows that the A&L definition of conditional expectation is appropriate for  $X \in L^2(\Omega, \mathcal{F}, P)$ .

**Theorem 38** *If  $EX^2 < \infty$  and  $\mathcal{G} \subset \mathcal{F}$  is a sub  $\sigma$ -algebra, then  $E(X|\mathcal{G})$  minimizes*

$$E(X - Y)^2$$

*over choices of  $\mathcal{G}$ -measurable random variables  $Y$ .*

(So under the hypotheses of Theorem 38,  $E(X|\mathcal{G})$  is indeed the projection of  $X$  onto the subspace of  $L^2(\Omega, \mathcal{F}, P)$  consisting of  $\mathcal{G}$ -measurable random variables.)

### 3.3 Conditional Analogues of Properties of Ordinary Expectation

Many properties of ordinary expectation have analogues for conditional expectations. For one, there is a conditional version of Jensen's inequality (that is 12.2.4 of A&L).

**Theorem 39** For  $-\infty \leq a < b \leq \infty$  suppose that  $X$  on  $(\Omega, \mathcal{F}, P)$  has  $P[a < X < b] = 1$  and that the function  $\phi$  is convex on  $(a, b)$  with  $E|\phi(X)| < \infty$ . Then for any sub  $\sigma$ -algebra  $\mathcal{G} \subset \mathcal{F}$ ,

$$\phi(E(X|\mathcal{G})) \leq E(\phi(X)|\mathcal{G}) \text{ a.s.}$$

And there are convergence theorems for conditional expectations like a conditional monotone convergence theorem (12.2.1 of A&L), a conditional version of the Fatou lemma (12.2.2 of A&L), and a conditional dominated convergence theorem (12.2.3 of A&L).

**Theorem 40** (*Monotone Convergence Theorem for Conditional Expectation*) If  $\{X_n\}$  is a sequence of non-negative random variables on  $(\Omega, \mathcal{F}, P)$  with

$$X_n \leq X_{n+1} \text{ a.s. and } X_n \rightarrow X \text{ a.s.}$$

then

$$E(X_n|\mathcal{G}) \rightarrow E(X|\mathcal{G}) \text{ a.s.}$$

**Theorem 41** (*Fatou's Lemma for Conditional Expectation*) If  $\{X_n\}$  is a sequence of non-negative random variables on  $(\Omega, \mathcal{F}, P)$ , then

$$\underline{\lim} E(X_n|\mathcal{G}) \geq E(\underline{\lim} X_n|\mathcal{G}) \text{ a.s.}$$

**Theorem 42** (*Dominated Convergence Theorem for Conditional Expectation*) Suppose  $\{X_n\}$  is a sequence of random variables on  $(\Omega, \mathcal{F}, P)$  and that  $X_n \rightarrow X$  a.s. If there  $\exists$  a random variable  $Z$  such that  $|X_n| \leq Z$  a.s. for all  $n$  and  $E|Z| < \infty$ , then

$$E(X_n|\mathcal{G}) \rightarrow E(X|\mathcal{G}) \text{ a.s.}$$

For statistical applications, it is Theorem 39 that is the most important of these analogues of standard results for ordinary expectations.

### 3.4 Conditional Probability

Consider the implications of Definition 31. It's possible to argue that for every countable set of disjoint  $A_i \in \mathcal{F}$ ,

$$\sum P(A_i|\mathcal{G}) = P(\cup A_i|\mathcal{G}) \text{ a.s.}$$

i.e. that  $\exists N_{\{A_i\}}$  such that  $P(N_{\{A_i\}}) = 0$  and

$$\sum P(A_i|\mathcal{G})(\omega) = P(\cup A_i|\mathcal{G})(\omega) \quad \forall \omega \in \Omega \setminus N_{\{A_i\}} \quad (3)$$

It's tempting to expect that this means that except for a set of  $P$  measure 0, I may think of

$$P(\cdot|\mathcal{G})(\omega)$$

as a probability measure. But (3) does not in general guarantee this. The problem is that there are typically uncountably many sets  $\{A_i\}$  and  $\cup N_{\{A_i\}}$  need not be  $\mathcal{F}$ -measurable, let alone have  $P$  measure 0. So the fact that (3) holds does not guarantee the existence of an object satisfying the next definition.

**Definition 43** For  $(\Omega, \mathcal{F}, P)$  a probability space and  $\mathcal{G} \subset \mathcal{F}$  and  $\mathcal{D} \subset \mathcal{F}$  two sub  $\sigma$ -algebras, a function

$$\mu : \mathcal{D} \times \Omega \rightarrow [0, 1]$$

is called **a regular conditional probability on  $\mathcal{D}$  given  $\mathcal{G}$**  provided

1.  $\forall A \in \mathcal{D}, \mu(A, \cdot) = P(A|\mathcal{G})(\cdot)$  a.s.,
2.  $\forall \omega \in \Omega, \mu(\cdot, \omega)$  is a probability measure on  $(\Omega, \mathcal{D})$ .

A regular conditional probability need not exist. (See the example on page 81 of Breiman's *Probability*.) When a regular conditional probability does exist, conditional expectations can be computed using it. That is, there is the following.

**Lemma 44** If  $\mu$  is a regular conditional probability on  $\mathcal{D}$  given  $\mathcal{G}$  and  $Y$  is  $\mathcal{D}$ -measurable, then

$$E(Y|\mathcal{G})(\omega) = \int Y d\mu(\cdot, \omega) \text{ for } P \text{ almost all } \omega$$

In the case where  $\mathcal{D} = \sigma(X)$ , a regular conditional probability on  $\mathcal{D}$  given  $\mathcal{G}$  is called a regular conditional distribution of  $X$  given  $\mathcal{G}$ . Provided that  $X$  takes values in a nice enough space, (like  $(\mathbb{R}^k, \mathcal{B}^k)$  or even  $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ ) it's possible to conclude that a regular conditional distribution of  $X$  given  $\mathcal{G}$  exists. This is 12.3.1 of A&L.

**Theorem 45** Suppose that  $(\Omega, \mathcal{F}, P)$  is a probability space and  $(\mathcal{X}, \mathcal{B})$  is a Polish space with Borel  $\sigma$ -algebra. Let  $X : \Omega \rightarrow \mathcal{X}$  be measurable. Then for any sub  $\sigma$ -algebra  $\mathcal{G} \subset \mathcal{F}$   $\exists$  a regular conditional distribution of  $X$  given  $\mathcal{G}$  (a regular conditional probability on  $\sigma(X)$  given  $\mathcal{G}$ ).

Under the conditions of Theorem 45, for any measurable  $h : \mathcal{X} \rightarrow \mathbb{R}$ , with  $E|h(X)| < \infty$ , for  $\mu$  the conditional distribution of  $X$  given  $\mathcal{G}$  guaranteed to exist by the theorem

$$E(h(X)|\mathcal{G})(\omega) = \int h(X) d\mu(\cdot, \omega) \text{ for } P \text{ almost all } \omega$$

and this obviously then works for indicator functions  $h$ . So for  $A \in \mathcal{B}$

$$\begin{aligned} P(X^{-1}(A) | \mathcal{G})(\omega) &= \mathbb{E}(I_{X^{-1}(A)} | \mathcal{G})(\omega) \\ &= \mathbb{E}(I_A(X) | \mathcal{G})(\omega) \\ &= \mu(\{\omega' | X(\omega') \in A\}, \omega) \\ &= \mu(X^{-1}(A), \omega) \text{ for } P \text{ almost all } \omega \end{aligned}$$

making use of Lemma 44 for the switch between conditional expectation and values of  $\mu$ .

A final topic to be considered in the discussion of conditional probability is conditional independence.

**Definition 46** Suppose  $\mathcal{G}$  is a sub  $\sigma$ -algebra of  $\mathcal{F}$  and for each  $l$  belonging to some index set  $\Lambda$ ,  $\mathcal{F}_l$  is a subset of  $\mathcal{F}$ . We will say that  $\{\mathcal{F}_l\}_{l \in \Lambda}$  is **conditionally independent given  $\mathcal{G}$**  if for any positive integer  $k$  and  $l_1, l_2, \dots, l_k \in \Lambda$

$$P(A_1 \cap A_2 \cap \dots \cap A_k | \mathcal{G}) = \prod_{i=1}^k P(A_i | \mathcal{G}) \text{ a.s.}$$

for all choices of  $A_i \in \mathcal{F}_{l_i}$ . A collection of random variables  $\{X_l\}_{l \in \Lambda}$  is called **conditionally independent given  $\mathcal{G}$**  if  $\{\sigma(X_l)\}_{l \in \Lambda}$  is conditionally independent given  $\mathcal{G}$ .

Next is the result of Problem 12.18 of A&L.

**Proposition 47** Suppose  $\mathcal{G}_1, \mathcal{G}_2$ , and  $\mathcal{G}_3$  are three sub  $\sigma$ -algebras of  $\mathcal{F}$ .  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are conditionally independent given  $\mathcal{G}_3$  if and only if

$$P(A | \sigma(\mathcal{G}_2 \cup \mathcal{G}_3)) = P(A | \mathcal{G}_3) \quad \forall A \in \mathcal{G}_1$$

And there is the result of Problem 12.19 of A&L.

**Proposition 48** Suppose  $\mathcal{G}_1, \mathcal{G}_2$ , and  $\mathcal{G}_3$  are three sub  $\sigma$ -algebras of  $\mathcal{F}$ . If  $\sigma(\mathcal{G}_1 \cup \mathcal{G}_3)$  is independent of  $\mathcal{G}_2$ , then  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are conditionally independent given  $\mathcal{G}_3$

## 4 Transition to Statistics

In the usual probability theory set-up, a (measurable) random observable  $X$  maps  $(\Omega, \mathcal{F}, P)$  to some measure space  $(\mathcal{X}, \mathcal{B})$  and thereby induces a probability measure  $P^X$  on  $(\mathcal{X}, \mathcal{B})$  (the distribution of the random observable). In statistical theory, the space  $(\Omega, \mathcal{F}, P)$  largely disappears from consideration, being replaced by consideration of  $(\mathcal{X}, \mathcal{B}, P^X)$ , and there is not just a single  $P^X$  to be considered but rather a whole family of such (giving different models for the observable) indexed by a parameter  $\theta$  belonging to some index set  $\Theta$ . It is completely standard to abuse notation somewhat and replace the  $P^X$  notation

with just  $P$  notation for the distribution of  $X$ , and thus consider a set of models for  $X$

$$\mathcal{P} \equiv \{P_\theta\}_{\theta \in \Theta}$$

We'll use a  $\theta$  subscript to indicate computations made under the  $P_\theta$  model. So, for example,

$$E_\theta g(X) = \int_{\mathcal{X}} g dP_\theta$$

The object is then to use  $X$  to learn about  $\theta$ , some function of  $\theta$ , or perhaps some as yet unobserved  $Y$  that has a distribution depending upon  $\theta$ . The reason to have learned measure-theoretic probability is that it now allows a logically coherent approach to statistical analysis in models more general than the completely discrete or completely continuous models of Stat 542/543.

Henceforth, we will suppose that  $\exists$  a  $\sigma$ -finite measure  $\mu$  such that

$$P_\theta \ll \mu \quad \forall \theta \in \Theta$$

(we will say that  $\mathcal{P}$  is dominated by  $\mu$ ). The Radon-Nikodym theorem then promises that for each  $\theta \in \Theta$ ,  $\exists f_\theta : \mathcal{X} \rightarrow \mathbb{R}$  such that

$$P_\theta(B) = \int_B f_\theta d\mu \quad \forall B \in \mathcal{B}$$

( $f_\theta = \frac{dP_\theta}{d\mu}$  is the R-N derivative of  $P_\theta$  with respect to  $\mu$ ).  $f_\theta$  basically tells one how to go from probabilities for one model for  $X$  to probabilities for another.

Then, for observable  $X$ ,  $f_\theta(X)$  is a random function of  $\theta$  that becomes **the likelihood function** in a properly general statistical set-up.

We note for future reference that a Bayesian approach to inference will add to these measure-theoretic model assumptions an assumption that  $\theta$  has some distribution  $G$  (on  $\Theta$  after it is given some  $\sigma$ -algebra). Notice then that  $X|\theta \sim P_\theta$  together with  $\theta \sim G$  then presumably produces some (*single*) joint distribution on  $\mathcal{X} \times \Theta$  that in turn typically can be thought of as producing some regular conditional distribution of  $\theta$  given  $\sigma(X)$  in the style of Section 3.4 (that serves as a **posterior distribution** of  $\theta|X$ ).

## 5 Sufficiency and Related Concepts

### 5.1 Sufficiency and the Factorization Theorem

Roughly,  $T(X)$  is sufficient for  $\mathcal{P} \equiv \{P_\theta\}_{\theta \in \Theta}$  (or for  $\theta$ ) provided "the conditional distribution of  $X$  given  $T(X)$  is the same for each  $\theta$ ," and this ought to be equivalent to the likelihood function being essentially the same for any two possible values of  $X$ , say  $x$  and  $x'$ , with  $T(x) = T(x')$ . This equivalence is the "factorization theorem."

A formal development that will produce the factorization theorem begins with a statistic  $T$ , a measurable map

$$T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$$

and the sub  $\sigma$ -algebra of  $\mathcal{B}$ ,

$$\mathcal{B}_0 = \mathcal{B}(T) \equiv \{T^{-1}(F)\}_{F \in \mathcal{F}} = \sigma(T)$$

**Definition 49**  $T$  (or  $\mathcal{B}_0$ ) is **sufficient for  $\mathcal{P}$**  (or for  $\theta$ ) if for every  $B \in \mathcal{B} \ni$  a  $\mathcal{B}_0$ -measurable function  $Y$  such that  $\forall \theta \in \Theta$

$$Y = P_\theta(B|\mathcal{B}_0) \text{ a.s. } P_\theta$$

This general definition of sufficiency is phrased in terms of conditional probabilities (conditional expectations of indicators) and says that for each  $B$ , there is a single  $Y$  that works as conditional probability for  $B$  given  $\mathcal{B}_0$  as one looks across all  $\theta$ . In the event that there is a single regular conditional distribution of  $X$  given  $\mathcal{B}_0 = \mathcal{B}(T) = \sigma(T)$  that works for all  $\theta \in \Theta$ , Definition 49 says that  $T$  is sufficient. (But the definition doesn't require the existence of such a regular conditional distribution, only the existence of the " $\theta$ -free" conditional expectations one  $B$  at a time.)

**Theorem 50** (*Halmos-Savage Factorization Theorem*) Suppose that  $\mathcal{P} \ll \mu$  where  $\mu$  is  $\sigma$ -finite. Then  $T$  is sufficient for  $\mathcal{P}$  iff  $\exists$  a nonnegative  $\mathcal{B}$ -measurable function  $h$  and nonnegative  $\mathcal{F}$ -measurable functions  $g_\theta$  such that  $\forall \theta$

$$\frac{dP_\theta}{d\mu}(x) = f_\theta(x) = g_\theta(T(x))h(x) \text{ a.s. } \mu$$

Theorem 50 says that  $T$  is sufficient iff  $x$  and  $x'$  with  $T(x) = T(x')$  have likelihood functions that are proportional.

