

THE FLORIDA STATE UNIVERSITY  
COLLEGE OF ARTS AND SCIENCES

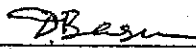
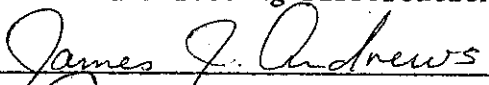
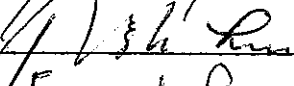
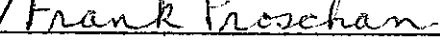
BAYESIAN SOLUTIONS TO SOME CLASSICAL  
PROBLEMS OF STATISTICS

by

CARLOS ALBERTO DE BRAGANCA PEREIRA

A Dissertation submitted to the  
Department of Statistics  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

Approved:

  
\_\_\_\_\_  
Professor Directing Dissertation  
  
\_\_\_\_\_  
  
\_\_\_\_\_  
  
\_\_\_\_\_

December, 1980

BAYESIAN SOLUTIONS TO SOME CLASSICAL  
PROBLEMS OF STATISTICS

(Publication No. \_\_\_\_\_)

Carlos Alberto de Bragança Pereira, Ph.D.

The Florida State University, 1980

Major Professor: Debrabrata Basu, Ph.D.

Three of the basic questions of Statistics may be stated as follows:

A. Which portion of the data  $X$  is actually informative about the parameter of interest  $\theta$ ?

B. How can all the relevant information about  $\theta$  provided by the data  $X$  be extracted?

C. What kind of information about  $\theta$  do the data  $X$  possess?

The perspective of this dissertation is that of a Bayesian.

Chapter I is essentially concerned with question A. The theory of conditional independence is explained and the relations between ancillarity, sufficiency, and statistical independence are discussed in depth. Some related concepts like specific sufficiency, bounded completeness, and splitting sets are also studied in some details. The language of conditional independence is used in the remaining Chapters.

Chapter II deals with question B for the particular problem of analysing categorical data with missing entries. It is demonstrated how a suitably chosen prior for the frequency parameters can streamline the analysis in the presence of missing entries due to non-response or other causes. The two cases where the data follow the Multinomial or the Multivariate Hypergeometric model are treated separately. In the first case it is adequate to restrict the prior (for the cell probabilities) to the class of Dirichlet distributions. In the Hypergeometric case it is convenient to select a prior (for the cell population frequencies) from the class of Dirichlet-Multinomial (DM) distributions. The DM distributions are studied in detail.

Chapter III is directly related to question C. Conditions on the likelihood function and on the prior distribution are presented in order to assess the effect of the sample on the posterior distribution. More specifically, it is shown that under certain conditions, the larger the observations obtained, the larger (stochastically in terms of the posterior distribution) is the appropriate parameter.

Finally, Chapter IV deals with the characterization of distributions in terms of Blackwell comparison of experiments. It is shown that a result (for the Hypergeometric model) obtained in Chapter II is actually a consequence of a property of complete families of distributions.

REFERENCES

	Page
ABSTRACT . . . . .	ii
ACKNOWLEDGMENTS . . . . .	vi
CHAPTER I - CONDITIONAL INDEPENDENCE IN STATISTICS . . . . .	1
1 - INTRODUCTION . . . . .	1
2 - NOTATION AND PRELIMINARIES . . . . .	5
3 - DEFINITION OF CONDITIONAL INDEPENDENCE . . . . .	9
4 - THE DROP/ADD PRINCIPLES AND OTHER PROPERTIES OF CONDITIONAL INDEPENDENCE . . . . .	12
5 - MARKOV CHAINS AND BAYESIAN INFERENCE . . . . .	23
6 - ON MEASURABLE SEPARABILITY OF RANDOM OBJECTS . . . . .	32
7 - BASU THEOREM . . . . .	36
REFERENCES . . . . .	41
CHAPTER II - ON THE BAYESIAN ANALYSIS OF CATEGORICAL DATA: THE PROBLEM OF NONRESPONSE . . . . .	44
1 - INTRODUCTION . . . . .	44
2 - NONRESPONSE: THE MULTINOMIAL MODEL . . . . .	49
3 - THE DIRICHLET-MULTINOMIAL DISTRIBUTION: PROPERTIES . . . . .	55
4 - THE DM DISTRIBUTION: A NATURAL FAMILY OF PRIORS FOR FINITE POPULATION STUDIES . . . . .	61
5 - NONRESPONSE: THE MULTIVARIATE HYPERGEOMETRIC MODEL . . . . .	64
6 - FINAL REMARKS . . . . .	69

	Page
REFERENCES . . . . .	71
APPENDIX . . . . .	72
CHAPTER III - THE INFLUENCE OF THE SAMPLE ON THE POSTERIOR DISTRIBUTION . . . . .	74
1 - INTRODUCTION . . . . .	74
2 - PRELIMINARIES . . . . .	75
3 - THEORETICAL RESULTS . . . . .	78
4 - APPLICATIONS . . . . .	85
5 - ACKNOWLEDGMENTS . . . . .	90
REFERENCES . . . . .	91
CHAPTER IV - ON THE CHARACTERIZATION OF DISTRIBUTIONS IN TERMS OF SUFFICIENCY AND COMPLETENESS . . . . .	92
1 - INTRODUCTION . . . . .	92
2 - CHARACTERIZATION OF THE HYPERGEOMETRIC MODELS . . . . .	94
3 - CHARACTERIZATION OF OTHER DISTRIBUTIONS . . . . .	98
REFERENCES . . . . .	106
VITA . . . . .	107

## ACKNOWLEDGMENTS

I am very grateful to my major professor, Dr. D. Basu, for his guidance and encouragement prior to and throughout the preparation of this dissertation. His investments of time and energy made this research possible.

Special appreciation goes to Dr. F. Proschan who painstakingly read this work and offered his corrections. His expert assistance was vital in Chapter III. The other members of my committee, Drs. P. E. Lin and J. J. Andrews, receive thanks for their willing service.

I also wish to acknowledge Universidade de Sao Paulo and CAPES for providing the financial support. Mrs. Kathy Strickland deserves my thanks for the precious typing job.

I am thankful to my Florida State colleagues, especially Wai and Tiwari, for the stimulating discussions with me. The moral support of my Brazilian friends, Julio, Severo, and Wagner, will never be forgotten.

Finally, my deepest gratitude goes to my parents, my wife, and my son for their love and encouragement throughout these years in graduate school.

## CHAPTER I. CONDITIONAL INDEPENDENCE IN STATISTICS

### 1 - INTRODUCTION

The notion of conditional independence is a central theme of Statistics. In a series of recent articles A. P. Dawid (1979a, b, 1980), J. P. Florens and M. Mouchart (1977), and M. Mouchart and J. M. Rolin (1978) have explained at length the grammar of conditional independence as a language of statistics. This chapter is a further elucidation on the subject and is generally of an expository nature. Several results that have already appeared elsewhere are amplified and their proofs simplified and unified. The only mathematical tool that is repeatedly used is that of conditioning operator. The language of conditional independence developed in this chapter will be fully utilized in Chapter 2.

The statistical perspective of this dissertation is that of a Bayesian. A problem begins with a parameter (state of nature)  $\theta$  with its prior probability model  $(\Theta, \mathcal{B}, \xi)$  that exists only in the mind of the investigator. There is an observable  $X$  with an associated statistical model  $(X, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$ . Writing  $\omega = (\theta, X)$ ,  $(\Omega, \mathcal{F}) = (\Theta \times X, \mathcal{B} \times \mathcal{A})$ , and  $\Pi$  for the joint distribution of  $(\theta, X)$ , there exists then a subjective probability model  $(\Omega, \mathcal{F}, \Pi)$  for  $\tilde{\omega}$ . Hidden behind the wings of the Bayesian probability model  $(\Omega, \mathcal{F}, \Pi)$  are the four models:

- (i) The prior model  $(\Theta, \mathcal{B}, \xi)$ ,
  - (ii) the statistical model  $(X, \mathcal{A}, \{P_\theta: \theta \in \Theta\})$
  - (iii) the posterior model  $(\Theta, \mathcal{B}, \{\xi_x: x \in X\})$ ,
- and (iv) the predictive model  $(X, \mathcal{A}, P)$ , where  $P$  is the marginal or predictive distribution of  $X$ .

In classical probability theory, the notion of conditional independence appears in a rather indirect fashion in the study of Markov chains and processes. A sequence of three random entities  $(X, Y, Z)$  is said to possess the Markov property if, given  $X$  and  $Y$ , the conditional distribution of  $Z$  depends on  $(X, Y)$  only through  $Y$ . An equivalent characterization of the Markov property may be stated in the symmetric form:  $X$  and  $Z$  are conditionally independent given  $Y$ . In Section 3 we make precise these two definitions of conditional independence in terms of the conditioning operator.

In Statistics the phenomenon of conditional independence manifests itself in a much more direct and natural fashion. The statistical model that is most commonly in use is that of a sequence  $\underline{X} = (X_1, X_2, \dots)$  of observables that are independently and identically distributed (i.i.d.) for each given value of  $\theta$ . It was DeFinetti (1937) who emphasized that, in view of the fact that  $\theta$  is not fully known, it is appropriate to regard the sequence of  $X_i$ 's not as i.i.d. random variables but as an exchangeable process. The fact that the  $X_i$ 's are conditionally i.i.d. implies that they are positively dependent - if we consider the (predictive) conditional distributions,



$X_2$  is stochastically increasing with  $X_1$ ,  $X_3$  is stochastically increasing with  $(X_1, X_2)$ , and so on. (See Chapter 3 for details in some concepts of dependence.)