Theorem 50 is not so easy to prove. A series of lemmas is needed. The first is Lemma 1.2 of Shao and the hint of problem 12.1 of A&L. (Below  $\mathcal{B}^k$  is the Borel  $\sigma$ -algebra on  $\mathbb{R}^k$ .)

**Lemma 51** (*Lehmann's Theorem*) Let  $T : \mathcal{X} \rightarrow \mathcal{T}$  be measurable and  $\phi : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}^k, \mathcal{B}^k)$  be  $\mathcal{B}(T)$ -measurable. Then  $\exists$  an  $\mathcal{F}$ -measurable function  $\psi : (\mathcal{T}, \mathcal{F}) \rightarrow (\mathbb{R}^k, \mathcal{B}^k)$  such that  $\phi(x) = \psi(T(x))$ .

Next is Lemma 2.1 of Shao.

**Lemma 52**  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure  $\mu$  iff  $\mathcal{P}$  is dominated by a probability measure  $\lambda$  of the form

$$\lambda = \sum_{i=1}^{\infty} c_i P_{\theta_i}$$

for some countable set  $\{P_{\theta_i}\} \subset \mathcal{P}$  and set of  $c_i \geq 0$  with  $\sum_{i=1}^{\infty} c_i = 1$ .

**Lemma 53** *Suppose that  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure  $\mu$  and  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$ . Then  $T$  is sufficient iff  $\exists$  nonnegative  $\mathcal{F}$ -measurable functions  $g_\theta$  such that  $\forall \theta$  (and with  $\lambda$  as in Lemma 52 fixed)*

$$\frac{dP_\theta}{d\lambda}(x) = g_\theta(T(x)) \quad \text{a.s. } \lambda$$

We'll take Lemmas 51 and 52 as given. An outline of the proof of Lemma 53 using Lemmas 51 and 52 is

( $\implies$ ) For  $P_\theta^0$  and  $\lambda^0$  the restrictions of  $P_\theta$  and  $\lambda$  to  $\mathcal{B}_0$ ,  $\frac{dP_\theta^0}{d\lambda^0} = g_\theta \circ T$  by Lehmann's Theorem. Then show  $g_\theta \circ T = \frac{dP_\theta}{d\lambda}$  using facts about  $P(B|\mathcal{B}_0)$

( $\impliedby$ ) Show  $E_\theta(I_B|\mathcal{B}_0) = E_\lambda(I_B|\mathcal{B}_0)$  using the representation of  $\frac{dP_\theta}{d\lambda}$ .

Then, an outline of the proof of Theorem 50 using Lemmas 51-53 is

( $\implies$ )  $\frac{dP_\theta}{d\mu}$  " = "  $\frac{dP_\theta}{d\lambda} \cdot \frac{d\lambda}{d\mu}$  plus a.e.'s for  $\frac{dP_\theta}{d\lambda} = g_\theta \circ T$  (Lemma 53) and  $\frac{d\lambda}{d\mu} = h$ .

( $\impliedby$ )  $\frac{dP_\theta}{d\mu}$  " = "  $\frac{dP_\theta}{d\lambda} \cdot \frac{d\lambda}{d\mu}$  plus a.e.'s for  $\frac{dP_\theta}{d\mu} = (g_\theta \circ T) \cdot h$  and  $\frac{d\lambda}{d\mu} = h \cdot \sum_{i=1}^{\infty} c_i (g_{\theta_i} \circ T)$ . This implies that  $\frac{dP_\theta}{d\lambda}$  is of the "right" form.

## 5.2 Minimal Sufficiency

A sufficient statistic  $T(X)$  is a reduction of data  $X$  that in some sense still carries all the information available about  $\mathcal{P}$ . A sensible question is what form the "most compact" sufficient reduction of  $X$  takes. This is the notion of minimal sufficiency.

**Definition 54** *A sufficient statistic  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$  is **minimal sufficient for  $\mathcal{P}$**  (or for  $\theta$ ) provided for every sufficient statistic  $S : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{S}, \mathcal{G}) \exists$  a function  $U : \mathcal{S} \rightarrow \mathcal{T}$  such that*

$$T = U \circ S \quad \text{a.s. } \mathcal{P}$$

(This definition is equivalent to or close to equivalent to  $\mathcal{B}(T) \subset \mathcal{B}(S)$ .)

The fundamental qualitative insight here is that "the 'shape' of the log-likelihood" is minimal sufficient. This is phrased in technical terms in the next result.

**Theorem 55** Suppose that  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure  $\mu$ , and that  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$  is sufficient for  $\mathcal{P}$ . Suppose further that  $\exists$  versions of the R-N derivatives

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

such that the existence of  $k(x, y) > 0$  such that  $f_\theta(y) = f_\theta(x) k(x, y) \forall \theta$  implies that  $T(x) = T(y)$ . Then  $T$  is minimal sufficient.

A more technical and perhaps less illuminating line of argument for establishing minimal sufficiency is based on the next two results. (Shao's Theorem 2.3ii is the  $k = \infty$  version of Theorem 56.)

**Theorem 56** Suppose that  $\mathcal{P} = \{P_i\}_{i=0}^k$  is finite and let  $f_i$  be the R-N derivative of  $P_i$  with respect to some dominating  $\sigma$ -finite measure  $\mu$ . Suppose that all  $k + 1$  densities  $f_i$  are positive everywhere on  $\mathcal{X}$ . Then

$$T(X) = \left( \frac{f_1(X)}{f_0(X)}, \frac{f_2(X)}{f_0(X)}, \dots, \frac{f_k(X)}{f_0(X)} \right)$$

is minimal sufficient for  $\mathcal{P}$ .

(The positivity assumption in Theorem 56 is isn't really necessary, but proof without that assumption is fairly unpleasant.) The next result is a version of Shao's Theorem 2.3i.

**Theorem 57** Suppose that  $\mathcal{P}$  is a family of distributions and  $\mathcal{P}_0 \subset \mathcal{P}$  dominates  $\mathcal{P}$  ( $P(B) = 0 \forall P \in \mathcal{P}_0$  implies that  $P(B) = 0 \forall P \in \mathcal{P}$ ). If  $T$  is sufficient for  $\mathcal{P}$  and minimal sufficient for  $\mathcal{P}_0$ , then it is minimal sufficient for  $\mathcal{P}$ .

### 5.3 Ancillarity and Completeness

Sufficiency of  $T(X)$  means roughly that "what is left over in  $X$  beyond the information in  $T(X)$  is of no additional use in inferences about  $\theta$ ." This notion is related to the concepts of ancillarity and completeness.

**Definition 58** A statistic  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$  is said to be **ancillary for**  $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$  if the distribution of  $T(X)$  (i.e. the measure  $P_\theta^T$  on  $(\mathcal{T}, \mathcal{F})$  defined by  $P_\theta^T(F) = P_\theta(T^{-1}(F))$ ) does not depend upon  $\theta$  (i.e. is the same for all  $\theta$ ).

**Definition 59** A statistic  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathbb{R}^1, \mathcal{B}^1)$  is said to be **1st order ancillary for**  $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$  if  $E_\theta T(X)$  does not depend upon  $\theta$  (i.e. is the same for all  $\theta$ ).

Two slightly different versions of the notion of completeness are next. They are both stated in the framework of Definition 58. That is, suppose  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$ , define  $P_\theta^T$  on  $(\mathcal{T}, \mathcal{F})$  by  $P_\theta^T(F) = P_\theta(T^{-1}(F))$ , and let  $\mathcal{P}^T = \{P_\theta^T\}$ .

**Definition 60** (A)  $\mathcal{P}^T$  (or  $T$ ) is *complete* (or *boundedly complete*) for  $\mathcal{P}$  or  $\theta$  if

- i)  $U : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathbb{R}^1, \mathcal{B}^1)$  is  $\mathcal{B}(T)$ -measurable (and bounded) and
  - ii)  $E_\theta U(X) = 0 \forall \theta$
- imply that  $U = 0$  a.s.  $P_\theta \forall \theta$ .

(B)  $\mathcal{P}^T$  (or  $T$ ) is *complete* (or *boundedly complete*) for  $\mathcal{P}$  or  $\theta$  if

- i)  $h : (\mathcal{T}, \mathcal{F}) \longrightarrow (\mathbb{R}^1, \mathcal{B}^1)$  is  $\mathcal{F}$ -measurable (and bounded) and
  - ii)  $E_\theta h(T(X)) = 0 \forall \theta$
- imply that  $h \circ T = 0$  a.s.  $P_\theta \forall \theta$ .

By virtue of Lehmann's Theorem, the two versions of completeness in Definition 60 are equivalent. They both say that there is no nontrivial (bounded) function of  $T$  that is an unbiased estimator of 0. This is equivalent to saying that there is no nontrivial (bounded) function of  $T$  that is first order ancillary. Bounded completeness is obviously a slightly weaker condition than completeness. Two simple results regarding completeness and sub-families of models are next.

**Proposition 61** If  $T$  is complete for  $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$  and  $\Theta' \subset \Theta$ ,  $T$  need not be complete for  $\mathcal{P}' = \{P_\theta\}_{\theta \in \Theta'}$ .

**Theorem 62** If  $\Theta \subset \Theta''$  and  $\Theta$  dominates  $\Theta''$  in the sense that  $P_\theta(B) = 0 \forall \theta \in \Theta$  implies that  $P_\theta(B) = 0 \forall \theta \in \Theta''$ , then  $T$  (boundedly) complete for  $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$  implies that  $T$  is (boundedly) complete for  $\mathcal{P}'' = \{P_\theta\}_{\theta \in \Theta''}$ .

Completeness of a statistic guarantees that there is no part of it (no nontrivial function of it) that is somehow irrelevant in and of itself to inference about  $\theta$ . That is reminiscent of minimal sufficiency. An important connection between the two concepts is next. The proof of part i) of the following is on pages 95-96 of Schervish. (Silvey page 30 has a heuristic argument for part ii).)

**Theorem 63** (Bahadur's Theorem) Suppose that  $T : (\mathcal{X}, \mathcal{B}) \longrightarrow (\mathcal{T}, \mathcal{F})$  is sufficient and boundedly complete. Then

- i) if  $(\mathcal{T}, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}^k)$  then  $T$  is minimal sufficient, and
- ii) if there is any minimal sufficient statistic, then  $T$  is minimal sufficient.

Another standard result involving completeness, sufficiency, and ancillarity is the following (from page 99 of Schervish).

**Theorem 64** (Basu's Theorem) If  $T$  is a boundedly complete sufficient statistic and  $U$  is ancillary, then for all  $\theta$  the random variables  $U$  and  $T$  are independent (according to  $P_\theta$ ).

## 6 Facts About Common Statistical Models

### 6.1 Bayes Models

The "Bayes" approach to statistical problems is to add model assumptions to the basic structure introduced thus far. To the standard assumptions concerning  $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ , "Bayesians" add a distribution on the parameter space  $\Theta$ . That is, suppose that

$G$  is a distribution on  $(\Theta, \mathcal{C})$

where it may be convenient to suppose that  $G$  is dominated by some  $\sigma$ -finite measure  $\nu$  and that

$$\frac{dG}{d\nu}(\theta) = g(\theta)$$

If considered as a function of both  $x$  and  $\theta$ ,  $f_\theta(x)$  is  $\mathcal{B} \times \mathcal{C}$ -measurable, then  $\exists$  a probability (a joint distribution for  $(X, \theta)$  on  $(\mathcal{X} \times \Theta, \mathcal{B} \times \mathcal{C})$ ) defined by

$$\pi^{X, \theta}(A) = \int_A f_\theta(x) d(\mu \times G)(x, \theta) = \int_A f_\theta(x) g(\theta) d(\mu \times \nu)(x, \theta)$$

This distribution has marginals

$$\pi^X(B) = \int_{B \times \Theta} f_\theta(x) d(\mu \times G)(x, \theta) = \int_B \int_\Theta f_\theta(x) dG(\theta) d\mu(x)$$

(so that  $\frac{d\pi^X}{d\mu}(x) = \int_\Theta f_\theta(x) dG(\theta) = \int_\Theta f_\theta(x) g(\theta) d\nu(\theta)$ ) and

$$\pi^\theta(C) = \int_{\mathcal{X} \times C} f_\theta(x) d(\mu \times G)(x, \theta) = \int_C \int_{\mathcal{X}} f_\theta(x) d\mu(x) dG(\theta) = G(C)$$

Further, the conditionals (regular conditional probabilities) are computable as

$$\pi^{X|\theta}(B|\theta) = \int_B f_\theta(x) d\mu(x) = P_\theta(B)$$

and

$$\pi^{\theta|X}(C|x) = \int_C \frac{f_\theta(x)}{\int_\Theta f_\theta(x) dG(\theta)} dG(\theta)$$

Note then that

$$\frac{d\pi^{\theta|X}}{dG}(\theta|x) = \frac{f_\theta(x)}{\int_\Theta f_\theta(x) dG(\theta)}$$

and thus that

$$\frac{d\pi^{\theta|X}}{d\nu}(\theta|x) = \frac{f_\theta(x) g(\theta)}{\int_\Theta f_\theta(x) g(\theta) d\nu(\theta)}$$

is the usual "posterior density" for  $\theta$  given  $X = x$ .

Bayesians use the posterior distribution or density (THAT DEPENDS ON ONE'S CHOICE OF  $G$ ) as their basis of inference for  $\theta$ . Pages 84-88 of Schervish concern a Bayesian formulation of sufficiency and argue (basically) that the Bayesian form is equivalent to the "classical" form.

## 6.2 Exponential Families

Many standard families of distributions can be treated with a single set of analyses by recognizing them to be of a common “exponential family” form. The following is a list of facts about such families. (See, for example, Schervish Sections 2.2.1 and 2.2.2, or scattered results in Shao or the old books by Lehmann (*TSH* and *TPE*) for details.)

**Definition 65** Suppose that  $\mathcal{P} = \{P_\theta\}$  is dominated by a  $\sigma$ -finite measure  $\mu$ .  $\mathcal{P}$  is called an **exponential family** if for some  $h(x) \geq 0$ ,

$$f_\theta(x) \doteq \frac{dP_\theta}{d\mu}(x) = \exp\left(a(\theta) + \sum_{i=1}^k \eta_i(\theta)T_i(x)\right) h(x) \quad \forall \theta \quad .$$

**Definition 66** A family of probability measures  $\mathcal{P} = \{P_\theta\}$  is said to be **identifiable** if  $\theta_1 \neq \theta_2$  implies that  $P_{\theta_1} \neq P_{\theta_2}$ .

**Claim 67** Let  $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_k)$ ,  $\Gamma = \{\boldsymbol{\eta} \in \mathcal{R}^k \mid \int h(x) \exp(\sum \eta_i T_i(x)) d\mu(x) < \infty\}$ , and consider also the family of distributions (on  $\mathcal{X}$ ) with R-N derivatives wrt  $\mu$  of the form

$$f_{\boldsymbol{\eta}}(x) = K(\boldsymbol{\eta}) \exp\left(\sum_{i=1}^k \eta_i T_i(x)\right) h(x) \quad , \quad \boldsymbol{\eta} \in \Gamma \quad .$$

Call this family  $\mathcal{P}^*$ . This set of distributions on  $\mathcal{X}$  is at least as large as  $\mathcal{P}$  (and this second parameterization is mathematically nicer than the first).  $\Gamma$  is called the natural parameter space for  $\mathcal{P}^*$ . It is a convex subset of  $\mathcal{R}^k$ . (*TSH*, page 57.) If  $\Gamma$  lies in a subspace of dimension less than  $k$ , then  $f_{\boldsymbol{\eta}}(x)$  (and therefore  $f_\theta(x)$ ) can be written in a form involving fewer than  $k$  statistics  $T_i$ . We will henceforth assume  $\Gamma$  to be fully  $k$  dimensional. Note that depending upon the nature of the functions  $\eta_i(\theta)$  and the parameter space  $\Theta$ ,  $\mathcal{P}$  may be a proper subset of  $\mathcal{P}^*$ . That is, defining  $\Gamma_\Theta = \{(\eta_1(\theta), \eta_2(\theta), \dots, \eta_k(\theta)) \in \mathcal{R}^k \mid \theta \in \Theta\}$ ,  $\Gamma_\Theta$  can be a proper subset of  $\Gamma$ .

**Claim 68** The "support" of  $P_\theta$ , defined as  $\{x \mid f_\theta(x) > 0\}$ , is clearly  $\{x \mid h(x) > 0\}$ , which is independent of  $\theta$ . The distributions in  $\mathcal{P}$  are thus mutually absolutely continuous.

**Claim 69** From the Factorization Theorem, the statistic  $\mathbf{T} = (T_1, T_2, \dots, T_k)$  is sufficient for  $\mathcal{P}$ .

**Claim 70**  $\mathbf{T}$  has induced measures  $\{P_\theta^{\mathbf{T}} \mid \theta \in \Theta\}$  which also form an exponential family.

**Claim 71** If  $\Gamma_\Theta$  contains an open rectangle in  $\mathcal{R}^k$ , then  $\mathbf{T}$  is complete for  $\mathcal{P}$ . (See pages 142-143 of TSH.)

**Claim 72** If  $\Gamma_\Theta$  contains an open rectangle in  $\mathcal{R}^k$  (and actually under the much weaker assumptions given on page 44 of TPE)  $\mathbf{T}$  is minimal sufficient for  $\mathcal{P}$ .

**Claim 73** If  $g$  is any measurable real valued function such that  $E_\eta|g(X)| < \infty$ , then

$$E_\eta g(X) = \int g(x) f_\eta(x) d\mu(x)$$

is continuous on  $\Gamma$  and has continuous partial derivatives of all orders on the interior of  $\Gamma$ . These can be calculated as

$$\frac{\partial^{\alpha_1 + \alpha_2 + \dots + \alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \dots \partial \eta_k^{\alpha_k}} E_\eta g(X) = \int g(x) \frac{\partial^{\alpha_1 + \alpha_2 + \dots + \alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \dots \partial \eta_k^{\alpha_k}} f_\eta(x) d\mu(x) \quad .$$

(See page 59 of TSH.)

**Claim 74** If for  $\mathbf{u} = (u_1, u_2, \dots, u_k)$ , both  $\boldsymbol{\eta}_0$  and  $\boldsymbol{\eta}_0 + \mathbf{u}$  belong to  $\Gamma$

$$E_{\boldsymbol{\eta}_0} \exp\{u_1 T_1(X) + u_2 T_2(X) + \dots + u_k T_k(X)\} = \frac{K(\boldsymbol{\eta}_0)}{K(\boldsymbol{\eta}_0 + \mathbf{u})} \quad .$$

Further, if  $\boldsymbol{\eta}_0$  is in the interior of  $\Gamma$ , then

$$E_{\boldsymbol{\eta}_0} (T_1^{\alpha_1}(X) T_2^{\alpha_2}(X) \dots T_k^{\alpha_k}(X)) = K(\boldsymbol{\eta}_0) \frac{\partial^{\alpha_1 + \alpha_2 + \dots + \alpha_k}}{\partial \eta_1^{\alpha_1} \partial \eta_2^{\alpha_2} \dots \partial \eta_k^{\alpha_k}} \left( \frac{1}{K(\boldsymbol{\eta})} \right) \Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \quad .$$

In particular,  $E_{\boldsymbol{\eta}_0} T_j(X) = \frac{\partial}{\partial \eta_j} (-\ln K(\boldsymbol{\eta})) \Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$ ,  $Var_{\boldsymbol{\eta}_0} T_j(X) = \frac{\partial^2}{\partial \eta_j^2} (-\ln K(\boldsymbol{\eta})) \Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$ ,

and  $Cov_{\boldsymbol{\eta}_0} (T_j(X), T_l(X)) = \frac{\partial^2}{\partial \eta_j \partial \eta_l} (-\ln K(\boldsymbol{\eta})) \Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$ .

**Claim 75** If  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  has iid components, with each  $X_i \sim P_\theta$ , then  $\mathbf{X}$  generates a  $k$  dimensional exponential family on  $(\mathcal{X}^n, \mathcal{B}^n)$  wrt  $\mu^n$ . The  $k$  dimensional statistic  $\sum_{i=1}^n \mathbf{T}(X_i)$  is sufficient for this family. Under the condition that  $\Gamma_\Theta$  contains an open rectangle, this statistic is also complete and minimal sufficient.

### 6.3 Measures of Statistical Information

Details for the development here can be found in Section 2.3 of Schervish.

### 6.3.1 Fisher Information

This is an attempt to quantify how much on average the likelihood  $f_{\theta}(X)$  tells one about  $\theta \in \mathbb{R}^k$ . To even begin, one must have enough technical conditions in place to make it possible to define statistical information.

**Definition 76** We will say that the model  $\mathcal{P}$  with dominating  $\sigma$ -finite measure  $\mu$  is **FI Regular at  $\theta_0 \in \Theta \subset \mathbb{R}^k$**  provided there is an open neighborhood of  $\theta_0$ , say  $\mathcal{O}$ , such that

1.  $f_{\theta}(x) > 0 \forall x \in \mathcal{X}$  and  $\theta \in \mathcal{O}$ ,
2.  $\forall x$ ,  $f_{\theta}(x)$  has first order partial derivatives at  $\theta_0$ , and
3.  $1 = \int f_{\theta}(x) d\mu(x)$  can be differentiated with respect to each  $\theta_i$  under the integral sign at  $\theta_0$ , that is  $0 = \int \frac{\partial}{\partial \theta_i} f_{\theta}(x) \Big|_{\theta=\theta_0} d\mu(x)$ .

**Definition 77** If the model  $\mathcal{P}$  with dominating  $\sigma$ -finite measure  $\mu$  is FI Regular at  $\theta_0 \in \Theta \subset \mathbb{R}^k$ , and

$$E_{\theta_0} \left( \frac{\partial}{\partial \theta_i} \ln f_{\theta}(X) \Big|_{\theta=\theta_0} \right)^2 < \infty \forall i$$

then the  $k \times k$  matrix

$$\mathbf{I}(\theta_0) = \left( E_{\theta_0} \frac{\partial}{\partial \theta_i} \ln f_{\theta}(X) \Big|_{\theta=\theta_0} \frac{\partial}{\partial \theta_j} \ln f_{\theta}(X) \Big|_{\theta=\theta_0} \right)$$

is called the **Fisher Information about  $\theta$  contained in  $X$  at  $\theta_0$** .

**Claim 78** Fisher information doesn't depend upon the dominating measure.

The Fisher Information certainly *does* depend upon the parameterization one uses. For example, suppose that the model  $\mathcal{P}$  is FI regular at  $\theta_0 \in \mathbb{R}^1$  and that for some "nice" function  $h$  (say 1-1 with derivative  $h^{-1}$ )

$$\eta = h(\theta)$$

Then for  $P_{\theta}$  with R-N derivative  $f_{\theta}$  with respect to  $\mu$ , the distributions  $Q_{\eta}$  (for  $\eta$  in the range of  $h$ ) are defined by

$$Q_{\eta} = P_{h^{-1}(\eta)}$$

and have R-N derivatives (again with respect to  $\mu$ )

$$g_{\eta} = \frac{dQ_{\eta}}{d\mu} = f_{h^{-1}(\eta)}$$

Then continuing to let "prime" denote differentiation with respect to  $\theta$  or  $\eta$ ,

$$g'_{\eta} = f'_{h^{-1}(\eta)} \frac{1}{h'(h^{-1}(\eta))}$$

and

$$I(\eta) = \left( \frac{1}{h'(h^{-1}(\eta))} \right)^2 J(h^{-1}(\eta))$$

where  $I$  is the information in  $X$  about  $\eta$  and  $J$  is the information in  $X$  about  $\theta$ .

Under appropriate conditions, there is a second form for the Fisher Information matrix.

**Theorem 79** *Suppose that the model  $\mathcal{P}$  is FI regular at  $\theta_0 \in \Theta$ . If  $\forall x, f_{\theta}(x)$  has continuous partials in the neighborhood  $\mathcal{O}$ , and  $1 = \int f_{\theta}(x) d\mu(x)$  can be differentiated twice with respect to each  $\theta_i$  under the integral sign at  $\theta_0$ , that is  $0 = \int \frac{\partial}{\partial \theta_i} f_{\theta}(x) \Big|_{\theta=\theta_0} d\mu(x)$  and  $0 = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_{\theta}(x) \Big|_{\theta=\theta_0} d\mu(x) \forall i$  and  $j$ , then*

$$\mathbf{I}(\theta_0) = -E_{\theta_0} \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_{\theta}(X) \Big|_{\theta=\theta_0} \right)$$

One reason Fisher Information is an attractive measure is that for independent observations, it is additive.

**Proposition 80** *If  $X_1, X_2, \dots, X_n$  are independent with  $X_i \sim P_{i,\theta}$ , then  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  carries Fisher Information  $\mathbf{I}(\theta) = \sum_{i=1}^n \mathbf{I}_i(\theta)$  (where in the obvious notation  $\mathbf{I}_i(\theta)$  is the Fisher information carried by  $X_i$ ).*

Fisher information also behaves "properly" in terms of what happens when one reduces  $X$  to some statistic  $T(X)$ . That is, there are the next two results, the second of which is like 2.86 of Schervish.

**Proposition 81** *Suppose the model  $\mathcal{P}$  is FI Regular at  $\theta_0 \in \Theta \subset \mathbb{R}^k$ . If the function  $T$  is 1-1, then the Fisher Information in  $T(X)$  is the same as the Fisher Information in  $X$ .*

**Proposition 82** *Suppose the model  $\mathcal{P}$  is FI Regular at  $\theta_0 \in \Theta \subset \mathbb{R}^k$ . (?And also suppose that  $\{P_{\theta}^T\}_{\theta \in \Theta}$  is FI regular at  $\theta_0$ ?) Then the matrix*

$$\mathbf{I}_X(\theta_0) - \mathbf{I}_{T(X)}(\theta_0)$$

*is non-negative definite. Further, if  $T(X)$  is sufficient, then  $\mathbf{I}_X(\theta_0) - \mathbf{I}_{T(X)}(\theta_0) = \mathbf{0}$ . Also,  $\mathbf{I}_X(\theta_0) - \mathbf{I}_{T(X)}(\theta_0) = \mathbf{0} \forall \theta_0$  implies that  $T(X)$  is sufficient.*

In exponential families, the Fisher Information has a very nice form.

**Proposition 83** *For a family of distributions as in Claim 67,*

$$\mathbf{I}(\eta_0) = \text{Var}(\mathbf{T}(X)) = \left( \frac{\partial^2}{\partial \eta_i \partial \eta_j} (-\ln K(\eta)) \Big|_{\eta=\eta_0} \right)$$

### 6.3.2 Kullback-Leibler Information

This is another measure of "information" that aims to measure how far apart two distributions are in the sense of likelihood, i.e. in the sense of how hard it will be to discriminate between them on the basis of  $X$ . It is a quantity that arises naturally in the asymptotics of maximum likelihood.

**Definition 84** *If  $P$  and  $Q$  are probability measures on  $\mathcal{X}$  with R-N derivatives with respect to a common dominating measure  $\mu$ ,  $p$  and  $q$  respectively, then the **Kullback-Leibler Information** is the  $P$  expected log likelihood ratio*

$$I(P, Q) = E_P \ln \left( \frac{p(X)}{q(X)} \right) = \int \ln \left( \frac{p(x)}{q(x)} \right) p(x) d\mu(x)$$

**Claim 85**  $I(P, Q)$  is not in general the same as  $I(Q, P)$ .

**Lemma 86**  $I(P, Q) \geq 0$  and there is equality only when  $P = Q$ .

Lemma 86 follows from the next (slightly more general) lemma (from page 75 of Silvey), upon noting that  $C(\cdot) = -\ln(\cdot)$  is strictly convex.

**Lemma 87** *Suppose that  $P$  and  $Q$  are two different probability measures on  $\mathcal{X}$  with R-N derivatives with respect to a common dominating measure  $\mu$ ,  $p$  and  $q$  respectively. Suppose further that a function  $C : (0, \infty) \rightarrow \mathbb{R}$  is convex. Then*

$$E_P C \left( \frac{q(x)}{p(x)} \right) \geq C(1)$$

with strict inequality if  $C$  is strictly convex.

K-L Information has a kind of additive property similar to that possessed by the Fisher Information.

**Theorem 88** *If  $X = (X_1, X_2)$  and under both  $P$  and  $Q$  the variables  $X_1$  and  $X_2$  are independent, then*

$$I_X(P, Q) = I_{X_1}(P^{X_1}, Q^{X_1}) + I_{X_2}(P^{X_2}, Q^{X_2})$$

K-L Information also has the kind of non-increasing property associated with reduction of  $X$  to a statistic  $T(X)$  possessed by the Fisher Information.

**Theorem 89**  $I_X(P, Q) \geq I_{T(X)}(P^{T(X)}, Q^{T(X)})$  and there is equality if and only if  $T(X)$  is sufficient for  $\{P, Q\}$ .

Finally, there is also an important connection between the notions of Fisher and K-L Information under the formulation of Section 6.3.1.

**Theorem 90** *Under the hypotheses of Theorem 79, if one can reverse the order of limits in all*

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} (E_{\theta_0} \ln f_{\theta}(X)) \Big|_{\theta=\theta_0}$$

to get

$$E_{\theta_0} \left( \frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln f_{\theta}(X) \Big|_{\theta=\theta_0} \right)$$

then

$$\left( \frac{\partial^2}{\partial\theta_i\partial\theta_j} I(P_{\theta_0}, P_{\theta}) \Big|_{\theta=\theta_0} \right) = I(\theta_0)$$

(Clearly, the first  $I$  is the K-L Information and the second is the Fisher Information).