Consider for example the particular case where  $X_1, \dots, X_n$  are i.i.d. with common distribution  $N(\mu, \sigma^2)$  with  $\theta = (\mu, \sigma^2)$  not fully known. In almost every textbook on Statistics it is proved that the statistic  $\bar{X} = n^{-1} \sum X_i$  is stochastically independent of  $S^2 = n^{-1} \sum (X_i - \bar{X})^2$ . Does it mean that  $\bar{X}$ , when observed, carries no information about  $S^2$ ? That the answer cannot be "yes" is easily seen as follows. Suppose that the sample size  $n = 25$  and that our partial knowledge about  $\theta = (\mu, \sigma^2)$  is as follows:  $\mu = 0$  or  $1$  and  $\sigma^2 = 1$  or  $100$  (that is,  $\theta = \{(0, 1), (1, 1), (0, 100), (1, 100)\}$ ). Suppose now that  $\bar{X}$  is observed and is equal to  $2.1$ . This observation generates the four likelihoods  $L(0, 1)$ ,  $L(1, 1)$ ,  $L(0, 100)$ , and  $L(1, 100)$  where  $L(0, 1) = \frac{5}{\sqrt{2\pi}} \exp\{-\frac{25}{2} (2.1)^2\}$  and so on. The relative likelihoods work out roughly as  $10^{-17}$ ,  $1$ ,  $2(10)^5$ , and  $3(10)^5$  respectively. Thus, it is intuitive that the observation  $\bar{X} = 2.1$  almost categorically rules out the points  $(0, 1)$  and  $(1, 1)$ . Therefore, the observation of  $\bar{X} = 2.1$  asserts that  $\sigma^2 = 100$  with a lot of emphasis and so we may conclude that  $S^2$  is of the order of  $100$ . Then  $\bar{X}$  and  $S^2$ , even though they are conditionally independent given  $\theta$ , are in effect highly dependent.

The three entities  $\theta = (\mu, \sigma^2)$ ,  $T = (\bar{X}, S^2)$ , and  $\underline{X} = (X_1, \dots, X_n)$ , in this order, have the Markov property in the sense that, given  $\theta$

and  $T$ , the conditional distribution of  $\underline{X}$  depends on  $(\theta, T)$  only through  $T$ . This is the sufficiency property of the statistic  $T$  as recognized by R. A. Fisher (1920, 1922). A. N. Kolmogorov (1942) gave a Bayesian characterization of the notion of sufficiency by noting that irrespective of the choice of the prior distribution  $\xi$  for the parameter  $\theta$ , the posterior distribution  $\xi_{\underline{X}}$  of  $\theta$  depends on  $\underline{X}$  only through  $T$ . In other words, the sequence  $(\underline{X}, T, \theta)$  have the Markov property; that is  $\underline{X}$  and  $\theta$  are conditionally independent given  $T$ . Note that the Fisher characterization of sufficiency is made only in terms of the statistical model for  $\underline{X}$  whereas the Kolmogorov characterization is made in terms of a large family of Bayesian models  $(\Omega, F, \Pi)$  for  $\omega = (\theta, \underline{X})$ . (See Basu (1977) and Cheng (1978) for further details on these characterizations.)

Fisher regarded a sufficient statistic  $T$  as one that summarizes in itself all the available relevant information in the sample  $\underline{X}$  about the parameter  $\theta$ . He called a statistic  $Y = Y(\underline{X})$  ancillary if the conditional distribution of  $Y$  given  $\theta$ , does not involve  $\theta$  (is the same for all values of  $\theta$ ). For example, the statistic  $\sum (x_i - \bar{X})^4 / S^4$  is ancillary. In a series of articles D. Basu (1955, 1958, 1959, 1964, 1967) studied the phenomena of sufficiency, ancillarity, and conditional independence from various angles. In these articles, Basu's viewpoint was non-Bayesian in the sense that he did not introduce a prior distribution  $\xi$  for the parameter  $\theta$ .

M. Mouchart and J. M. Rolin (1978) studied in depth the familiar Basu theorems on sufficiency, ancillarity, and conditional independence from the viewpoint of a Bayesian model  $(\Omega, \mathcal{F}, \Pi)$ . In Sections 5, 6, and 7 we review Basu's results from the Bayesian perspective. This is done mainly as an exercise in the use of the language of conditional independence developed earlier.

## 2 - NOTATION AND PRELIMINARIES

Let  $(\Omega, \mathcal{F}, \Pi)$  be the basic probability space. By a "random object"  $X$  we mean a measurable map  $\omega \rightarrow X(\omega)$  of  $(\Omega, \mathcal{F})$  into another measurable space  $(X, \mathcal{A})$ . The sub- $\sigma$ -algebra (to be called subfield) of  $X$ -events  $\{X^{-1}A; A \in \mathcal{A}\}$  will be denoted by  $\mathcal{F}_X$ . The two probability spaces  $(\Omega, \mathcal{F}_X, \Pi)$ , and  $(X, \mathcal{A}, \Pi^{-1})$  are undistinguishable in a sense, and so we shall, as a rule, identify a random object  $X$  with the induced subfield  $\mathcal{F}_X$  of  $\mathcal{F}$ . In that way, one could say that random objects are generators of subfields. Examples of random objects include random variables, random vectors, and any collection of random variables (stochastic processes).

For any two subfields  $\mathcal{F}'$  and  $\mathcal{F}''$  of  $\mathcal{F}$ ,  $\mathcal{F}' \vee \mathcal{F}''$  denotes the smallest subfield of  $\mathcal{F}$  that contains both  $\mathcal{F}'$  and  $\mathcal{F}''$ . The smallest subfield that contains all null sets of  $\mathcal{F}$  (a set  $N$  is null if  $\Pi(N) = 0$ ) is denoted by  $\overline{\mathcal{F}}_0$ ; that is,

$$\overline{\mathcal{F}}_0 = \{F; F \in \mathcal{F} \text{ and } \Pi(F) = 0 \text{ or } \Pi(F) = 1\}$$

and write  $F_0 = \{\phi, \Omega\}$ , the trivial subfield.

A subfield of  $F$  is said to be completed if it contains  $\bar{F}_0$ .

For any subfield  $F'$  of  $F$  its completion is defined by:

$$\bar{F}' = F' \vee \bar{F}_0.$$

For a random object  $X$ , the notation  $X \in F'$  indicates that  $F_X \subset \bar{F}'$  and  $X$  is said to be essentially  $F'$ -measurable or  $X$  is ess- $F'$ -measurable. A random variable is a random object with range  $(R_1, B_1)$  where  $R_1$  is the real line and  $B_1$  is the Borel  $\sigma$ -algebra. A random variable  $f$  is said to be bounded if  $\exists a \in R_1$  such that  $\Pi\{\omega; |f(\omega)| \leq a\} = 1$ . In the sequel, all random variables shall be regarded as bounded unless stated otherwise and the use of small letters shall be restricted to their representation. The notation  $f \subset X$  indicates that the random variable  $f$  is ess- $F_X$ -measurable. In the same spirit, for two random objects  $X$  and  $Y$ , we write  $X \subset Y$  to indicate that  $\bar{F}_X \subset \bar{F}_Y$ . If  $\bar{F}_X = \bar{F}_Y$  we write  $X \equiv Y$  to indicate the essential equivalence between  $X$  and  $Y$ . The class of all bounded random variables on  $(\Omega, F, \Pi)$  is denoted by  $L_\infty$  and  $L_\infty(X)$  denotes the class of all ess- $F_X$ -measurable random variables. Here and for the rest of this chapter, equality of two random variables means essential equality; that is,  $f = g$  means that  $\{\omega; f(\omega) \neq g(\omega)\}$  is a null set.

DEFINITION 1

The conditional expectation of  $f$ , given a random object  $X$ , is a random variable  $f^{*X} \in L_{\infty}(X)$  such that

$$\int fg d\Pi = \int f^{*X} g d\Pi \quad \forall g \in L_{\infty}(X).$$

Another notation for  $f^{*X}$  is  $E\{f|X\}$ . When the conditioning random object  $X$  is implicit in the context,  $f^*$  is substituted for  $f^{*X}$ .

The map  $f \rightarrow f^*$  of  $L_{\infty}$  to  $L_{\infty}(X)$  is linear, constant preserving, monotone, idempotent, and is a contraction in the  $L_p$  norm if  $p \geq 1$ .

The following proposition, known as smoothing theorem, is widely used in this chapter. Here,  $*$  is substituted for  $*X$  and  $\dagger$  is substituted for  $*Y$ .

PROPOSITION 1

If two random objects  $X$  and  $Y$  are such that  $X \subset Y$ , then  
 $\forall f \in L_{\infty}$

- (i)  $E\{f^*|Y\} = (f^*)^{\dagger} = f^*$
- (ii)  $E\{f^{\dagger}|X\} = (f^{\dagger})^* = f^*$
- (iii)  $f^{\dagger} \subset X \rightarrow f^{\dagger} = f^*$

The following result which is a restatement of the property of self-adjointness of the  $*$ -operator will be repeatedly used in the sequel.

PROPOSITION 2

If  $f \in L_{\infty}$ ,  $g \in L_{\infty}$ , and  $h \in L_{\infty}(X)$ , then

$$E\{f*gh\} = E\{fg*h\} = E\{f*g*h\}. \quad (* \text{ stands for } *X.)$$

The proof follows from the observation that  $(f*gh)^* = f*g*h$  and that  $E\{f\} = E\{f^*\}$  for every  $f \in L_\infty$ .

This proposition together with the fact that the  $*$ -operator is idempotent (that is,  $(f^*)^* = f^*$ ) implies that the  $*$ -operator is a projection of  $L_\infty$  in  $L_\infty(X)$  when the  $L_2$  - norm is considered.

Given two random objects  $X$  and  $Y$ , the random object  $(X, Y): \Omega \rightarrow X \times Y$  generates the subfield  $F_X \vee F_Y$  and may be identified with its completion; that is,  $\overline{F}_{(X,Y)} = \overline{F_X \vee F_Y}$ . A random object that essentially generates the subfield  $\overline{F}_X \cap \overline{F}_Y$  will be denoted in this chapter by  $X \wedge Y$  despite the fact that it does not have a neat representation in terms of  $X$  and  $Y$  as in the case of  $(X, Y)$ .

#### REMARK

Given any two subfields  $F'$  and  $F''$  of  $F$ , the following are well known relations among completed subfields:

$$(i) \quad \overline{F' \vee F''} = \overline{F'} \vee \overline{F''} = \overline{F'} \vee \overline{F''}.$$

$$(ii) \quad \overline{F' \cap F''} \subset \overline{F'} \cap \overline{F''} = \overline{F' \cap F''}.$$

The following definition and theorem due to Dynkin are of great importance. They enable us to present simple proofs of some of the results stated in the sequel.

#### DEFINITION 2

Let  $\mathcal{D}$  be a class of subsets of  $\Omega$ .  $\mathcal{D}$  is said to be a D-system (D for Dynkin) if the following conditions hold:

- (i)  $\Omega \in \mathcal{D}$ .
- (ii) If  $B, A \in \mathcal{D}$ ,  $B \subset A$  then  $A - B \in \mathcal{D}$ .
- (iii) If  $A_1, A_2, \dots \in \mathcal{D}$  and  $A_n \uparrow A$  then  $A \in \mathcal{D}$ .

### THEOREM 1

Let  $C$  be a class of subsets of  $\Omega$  and assume that  $C$  is closed under finite intersections. If  $\mathcal{D}$  is a  $\mathcal{D}$ -system such that  $C \subset \mathcal{D}$  then  $\sigma(C) \subset \mathcal{D}$ . ( $\sigma(C)$  is the smallest  $\sigma$ -field that contains  $C$ .)

For a proof of this result we refer to Ash (1972) pp. 168-169. For applications see Basu (1967).

In the next section we discuss the concept of conditional independence.

### 3 - DEFINITION OF CONDITIONAL INDEPENDENCE

In this section, the two most popular definitions of conditional independence (c.i.) are discussed. They are called here Intuitive and Symmetric. A simple proof of the equivalence between them is presented. Further characterization of the concept of c.i. will be presented in Section 4.

Three random objects  $X$ ,  $Y$ , and  $Z$  are being considered and, in this section,  $*$  stands for the  $*Z$ -operator.

#### DEFINITION 3 - (Intuitive)

The random objects  $X$  and  $Y$  are conditionally independent given  $Z$  (in symbols  $X \perp\!\!\!\perp Y|Z$ ) if for any  $f \in L_\infty(X)$

$$E\{f|(Y, Z)\} = f^*(Y, Z) = f^*.$$

Note that if  $X$ ,  $Y$ , and  $Z$  are random variables, then to say that  $X \perp\!\!\!\perp Y|Z$  is equivalent to say that  $X|(Y, Z)$  has the same conditional distribution as does  $X|Z$ . This is the intuition behind Definition 3. Frequently we will use the notation  $X|(Y, Z) \sim X|Z$  for  $X \perp\!\!\!\perp Y|Z$ .

An equivalent way to define c.i. is to say that the map  $f \rightarrow f^*(Y, Z)$  from  $L_\infty$  to  $L_\infty(Y, Z)$  has its range restricted to  $L_\infty(Z)$ . Particularly, if  $Z$  is essentially a generator of  $F_0$  (the trivial subfield), then  $(Y, Z) \subset Y$  and the usual concept of independence is attained since  $L_\infty(Z)$  becomes the class of all essentially constant functions. In this case the notation is  $X \perp\!\!\!\perp Y$ .

#### DEFINITION 3a (SYMMETRIC)

The random objects  $X$  and  $Y$  are conditionally independent given  $Z$  if for any  $f \in L_\infty(X)$  and  $g \in L_\infty(Y)$ ,

$$(fg)^* = f^*g^*.$$

The following theorem gives the equivalence of the two definitions showing that  $X \perp\!\!\!\perp Y|Z$  implies  $Y \perp\!\!\!\perp X|Z$  which is not clear by looking only at Definition 3.

#### THEOREM 2

Definitions 3 and 3a are equivalent.



PROOF3  $\rightarrow$  3a

By using Proposition 1 and the linearity of the  $*$ -operator we have:

$$\begin{aligned} (fg)^* &= E\{E\{fg|(Y, Z)\}|Z\} = E\{gE\{f|(Y, Z)\}|Z\} \\ &= E\{gE\{f|Z\}|Z\} = (gf^*)^* = f^*g^* \end{aligned}$$

3a  $\rightarrow$  3

We wish to prove that for any  $f \in X$  and  $g \in Y$ ,  $(fg)^* = f^*g^*$  implies  $E\{f|(Y, Z)\} = f^*$ .

Let  $E$  be a class of subsets defined as  $E = \{E; E \in \bar{F}_Y \vee \bar{F}_Z \text{ and } \int_E fd = \int_E f^*d\pi \vee f \in X\}$ . Clearly  $E$  is a D-system since  $\Omega \in E$ ,  $E$  is a monotone class (by monotone convergence theorem) and for  $A, B \in E$  with  $A \subset B$  we have  $B - A \in E$ .

Now take any two sets  $C$  and  $D$  with  $C \in \bar{F}_Y$  and  $D \in \bar{F}_Z$ . Clearly  $CD \in \bar{F}_Y \vee \bar{F}_Z$  and

$$\int_{CD} fd\pi = E\{I_C I_D f\} = E\{(I_C I_D f)^*\} = E\{I_D (I_C f)^*\}.$$

But by Proposition 2 and by hypothesis, we have

$$E\{I_D (I_C f)^*\} = E\{I_D I_C^* f^*\} = E\{I_D I_C f^*\} = \int_{CD} f^*d\pi.$$

Thus,  $E' \subset E$  where

$$E' = \{CD; C \in \bar{F}_Y \text{ and } D \in \bar{F}_Z\}.$$

Since  $E'$  is closed under finite intersections, and  $\sigma(E') = \bar{F}_Y \vee \bar{F}_Z$  we conclude, by Theorem 1, that  $\bar{F}_Y \vee \bar{F}_Z \subset E$ ; that is,  $f^* = E\{f|(Y, Z)\} \vee f \subset X$ .  $\square$

An important case of c.i. is  $X \perp\!\!\!\perp Y|X'$  where  $X' \subset X$ . Note that the meaning of this relation is better understood when stated as

$$\forall g \subset Y, E\{g|X\} = E\{g|X'\}$$

since  $X \equiv (X, X')$ . In Bayesian inference, if  $X$  represents the sample, and  $Y$  the parameter then  $X'$  is said to be sufficient for  $X$ .

Some applications of the concept of c.i. are presented in the sequel and emphasis is given to the Bayesian framework.

#### 4 - THE DROP/ADD PRINCIPLES AND OTHER PROPERTIES OF CONDITIONAL INDEPENDENCE

The concept of c.i. gives rise to many questions. Among them are questions involving the DROP and ADD (DROP/ADD) principles. Suppose that  $X, Y, Z, W, X_1$ , and  $Z_1$  are random objects such that  $X \perp\!\!\!\perp Y|Z, X_1 \subset X$ , and  $Z_1 \subset Z$ . What can be said about the relation  $\perp\!\!\!\perp$  if  $X_1$  is substituted for  $X$ ,  $Z_1$  for  $Z$ ,  $(Y, W)$  for  $Y$ , or  $(Z, W)$  for  $Z$ ? In other words, can  $F_X, F_Y$ , or  $F_Z$  be essentially reduced or enlarged without destroying the c.i. relation? In general, the answer is no. However, for certain kinds of reductions and enlargements, the relationship will be preserved. To indicate that the relation  $\perp\!\!\!\perp$  does not hold we write  $\not\perp\!\!\!\perp$ .

The following simple examples show that arbitrary enlargements of  $F_X$ ,  $F_Y$ , or  $F_Z$  may destroy the c.i. property. For a set  $A \subset \Omega$ ,  $I_A(\omega)$  is the indicator function of  $A$ .

EXAMPLE 1

Let  $\Omega = \{1, 2, 3, 4\}$ ,  $F$  be the power set, and  $\Pi\{i\} = 1/4$ . Let  $X = I_{\{1,2\}}$ ,  $Y = I_{\{1,3\}}$ ,  $Z = \text{constant}$ , and  $W = I_{\{1,4\}}$ . Clearly,  $X \perp\!\!\!\perp Y$  and  $X \perp\!\!\!\perp W$  but  $X \not\perp\!\!\!\perp (Y, W)$ .  $\square$

EXAMPLE 2

Let  $\Omega = \{0, 1\} \times \{0, 1\} \times \{0, 1\}$ ,  $F$  be the power set, and for  $i \neq j$  ( $i, j = 0, 1$ )  $\Pi\{(i, i, i)\} = .15$ ,  $\Pi\{(i, i, j)\} = .10$  and  $\Pi\{(i, j, i)\} = .25$ . If  $X, Y, Z$ , and  $W$  are such that  $X(x, y, w) = x$ ,  $Y(x, y, w) = y$ ,  $W(x, y, w) = w$ , and  $Z$  is a constant in  $\Omega$ , then  $X \perp\!\!\!\perp Y$  and  $X \perp\!\!\!\perp Y|W$ . This is clear since we obtain the following probability functions (p.f.):

		Y		
		0	1	
X	0	.3	.5	.8
	1	0	.2	.2
		.3	.7	1

p.f. of  $(X, Y)$  given  $W = 0$

		Y		
		0	1	
X	0	.2	0	.2
	1	.5	.3	.8
		.7	.3	1

p.f. of  $(X, Y)$  given  $W = 1$

		Y		
		0	1	
X	0	.25	.25	.5
	1	.25	.25	.5
		.5	.5	1

p.f. of (X, Y).

□

### EXAMPLE 3

Suppose that X and Y are two independent random variables with the same distribution  $N(0, 1)$ . Clearly,  $(X - Y) \perp\!\!\!\perp (X + Y) | Y$ .

However, it is well known that  $(X - Y) \not\perp\!\!\!\perp (X + Y)$ . □

Looking at the problem from the opposite direction, we present the following similar examples which show that arbitrary reductions of the conditioning subfield may destroy the c.i. relation.

### EXAMPLE 4

In Example 2 consider  $\Pi$  as follows:

$$\Pi\{(0, 0, 0)\} = \Pi\{(0, 0, 1)\} = \Pi\{(0, 1, 0)\} = \Pi\{(1, 0, 1)\} = .10, \text{ and}$$

$$\Pi\{(1, 1, 1)\} = \Pi\{(1, 1, 0)\} = \Pi\{(1, 0, 0)\} = \Pi\{(0, 1, 1)\} = .15.$$

Here, we conclude that  $X \perp\!\!\!\perp Y | W$ , but  $X \not\perp\!\!\!\perp Y$ . The probability functions in this case are:

		Y		
		0	1	
X	0	.2	.2	.4
	1	.3	.3	.6
		.5	.5	1

p.f. of (X, Y) given W = 0.

		Y		
		0	1	
X	0	.2	.3	.5
	1	.2	.3	.5
		.4	.6	1

p.f. of (X, Y) given W = 1.

		Y		
		0	1	
X	0	.20	.25	.45
	1	.25	.30	.55
		.45	.55	1

p.f. of (X, Y).