As a bit of an aside, it is worth considering what would establish the legitimacy of exchange of order of differentiation and integration (e.g. required in the hypotheses of Theorem 90). For any function  $h(\theta) : \mathbb{R}^1 \rightarrow \mathbb{R}^1$ , define

$$D_{\Delta}^1(h)(\theta) = h(\theta + \Delta) - h(\theta)$$

and

$$D_{\Delta}^2(h)(\theta) = h(\theta + \Delta) + h(\theta - \Delta) - 2h(\theta)$$

Then

$$\lim_{\Delta \rightarrow 0} \frac{1}{\Delta} D_{\Delta}^1(h)(\theta) = h'(\theta)$$

and

$$\lim_{\Delta \rightarrow 0} \frac{1}{\Delta^2} D_{\Delta}^2(h)(\theta) = h''(\theta)$$

Consider  $g(\theta, x) : \mathbb{R}^1 \times \mathcal{X} \rightarrow \mathbb{R}^1$ . Then, for example,

$$\begin{aligned} & \int \left( \frac{1}{\Delta^2} D_{\Delta}^2(g(\cdot, x))(\theta_0) \right) f_{\theta_0}(x) d\mu(x) \\ &= \frac{1}{\Delta^2} \int [g(\theta_0 + \Delta, x) + g(\theta_0 - \Delta, x) - 2g(\theta_0, x)] f_{\theta_0}(x) d\mu(x) \\ &= \frac{1}{\Delta^2} D_{\Delta}^2 \left( \int g(\cdot, x) f_{\theta_0}(x) d\mu(x) \right) (\theta_0) \end{aligned}$$

Now the limit of the last of these (as  $\Delta \rightarrow 0$ ) is

$$\frac{d^2}{d\theta^2} \left( \int g(\theta, x) f_{\theta_0}(x) d\mu(x) \right) \Big|_{\theta=\theta_0}$$

So the question of interchange of expectation and second derivative is that of whether the  $(\mu)$  integral of the limit of

$$\left( \frac{1}{\Delta^2} D_{\Delta}^2(g(\cdot, x))(\theta_0) \right) f_{\theta_0}(x)$$

is the limit of the  $(\mu)$  integral. This might be established using the dominated convergence theorem.

## 7 Statistical Decision Theory

This is a formal framework for balancing of various kinds of risks under uncertainty, and can sometimes be used look for "good" statistical procedures. It finds modern uses in the search for good data mining and machine learning algorithms.

### 7.1 Basic Framework and Concepts

To the usual statistical modeling framework of Section 4 we now add the following elements:

1. some "action space"  $\mathcal{A}$  with  $\sigma$ -algebra  $\mathcal{E}$ ,
2. a suitably measurable "loss function"  $L : \Theta \times \mathcal{A} \rightarrow [0, \infty)$ , and
3. (non-randomized) decision rules  $\delta : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathcal{A}, \mathcal{E})$ .

(For data  $X$ ,  $\delta(X)$  is the action taken on the basis of  $X$ .)

**Definition 91**  $R(\theta, \delta) \equiv E_{\theta} L(\theta, \delta(X)) = \int L(\theta, \delta(x)) dP_{\theta}(x)$  mapping  $\Theta \rightarrow [0, \infty)$  is called the **risk function** for  $\delta$ .

**Definition 92**  $\delta$  is **at least as good as**  $\delta'$  if  $R(\theta, \delta) \leq R(\theta, \delta') \forall \theta$ .

**Definition 93**  $\delta$  is **better than (dominates)**  $\delta'$  if  $R(\theta, \delta) \leq R(\theta, \delta') \forall \theta$  with strict inequality for at least one  $\theta \in \Theta$ .

**Definition 94**  $\delta$  and  $\delta'$  are **risk equivalent** if  $R(\theta, \delta) = R(\theta, \delta') \forall \theta$ .

**Definition 95**  $\delta$  is **best in a class of decision rules**  $\Delta$  if  $\delta \in \Delta$  and  $\delta$  is at least as good as any other  $\delta' \in \Delta$ .

**Definition 96**  $\delta$  is **inadmissible** in  $\Delta$  if  $\exists \delta' \in \Delta$  which is better than (dominates)  $\delta$ .

**Definition 97**  $\delta$  is **admissible** in  $\Delta$  if it is not inadmissible in  $\Delta$  (if there is no  $\delta'$  which is better).

We will use the notation

$$\mathcal{D} = \{\delta\} = \text{the class of (non-randomized) decision rules}$$

It is technically useful to extend the notion of decision procedures to include the possibility of randomizing in various ways.

**Definition 98** If for each  $x \in \mathcal{X}$ ,  $\phi_x$  is a distribution on  $(\mathcal{A}, \mathcal{E})$ , then  $\phi_x$  is called a **behavioral decision rule**.

The notion of a behavioral decision rule is that one observes  $X = x$  and then makes a random choice of an element of  $\mathcal{A}$  using distribution  $\phi_x$ . We'll let

$$\mathcal{D}^* = \{\phi_x\} = \text{the class of behavioral decision rules}$$

It's possible to think of  $\mathcal{D}$  as a subset of  $\mathcal{D}^*$  by associating with  $\delta \in \mathcal{D}$  a behavioral decision rule  $\phi_x^\delta$  that is a point mass distribution on  $\mathcal{A}$  concentrated at  $\delta(x)$ . The natural definition of the risk function of a behavioral decision rule is (abusing notation and using "R" here too)

$$R(\theta, \phi) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) d\phi_x(a) dP_\theta(x)$$

A second (less intuitively appealing) notion of randomizing decisions is one that might somehow pick an element of  $\mathcal{D}$  at random (and then plug in  $X$ ). Let  $\mathcal{F}$  be a  $\sigma$ -algebra on  $\mathcal{D}$  that contains all singleton sets.

**Definition 99** A *randomized decision function (or rule)*  $\psi$  is a probability measure on  $(\mathcal{D}, \mathcal{F})$ .

$\delta$  with distribution  $\psi$  becomes a random object. Let

$$\mathcal{D}_* = \{\psi\} = \text{the class of randomized decision rules}$$

It's possible to think of  $\mathcal{D}$  as a subset of  $\mathcal{D}_*$  by associating with  $\delta \in \mathcal{D}$  a randomized decision rule  $\psi_\delta$  placing mass 1 on  $\delta$ . The natural definition of the risk function of a behavioral decision rule is (yet again abusing notation and using "R" here too)

$$\begin{aligned} R(\theta, \psi) &= \int_{\mathcal{D}} R(\theta, \delta) d\psi(\delta) \\ &= \int_{\mathcal{D}} \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x) d\psi(\delta) \end{aligned}$$

(assuming that  $R(\theta, \delta)$  is properly measurable).

The behavioral decision rules are most natural, while the randomized decision rules are easiest to deal with in some proofs. So it is a reasonably important question when  $\mathcal{D}^*$  and  $\mathcal{D}_*$  are equivalent in the sense of generating the same set of risk functions. Some properly qualified version of the following is true.

**Proposition 100** *If  $\mathcal{A}$  is a complete separable metric space with  $\mathcal{E}$  the Borel  $\sigma$ -algebra and ??? regarding the distributions  $P_\theta$  and ???, then  $\mathcal{D}^*$  and  $\mathcal{D}_*$  are equivalent in terms of generating the same set of risk functions.*

$\mathcal{D}^*$  and  $\mathcal{D}_*$  are clearly more complicated than  $\mathcal{D}$ . A sensible question is when they really provide anything  $\mathcal{D}$  doesn't provide. One kind of negative answer can be given for the case of convex loss. The following is like 2.5a of Shao, page 151 of Schervish, page 40 of Berger, or page 78 of Ferguson.

**Lemma 101** Suppose that  $\mathcal{A}$  is a convex subset of  $\mathbb{R}^d$  and  $\phi_x$  is a behavioral decision rule. Define a non-randomized decision rule by

$$\delta(x) = \int_{\mathcal{A}} a d\phi_x(a)$$

(In the case that  $d > 1$ , interpret  $\delta(x)$  as vector-valued, the integral as a vector of integrals over  $d$  coordinates of  $a \in \mathcal{A}$ .) Then

1. if  $L(\theta, \cdot) : \mathcal{A} \rightarrow [0, \infty)$  is convex, then

$$R(\theta, \delta) \leq R(\theta, \phi)$$

and

2. if  $L(\theta, \cdot) : \mathcal{A} \rightarrow [0, \infty)$  is strictly convex,  $R(\theta, \phi) < \infty$  and  $P_{\theta}(\{x | \phi_x \text{ is non-degenerate}\}) > 0$ , then

$$R(\theta, \delta) < R(\theta, \phi)$$

**Corollary 102** Suppose that  $\mathcal{A}$  is a convex subset of  $\mathbb{R}^d$ ,  $\phi_x$  is a behavioral decision rule and

$$\delta(x) = \int_{\mathcal{A}} a d\phi_x(a)$$

Then

1. if  $L(\theta, a)$  is convex in  $a \forall \theta$ ,  $\delta$  is at least as good as  $\phi$ ,
2. if  $L(\theta, a)$  is convex in  $a \forall \theta$  and for some  $\theta_0$  the function  $L(\theta_0, a)$  is strictly convex in  $a$ ,  $R(\theta_0, \phi) < \infty$  and  $P_{\theta_0}(\{x | \phi_x \text{ is non-degenerate}\}) > 0$ , then  $\delta$  is better than  $\phi$ .

The corollary shows, e.g., that for squared error loss estimation, averaging out over non-trivial randomization in a behavioral decision rule will in fact improve that estimator.

## 7.2 (Finite Dimensional) Geometry of Decision Theory

A very helpful device for understanding some of the basics of decision theory is the geometry associated with cases where  $\Theta$  is finite. Let

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$$

Assume that  $R(\theta, \psi) < \infty \forall \theta \in \Theta$  and  $\psi \in \mathcal{D}_*$  and note that in this case  $R(\cdot, \psi) : \Theta \rightarrow [0, \infty)$  can be thought of as a  $k$ -vector. Let

$$\begin{aligned} \mathcal{S} &= \{y_{\psi} = (y_1, y_2, \dots, y_k) \in \mathbb{R}^k \mid y_i = R(\theta_i, \psi) \text{ for all } i \text{ and some } \psi \in \mathcal{D}_*\} \\ &= \text{the set of all randomized risk vectors} \end{aligned}$$

**Theorem 103**  $\mathcal{S}$  is convex.

In fact, it turns out that if

$$\mathcal{S}^0 = \{y_\delta = (y_1, y_2, \dots, y_k) \in \mathbb{R}^k \mid y_i = R(\theta_i, \delta) \text{ for all } i \text{ and some } \delta \in \mathcal{D}\}$$

$\mathcal{S}$  is the convex hull of  $\mathcal{S}^0$ . This is the smallest convex set containing  $\mathcal{S}^0$ , the set of all convex combinations of points of  $\mathcal{S}^0$ , the intersection of all convex sets containing  $\mathcal{S}^0$ . (By the definition of randomized decision rules, each  $\psi$  produces a  $y_\psi$  that is a convex combination of points in  $\mathcal{S}^0$ , so it is at least obvious that  $\mathcal{S}$  is a subset of the convex hull of  $\mathcal{S}^0$ .)

**Definition 104** For  $x \in \mathbb{R}^k$ , the **lower quadrant** of  $x$  is

$$Q_x = \{z \mid z_i \leq x_i \ \forall i\}$$

**Theorem 105**  $y \in \mathcal{S}$  (or the decision rule giving rise to  $y$ ) is admissible if and only if  $Q_y \cap \mathcal{S} = \{y\}$ .

**Definition 106** For  $\bar{\mathcal{S}}$  the closure of  $\mathcal{S}$ , the **lower boundary** of  $\mathcal{S}$  is

$$\lambda(\mathcal{S}) = \{y \mid Q_y \cap \bar{\mathcal{S}} = \{y\}\}$$

**Definition 107**  $\mathcal{S}$  is **closed from below** if  $\lambda(\mathcal{S}) \subset \mathcal{S}$ .

Let  $A(\mathcal{S}) = \{y \mid Q_y \cap \mathcal{S} = \{y\}\}$  be the set of admissible risk points.

**Theorem 108** If  $\mathcal{S}$  is closed  $A(\mathcal{S}) = \lambda(\mathcal{S})$ .

Theorem 108 is very easy to prove. A better theorem (one with weaker hypotheses) that is perhaps surprisingly harder to prove is next.

**Theorem 109** If  $\mathcal{S}$  is closed from below  $A(\mathcal{S}) = \lambda(\mathcal{S})$ .

Now drop the finiteness assumption.

### 7.3 Complete Classes of Decision Rules

**Definition 110** A class of decision rules  $\mathcal{C} \subset \mathcal{D}^*$  is a **complete class** if for any  $\phi \notin \mathcal{C}$ ,  $\exists \phi' \in \mathcal{C}$  such that  $\phi'$  is better than  $\phi$ .

(One can do better in the complete class than anywhere outside of the complete class in terms of risk functions.)

**Definition 111** A class of decision rules  $\mathcal{C} \subset \mathcal{D}^*$  is an **essentially complete class** if for any  $\phi \notin \mathcal{C}$ ,  $\exists \phi' \in \mathcal{C}$  such that  $\phi'$  is at least as good as  $\phi$ .

**Definition 112** A class of decision rules  $\mathcal{C} \subset \mathcal{D}^*$  is a **minimal complete class** if it is complete and is a subset of any other complete class.

Let  $A(\mathcal{D}^*)$  be the set of admissible rules in  $\mathcal{D}^*$ .

**Theorem 113** *If a minimal complete class  $\mathcal{C}$  exists, then  $\mathcal{C} = A(\mathcal{D}^*)$ .*

**Theorem 114** *If  $A(\mathcal{D}^*)$  is complete, then it is minimal complete.*

Theorems 113 and 114 are properly qualified versions of the rough (not quite true) statement

" $A(\mathcal{D}^*)$  is the minimal complete class."

## 7.4 Sufficiency and Decision Theory

Presumably one can typically do as well in a decision problem based on a sufficient statistic  $T(X)$  as one can do based on  $X$ .

**Theorem 115** *If  $T$  is sufficient for  $\mathcal{P}$  and  $\phi$  is a behavioral decision rule, then  $\exists$  another behavioral decision rule  $\phi'$  that is a function of  $T$  and has the same risk function as  $\phi$ .*

(That  $\phi'$  is a function of  $T$  means that  $T(x) = T(y)$  implies that  $\phi'_x$  and  $\phi'_y$  are the same distributions on  $\mathcal{A}$ .)

Theorem 115 shows that the class of rules that are functions of a sufficient statistic is essentially complete. Theorem 118 shows that under a convex loss, conditioning on a sufficient statistic can sometimes actually *improve* a decision rule. We state two small results that lead up to this.

**Lemma 116** *Suppose  $\mathcal{A} \subset \mathbb{R}^d$  is convex and  $\delta_1$  and  $\delta_2$  are two non-randomized decision rules. Then*

$$\delta = \frac{1}{2}(\delta_1 + \delta_2)$$

*is also a decision rule. Further,*

1. *if  $L(\theta, a)$  is convex in  $a$  and  $R(\theta, \delta_1) = R(\theta, \delta_2)$ , then  $R(\theta, \delta) \leq R(\theta, \delta_1)$ , and*
2. *if  $L(\theta, a)$  is strictly convex in  $a$ ,  $R(\theta, \delta_1) = R(\theta, \delta_2) < \infty$ , and  $P_\theta(\delta_1(X) \neq \delta_2(X)) > 0$ , then  $R(\theta, \delta) < R(\theta, \delta_1)$ .*

**Corollary 117** *Suppose  $\mathcal{A} \subset \mathbb{R}^d$  is convex and  $\delta_1$  and  $\delta_2$  are two nonrandomized decision rules with identical risk functions. If  $L(\theta, a)$  is convex in  $a \forall \theta$ , the existence of a  $\theta \in \Theta$  for which  $L(\theta, a)$  is strictly convex in  $a$ ,  $R(\theta, \delta_1) = R(\theta, \delta_2) < \infty$  and  $P_\theta(\delta_1(X) \neq \delta_2(X)) > 0$ , implies that  $\delta_1$  and  $\delta_2$  are inadmissible.*

Finally, we have the general form of the Rao-Blackwell Theorem (that usually appears in its SELE form in Stat 543 and is 2.5.6 of Shao, page 152 of Schervish, Berger page 41, and Ferguson page 121).