□

EXAMPLE 5

In example 3 consider an additional random variable Z such that  $Z \perp\!\!\!\perp (X - Y)$  and  $Z \perp\!\!\!\perp (X + Y)$ . Obviously,  $(X - Y + Z) \perp\!\!\!\perp (X + Y + Z) | Z$  but  $(X - Y + Z) \not\perp\!\!\!\perp (X + Y + Z)$ . □

Examples 2 to 5 can be viewed as cases of Simpson's paradox (Dawid [1979a]). The paradox, however, is much stronger. For instance, let Z and W be two independent normal variables with zero means. Define  $X = Z + W$  and  $Y = Z - W$ . The correlation between X and Y is given by  $\rho(X, Y) = \frac{1 - \delta}{1 + \delta}$  where  $\delta = \frac{\text{Var}(W)}{\text{Var}(Z)}$ . Given Z, the

conditional correlation  $\rho(X, Y|Z)$  is clearly equal to  $-1$ . On the other hand,  $\delta$  may be taken very small in order to make  $\rho(X, Y)$  close to  $1$ . This shows that we can have a case where  $X$  and  $Y$  are strongly positive (negative) dependent but, when  $Z$  is given,  $X$  and  $Y$  turn to be strongly negative (positive) dependent.

The essence of DROP/ADD principles for conditional independence is contained in the following proposition and corollaries.

PROPOSITION 3

If  $X \perp\!\!\!\perp Y|Z$  then for every  $X' \subset X$  we have:

- (i)  $X' \perp\!\!\!\perp Y|Z$ .
- (ii)  $X \perp\!\!\!\perp Y|(Z, X')$ .

PROOF

(i) Since  $X' \subset X$ ,  $\forall f \in X' \rightarrow f \in X$ . Then, for every  $f \in X'$ , since  $X \perp\!\!\!\perp Y|Z$ ,  $E\{f|(Y, Z)\} = E\{f|Z\}$ .

(ii) Clearly,  $(Z, X', X) \equiv (Z, X)$  then, for every  $g \in Y$ ,

$$E\{g|(Z, X', X)\} = E\{g|(Z, X)\} = E\{g|Z\} = g^*.$$

On the other hand, by Proposition 1,

$$E\{g|(Z, X')\} = E\{E\{g|(Z, X', X)\} | (Z, X')\} = E\{g^* | (Z, X')\} = g^*$$

Thus,  $\forall g \in Y$   $E\{g|(Z, X')\} = E\{g|(Z, X', X)\}$ .  $\square$

COROLLARY 1

For any  $Z' \subset Z$ ,

$$X \perp\!\!\!\perp Y|Z \text{ if and only if } X \perp\!\!\!\perp (Y, Z')|Z.$$

COROLLARY 2

If  $X \perp\!\!\!\perp Y|Z$  then, for any  $W_1 \subset (X, Z)$  and  $W_2 \subset (Y, Z)$ , we have:

$$(i) \quad W_1 \perp\!\!\!\perp W_2|Z$$

$$(ii) \quad X \perp\!\!\!\perp Y|(Z, W_1, W_2).$$

By way of explanation, if  $X \perp\!\!\!\perp Y|Z$  then the relation  $\perp\!\!\!\perp$  is preserved when (i)  $X$  and  $Y$  is increased (ADD) by any essential part of  $Z$ , (ii)  $Z$  is increased (ADD) by any essential part of  $X$  or of  $Y$ , and (iii)  $X$  and  $Y$  are arbitrarily reduced (DROP).

The following interesting result, in one direction, has its version in classical statistics. If  $X_0$  is sufficient for  $X$  then, for every statistic  $f$ , there is a corresponding function  $g$  of  $X_0$  with the same mean of  $f$ .

PROPOSITION 4

Let  $X'$ ,  $X$ , and  $Y$  be three random objects such that  $X' \subset X$ . The following condition is necessary and sufficient to have  $X \perp\!\!\!\perp Y|X'$ :

$$\forall f \subset X, E\{f^*|Y\} = E\{f|Y\}, \text{ where } f^* = E\{f|X'\}.$$

PROOF

Here,  $*$  stands for  $*X'$  and  $\dagger$  for  $*Y$ .

(i) Necessity.

Since  $\forall f \subset X$ ,  $f^* = E\{f|(Y, X')\}$ , by Proposition 1 we conclude that  $\forall f \subset X$ ,  $(f^*)^\dagger = f^\dagger$ .

(ii) Sufficiency.

Let  $f \subset X$ ,  $g \subset Y$ , and  $f' \subset X'$ . Clearly  $ff' \subset X$ . Note that

$$(fgf')^\dagger = g(ff')^\dagger = g(ff')^{*\dagger} = g(f^*f')^\dagger = (f^*gf')^\dagger.$$

Since  $E\{(fgf')^\dagger\} = E\{fgf'\}$ , by Proposition 2 we can write

$$E\{fgf'\} = E\{f^*gf'\} = E\{f^*g^*f'\}$$

Then  $(fg)^* = f^*g^*$ .  $\square$

An equivalent result introduced by Mouchart and Rolin (1978), which is stated below, is a characterization of c.i..

COROLLARY 3

The following condition is necessary and sufficient to have  $X \perp\!\!\!\perp Y|Z$ :

$$\forall f \subset (X, Z), E\{f^*|Y\} = E\{f|Y\}, \text{ where } f^* = E\{f|Z\}.$$

The equivalence of this result with Proposition 4 follows directly from Corollary 1.



that if  $W$  is a sufficient statistic and  $T$  is a statistic "marginally independent" of  $W$  ( $T \perp\!\!\!\perp W$ ), then  $T$  is ancillary ( $T \perp\!\!\!\perp Y$ ) and is "independent" of  $W$  ( $T \perp\!\!\!\perp W|Y$ ).

Now we extend the concept of conditional independence for a set of random objects. Let  $Z$  be a random object,  $\tau$  be a set of indices, and  $\{X_t; t \in \tau\}$  be a collection of random objects.

#### DEFINITION 4

The set  $\{X_t; t \in \tau\}$  is said to be mutually conditionally independent given  $Z$  if, for any partition  $(\tau_1, \tau_2)$  of  $\tau$ , the two random objects  $\{X_t; t \in \tau_1\}$  and  $\{X_t; t \in \tau_2\}$  are conditionally independent given  $Z$ .

For example,  $X_1, X_2$  and  $X_3$  are mutually conditionally independent given  $Z$  if  $X_1 \perp\!\!\!\perp (X_2, X_3)|Z$ ,  $X_2 \perp\!\!\!\perp (X_1, X_3)|Z$ , and  $X_3 \perp\!\!\!\perp (X_1, X_2)|Z$ .

The next result is called here the transfer principle for c.i. It shows that, for finite sets of random objects, to check Definition 4 we do not have to study all partitions.

#### PROPOSITION 6

If  $X_1 \perp\!\!\!\perp X_2|Z$  and  $(X_1, X_2) \perp\!\!\!\perp X_3|Z$ , then  $X_1 \perp\!\!\!\perp (X_2, X_3)|Z$ .

#### PROOF

By DROP/ADD principles

$$(X_1, X_2) \perp\!\!\!\perp X_3|Z \rightarrow X_1 \perp\!\!\!\perp X_3|(Z, X_2).$$

A useful result in statistical applications by Dawid (1979a), is stated as follows:

PROPOSITION 5

The following properties are equivalent:

(i)  $X \perp\!\!\!\perp Y|Z$  and  $X \perp\!\!\!\perp W|(Y, Z)$ .

(ii)  $X \perp\!\!\!\perp (Y, W)|Z$ .

PROOF

(i)  $\rightarrow$  (ii)

From (i) we have that  $X|(W, Y, Z) \sim X|(Y, Z) \sim X|Z$ . Then,  $X|(W, Y, Z) \sim X|Z$  or equivalently,  $X \perp\!\!\!\perp (W, Y)|Z$ .

(ii)  $\rightarrow$  (i)

By Proposition 3, we conclude that  $X \perp\!\!\!\perp Y|Z$  and  $X \perp\!\!\!\perp (Y, W)|(Z, Y)$  which implies  $X \perp\!\!\!\perp W|(Z, Y)$ .  $\square$

Note that since property (ii) is symmetric ( $Y$  and  $W$  may commute), the inclusion of the following property is implicit: (iii)  $X \perp\!\!\!\perp W|Z$  and  $X \perp\!\!\!\perp Y|(Z, W)$ . The corollary below is an example of a kind of result we may prove by using the equivalence between (i) and (iii).

COROLLARY 4

For  $T \subset (X, Z, W)$ , if  $X \perp\!\!\!\perp Y|(Z, W)$  and  $T \perp\!\!\!\perp W|Z$ , then  $T \perp\!\!\!\perp Y|Z$  and  $T \perp\!\!\!\perp W|(Y, Z)$ .

This result is better understood in the Bayesian context when  $X$  represents the sample,  $(T, W) \subset X$ ,  $Y$  represents the parameter, and  $Z$  is essentially a constant ( $Z$  is a generator of  $F_0$ ). We might say

a)  $\Omega \in E$

b) For  $E_1, E_2 \subset E$ , if  $E_1 \subset E_2$  then

$$\begin{aligned} (I_{(E_2-E_1)} f)^* &= (I_{E_2} f - I_{E_1} f)^* = I_{E_2}^* f^* - I_{E_1}^* f^* = (I_{E_2} - I_{E_1})^* f^* \\ &= I_{(E_2-E_1)}^* f^*. \end{aligned}$$

That is,  $E_2 - E_1 \in E$ .

c) For any monotone sequence  $E_1, E_2, \dots$ , of  $E$ , we have that

$I_{\lim E_n} = \lim I_{E_n}$  and by the dominated convergence theorem for conditional expectation,  $(\lim I_{E_n} f)^* = \lim (I_{E_n} f)^*$ . Since  $E_n \in E$ ,  $(\lim I_{E_n} f)^* = \lim I_{E_n}^* f^* = I_{\lim E_n}^* f^*$ . That is,  $\lim E_n \in E$ .

To conclude the proof recall that, by hypothesis,

$\bigcup_{n=1}^{\infty} F_n \subset E$  and then by Theorem 1,

$$\bigcap_{n=1}^{\infty} F_n \subset E. \quad \square$$

To conclude this section we extend Proposition 6 to the countable case.

#### PROPOSITION 7

Let  $Z, X_1, X_2, \dots$  be a sequence of random objects such that, for each  $n = 1, 2, \dots$ ,  $(X_1, \dots, X_n) \perp\!\!\!\perp X_{n+1} | Z$ . Then  $\{X_1, X_2, \dots\}$  is mutually conditionally independent given  $Z$ .

By Proposition 5,  $X_1 \perp\!\!\!\perp X_2 | Z$  and  $X_1 \perp\!\!\!\perp X_3 | (Z, X_2)$  hold if and only if  $X_1 \perp\!\!\!\perp (X_2, X_3) | Z$ .  $\square$

It is clear now that to check Definition 4 for a finite set of random objects, say  $X_1, \dots, X_n$ , we need only check that

$$(X_1, \dots, X_k) \perp\!\!\!\perp X_{k+1} | Z$$

for every  $k = 1, 2, \dots, n - 1$ .

To extend this result to the countable case, we prove the following theorem which is called the limiting property of c.i.. It will be applied in a characterization of Markov Chains presented in Section 5.

### THEOREM 3

Let  $Z, X, Y_1, Y_2, \dots$  be random objects such that  $X \perp\!\!\!\perp (Y_1, Y_2, \dots, Y_n) | Z$  for every  $n = 1, 2, \dots$ . Then,  $X \perp\!\!\!\perp (Y_1, Y_2, \dots) | Z$  where  $(Y_1, Y_2, \dots)$  essentially is the generator of  $\bigcap_{n=1}^{\infty} F_n = \sigma(\bigcup_{n=1}^{\infty} F_n)$ . (Here,  $F_n \equiv F_{Y_n}$ .)

### PROOF

Since  $\bigcup F_n$  is a field, it is closed under finite intersections.

Let  $*$  stand for  $*Z$  and consider the set

$$E = \{E; E \in \bigcap_{n=1}^{\infty} F_n \text{ and } (I_E f)^* = I_E^* f^* \forall f \in X\}.$$

The following conditions show that  $E$  is a D-system:

PROOF

Let  $(\{i_1, i_2, \dots\}, \{j_1, j_2, \dots\})$  be a partition of the set  $\{1, 2, \dots\}$ . We wish to prove that the relation

$(X_{i_1}, X_{i_2}, \dots) \perp\!\!\!\perp (X_{j_1}, X_{j_2}, \dots) | Z$  holds. Note that, for any

$k, \ell \in \{1, 2, \dots\}$  the finite relation

$(X_{i_1}, X_{i_2}, \dots, X_{i_k}) \perp\!\!\!\perp (X_{j_1}, X_{j_2}, \dots, X_{j_\ell}) | Z$  holds. This

follows from the discussion after Proposition 6 and from the fact

that  $(X_1, \dots, X_m) \perp\!\!\!\perp X_{m+1} | Z \forall m = 1, 2, \dots, v$ , where

$v = \max(i_1, \dots, i_k, j_1, \dots, j_\ell)$ . By Theorem 3 it follows that

$(X_{i_1}, \dots, X_{i_k}) \perp\!\!\!\perp (X_{j_1}, X_{j_2}, \dots) | Z$ . Finally, applying again

Theorem 3 we prove our claim.  $\square$

We write  $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \dots | Z$  or  $\prod_{n=1}^{\infty} X_n | Z$  to indicate that the sequence  $(X_1, X_2, \dots)$  is mutually conditionally independent given  $Z$ .

The next section presents some applications of c.i. in Bayesian statistics and in a characterization of Markov chains.

## 5 - MARKOV CHAINS AND BAYESIAN INFERENCE

As discussed in Dawid (1979a, 1980), many of the important statistical concepts are simply manifestations of the concept of conditional independence. In this section we use the framework of c.i. to study a well known characterization of the Markov Chain property and to describe the Bayesian version of those statistical concepts and their properties.

The following is the usual definition of Markov Chain.

DEFINITION 5

A sequence of random objects,  $X_1, X_2, \dots$  is said to form a Markov Chain if,

$$(5.1) \quad \forall n \geq 1, (X_1, \dots, X_n) \perp\!\!\!\perp X_{n+2} | X_{n+1}.$$

This concept is better understood when the relations (5.1) are replaced by,

$$(5.2) \quad \forall n \geq 1, (X_1, \dots, X_n) \perp\!\!\!\perp (X_{n+2}, X_{n+3}, \dots) | X_{n+1}.$$

Here, if the indices represent time we might say that the past is independent of the future given the present. The following proposition states the equivalence among (5.1) and (5.2).

PROPOSITION 8

The sequence  $X_1, X_2, \dots$  of random objects forms a Markov Chain if and only if the relations (5.2) are satisfied.

PROOF

(5.2)  $\rightarrow$  (5.1) Follows directly from Proposition 3.

(5.1)  $\rightarrow$  (5.2)

Step 1 - First we wish to prove that

$$\forall n \geq 1, (X_1, \dots, X_n) \perp\!\!\!\perp (X_{n+2}, X_{n+3}) | X_{n+1}.$$

But, using DROP/ADD principles, (5.1) implies that

$\forall n \geq 1, (X_1, \dots, X_n) \perp\!\!\!\perp X_{n+2} | X_{n+1}$  and

$(X_1, \dots, X_n) \perp\!\!\!\perp X_{n+3} | (X_{n+1}, X_{n+2})$ .

The conclusion of step 1 follows now directly from Proposition 5.

Step 2 - Now we wish to prove that

$\forall n \geq 1$  and  $k \geq 2, (X_1, \dots, X_n) \perp\!\!\!\perp (X_{n+2}, \dots, X_{n+k}) | X_{n+1}$ .

By induction (in  $k$ ), suppose that

$\forall n \geq 1, (X_1, \dots, X_n) \perp\!\!\!\perp (X_{n+2}, \dots, X_{n+k-1}) | X_{n+1}$  and

$(X_1, \dots, X_{n+1}) \perp\!\!\!\perp (X_{n+3}, \dots, X_{n+k}) | X_{n+2}$ .

With the same argument as in step 1, we conclude step 2.

Step 3 - Finally, we wish to prove (5.2). But (5.2) follows directly from step 2 and Theorem 3.  $\square$

Having established the concept of Markov Chains, the following properties are immediately stated:

A - If  $(X_1, X_2, \dots)$  forms a Markov Chain, so does  $(\dots, X_2, X_1)$ . [From Definition 3a.]

B - Any subsequence of a Markov Chain is a Markov Chain. [From DROP/ADD principles.]

C - If  $(X_1, X_2, \dots)$  forms a Markov Chain, then  $\forall n \geq m + 2, m \geq 1$

$(X_1, \dots, X_m) \perp\!\!\!\perp (X_{m+2}, \dots, X_n) \perp\!\!\!\perp (X_{n+2}, \dots) | (X_{m+1}, X_{n+1})$ .

To prove property C, it is enough to have

PROPOSITION 9

If  $(X_1, X_2, X_3, X_4, X_5)$  forms a Markov Chain, then

$$X_1 \perp\!\!\!\perp X_3 \perp\!\!\!\perp X_5 \mid (X_2, X_4).$$

PROOF

By hypothesis,  $X_1 \perp\!\!\!\perp (X_3, X_4, X_5) \mid X_2$  and  $(X_1, X_2, X_3) \perp\!\!\!\perp X_5 \mid X_4$ . By DROP/ADD principles this implies that  $X_1 \perp\!\!\!\perp (X_3, X_4) \mid (X_2, X_5)$  and  $X_3 \perp\!\!\!\perp X_5 \mid (X_2, X_4)$ . The conclusion follows directly from Proposition 6.  $\square$

To conclude our discussion on the concept of Markov Chains, we notice that Definition 5 can be generalized by considering an additional random object  $Z$  in the conditioning random objects of (5.1). That is, in the place of (5.1) consider the relations  $\forall n \geq 1, (X_1, \dots, X_n) \perp\!\!\!\perp X_{n+2} \mid (Z, X_{n+1})$ . In this case, we say that  $(X_1, X_2, \dots)$  form a conditional Markov Chain given  $Z$ . It is clear that we could have a similar discussion for this general concept. Finally, we notice that if  $(X_1, X_2, \dots)$  forms a conditional Markov Chain given  $Z$  and  $\forall n \geq 1, X_{n+2} \perp\!\!\!\perp Z \mid X_{n+1}$  then,  $(X_1, X_2, \dots)$  forms a Markov Chain. This is a direct application of Proposition 5.

In order to focus our attention on applications in Bayesian statistics, it is important to review some of the structures involved.

Let  $(X, A)$  be the usual sample space and  $\{P_\theta; \theta \in \Theta\}$  be a family of probability measures on  $(X, A)$  where  $\Theta$  is the usual



parameter "space". In addition, the Bayesians consider a (prior) probability space  $(\Theta, \mathcal{B}, \xi)$  where  $\mathcal{B}$  is a  $\sigma$ -algebra of subsets of  $\Theta$  such that  $P_\theta(A)$  is a  $\mathcal{B}$ -measurable function for every fixed  $A \in \mathcal{A}$ . Clearly, the choice of the prior model is not completely arbitrary, since it has to match the statistical structure on the  $\mathcal{B}$ -measurability of  $P_\theta(A)$ .

After all these considerations, it becomes clear that we can restrict ourselves to the probability space  $(\Omega, \mathcal{F}, \Pi)$ , where now  $\Omega = \Theta \times X$ ,  $\mathcal{F} = \mathcal{B} \times \mathcal{A}$  and  $\Pi$  is defined as

$$\Pi(F) = \int_{\Theta} P_\theta(F[\theta]) \xi(d\theta)$$

for every  $F \in \mathcal{F}$  where  $F[\theta] = \{x \in X; (\theta, x) \in F\}$ . Note that if  $A \in \mathcal{A}$  and  $B \in \mathcal{B}$ , then

$$\Pi(B \times A) = \int_B P_\theta(A) \xi(d\theta).$$

The uniqueness of  $\Pi$  and the fact that  $\Pi$  is a probability measure are included in Theorem 2.6.2 of Ash (1972). Now, we can define a (marginal) probability measure  $P$  on  $(X, \mathcal{A})$  in the following way:

$$P(A) = \Pi(\Theta \times A)$$

for every  $A \in \mathcal{A}$ .