**Theorem 118** (*Rao-Blackwell Theorem*) Suppose  $\mathcal{A} \subset \mathbb{R}^d$  is convex and  $\delta$  is a nonrandomized decision function with  $E_\theta \|\delta(X)\| < \infty \forall \theta$ . Suppose further that  $T$  is sufficient for  $\theta$  and with  $\mathcal{B}_0 = \mathcal{B}(T)$  let

$$\delta_0(x) \equiv E(\delta | \mathcal{B}_0)(x)$$

Then  $\delta_0$  is a non-randomized decision rule. Further,

1. if  $L(\theta, \delta)$  is convex in  $a$ , then  $R(\theta, \delta_0) \leq R(\theta, \delta)$ , and
2. for any  $\theta$  for which  $L(\theta, a)$  is strictly convex in  $a$ ,  $R(\theta, \delta) < \infty$ , and  $P_\theta(\delta_0(X) \neq \delta(X)) > 0$ , it is the case that  $R(\theta, \delta_0) < R(\theta, \delta)$ .

## 7.5 Bayes Decision Rules

The Bayes approach to decision theory is one means of reducing risk functions to numbers so that they can be compared in a straightforward fashion. Let  $G$  be a distribution on  $(\Theta, \mathcal{C})$ .

**Definition 119** The **Bayes risk** of  $\phi \in \mathcal{D}^*$  with respect to the prior  $G$  is (abusing notation again and once more using the symbols  $R(\cdot, \cdot)$ )

$$R(G, \phi) = \int R(\theta, \phi) dG(\theta)$$

Further (again abusing notation) the "minimum" Bayes risk is

$$R(G) \equiv \inf_{\phi' \in \mathcal{D}^*} R(G, \phi')$$

**Definition 120**  $\phi \in \mathcal{D}^*$  is said to be **Bayes with respect to  $G$**  if

$$R(G, \phi) = R(G)$$

**Definition 121**  $\phi \in \mathcal{D}^*$  is said to be  **$\epsilon$ -Bayes with respect to  $G$**  provided

$$R(G, \phi) \leq R(G) + \epsilon$$

Bayes rules are often admissible, at least if the prior involved "spreads its mass around sufficiently."

**Theorem 122** If  $\Theta$  is countable and  $G$  is a prior with  $G(\theta) < \infty \forall \theta$  and  $\phi$  is Bayes with respect to  $G$ , then  $\phi$  is admissible.

**Theorem 123** Suppose  $\Theta \subset \mathbb{R}^k$  is such that every neighborhood of any point  $\theta \in \Theta$  has a non-empty intersection with the interior of  $\Theta$ . Suppose further that  $R(\theta, \phi)$  is continuous in  $\theta$  for all  $\phi \in \mathcal{D}^*$ . Let  $G$  be a prior distribution that has support  $\Theta$  in the sense that every open ball that is a subset of  $\Theta$  has positive  $G$  probability. Then if  $R(G) < \infty$  and  $\phi$  is a Bayes rule with respect to  $G$ ,  $\phi$  is admissible.

**Theorem 124** *If every Bayes rule with respect to  $G$  has the same risk function, they are all admissible.*

**Corollary 125** *In a problem where Bayes rules exist and are unique, they are admissible.*

A converse of Theorem 122 (for finite  $\Theta$ ) is given in Theorem 127. Its proof requires the use of a piece of  $k$ -dimensional analytical geometry (called the separating hyperplane theorem) that is stated next.

**Theorem 126** (*Separating Hyperplane Theorem*) *Let  $\mathcal{S}_1$  and  $\mathcal{S}_2$  be two disjoint convex subsets of  $\mathbb{R}^k$ . Then  $\exists$  a  $p \in \mathbb{R}^k$  such that  $p \neq 0$  and  $\sum p_i x_i \leq \sum p_i y_i$   $\forall y \in \mathcal{S}_1$  and  $\forall x \in \mathcal{S}_2$ .*

**Theorem 127** *If  $\Theta$  is finite and  $\phi$  is admissible, then  $\phi$  is Bayes with respect to some prior.*

For non-finite  $\Theta$ , admissible rules are "usually" Bayes or "limits" of Bayes rules. See Section 2.10 of Ferguson or page 546 of Berger.

As the next result shows, Bayesians don't need randomized decision rules, at least for purposes of achieving minimum Bayes risk,  $R(G)$ . See also page 147 of Schervish.

**Proposition 128** *Suppose that  $\psi \in \mathcal{D}_*$  is Bayes versus  $G$  and  $R(G) < \infty$ . Then there exists a nonrandomized rule  $\delta \in \mathcal{D}$  that is also Bayes versus  $G$ .*

There remain the questions of 1) when Bayes rules exist and 2) what they look like when they do exist. These are addressed next.

**Theorem 129** *If  $\Theta$  is finite,  $\mathcal{S}$  is closed from below, and  $G$  assigns positive probability each  $\theta \in \Theta$ , there is a rule Bayes versus  $G$ .*

**Definition 130** *A formal nonrandomized Bayes rule versus a prior  $G$  is a rule  $\delta(x)$  such that  $\forall x \in \mathcal{X}$ ,*

$$\delta(x) \text{ is an } a \in \mathcal{A} \text{ minimizing } \int_{\Theta} L(\theta, a) \left( \frac{f_{\theta}(x)}{\int_{\Theta} f_{\theta}(x) dG(\theta)} \right) dG(\theta)$$

**Definition 131** *If  $G$  is a  $\sigma$ -finite measure, a formal nonrandomized generalized Bayes rule versus  $G$  is a rule  $\delta(x)$  such that that  $\forall x \in \mathcal{X}$ ,*

$$\delta(x) \text{ is an } a \in \mathcal{A} \text{ minimizing } \int_{\Theta} L(\theta, a) f_{\theta}(x) dG(\theta)$$

## 7.6 Minimax Decision Rules

An alternative to the Bayes reduction of  $R(\theta, \phi)$  to the number  $R(G, \phi) = \int R(\theta, \phi) dG(\theta)$  is to reduce  $R(\theta, \phi)$  to the number  $\sup_{\theta} R(\theta, \phi)$ .

**Definition 132** A decision rule  $\phi \in \mathcal{D}^*$  is said to be **minimax** if

$$\sup_{\theta} R(\theta, \phi) = \inf_{\phi'} \sup_{\theta} R(\theta, \phi')$$

It is intuitively plausible that if one is trying to produce a minimax rule by pushing down "the highest peak" in  $R(\theta, \phi)$  (considered as a function of  $\theta$ ), that will tend to direct one toward rules with fairly "flat" risk functions. So the notion of "equalizer rules" has a central place in minimax theory.

**Definition 133** If a decision rule  $\phi \in \mathcal{D}^*$  has a constant risk function, it is called an **equalizer rule**.

**Theorem 134** If  $\phi \in \mathcal{D}^*$  is an equalizer rule and is admissible, then it is minimax.

**Theorem 135** Suppose that  $\{\phi_i\}$  is a sequence of decision rules, each  $\phi_i$  Bayes versus a prior  $G_i$ . If  $R(G_i, \phi_i) \rightarrow C < \infty$  and  $\phi$  is a decision rule with  $R(\theta, \phi) \leq C \forall \theta$ , then  $\phi$  is minimax.

**Corollary 136** If  $\phi \in \mathcal{D}^*$  is an equalizer rule and is Bayes with respect to a prior  $G$ , then it is minimax.

**Corollary 137** If  $\phi \in \mathcal{D}^*$  is Bayes versus  $G$  and  $R(\theta, \phi) \leq R(G) \forall \theta$ , then it is minimax.

In light of Corollaries 136 and 137, a reasonable question is "How might I identify a prior that might produce a Bayes rule that is minimax?" An answer to this question is "Look for a 'least favorable' prior distribution."

**Definition 138** A prior distribution  $G$  is said to be **least favorable** if

$$R(G) = \sup_{G'} R(G')$$

**Theorem 139** If  $\phi$  is Bayes versus  $G$  and  $R(\theta, \phi) \leq R(G) \forall \theta$ , then  $G$  is least favorable.

## 7.7 Invariance/Equivalence Theory

In decision problems with sufficient symmetries, it may make sense to restrict attention to decision rules that have corresponding symmetries. It then can happen that risk functions of such rules take relatively few values ... are potentially even constant. Then when comparing (constant) risk functions, one is in a position to identify a best (symmetric) rule. Invariance/equivariance theory is the formalization of this line of argument.

### 7.7.1 Location Parameter Estimation and Invariance/Equivariance

**Definition 140** When  $\mathcal{X} = \mathbb{R}^k$  and  $\Theta \subset \mathbb{R}^k$ , to say that  $\theta$  is a **location parameter** means that under  $P_\theta$ ,  $X - \theta$  has the same distribution for all  $\theta$ .

**Definition 141** When  $\mathcal{X} = \mathbb{R}^k$  and  $\Theta \subset \mathbb{R}$ , to say that  $\theta$  is a **1-dimensional location parameter** means that under  $P_\theta$ ,  $X - \theta \mathbf{1}$  has the same distribution for all  $\theta$ .

Until further notice, assume that  $\Theta = \mathcal{A} = \mathbb{R}$ .

**Definition 142** A loss function is called **location invariant** if  $L(\theta, a) = \rho(\theta - a)$  for some non-negative function  $\rho$ . If  $\rho(z)$  is increasing in  $z$  for  $z > 0$  and decreasing in  $z$  for  $z < 0$ , the corresponding decision problem is called a **location estimation problem**.

**Definition 143** A non-randomized decision rule  $\delta$  is called **location equivariant** if

$$\delta(x + c\mathbf{1}) = \delta(x) + c$$

**Theorem 144** If  $\theta$  is a 1-dimensional location parameter,  $L(\theta, a)$  is location invariant and  $\delta$  is location equivariant, then  $R(\theta, \delta)$  is constant in  $\theta$ .

**Definition 145** A function  $g: \mathbb{R}^k \rightarrow \mathbb{R}$  is called **location invariant** if  $g(x + c\mathbf{1}) = g(x)$ .

A location invariant function is "constant on '45° lines.'"

**Lemma 146** A function  $u$  is location invariant if and only if  $u$  depends upon  $x$  only through

$$y(x) = (x_1 - x_k, x_2 - x_k, \dots, x_{k-1} - x_k)$$

**Lemma 147** Suppose that a non-randomized decision rule  $\delta_0$  is location equivariant. Then  $\delta_1$  is location equivariant if and only if  $\exists$  a location invariant function  $u$  such that

$$\delta_1 = \delta_0 + u$$

**Theorem 148** Suppose that  $L(\theta, a) = \rho(\theta - a)$  is location invariant and  $\exists$  a location equivariant estimator  $\delta_0$  for  $\theta$  with finite risk. Let  $\rho'(z) = \rho(-z)$ . Suppose that for each  $y \exists$  a number  $v^*(y)$  that minimizes

$$E_{\theta=0}[\rho'(\delta_0 - v) | Y = y]$$

over choices of  $v$ . Then

$$\delta^*(x) = \delta_0(x) - v^*(y(x))$$

is a best location equivariant estimator of  $\theta$ .

In the case of squared error loss, the estimator of Theorem 148 is "Pittman's estimator" and can be written out more explicitly in some situations.

**Theorem 149** *Suppose that  $L(\theta, a) = (\theta - a)^2$  and that the distribution of  $X = (X_1, \dots, X_k)$  ( $P_\theta$ ) has R-N derivative with respect to Lebesgue measure on  $\mathbb{R}^k$*

$$f_\theta(x) = f(x_1 - \theta, \dots, x_k - \theta) = f(x - \theta \mathbf{1})$$

and that  $f$  has second moments. Then the best location equivariant estimator of  $\theta$  is

$$\delta^*(x) = \frac{\int_{-\infty}^{\infty} \theta f(x - \theta \mathbf{1}) d\theta}{\int_{-\infty}^{\infty} f(x - \theta \mathbf{1}) d\theta}$$

Pittman's estimator is the formal generalized Bayes estimator versus Lebesgue measure on  $\mathbb{R}$ .

**Lemma 150** *For the case of  $L(\theta, a) = (\theta - a)^2$  a best location equivariant estimator of  $\theta$  must be unbiased for  $\theta$ .*

## 7.7.2 Generalities of Invariance/Equivariance Theory

A somewhat more general story about invariance leads to a generalization of Theorem 144. So, suppose that  $\mathcal{P} = \{P_\theta\}$  is identifiable and that  $L(\theta, a) = L(\theta, a') \forall \theta$  implies that  $a = a'$ . Invariance/equivariance theory is phrased in terms of transformations  $\mathcal{X} \rightarrow \mathcal{X}$ ,  $\Theta \rightarrow \Theta$ , and  $\mathcal{A} \rightarrow \mathcal{A}$  and how decision rules relate to them.

**Definition 151** *For 1-1 transformations of  $\mathcal{X}$  onto  $\mathcal{X}$ ,  $g_1$  and  $g_2$ , the **composition** of  $g_1$  and  $g_2$  is another 1-1 transformations of  $\mathcal{X}$  onto  $\mathcal{X}$  defined by*

$$g_2 \circ g_1(x) = g_2(g_1(x)) \quad \forall x$$

**Definition 152** *For  $g$  a 1-1 transformation of  $\mathcal{X}$  onto  $\mathcal{X}$ , if  $h$  is another such transformation such that  $h \circ g(x) = x \quad \forall x$ ,  $h$  is called the **inverse** of  $g$  and denoted as  $h = g^{-1}$ .*

It's an easy result that both  $g^{-1} \circ g(x) = x \quad \forall x$  and  $g \circ g^{-1}(x) = x \quad \forall x$ .

**Definition 153** *The **identity** transformation on  $\mathcal{X}$  is defined by  $e(x) = x \quad \forall x$ .*

Another easy fact is then that  $e \circ g = g \circ e = g$  for  $g$  any 1-1 transformation of  $\mathcal{X}$  onto  $\mathcal{X}$ .

**Proposition 154** *If  $\mathcal{G}$  is a class of 1-1 transformations of  $\mathcal{X}$  onto  $\mathcal{X}$  that is closed under composition and inversion, then with the operation " $\circ$ "  $\mathcal{G}$  is a group in the usual mathematical sense.*

To go anywhere with this theory, we need transformations on  $\Theta$  that fit with the ones on  $\mathcal{X}$ . At a minimum, we don't want the transformations on  $\mathcal{X}$  to move us outside of  $\mathcal{P}$ .

**Definition 155** If  $g$  is a 1-1 transformation of  $\mathcal{X}$  onto  $\mathcal{X}$  such that

$$\left\{ P_{\theta}^{g(X)} \right\}_{\theta \in \Theta} = \mathcal{P}$$

the transformation is said to **leave the model invariant**. If each element of a group of transformations  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant, we'll say that the group leaves the model invariant.