Let  $X$  and  $Y$  be two random objects on  $(\Omega, \mathcal{F})$ . We say that  $X$  represents the sample and  $Y$  represents the parameter if

$$F_X \equiv \{\theta \times A; A \in \mathcal{A}\} \text{ and}$$

$$F_Y \equiv \{B \times X; B \in \mathcal{B}\}.$$

In addition to  $X$  and  $Y$  as defined above, consider two random objects  $X_1$  and  $X_2$  such that  $(X_1, X_2) \subset X$ . The Bayesian version of the concepts of sufficiency and ancillarity is contained in the following.

DEFINITION 6

a) If  $X \perp\!\!\!\perp Y|X_1$  we say that  $X_1$  is sufficient for  $X$  with respect to  $Y$ .

b) If  $X_2 \perp\!\!\!\perp Y$  we say that  $X_2$  is ancillary with respect to  $Y$ .

The classical concept of statistical independence between  $X_1$  and  $X_2$  has its Bayesian version as:

c)  $X_1 \perp\!\!\!\perp X_2|Y$ .

Basu (1955, 1958) speculates under what conditions two of the three relations a), b), and c) imply the third. One of the objectives of this chapter is to study Basu's theorems under the Bayesian framework. The next result which is Basu's first conjecture presents conditions to have b) and c) implying a).

PROPOSITION 10

If in addition to  $X_2 \perp\!\!\!\perp Y$  and  $X_1 \perp\!\!\!\perp X_2|Y$  we have  $X \perp\!\!\!\perp Y|(X_1, X_2)$ , then  $X \perp\!\!\!\perp Y|X_1$ .

COROLLARY 4

Let  $X$ ,  $Y$ , and  $Z$  be three random objects such that  $X \perp\!\!\!\perp Y|Z$  and  $X \perp\!\!\!\perp Z|Y$ . Then,  $X \perp\!\!\!\perp (Y, Z)|Y \wedge Z$ .

Note that Corollary 4 shows that relations a) and c) imply b) if  $X_1 \wedge Y$  is essentially constant on  $\Omega$  (that is, essentially generates  $F_0$ ). This condition will be studied in Section 6 in connection with Basu's second result.

To end this section we present an extreme case of DROP/ADD principles for the conditioning random object. It appears in Dawid (1980) and it was originally introduced by G. Udney Yule in terms of collapsibility of contingency tables. It must clarify the problems with Simpson's paradox in Examples 2 and 4.

PROPOSITION 12

Let  $X$ ,  $Y$ , and  $Z$  be three random objects such that  $F_Z \equiv \{\phi, \Omega, A, A^c\}$  with  $0 < \Pi(A) < 1$ . If  $X \perp\!\!\!\perp Y$  and  $X \perp\!\!\!\perp Y|Z$ , then either  $X \perp\!\!\!\perp Z$  or  $Y \perp\!\!\!\perp Z$ .

The proof becomes simple when we recognize the following general result:

LEMMA 1

If  $X \perp\!\!\!\perp Y$  and  $X \perp\!\!\!\perp Y|Z$ , then for every atom  $A$  of  $Z$  with  $\Pi(A) > 0$ , we have

$$E\{I_A|(X, Y)\} = [\Pi(A)]^{-1}E\{I_A|X\}E\{I_A|Y\}.$$

PROOF

By Proposition 5 we have that:

i)  $X_2 \perp\!\!\!\perp Y$  and  $X_1 \perp\!\!\!\perp X_2|Y$  if and only if  $X_1 \perp\!\!\!\perp X_2$  and  $X_2 \perp\!\!\!\perp Y|X_1$ .

ii)  $X_2 \perp\!\!\!\perp Y|X_1$  and  $X \perp\!\!\!\perp Y|(X_1, X_2)$  if and only if  $X \perp\!\!\!\perp Y|X_1$  since  $(X_2, X) \equiv X$ .  $\square$

Looking at the above proof, we see that if  $X_1 \perp\!\!\!\perp X_2$ , then a) implies b) and c). The meaning of the relation  $X_1 \perp\!\!\!\perp X_2$  in classical statistics, however, is void.

Note that Proposition 10 gives conditions for reducing (DROP) the conditioning random object. Actually, all of Basu's theorems are cases of DROP/ADD principles. Basu's other theorems are discussed in the next sections of this chapter.

Another type of reduction of the conditioning random object is presented in the proposition below which is a Bayesian version of a theorem introduced by Burkholder (1961).

PROPOSITION 11

Let  $X_0$  and  $X_1$  be two random objects such that  $(X_0, X_1) \subset X$ ,  $X \perp\!\!\!\perp Y|X_0$  and  $X \perp\!\!\!\perp Y|X_1$ . Then  $X \perp\!\!\!\perp Y|X_0 \wedge X_1$ . (If  $X_0$  and  $X_1$  are sufficient for  $X$ , then so is  $X_0 \wedge X_1$ .)

The proof follows directly from the definition of c.i.. As an important consequence of this proposition we have the following result which was introduced by Dawid (1979b).

PROOF OF LEMMA 1

Here we use  $*$  for  $*X$  and  $\dagger$  for  $*Y$ . Let  $B$  and  $C$  be two sets such that  $I_B \subset X$  and  $I_C \subset Y$ . Using the properties of conditional expectation and the fact that  $X \perp\!\!\!\perp Y$  we have that,

$$\int_{BC} I_A^* I_A^\dagger d\Pi = \int_{I_B} I_A^* I_C I_A^\dagger d\Pi = \int I_{AB}^* I_{AC}^\dagger d\Pi =$$

$$[\int I_{AB}^* d\Pi][\int I_{AC}^\dagger d\Pi] = [\int I_{AB} d\Pi][\int I_{AC} d\Pi].$$

That is, since  $\Pi(A) > 0$ ,

$$\int_{BC} I_A^* I_A^\dagger d\Pi = \Pi(AB)\Pi(AC) = [\Pi(A)]^2 \Pi(B|A)\Pi(C|A),$$

where  $\Pi(B|A) = \frac{\Pi(AB)}{\Pi(A)}$ . We notice now that on the atom  $A$ , the functions  $I_B^{*Z}$  and  $I_C^{*Z}$  are constants and equal respectively to  $\Pi(B|A)$  and  $\Pi(C|A)$ . Analogously, the function  $I_{BC}^{*Z}$  is equal to  $\Pi(BC|A)$  on  $A$ . On the other hand since  $X \perp\!\!\!\perp Y|Z$ ,  $I_B^{*Z} I_C^{*Z} \equiv I_{BC}^{*Z}$ ; thus  $\Pi(B|A)\Pi(C|A) = \Pi(BC|A)$ . This shows that

$$\int_{BC} I_A^* I_A^\dagger d\Pi = \Pi(A)\Pi(ABC).$$

To conclude the proof we must prove that  $\Pi(A) \int_D I_A d\Pi = \int_D I_A^* I_A^\dagger d\Pi$  for every  $D$  such that  $I_D \subset (X, Y)$ . Following the same technique used in Theorem 2 and 3 we obtain this as a consequence of Theorem 1.  $\square$

PROOF OF PROPOSITION 12

Let  $p = \Pi(A)$ ,  $I^* = E\{I_A | X\}$ , and  $I^\dagger = E\{I_A | Y\}$ . From Lemma 1 we have that

$$E\{I_A | (X, Y)\} = \frac{I^* I^\dagger}{p} \text{ and } E\{I_{A^c} | (X, Y)\} = \frac{(1 - I^*)(1 - I^\dagger)}{1 - p}.$$

Clearly

$$\frac{I^* I^\dagger}{p} + \frac{(1 - I^*)(1 - I^\dagger)}{1 - p} = 1;$$

that is,

$$\left(1 - \frac{I^*}{p}\right) \left(1 - \frac{I^\dagger}{p}\right) = 0.$$

Since  $X \perp\!\!\!\perp Y$ , this last equation holds if and only if either  $\frac{I^*}{p} \equiv 1$  or  $\frac{I^\dagger}{p} \equiv 1$  almost surely.  $\square$

#### REMARK

1 - Let  $Y \equiv (Y_1, Y_2)$  represent the parameter and  $X$  represent the sample. If  $Y_1$  and  $Y_2$  are independent a priori and a posteriori (i.e.,  $Y_1 \perp\!\!\!\perp Y_2$  and  $Y_1 \perp\!\!\!\perp Y_2 | X$ ), then from Lemma 1,

$$(5.3) \quad E\{I_A | Y\} = [\pi(A)]^{-1} E\{I_A | Y_1\} E\{I_A | Y_2\},$$

where  $A$  is a positive atom of  $X$ . Note that if  $Y_1$  and  $Y_2$  are independent a priori, and  $X$  is a discrete random variable, then  $Y_1$  and  $Y_2$  are independent a posteriori if and only if (5.3) holds and (5.3) defines the likelihood function. This result is the discrete case of the theorem introduced in Section 9 of Basu (1977).

#### 6 - ON MEASURABLE SEPARABILITY OF RANDOM OBJECTS

Basu (1955) stated that any statistic independent of a sufficient statistic is ancillary. Later on Basu (1958) presented a

counter-example and recognized the necessity of an additional condition (connectedness) on the family  $\{P_\theta: \theta \in \Theta\}$  of probability measures. Koehn and Thomas (1975) strengthened this result by introducing a necessary and sufficient condition on the family. More recently Basu and Cheng (1979), generalizing results of Pathak (1975), showed the equivalence between these two conditions in Coherent Models.

In the scope of the present work, this question will be stated in terms of random objects. Suppose that  $X$  represents the sample and  $Y$  the parameter. The following theorem is a Bayesian version of the result of Koehn and Thomas (1975).

#### THEOREM 4

Let  $X_1 \subset X$  be a sufficient random object (i.e.,  $X \perp\!\!\!\perp Y|X_1$ ). The random object  $Y \wedge X_1$  is essentially a constant (i.e.,  $F_{Y \wedge X_1} \equiv F_0$ ) if and only if  $X_2 \perp\!\!\!\perp Y$  whenever  $X_2 \subset X$  and  $X_1 \perp\!\!\!\perp X_2|Y$  (i.e.,  $X_2$  is ancillary if  $X_1$  and  $X_2$  are statistically independent).

#### PROOF

→ See discussion following Corollary 4.

← Take  $X_2$  such that  $X_2 \equiv Y \wedge X_1$ . Since  $X_2 \subset Y$ ,  $X_1 \perp\!\!\!\perp X_2|Y$ . Then by hypothesis  $X_2 \perp\!\!\!\perp Y$ , which implies that  $X_2 \perp\!\!\!\perp X_1$  since  $X_2 \subset Y$ ; that is,  $X_2 \equiv Y \wedge X_1$  is essentially a constant.  $\square$

REMARKS

2 - The condition introduced by Koehn and Thomas (1975) is the non-existence of a splitting set. A set  $A$  in the sample space (i.e.,  $A \in \mathcal{A}$ ) is a splitting set if  $P_\theta(A) = 0$  or  $1$  for all  $\theta \in \Theta$  and at least for a pair  $\{\theta_1, \theta_2\} \subset \Theta$ ,  $P_{\theta_1}(A) = P_{\theta_2}(A^c) = 1$ . In the Bayesian framework, since  $X$  represents the sample and  $Y$  the parameter, an analogous definition is as follows: A set  $A$  such that  $I_A \subset X$  is a splitting set if  $0 < \Pi(A) < 1$  and  $E\{I_A|Y\} = E^2\{I_A|Y\}$ . Let  $I_A^* = E\{I_A|Y\}$  and note that  $\{(I_A - I_A^*)^2\}^* = I_A^* - (I_A^*)^2$ . Thus, if  $A$  is a splitting set,  $E\{(I_A - I_A^*)^2\} = 0$ ; that is,  $I_A = I_A^*$ . Then  $I_A \subset Y$  or equivalently  $I_A \subset Y \wedge X$ . We conclude that the non-existence of a splitting set is equivalent to  $Y \wedge X$  being essentially a constant.

3 - Let  $X, Y$ , and  $Z$  be three random objects such that  $X \perp\!\!\!\perp Y|Z$ . Since this is equivalent to  $(X, Z) \perp\!\!\!\perp (Y, Z)|Z$ , with the same argument we use in the proof of Theorem 4, we can easily show that  $(X, Z) \wedge (Y, Z) \equiv Z$ . Intuitively we would say that if  $X \perp\!\!\!\perp Y|Z$ , then  $Z$  possesses all common information contained in both  $X$  and  $Y$ .

The following result is a Bayesian solution for a two-parameter problem in inference. Suppose that the parameter  $Y$  is such that  $Y \equiv (Y_1, Y_2)$ . Let  $X$  represent the sample,  $X_1 \subset X$  be specific sufficient with respect to  $Y_2$ , and  $X_2 \subset X$  be specific sufficient with respect to  $Y_1$ . That is,  $X \perp\!\!\!\perp Y_2|(X_1, Y_1)$  and



$X \perp\!\!\!\perp Y_1 | (X_2, Y_2)$ . (See Basu (1978) for details on the notion of specific sufficiency.) The question here is under what conditions does the specific sufficiency of  $X_1$  and  $X_2$  imply the sufficiency of  $(X_1, X_2)$ ?

PROPOSITION 13

If  $(X_1, Y_1) \wedge (X_2, Y_2) \subset (X_1, X_2)$ , then  $X \perp\!\!\!\perp Y_2 | (X_1, Y_1)$  and  $X \perp\!\!\!\perp Y_1 | (X_2, Y_2)$  imply  $X \perp\!\!\!\perp Y | (X_1, X_2)$ .

PROOF

From DROP/ADD principles we have that  $X \perp\!\!\!\perp Y | (X_1, Y_1)$  and  $X \perp\!\!\!\perp Y | (X_2, Y_2)$ . Thus, by Proposition 11,

$$X \perp\!\!\!\perp Y | (X_1, Y_1) \wedge (X_2, Y_2),$$

and since  $(X_1, X_2) \subset X$ , the result follows.  $\square$

The following related result is a direct consequence of Proposition 5.

PROPOSITION 14

If  $X \perp\!\!\!\perp Y_2 | (X_1, Y_1)$  and  $X \perp\!\!\!\perp Y_1 | (X_2, Y_2)$ , then  $X \perp\!\!\!\perp Y | (X_1, X_2)$  if and only if  $X \perp\!\!\!\perp Y_1 | (X_1, X_2)$  [equivalently  $X \perp\!\!\!\perp Y_2 | (X_1, X_2)$ ].

Note that the condition  $X \perp\!\!\!\perp Y_1 | (X_1, X_2)$  does not have an interpretation in classical statistics since distributions depend on both parameters  $Y_1$  and  $Y_2$ . Our conjecture for a future work is that specific sufficiency of  $X_1$  and  $X_2$  implies sufficiency of

$(X_1, X_2)$  if  $Y_1$  and  $Y_2$  are variation independent (i.e., the parameter space is the cartesian product of the domain of  $Y_1$  by the domain of  $Y_2$ ). (See Basu (1977) and Barndorff-Nielsen (1978) for details on the notion of variation independence.) Dawid (1979b) presented an example where  $(X_1, X_2)$  is not sufficient even though  $X_1$  and  $X_2$  are specific sufficient. In this example, however, the parameters are not variation independent.

The title of this section was motivated by the following:

#### DEFINITION 7

The random objects  $X$  and  $Y$  are said to be measurably separated conditionally on  $Z$  if  $(X, Z) \wedge (Y, Z) \equiv Z$ . When  $Z$  is essentially a constant we simply say that  $X$  and  $Y$  are measurably separated.

A large list of results related with this concept appears in Mouchart and Rolin (1978).

#### 7 - BASU THEOREM

Basu (1955) proved that any ancillary statistic is statistically independent of any bounded complete sufficient statistic. The Bayesian analogous concept of boundedly completeness is the concept of strong identifiability (Dawid [1980] and Mouchart, and Rolin [1978]). The main objective of this section is to study this concept and present Basu's result under the Bayesian framework.

Let  $X$  and  $Y$  be two random objects. As before, we study some aspects of the linear maps  $L_{\infty}(Y) \xrightarrow{*} L(X)$  and  $L_{\infty}(X) \xrightarrow{\dagger} L_{\infty}(Y)$ , where  $*$  is for  $*X$ , and  $\dagger$  is for  $*Y$ . Recall that for two random variables  $f_1$  and  $f_2$ , by  $f_1 \neq f_2$  we mean that  $\Pi\{\omega; f_1(\omega) \neq f_2(\omega)\} > 0$ .

DEFINITION 8

The map  $L_{\infty}(X) \xrightarrow{\dagger} L_{\infty}(Y)$  is essentially one-one if  $f_1^{\dagger} \neq f_2^{\dagger}$  whenever  $(f_1, f_2) \in X$  and  $f_1 \neq f_2$ . In this case we say that  $X$  is strongly identified by  $Y$  and write  $X \ll Y$ .

Clearly,  $X \ll Y$  if and only if for  $f \in X$ ,  $f^{\dagger} = 0$  implies  $f = 0$ . This shows intuitively that when  $Y$  represents the parameter and  $X$  the sample, Definition 8 is the Bayesian version of the concept of bounded completeness.

DEFINITION 9

The map  $L_{\infty}(Y) \xrightarrow{*} L_{\infty}(X)$  is essentially onto if for every  $f \in X$  there is a  $g \in Y$  such that  $g^* = f$ .

The following result relates these two definitions.

PROPOSITION 15

If the map  $L_{\infty}(Y) \xrightarrow{*} L_{\infty}(X)$  is essentially onto, then  $X \ll Y$ .

PROOF

Let  $(f, h) \in X$  and  $f^{\dagger} = 0$ . Since  $*$  is essentially onto  $\exists g \in Y$  s.t.  $g^* = h$ . Then

$$E\{fh\} = E\{fg^*\} = E\{fg\} = E\{f^{\dagger}g\} = 0.$$

Since  $h$  is arbitrary,  $f = 0$ .  $\square$

Let  $X_{[Y]}$  be the random object that generates the smallest subfield that contains all functions  $g^*$  where  $g \in Y$ . Note that  $X_{[Y]} \subset X$ . The following result shows that  $X_{[Y]}$  may be viewed as the Bayesian minimal sufficient statistic.

PROPOSITION 16

(i)  $X \perp\!\!\!\perp Y | X_{[Y]}$

(ii) If  $X_1 \subset X$  is such that  $X \perp\!\!\!\perp Y | X_1$ , then  $X_{[Y]} \subset X_1$ .PROOF(i)  $\forall g \in Y$ ,  $E\{g | (X, X_{[Y]})\} = g^* \in X_{[Y]}$  by definition.(ii)  $\forall g \in Y$ ,  $E\{g | (X, X_1)\} = E\{g | X\} = E\{g | X_1\}$ .

Then for every  $g \in Y$ ,  $g^* \in X_1$ . Since  $X_{[Y]}$  is the generator of the smallest subfield containing the functions  $g^*$ ,  $X_{[Y]} \subset X_1$ .  $\square$

When  $X_{[Y]} \equiv X$ ,  $X$  is said to be identified by  $Y$  (Dawid [1980], and Mouchart and Rolin [1978]). The name strong identification was motivated by the following result:

PROPOSITION 17If  $X \ll Y$ , then  $X_{[Y]} \equiv X$ .PROOFNote that  $X \perp\!\!\!\perp Y | X_{[Y]}$ . Thus,

$$\forall f \in X, \quad E\{E\{f | (Y, X_{[Y]})\} | Y\} = E\{E\{f | X_{[Y]}\} | Y\}.$$

For  $f^\dagger = E\{f | X_{[Y]}\}$  since  $X \ll Y$ , we have that

$$E\{(f - f^\dagger) | Y\} = 0 \rightarrow f = f^\dagger. \quad \text{Then } \forall f \in X, f \in X_{[Y]} \text{ and } X \equiv X_{[Y]}. \quad \square$$

The Bayesian version of the Basu theorem is contained in the result below.