One circumstance in which  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant is that where the model  $\mathcal{P}$  can be thought of as "generated by  $\mathcal{G}$ " in the sense that for some  $U \sim P$  (for a fixed  $P$ ) and  $\mathcal{P} = \{ P^{g(U)} | g \in \mathcal{G} \}$ .

If a group of transformations on  $\mathcal{X}$  (say  $\mathcal{G}$ ) leaves  $\mathcal{P}$  invariant, there is a natural corresponding group of transformations on  $\Theta$ . That is, for  $g \in \mathcal{G}$

$$X \sim P_{\theta} \Rightarrow g(X) \sim P_{\theta'} \text{ for some } \theta' \in \Theta$$

(and in fact, because of the indentifiability assumption, there is only one such  $\theta'$ ). So one might adopt the next definition.

**Definition 156** If  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant, for each  $g \in \mathcal{G}$ , define  $\bar{g} : \Theta \rightarrow \Theta$  by

$$\bar{g}(\theta) = \text{the element } \theta' \in \Theta \text{ such that } X \sim P_{\theta} \Rightarrow g(X) \sim P_{\bar{g}(\theta)}$$

**Proposition 157** For  $\mathcal{G}$  a group of 1-1 transformation of  $\mathcal{X}$  onto  $\mathcal{X}$  leaving  $\mathcal{P}$  invariant, the set

$$\bar{\mathcal{G}} = \{ \bar{g} | g \in \mathcal{G} \}$$

with the operation "function composition" is a group of 1-1 transformations of  $\Theta$  onto  $\Theta$ .

( $\bar{\mathcal{G}}$  is the homomorphic image of  $\mathcal{G}$  under the map  $g \rightarrow \bar{g}$ .)

Now we need transformations on  $\mathcal{A}$  that fit together nicely with those on  $\mathcal{X}$  and  $\Theta$ .

**Definition 158** A **loss function**  $L(\theta, a)$  is said to be **invariant** under the group of transformations  $\mathcal{G}$ , provided for every  $g \in \mathcal{G}$  and  $a \in \mathcal{A}$ ,  $\exists$  a unique  $a' \in \mathcal{A}$  such that

$$L(\theta, a) = L(\bar{g}(\theta), a') \quad \forall \theta$$

**Definition 159** If  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant and the loss function  $L(\theta, a)$  is invariant, define  $\tilde{g} : \mathcal{A} \rightarrow \mathcal{A}$  by

$$\tilde{g}(a) = \text{the element } a' \text{ of } \mathcal{A} \text{ such that } L(\theta, a) = L(\bar{g}(\theta), a') \quad \forall \theta$$

According to the notations in Definitions 156 and 159, for  $\mathcal{G}$  leaving  $\mathcal{P}$  invariant and  $L(\theta, a)$  invariant,

$$L(\theta, a) = L(\bar{g}(\theta), \tilde{g}(a))$$

**Proposition 160** For  $\mathcal{G}$  a group of 1-1 transformations  $\mathcal{X}$  onto  $\mathcal{X}$  leaving  $\mathcal{P}$  invariant and invariant loss function  $L(\theta, a)$ ,

$$\tilde{\mathcal{G}} = \{\tilde{g}|g \in \mathcal{G}\}$$

with the operation "function composition" is a group of 1-1 transformations of  $\mathcal{A}$  onto  $\mathcal{A}$ .

( $\tilde{\mathcal{G}}$  is the homomorphic image of  $\mathcal{G}$  under the map  $g \rightarrow \tilde{g}$ .)

Finally, we come to the business of equivariance of decision functions. Here restrict attention to non-randomized decision rules.

**Definition 161** In an invariant decision problem (one where  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant and the loss function  $L(\theta, a)$  is invariant) a non-randomized rule  $\delta$  is said to be **equivariant** provided

$$\delta(g(x)) = \tilde{g}(\delta(x)) \quad \forall x \in \mathcal{X} \text{ and } g \in \mathcal{G}$$

We now have enough structure to state the promised generalization of Theorem 144.

**Theorem 162** If  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant, the loss function  $L(\theta, a)$  is invariant, and the non-randomized rule  $\delta$  is equivariant,

$$R(\theta, \delta) = R(\bar{g}(\theta), \delta) \quad \forall \theta \in \Theta \text{ and } g \in \mathcal{G}$$

**Definition 163** For  $\theta \in \Theta$ ,  $\mathcal{O}_\theta = \{\theta' \in \Theta | \theta' = \bar{g}(\theta) \text{ for some } g \in \mathcal{G}\}$  is called the **orbit** in  $\Theta$  containing  $\theta$ .

The distinct orbits  $\mathcal{O}_\theta$  partition  $\Theta$ , and Theorem 162 says that an equivariant non-randomized rule in an invariant problem has a risk function that is constant on orbits. Comparing risk functions for equivariant non-randomized rules in an invariant problem reduces to comparing risks on orbits. This, of course, is most helpful when there is only one orbit.

**Definition 164** A group of 1-1 transformations of a space onto itself is called **transitive** if for any two points in the space there is a transformation in the group taking the first point to the second.

**Corollary 165** If  $\mathcal{G}$  leaves  $\mathcal{P}$  invariant, the loss function  $L(\theta, a)$  is invariant, and  $\tilde{\mathcal{G}}$  is transitive over  $\Theta$ , then every equivariant non-randomized decision rule has constant risk.

**Definition 166** In an invariant decision problem, a decision rule  $\delta$ , an equivariant decision rule  $\delta$  is said to be a **best (non-randomized) equivariant** (or **MRE: minimum risk equivariant**) **decision rule** provided that for any other (non-randomized) equivariant rule  $\delta'$

$$R(\theta, \delta) \leq R(\theta, \delta') \quad \forall \theta$$

(and it suffices to check that  $R(\theta, \delta) \leq R(\theta, \delta')$  for any one  $\theta$ ).

## 8 Asymptotics of Likelihood Inference

As usual, write  $\mathcal{P} = \{P_\theta\}$ . Suppose that  $\mathcal{P}$  is identifiable and dominated by a  $\sigma$ -finite measure,  $\mu$ , and write

$$f_\theta = \frac{dP_\theta}{d\mu}$$

We'll here consider asymptotics for inference based on the likelihood function  $f_\theta(X)$  (a random function of  $\theta$ ).

We'll concentrate on the iid (one sample) case. Here is some notation we'll use in the discussion. The basic one-observation/one-dimensional model is  $(\mathcal{X}, \mathcal{B}, P_\theta)$  with parameter space  $\Theta \subset \mathbb{R}^k$ . Based on this we'll take

$X^n = (X_1, X_2, \dots, X_n)$	the observable through stage $n$
$\mathcal{X}^n$	the observation space for $X^n$
$\mathcal{B}^n$	the $n$ -fold product $\sigma$ -algebra
$P_\theta^n$	the distribution of $X^n$
$\mu^n$	the dominating (product) measure on $\mathcal{X}^n$
$f_\theta^n = \frac{dP_\theta^n}{d\mu^n}$	

Notice that if one wants to prove almost sure convergence results, one needs a fixed probability space in which to operate for all  $n$ . In that case, one could use a big space involving  $\Omega = \mathcal{X}^\infty, \mathcal{C} = \mathcal{B}^\infty, P_\theta^\infty, \mu^\infty$  and think of  $X_i$  in  $X^n = (X_1, X_2, \dots, X_n)$  as

$$X_i(\omega) = \text{the } i\text{th coordinate of } \omega$$

### 8.1 Asymptotics of Likelihood Estimation

#### 8.1.1 Consistency of Likelihood-Based Estimators

**Definition 167** A statistic  $T : \mathcal{X}^n \rightarrow \Theta$  is called a *maximum likelihood estimator* of  $\theta$  if

$$f_{T(x^n)}^n(x^n) = \sup_{\theta \in \Theta} f_\theta^n(x^n) \quad \forall x^n \in \mathcal{X}^n$$

For a fixed  $x^n \in \mathcal{X}^n$  a  $\theta \in \Theta$  maximizing  $f_\theta^n(x^n)$  might be called a maximum likelihood estimate even when some  $x^n$ 's might not have corresponding maximizers.

**Definition 168** Suppose that  $\{T_n\}$  is a sequence of statistics,  $T_n : \mathcal{X}^n \rightarrow \Theta \subset \mathbb{R}^k$

1.  $\{T_n\}$  is (weakly) **consistent** at  $\theta_0 \in \Theta$  if for every  $\epsilon > 0$ ,

$$P_{\theta_0}^n [ \|T_n - \theta_0\| > \epsilon ] \rightarrow 0 \text{ as } n \rightarrow \infty$$

2.  $\{T_n\}$  is **strongly consistent** at  $\theta_0 \in \Theta$  if  $T_n \rightarrow \theta_0$  a.s.  $P_{\theta_0}^\infty$

Let

$$\begin{aligned} z_{\theta_0}(\theta) &= \mathbf{E}_{\theta_0} \ln f_\theta(X) \\ &= -I(\theta_0, \theta) + \int \ln(f_{\theta_0}(x)) f_{\theta_0}(x) d\mu(x) \end{aligned}$$

It is a corollary of Lemma 86 that for  $\theta \neq \theta_0$

$$z_{\theta_0}(\theta) < z_{\theta_0}(\theta_0)$$

Consider the loglikelihood through  $n$  observations

$$l_n(x^n, \theta) = \ln f_\theta^n(x^n) = \sum_{i=1}^n \ln f_\theta(x_i)$$

Note that

$$z_{\theta_0}(\theta) = \mathbf{E}_{\theta_0} \frac{1}{n} l_n(X^n, \theta)$$

and in fact that the LLN promises that for every  $\theta$

$$\frac{1}{n} l_n(X^n, \theta) \rightarrow z_{\theta_0}(\theta) \quad \text{in } \theta_0 \text{ probability}$$

So assuming enough regularity to make the convergence uniform in  $\theta$  there is perhaps hope that a single maximizer of  $\frac{1}{n} l_n(X^n, \theta)$  will exist and be close to the maximizer of  $z_{\theta_0}(\theta)$  (namely  $\theta_0$ ), i.e. that an MLE might exist and be consistent. To make this kind of argument precise is quite hard. See Theorems 7.49 and 7.54 of Schervish. We'll take a different, easier, and much more common line of development.

Where the  $f_\theta$  are differentiable in  $\theta$ , an MLE must satisfy the likelihood equations.

**Definition 169** In the case where  $\Theta \subset \mathbb{R}^k$  and the  $f_\theta$  are differentiable in  $\theta$ , the **likelihood equations** are

$$\frac{\partial}{\partial \theta_i} \ln f_\theta^n(x^n) = 0 \quad \text{for } i = 1, 2, \dots, k$$

Much of the relatively simple theory of (maximum) likelihood revolves more around solutions of the likelihood equations than it does maximizers of the likelihood. For a one-dimensional parameter there is the following result.

**Theorem 170** Suppose that  $k = 1$  and  $\exists$  an open neighborhood of  $\theta_0$ , say  $\mathcal{O}$ , such that

1.  $f_\theta(x) > 0 \forall x$  and  $\forall \theta \in \mathcal{O}$ ,

2.  $\forall x, f_\theta(x)$  is differentiable at every point  $\theta \in \mathcal{O}$ , and
3.  $E_{\theta_0} \ln f_\theta(X)$  exists for all  $\theta \in \mathcal{O}$  and  $E_{\theta_0} \ln f_{\theta_0}(X)$  is finite.

Then if  $\epsilon > 0$  and  $\delta > 0 \exists n$  such that  $\forall m > n$  the  $\theta_0$  probability that the equation

$$\frac{d}{d\theta} l_m(X^m, \theta) = 0$$

(the likelihood equation) has a root within  $\epsilon$  of  $\theta_0$  is at least  $1 - \delta$ .

**Corollary 171** Under the hypotheses of Theorem 170, suppose in addition that  $\Theta$  is open and  $f_\theta(x)$  is differentiable at every point  $\theta \in \Theta$ , so that the likelihood equation

$$\frac{d}{d\theta} l_n(X^n, \theta) = \frac{d}{d\theta} \sum_{i=1}^n \ln f_\theta(X_i) = 0$$

makes sense at all  $\theta \in \Theta$ . Define  $\rho_n$  to be the root of the likelihood equation when there is exactly one (and otherwise adopt any definition for  $\rho_n$ ). If with  $\theta_0$  probability approaching 1 the likelihood equation has a single root, then

$$\rho_n \rightarrow \theta_0 \text{ in } \theta_0 \text{ probability}$$

**Corollary 172** Under the hypotheses of Theorem 170, if  $\{T_n\}$  is a sequence of estimators consistent at  $\theta_0$  and

$$\hat{\theta}_n = \begin{cases} T_n & \text{if the likelihood} \\ & \text{equation has no roots} \\ \text{the root of the likelihood} & \text{otherwise} \\ \text{equation closest to } T_n & \end{cases}$$

then

$$\hat{\theta}_n \rightarrow \theta_0 \text{ in } \theta_0 \text{ probability}$$

Actually implementing the prescription for  $\hat{\theta}_n$  in Corollary 172 is not without its practical and philosophical difficulties. Another line of attack for possibly improving a consistent estimator is to adopt a "1-step Newton improvement" on it. Writing

$$L_n(\theta) = l_n(X^n, \theta) = \sum_{i=1}^n \ln f_\theta(X_i)$$

this is

$$\tilde{\theta}_n = T_n - \frac{L'_n(T_n)}{L''_n(T_n)}$$

(of course provided  $L''_n(T_n) \neq 0$ ). In the event that  $k > 1$ , this becomes

$$\tilde{\theta}_n = T_n - (\mathbf{L}''_n(T_n))^{-1} \mathbf{L}'_n(T_n)$$

for  $\mathbf{L}'_n(T_n)$  the  $k \times 1$  vector of first partials of  $L_n(\theta)$  and  $\mathbf{L}''_n(T_n)$  the  $k \times k$  matrix of second partials of  $L_n(\theta)$ . It is plausible that estimators of this type end up having asymptotic behavior similar to that of real roots of the likelihood equations. See, e.g., Schervish's development around his Theorem 7.75 for a more general ("M-estimation") version of this.

### 8.1.2 Asymptotic Normality of Likelihood-Based Estimators

An honest version of the Stat 543 "MLE's are asymptotically Normal" is next.

**Theorem 173** *Suppose that  $k = 1$  and  $\exists$  an open neighborhood of  $\theta_0$ , say  $\mathcal{O}$ , such that*

1.  $f_\theta(x) > 0 \forall x$  and  $\forall \theta \in \mathcal{O}$ ,
2.  $\forall x$ ,  $f_\theta(x)$  is three times differentiable at every point of  $\mathcal{O}$ ,
3.  $\exists M(x) \geq 0$  with  $E_{\theta_0}M(X) < \infty$  and

$$\left| \frac{d^3}{d\theta^3} \ln f_\theta(x) \right| \leq M(x) \quad \forall x \text{ and } \forall \theta \in \mathcal{O},$$

4.  $1 = \int f_\theta(x) d\mu(x)$  can be differentiated twice with respect to  $\theta$  under the integral at  $\theta_0$ , and
5.  $I_1(\theta) \in (0, \infty) \forall \theta \in \mathcal{O}$ .

If with  $\theta_0$  probability approaching 1,  $\hat{\theta}_n$  is a root of the likelihood equation and  $\hat{\theta}_n \rightarrow \theta_0$  in  $\theta_0$  probability, then under  $\theta_0$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

**Corollary 174** *Under the hypotheses of Theorem 173, if  $I_1(\theta)$  is continuous at  $\theta_0$ , then under  $\theta_0$*

$$\sqrt{nI_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1)$$

**Corollary 175** *Under the hypotheses of Theorem 173, under  $\theta_0$*

$$\sqrt{-L_n''(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1)$$

What is often a more practically useful result (parallel to Theorem 173) concerns "one-step Newton improvements" on " $\sqrt{n}$ -consistent" estimators. (The following is a special case of Schervish's Theorem 7.75.)

**Theorem 176** *Under the hypotheses 1-5 of Theorem 173, suppose that under  $\theta_0$  estimators  $T_n$  are  $\sqrt{n}$ -consistent, that is  $\sqrt{n}(T_n - \theta_0)$  converges in distribution (or more generally, is  $O(1)$  in  $\theta_0$  probability). Then with*

$$\tilde{\theta}_n = T_n - \frac{L_n'(T_n)}{L_n''(T_n)}$$

under  $\theta_0$

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

It is then obvious that versions of Corollaries 174 and 175 hold where  $\hat{\theta}_n$  is replaced by  $\tilde{\theta}_n$ .

## 8.2 Asymptotics of LRT-like Tests and Competitors and Related Confidence Regions

This primarily concerns the most common "general purpose" means of cooking up tests (and related confidence sets) in non-standard statistical models. For testing  $H_0: \theta \in \Theta_0$  versus  $H_a: \theta \in \Theta_1 \equiv \Theta - \Theta_0$  we might consider a statistic

$$LR(x) = \frac{\sup_{\theta \in \Theta_1} f_{\theta}(x)}{\sup_{\theta \in \Theta_0} f_{\theta}(x)}$$

or perhaps

$$\lambda(x) = \max(LR(x), 1) = \frac{\sup_{\theta \in \Theta} f_{\theta}(x)}{\sup_{\theta \in \Theta_0} f_{\theta}(x)}$$

where rejection is appropriate for large  $\lambda(x)$ . Notice that if there is an MLE of  $\theta$ , say  $\hat{\theta}$ ,

$$\sup_{\theta \in \Theta} f_{\theta}(x) = f_{\hat{\theta}}(x)$$

and there is thus the promise of using "MLE-type" asymptotics to establish limiting distributions for likelihood ratio-type test statistics. Corresponding to the case of a point null hypothesis ( $H_0: \theta = \theta_0$ ) in the iid (one-sample) context there is the following. (This is the  $k = 1$  version. Similar results hold for  $k > 1$  with  $\chi_k^2$  limits.)

**Theorem 177** *Under the hypotheses of Theorem 173, if*

$$\Lambda_n \equiv 2 \ln \left( \frac{f_{\hat{\theta}_n}(X^n)}{f_{\theta_0}(X^n)} \right) = 2 \left( L_n(\hat{\theta}_n) - L_n(\theta_0) \right)$$

then under  $\theta_0$

$$\Lambda_n \xrightarrow{d} \chi_1^2$$

Theorem 177 and its  $k > 1$  generalizations are relevant to testing of point null hypotheses and corresponding confidence set making for entire parameter vectors. There are also versions of  $\chi^2$  limit theorems for composite null hypotheses relevant to confidence set making for parts of a  $k$ -dimensional parameter vector. (See, for example, Schervish page 459 for the complete statement of a theorem of this type.) In vague terms, one can often prove things like what is indicated in the next claim.

**Claim 178** *Suppose that  $\Theta \subset \mathbb{R}^p$  and appropriate regularity conditions hold in the iid problem. If  $k < p$  and*

$$\lambda_n = \frac{\sup_{\theta \in \Theta} f_{\theta}^n(X^n)}{\sup_{\substack{\theta \text{ s.t. } \theta_i = \theta_i^0 \\ i = 1, 2, \dots, k}} f_{\theta}^n(X^n)}$$

then under any  $\theta \in \Theta$  for which  $\theta_i = \theta_i^0$  for  $i = 1, 2, \dots, k$

$$2 \ln \lambda_n \xrightarrow{d} \chi_k^2$$

Claim 178 is relevant to testing  $H_0: \theta_i = \theta_i^0$  for  $i = 1, 2, \dots, k$  and for making confidence sets for  $(\theta_1, \theta_2, \dots, \theta_k)$ . A way of thinking about confidence sets for part of a parameter vector that come from the inversion of likelihood ratio tests is in terms of "profile loglikelihood" functions.

**Definition 179** In the (iid) context where  $\Theta \subset \mathbb{R}^p$  and  $k < p$ , the function of  $(\theta_1, \theta_2, \dots, \theta_k)$

$$L_n^*(\theta_1, \theta_2, \dots, \theta_k) = \sup_{(\theta_{k+1}, \dots, \theta_p)} L_n(\theta)$$

is called the profile loglikelihood function for  $(\theta_1, \theta_2, \dots, \theta_k)$ .

With the notation/jargon of Definition 179 and results like Claim 178, large  $n$  confidence sets for  $(\theta_1, \theta_2, \dots, \theta_k)$  become

$$\left\{ (\theta_1, \theta_2, \dots, \theta_k) \mid L_n^*(\theta_1, \theta_2, \dots, \theta_k) > \sup_{(\theta_1, \theta_2, \dots, \theta_k)} L_n^*(\theta_1, \theta_2, \dots, \theta_k) - \frac{1}{2} \chi_k^2 \right\}$$

(for  $\chi_k^2$  a small upper percentage point of the  $\chi_k^2$  distribution).

A generalization of the hypothesis  $H_0: \theta_i = \theta_i^0$  for  $i = 1, 2, \dots, k$  is the hypothesis  $H_0: h_i(\theta) = c_i$  for  $i = 1, 2, \dots, k$  for some set of functions  $\{h_i(\theta)\}$  (that could be  $h_i(\theta) = \theta_i$ ) and constants  $\{c_i\}$ . The obvious LRTs based on

$$\lambda_n^{c_1, c_2, \dots, c_k} = \frac{\sup_{\theta \in \Theta} f_{\theta}^n(X^n)}{\sup_{\substack{\theta \text{ s.t. } h_i(\theta) = c_i \\ i = 1, 2, \dots, k}} f_{\theta}^n(X^n)}$$

could be defined,  $\chi_k^2$  limit distributions identified under appropriate conditions on the  $h_i(\theta)$ , and large sample confidence sets for  $(h_1(\theta), h_2(\theta), \dots, h_k(\theta))$  of the form

$$\{(c_1, c_2, \dots, c_k) \mid 2 \ln \lambda_n^{c_1, c_2, \dots, c_k} < \chi_k^2\}$$

made (for  $\chi_k^2$  a small upper percentage point of the  $\chi_k^2$  distribution).

Tests (and associated confidence sets) sometimes discussed as competitors for likelihood ratio tests (and regions) are so-called "Wald" and "score" procedures. So consider next Wald tests for  $H_0: h_i(\theta) = c_i$  for  $i = 1, 2, \dots, k$  in the (iid) context where  $\Theta \subset \mathbb{R}^p$  and  $k < p$ .

Define

$$h(\theta) = (h_1(\theta), h_2(\theta), \dots, h_k(\theta))'$$

and for  $\hat{\theta}_n$  such that under  $\theta_0$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \text{MVN}_p(0, I_1^{-1}(\theta_0))$$

consider the estimator of  $h(\theta)$

$$h(\hat{\theta}_n)$$

For differentiable  $h(\theta)$  let

$$\mathbf{H}(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_j} h_i(\theta) \end{pmatrix}_{k \times p}$$

The  $\Delta$ -method shows that under  $\theta_0$

$$\sqrt{n} \left( h(\hat{\theta}_n) - h(\theta_0) \right) \xrightarrow{d} \text{MVN}_k \left( 0, \mathbf{H}(\theta_0) I_1^{-1}(\theta_0) \mathbf{H}(\theta_0)' \right)$$

Let  $\mathbf{B}(\theta_0) = \mathbf{H}(\theta_0) I_1^{-1}(\theta_0) \mathbf{H}(\theta_0)'$ . Provided  $\mathbf{H}(\theta_0)$  is full rank so  $\mathbf{B}(\theta_0)$  is invertible and  $h(\theta_0) = (c_1, c_2, \dots, c_k)'$ , under  $\theta_0$

$$n \left( h(\hat{\theta}_n) - \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \right)' \mathbf{B}(\theta_0)^{-1} \left( h(\hat{\theta}_n) - \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \right) \xrightarrow{d} \chi_k^2$$

So, supposing that  $\mathbf{B}(\theta)$  is continuous at  $\theta_0$ , an "expected information" version of a Wald test of  $H_0: h_i(\theta) = c_i$  for  $i = 1, 2, \dots, k$  can be based on the statistic

$$W_n^{c_1, c_2, \dots, c_k} = n \left( h(\hat{\theta}_n) - \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \right)' \mathbf{B}(\hat{\theta}_n)^{-1} \left( h(\hat{\theta}_n) - \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \right)$$

and a limiting  $\chi_k^2$  null distribution. Or, an "observed information" version of  $W_n$  can be had by writing

$$\hat{\mathbf{B}}_n = \mathbf{H}(\hat{\theta}_n) \left( -\frac{1}{n} \mathbf{D}_n(\hat{\theta}_n) \right)^{-1} (\theta_0) \mathbf{H}(\hat{\theta}_n)'$$

for

$$\mathbf{D}_n(\theta) \equiv \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} L_n(\theta) \right)$$

and then replacing  $\mathbf{B}(\hat{\theta}_n)^{-1}$  in  $W_n^{c_1, c_2, \dots, c_k}$  with  $\hat{\mathbf{B}}_n^{-1}$ . Note finally that a ( $k$ -dimensional) large sample ellipsoidal confidence set for  $h(\theta)$  is then of the form

$$\left\{ (c_1, c_2, \dots, c_k)' \mid W_n^{c_1, c_2, \dots, c_k} < \chi_k^2 \right\}$$

(for  $\chi_k^2$  a small upper percentage point of the  $\chi_k^2$  distribution).

The "score test" competitor for LRT and Wald tests of  $H_0: h_i(\theta) = c_i$  for  $i = 1, 2, \dots, k$  is motivated by the notion that if  $H_0$  is true and  $\hat{\theta}_n$  maximizes the loglikelihood  $L_n(\theta)$  subject to the constraints imposed by the null hypothesis, then  $\hat{\theta}_n$  should typically nearly be an absolute maximizer of  $L_n(\theta)$ , and therefore

the gradient of  $L_n(\theta)$  at  $\tilde{\theta}_n$  should be small. That is, define the  $p \times 1$  gradient vector

$$\nabla L_n(\tilde{\theta}_n) = \left( \frac{\partial}{\partial \theta_i} L_n(\theta) \Big|_{\theta = \tilde{\theta}_n} \right)$$

(this is the score function evaluated at the restricted MLE,  $\tilde{\theta}_n$ ) and under  $H_0$  we expect this to be close to 0. The score test statistic is

$$V_n^{c_1, c_2, \dots, c_k} = \frac{1}{n} \nabla L_n(\tilde{\theta}_n)' I_1^{-1}(\tilde{\theta}_n) \nabla L_n(\tilde{\theta}_n)$$

As it turns out, it is typically the case that under  $H_0$ ,  $V_n^{c_1, c_2, \dots, c_k}$  differs from  $2 \ln \lambda_n^{c_1, c_2, \dots, c_k}$  (the version of the corresponding LRT statistic with the standard limiting null distribution) by a quantity tending to 0 in probability. Hence a  $\chi_k^2$  limiting null distribution is appropriate for a score test using  $V_n^{c_1, c_2, \dots, c_k}$ . Note too that another large sample confidence set for  $h(\theta)$  is then of the form

$$\{(c_1, c_2, \dots, c_k)' | V_n^{c_1, c_2, \dots, c_k} < \chi_k^2\}$$

(for  $\chi_k^2$  a small upper percentage point of the  $\chi_k^2$  distribution).

It is common to point out that when dealing with a composite null hypothesis, in order to use an LRT one must find both unrestricted and restricted MLEs, while to use a Wald test only an unrestricted MLE is needed, and to use a score test only a restricted MLE is needed.

### 8.3 Asymptotic Shape of the Loglikelihood and Related Bayesian Asymptotics

We consider some simple/crude/easy implications of the foregoing to questions of the large sample shape of random functions that are the loglikelihood and Bayes posterior densities. (Much finer/harder results are available and some are discussed in Schervish. See his Section 7.4.) Standard statistical folklore is that for large  $n$ , under  $\theta_0$

1. loglikelihoods are steep with a peak near  $\theta_0$  and a locally quadratic shape in continuous  $\Theta$  contexts, and
2. "posteriors are consistent" if a prior spreads some of its mass around at/near  $\theta_0$ , and in continuous  $\Theta$  contexts, the corresponding posterior density will be approximately normal with mean an MLE of  $\theta$ .

In the following discussion, continue the iid (one sample) context.

**Lemma 180** For  $\theta \neq \theta_0$

$$L_n(\theta_0) - L_n(\theta) \rightarrow \infty \text{ in } \theta_0 \text{ probability}$$

**Claim 181** If  $G$  is a prior on  $\Theta$  with density  $g$  with respect to  $\nu$  and the posterior density is therefore

$$g(\theta|X^n) = \frac{g(\theta) f_\theta^n(X^n)}{\int g(\theta) f_\theta^n(X^n) d\nu(\theta)}$$

then for  $\theta \neq \theta_0$ ,

$$\frac{g(\theta|X^n)}{g(\theta_0|X^n)} \rightarrow 0 \text{ in } \theta_0 \text{ probability}$$

**Corollary 182** If  $\Theta$  is finite,  $\nu$  is counting measure, and  $g(\theta) > 0 \forall \theta$ , the posterior  $\pi^{\theta|X^n}$  is consistent in the sense that for any  $A \subset \Theta$

$$\pi^{\theta|X^n}(A) \rightarrow I[\theta_0 \in A] \text{ in } \theta_0 \text{ probability}$$

A (crude) result giving a hint about shape for a loglikelihood is next.

**Lemma 183** Under the hypotheses of Theorem 173, suppose that  $\hat{\theta}_n$  is an MLE of  $\theta$  consistent at  $\theta_0$ . Then

$$Q_n(\Delta) \equiv L_n(\hat{\theta}_n) - L_n\left(\hat{\theta}_n + \frac{\Delta}{\sqrt{n}}\right) \rightarrow \frac{1}{2}\Delta^2 I_1(\theta_0) \text{ in } \theta_0 \text{ probability}$$

Lemma 183 and the phenomenon it represents have implications for the nature of a posterior distribution. Schervish has a very careful and very hard development. What we can get easily is the following.

**Corollary 184** Under the hypotheses of Theorem 173, suppose that a prior  $G$  has density  $g$  with respect to Lebesgue measure on  $\Theta$ ,  $g(\theta_0) > 0$  and  $g$  is continuous at  $\theta_0$ . If  $\hat{\theta}_n$  is an MLE of  $\theta$  consistent at  $\theta_0$ , then the posterior density  $g(\theta|X^n)$  has the property that

$$R(\Delta) \equiv \ln \left( \frac{g(\hat{\theta}_n|X^n)}{g\left(\hat{\theta}_n + \frac{\Delta}{\sqrt{n}}|X^n\right)} \right) \rightarrow \frac{1}{2}\Delta^2 I_1(\theta_0) \text{ in } \theta_0 \text{ probability}$$

Corollary 184 suggests that under  $\theta_0$  one might roughly speaking expect posterior densities to look approximately normal with mean  $\hat{\theta}_n$  and variance  $1/nI_1(\theta_0)$ . So a practical approximate form for  $g(\theta|X^n)$  might be

$$N\left(\hat{\theta}_n, \frac{1}{nI_1(\hat{\theta}_n)}\right) \text{ or } N\left(\hat{\theta}_n, \frac{-1}{L_n''(\hat{\theta}_n)}\right)$$

Schervish proves that under appropriate conditions the posterior density of

$$\sqrt{-L_n''(\hat{\theta}_n)}(\theta - \hat{\theta}_n)$$

(a random function of  $\theta$ ) converges in an appropriate sense to the standard normal density. It is another (harder) matter to prove that approximate posterior probabilities for this are in some appropriate sense close to real posterior probabilities.

## 9 Optimality in Finite Sample Point Estimation

### 9.1 Unbiasedness

**Definition 185** The **bias** of an estimator  $\delta$  for  $\gamma(\theta) \in \mathbb{R}$  is

$$E_{\theta}(\delta(X) - \gamma(\theta)) = \int_{\mathcal{X}} (\delta(x) - \gamma(\theta)) dP_{\theta}(x)$$

**Definition 186** An estimator  $\delta$  is **unbiased** for  $\gamma(\theta) \in \mathbb{R}$  provided

$$E_{\theta}(\delta(X) - \gamma(\theta)) = 0 \quad \forall \theta \in \Theta$$

The following suggests that often, no unbiased estimator can be much good.

**Theorem 187** If  $L(\theta, a) = w(\theta)(\gamma(\theta) - a)^2$  for  $w(\theta) > 0 \quad \forall \theta$  and  $\delta$  is both Bayes with respect to  $G$  with finite risk function and unbiased for  $\gamma(\theta)$ , then (according to the joint distribution of  $(X, \theta)$ )

$$\delta(X) = \gamma(\theta) \quad \text{a.s.}$$

**Corollary 188** Suppose  $\Theta \subset \mathbb{R}$  is nontrivial (has at least 2 elements) and  $\{P_{\theta}\}$  has elements that are mutually absolutely continuous. Then if  $L(\theta, a) = (\theta - a)^2$ , no Bayes estimator with finite variance can be unbiased.

**Definition 189**  $\delta$  is **best unbiased** for  $\gamma(\theta) \in \mathbb{R}$  if it is unbiased and is at least as good as any other unbiased estimator.

**Theorem 190** Suppose that  $\mathcal{A} \subset \mathbb{R}$  is convex,  $L(\theta, a)$  is convex in  $a \quad \forall \theta$  and  $\delta_1$  and  $\delta_2$  are two best unbiased estimators of  $\gamma(\theta)$ . Then for any  $\theta$  for which  $L(\theta, a)$  is strictly convex in  $a$  and  $R(\theta, \delta_1) < \infty$

$$P_{\theta}[\delta_1(X) = \delta_2(X)] = 1$$

The special case of SELE of  $\gamma(\theta)$  is of most importance. For this case  $E_{\theta}\delta^2(X) < \infty$  and  $\delta$  unbiased for  $\gamma(\theta)$  implies that

$$R(\theta, \delta) = \text{Var}_{\theta}\delta(X)$$

**Definition 191**  $\delta$  is called a **UMVUE** (uniformly minimum variance unbiased estimator) of  $\gamma(\theta) \in \mathbb{R}$  if it is unbiased and

$$\text{Var}_{\theta}\delta(X) \leq \text{Var}_{\theta}\delta'(X) \quad \forall \theta$$

for  $\delta'$  any other unbiased estimator of  $\gamma(\theta)$ .

The primary tool of UMVUE theory is the famous Lehmann-Scheffé Theorem.

**Theorem 192** (Lehmann-Scheffé) Suppose that  $\mathcal{A} \subset \mathbb{R}$  is convex,  $T$  is sufficient for  $\mathcal{P} = \{P_\theta\}$ , and  $\delta$  is an unbiased estimator of  $\gamma(\theta)$ . Let

$$\delta_0 \equiv E[\delta|T] = \phi \circ T$$

(the latter representation guaranteed by Lehmann's Theorem). Then  $\delta_0$  is an unbiased estimator of  $\gamma(\theta)$ . If in addition,  $T$  is complete for  $\mathcal{P}$

1. if  $\phi' \circ T$  is unbiased for  $\gamma(\theta)$ , then  $\phi' \circ T = \phi \circ T$  a.s.  $P_\theta \forall \theta$ , and
2.  $L(\theta, a)$  convex in  $a \forall \theta$  implies that

- (a)  $\delta_0$  is best unbiased, and
- (b) if  $\delta'$  is any other best unbiased estimator of  $\gamma(\theta)$

$$\delta' = \delta_0 \quad \text{a.s. } P_\theta$$

for any  $\theta$  for which  $R(\theta, \delta_0) < \infty$  and  $L(\theta, a)$  is strictly convex in  $a$ .

## 9.2 "Information" Inequalities

The topic of discussion here is inequalities about variances of estimators, some of which involve "information" measures. The basic tool is the Cauchy-Schwarz inequality.

**Lemma 193** (Cauchy-Schwarz) If  $U$  and  $V$  are random variables with  $EU^2 < \infty$  and  $EV^2 < \infty$ , then

$$EU^2 \cdot EV^2 \geq (EUV)^2$$

**Corollary 194** If  $X$  and  $Y$  have finite second moments

$$(\text{Cov}(X, Y))^2 \leq \text{Var}X \cdot \text{Var}Y$$

(so that  $|\text{Corr}(X, Y)| \leq 1$ ).

**Corollary 195** Suppose that  $\delta(Y)$  takes values in  $\mathbb{R}$  and  $E\delta^2(Y) < \infty$ . If  $g$  is a measurable function such that  $Eg(Y) = 0$  and  $0 < Eg^2(Y) < \infty$ , defining  $C \equiv E\delta(Y)g(Y)$

$$\text{Var}\delta(Y) \geq \frac{C^2}{Eg^2(Y)}$$

A more statistical-looking version of Corollary 195 is next.

**Corollary 196** Suppose that  $\delta(X)$  takes values in  $\mathbb{R}$  and  $E_\theta\delta^2(X) < \infty$ . If  $g_\theta$  is a measurable function such that  $E_\theta g_\theta(X) = 0$  and  $0 < E_\theta g_\theta^2(X) < \infty$ , defining  $C(\theta) \equiv E_\theta\delta(X)g_\theta(X)$

$$\text{Var}_\theta\delta(X) \geq \frac{C^2(\theta)}{E_\theta g_\theta^2(X)}$$

Corollary 196 is the source of several interesting inequalities as one chooses different interesting  $g_\theta$ 's. The most famous derives from the choice

$$g_\theta(x) = \frac{d}{d\theta} \ln f_\theta(x)$$

and is stated next.

**Theorem 197 (Cramér-Rao)** *Suppose that the model  $\mathcal{P} = \{P_\theta\}$  (for  $\Theta \subset \mathbb{R}$ ) is FI regular at  $\theta_0$  and that  $0 < I(\theta_0) < \infty$ . If  $E_\theta \delta(X) = \int \delta(x) f_\theta(x) dP_\theta(x)$  can be differentiated under the integral sign at  $\theta_0$ , i.e. if*

$$\left. \frac{d}{d\theta} E_\theta \delta(X) \right|_{\theta=\theta_0} = \int \delta(x) \left. \frac{d}{d\theta} f_\theta(x) \right|_{\theta=\theta_0} d\mu(x)$$

then

$$\text{Var}_{\theta_0} \delta(X) \geq \frac{\left( \left. \frac{d}{d\theta} E_\theta \delta(X) \right|_{\theta=\theta_0} \right)^2}{I(\theta_0)}$$

Then, consider the choice

$$g_\theta(x) = \frac{f_\theta(x) - f_{\theta'}(x)}{f_\theta(x)}$$

Provided  $P_{\theta'} \ll P_\theta$ , Corollary 196 implies that

$$\text{Var}_\theta \delta(X) \geq \frac{(E_\theta \delta(X) - E_{\theta'} \delta(X))^2}{E_\theta \left( \frac{f_\theta(X) - f_{\theta'}(X)}{f_\theta(X)} \right)^2}$$

This produces the so-called Chapman-Robbins inequality.

**Theorem 198 (Chapman-Robbins)**

$$\text{Var}_\theta \delta(X) \geq \sup_{\theta' \text{ s.t. } P_{\theta'} \ll P_\theta} \frac{(E_\theta \delta(X) - E_{\theta'} \delta(X))^2}{E_\theta \left( \frac{f_\theta(X) - f_{\theta'}(X)}{f_\theta(X)} \right)^2}$$

## 10 Optimality in Finite Sample Testing

We consider the classical hypothesis testing problem. Write  $\Theta = \Theta_0 \cup \Theta_1$  where  $\Theta_0 \cap \Theta_1 = \emptyset$  and we must decide between  $H_0: \theta \in \Theta_0$  and  $H_a: \theta \in \Theta_1$  based on  $X \sim P_\theta$ . This can be thought of in decision theoretic terms as a two-action decision problem, i.e. where  $\mathcal{A} = \{0, 1\}$ . One possible loss for such an approach is (0-1 loss)

$$L(\theta, a) = I[\theta \in \Theta_0] I[a = 1] + I[\theta \in \Theta_1] I[a = 0]$$

While a completely decision-theoretic approach is possible, the subject is older than formal decision theory and has its own set of emphases and peculiar terminology.

**Definition 199** A non-randomized decision function  $\delta : \mathcal{X} \rightarrow \mathcal{A} = \{0, 1\}$  is called an *hypothesis test*.

**Definition 200** The *rejection region* for a non-randomized test is  $\{x | \delta(x) = 1\}$ .

**Definition 201** A *Type 1 error* in a testing problem is taking action 1 when  $\theta \in \Theta_0$ . A *Type 2 error* is taking action 0 when  $\theta \in \Theta_1$ .

In a two-action decision problem a behavioral decision rule is for each  $x$  a distribution over  $\mathcal{A} = \{0, 1\}$ , which is clearly characterized by  $\phi_x(\{1\}) \in [0, 1]$ . In hypothesis testing terminology one has

**Definition 202** A *randomized test* is a function  $\phi : \mathcal{X} \rightarrow [0, 1]$  with the interpretation that if  $\phi(x) = \pi$ , one rejects  $H_0$  with probability  $\pi$ .

For 0-1 loss, for  $\theta \in \Theta_0$

$$R(\theta, \phi) = E_{\theta} I[a = 1] = E_{\theta} \phi(X)$$

while for  $\theta \in \Theta_1$

$$R(\theta, \phi) = E_{\theta} I[a = 0] = 1 - E_{\theta} \phi(X)$$

(the  $a$  the action randomly chosen by  $\phi$ ) and it is thus clear that the function of  $\theta$ ,  $E_{\theta} \phi(X)$  will be of central importance to the theory of hypothesis testing.

**Definition 203** The *power function* of a (possibly randomized) test  $\phi$  is

$$\beta_{\phi}(\theta) = E_{\theta} \phi(X) = \int \phi(x) dP_{\theta}(x)$$

**Definition 204** The *size or level* of the test  $\phi$  is  $\sup_{\theta \in \Theta_0} \beta_{\phi}(\theta)$ .

## 10.1 Simple versus Simple Testing

Consider the case of testing where  $\Theta_0 = \{\theta_0\}$  and  $\Theta_1 = \{\theta_1\}$ . We may for each test  $\phi$  define the (0-1 loss) risk vector

$$(R(\theta_0, \phi), R(\theta_1, \phi)) = (\beta_{\phi}(\theta_0), 1 - \beta_{\phi}(\theta_1))$$

and consider the standard decision theoretic two-dimensional risk set  $\mathcal{S}$ . Fairly obviously, a completely equivalent device is to instead consider the set  $\mathcal{V}$  of points

$$(\beta_{\phi}(\theta_0), \beta_{\phi}(\theta_1))$$

**Lemma 205** For a simple versus simple testing problem with

$$\mathcal{S} = \{(\beta_{\phi}(\theta_0), 1 - \beta_{\phi}(\theta_1)) | \phi \text{ is a test}\} \text{ and } \mathcal{V} = \{(\beta_{\phi}(\theta_0), \beta_{\phi}(\theta_1)) | \phi \text{ is a test}\}$$

1. both  $\mathcal{S}$  and  $\mathcal{V}$  are convex,
2. both  $\mathcal{S}$  and  $\mathcal{V}$  are symmetric with respect to the point  $(\frac{1}{2}, \frac{1}{2})$ ,

3.  $(0, 1) \in \mathcal{S}$ ,  $(1, 0) \in \mathcal{S}$ , and  $(0, 0) \in \mathcal{V}$ ,  $(1, 1) \in \mathcal{V}$ , and
4. both  $\mathcal{S}$  and  $\mathcal{V}$  are closed.

**Definition 206** For testing  $H_0:\theta = \theta_0$  versus  $H_a:\theta = \theta_1$  a test is called **most powerful of size  $\alpha$**  if it is of size  $\alpha$  and for any other test  $\phi^*$  with  $\beta_{\phi^*}(\theta_0) \leq \alpha$ ,  $\beta_{\phi^*}(\theta_1) \leq \beta_{\phi}(\theta_1)$

The fundamental theorem of all testing theory is the Neyman-Pearson Lemma. It has implications that reach far beyond the present case of simple versus simple testing, but is stated in the present context.

**Theorem 207 (Neyman-Pearson)** Suppose that  $\mathcal{P} = \{P_{\theta_0}, P_{\theta_1}\} \ll \mu$  a sigma-finite measure and let  $f_i = \frac{dP_{\theta_i}}{d\mu}$ .

1. (Sufficiency) Any test of the form

$$\phi(x) = \begin{cases} 1 & \text{if } f_1(x) > kf_0(x) \\ \nu(x) & \text{if } f_1(x) = kf_0(x) \\ 0 & \text{if } f_1(x) < kf_0(x) \end{cases} \quad (*)$$

for some  $k \in [0, \infty)$  and  $\nu(x) : \mathcal{X} \rightarrow [0, 1]$  is most powerful of its size. Further (corresponding to the  $k = \infty$  case), the test

$$\phi(x) = \begin{cases} 1 & \text{if } f_0(x) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (**)$$

is most powerful of its size  $\alpha = 0$ .

2. (Existence) For every  $\alpha \in [0, 1] \exists$  a test of the form (\*) with  $\nu(x) = \nu$  (a constant in  $[0, 1]$ ) or a test of form (\*\*) with  $\beta_{\phi}(\theta_0) = \alpha$ .
3. (Uniqueness) If  $\phi$  is a most powerful test of size  $\alpha$ , then it is of form (\*) or (\*\*) except possibly for  $x \in N$  with  $P_{\theta_0}(N) = P_{\theta_1}(N) = 0$ .