PROOF

From Proposition 16,  $X_{[Y]} \subset X_1$  and  $X \perp\!\!\!\perp Y|X_{[Y]}$ . Using Proposition 3 we can write (i)  $X_{[Y]} \perp\!\!\!\perp Y|X_1$ , (ii)  $X_1 \perp\!\!\!\perp Y|X_{[Y]}$  and (iii)  $X_1 \ll Y$ . Let  $f \in X_1$ , and note that from (ii) and Proposition 1 we have

$$E\{f|Y\} = E\{E\{f|X_{[Y]}\}|Y\}.$$

Since  $X_{[Y]} \subset X_1$ ,  $E\{f|X_{[Y]}\} \in X_1$ . From (iii) we conclude that  $f = E\{f|X_{[Y]}\} \in X_{[Y]}$ . Then every  $f \in X_1$  implies  $f \in X_{[Y]}$  which implies that  $X_1 \subset X_{[Y]}$ .  $\square$

REMARK

4 - The concept of strong identifiability may be generalized as follows:  $X$  is strongly identified by  $Y$  conditionally on  $Z$  ( $X \ll Y|Z$ ) if for every  $f \in (X, Z)$ ,  $E\{f|(Y, Z)\} = 0$  implies  $f = 0$ . Analogously,  $X$  is identified by  $Y$  conditionally on  $Z$  if

$$(X, Z)_{[Y, Z]} \equiv (X, Z).$$

All the results of this section may be easily generalized by introducing a conditioning random object  $Z$  to each relation stated. For our future work we intend to relate these general results with the work of Dawid (1979c), Ferreira (1980), and Godambe (1980).

THEOREM 5

Let  $X$ ,  $Y$ , and  $Z$  be three random objects. If  $X \perp\!\!\!\perp Y$ ,  $X \perp\!\!\!\perp Y|Z$ , and  $Z \ll Y$ , then  $X \perp\!\!\!\perp Z|Y$ .

PROOF

Since  $X \perp\!\!\!\perp Y|Z \forall f \in X$ ,  $E\{f|(Y, Z)\} = E\{f|Z\}$ . On the other hand, since  $X \perp\!\!\!\perp Y$ ,  $E\{f|Y\} = E\{f\}$  but, by Proposition 1,  $E\{f|Y\} = E\{E\{f|(Y, Z)\}|Y\} = E\{E\{f|Z\}|Y\}$ . Then  $E\{f\} = E\{E\{f|Z\}|Y\}$  which implies that  $E\{[E\{f|Z\} - E\{f\}]|Y\} = 0$ . Since  $Z \ll Y$ ,  $E\{f\} = E\{f|Z\}$  for every  $f \in X$ . That is, if  $Z \ll Y$ , then  $X \perp\!\!\!\perp Y$  and  $X \perp\!\!\!\perp Y|Z$  implies  $X \perp\!\!\!\perp Z$ . Now, by Proposition 5 we have that  $X \perp\!\!\!\perp Y|Z$  and  $X \perp\!\!\!\perp Z$  is equivalent to  $X \perp\!\!\!\perp Z|Y$  and  $X \perp\!\!\!\perp Y$ .  $\square$

Note that to obtain the Basu Theorem we consider  $X$  as the sample,  $Y$  as the parameter, and  $X_0$  and  $X_1$  two random objects such that  $(X_0, X_1) \in X$ ,  $X_0 \perp\!\!\!\perp Y$ ,  $X \perp\!\!\!\perp Y|X_1$  and  $X_1 \ll Y$ . Clearly  $X_0 \perp\!\!\!\perp Y|X_1$  and the result  $X_0 \perp\!\!\!\perp X_1|Y$  follows.

Lehmann and Scheffé (1950) proved that if a sufficient statistic is boundedly complete, then it is a minimal sufficient statistic. The Proposition below is the Bayesian version of this result.

PROPOSITION 18

Let  $X_1$ ,  $X$ , and  $Y$  be three random objects such that  $X_1 \in X$  and  $X \perp\!\!\!\perp Y|X_1$ . If  $X_1 \ll Y$  then  $X_1 \equiv X_{[Y]}$ .

REFERENCES

1. ASH, R. B. (1972). Real Analysis and Probability. Academic Press, New York.
2. BAHADUR, R. R. (1955). Measurable Subspaces and Subalgebras. Proc. Amer. Math. Soc., 6, 565-70.
3. BARNDORFF-NIELSEN, O. (1978). Information and Exponential Families in Statistical Theory. John Wiley, New York.
4. BASU, D. (1955). On Statistics Independent of a Complete Sufficient Statistic. Sankhya A, 15, 377-80.
5. BASU, D. (1958). On Statistics Independent of a Sufficient Statistic. Sankhya A, 20, 223-26.
6. BASU, D. (1959). The Family of Ancillary Statistics. Sankhyā A, 21, 247-56.
7. BASU, D. (1964). Recovery of Ancillary Information. Sankhyā A, 26, 3-16.
8. BASU, D. (1967). Problems Relating to the Existence of Maximal and Minimal Elements in Some Families of Statistics (Subfields). Proc. Fifth Berkeley Sym. Math. Statist. Prob., 1, 41-50.
9. BASU, D. (1977). On the Elimination of Nuisance Parameters. JASA, 72, 355-66.
10. BASU, D. (1978). On Partial Sufficiency: A Review. J. Statist. Plan. Inf., 2, 1-13.
11. BASU, D. and CHENG, S. C. (1979). A Note on Sufficiency in Coherent Models, Int. J. Math. Math. Sci. To appear.
12. BURKHOLDER, D. L. (1961). Sufficiency in the Undominated Case. Ann. Math. Statist., 32, 1191-200.
13. CHENG, S. C. (1978). A Mathematical Study of Sufficiency and Adequacy in Statistical Theory. Ph.D. Dissertation, FSU, Florida.

14. DAWID, A. P. (1979a). Conditional Independence in Statistical Theory. JRSS, B, 41, 1-31.
15. DAWID, A. P. (1979b). Some Misleading Arguments Involving Conditional Independence. JRSS, B, 41, 249-52.
16. DAWID, A. P. (1979c). A Bayesian Look at Nuisance Parameters. Trabajos de Estadística. To appear.
17. DAWID, A. P. (1980). Conditional Independence for Statistical Operations. Ann. Statist., 8, 598-617.
18. DE FINETTI, B. (1937). Foresight: Its Logical Laws, Its Subjective Sources. Translated edition 1964 in Studies in Subjective Probability (H. E. Kyburg and H. E. Smokler, editors). John Wiley, New York.
19. DE FINETTI, B. (1970). Theory of Probability, Vols. 1 and 2. Translated edition, 1974. John Wiley, London.
20. DOOB, J. L. (1953). Stochastic Processes. John Wiley, New York.
21. FERREIRA, P. E. (1980). Comments on Berkson's Paper "In Dispraise of ...". Unpublished report.
22. FISCHER, R. A. (1920). A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error. Mon. Not. Roy. Ast. Soc., 80, 758-70.
23. FISCHER, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. Phil. Trans. A, 222, 309-68.
24. FLORENS, J. P. and MOUCHART, M. (1977). Reduction of Bayesian Experiments. CORE, Discussion Paper 7737.
25. GODAMBE, V. P. (1979). On Sufficiency and Ancillarity in Presence of Nuisance Parameter. Unpublished report.
26. HALL, W. J., WIJSMAN, R. A., and GOSH, J. K. (1965). The Relationship Between Sufficiency and Invariance. Ann. Math. Statist., 36, 375-614.
27. KOEHN, U. and THOMAS, D. L. (1975). On Statistics Independent of a Sufficient Statistic: Basu's Lemma. Amer. Statist., 29, 40-2.



28. KOLMOGOROV, A. N. (1942). Determination of the Center of Dispersion and Degree of Accuracy for a Limited Number of Observations (in Russian). Izvestija Akademii Nauk, Ser. Mat. 6, 3-32.
29. LEHMAN, E. L. and SCHEFFÉ, H. (1950). Completeness, Similar Regions, and Unbiased Estimation, Part I. Sankhya A, 10, 305-40.
30. MOUCHART, M. and ROLIN, J. M. (1978). A Note on Conditional Independence. Unpublished report.
31. MOY, S. T. C. (1954). Characterization of Conditional Expectation as a Transformation on Function Spaces. Pacific J. Math, 4, 47-63.
32. PATHAK, P. K. (1975). Note on Basu's Lemma. Unpublished report.
33. PICCI, G. (1977). Some Connections Between the Theory of Sufficient Statistics and the Identifiability Problem. SIAM J. Appl. Math., 33, 383-98.

CHAPTER II. ON THE BAYESIAN ANALYSIS OF CATEGORICAL DATA:  
THE PROBLEM OF NONRESPONSE

1 - INTRODUCTION

The simplest case of the problem of nonresponse is as follows. Let  $\pi_1$  be the unknown proportion of individuals in a certain population,  $P$ , that belong to a particular category  $A_1$ . With  $\pi_1$  as the only parameter of interest, a survey is conducted using a simple random sample of size  $n$ . Of the  $n$  individuals surveyed,  $n_1$  respond to the question "Do you belong to category  $A_1$ ?" with a yes/no answer, but  $n_2 = n - n_1$  individuals do not respond. Denoting the category of respondents by  $R$ , and the complementary category by  $R'$ , the survey data may be summarized as:

(1.1)

	R	R'	
$A_1$	$x_1$	$n_2$	
$A_2$	$x_2$		
	$n_1$	$n_2$	$n$

with  $A_2$  being the complement of  $A_1$ .

In many practical problems, it is understood that the non-response of an individual is highly dependent on the value of the measurement under study. For example, suppose that one is surveying a population of students in order to estimate the proportion of cannabis smokers. In this case, it should be expected that a student who smokes has a higher chance of being a nonrespondent than one who does not. In this instance, at least, a nonresponse is a strong source of information.

The above understanding of the problem suggests that the population must also be partitioned into the categories R and R'; that is, the class of elements which would respond to the question, if selected, and its complement. The population proportions may be displayed in a  $2 \times 2$  - tabular form as:

(1.2)

	R	R'		
A <sub>1</sub>	p <sub>11</sub>	p <sub>12</sub>	π <sub>1</sub>	π <sub>2</sub> = 1 - π <sub>1</sub>
A <sub>2</sub>	p <sub>21</sub>	p <sub>22</sub>	π <sub>2</sub>	q = p <sub>11</sub> + p <sub>21</sub>
	q	1 - q	1	

How can the data (1.1) be analysed vis-à-vis the parameter of interest  $\pi_1 = p_{11} + p_{12}$ ?

If the population of size N is regarded to be infinitely large compared to the sample size n; that is, if a multinomial model for the data is adopted, then the likelihood function is:

$$(1.3) \quad L = p_{11}^{x_1} p_{21}^{x_2} (1 - q)^{n_2}.$$

We represent the data by  $X = (x_1, x_2, n_2)$  with  $n_2 = n - (x_1 + x_2)$ .

Since  $p_{12}$  cannot be defined in terms of the sampling distribution of  $X$ , an orthodox non-Bayesian would characterize  $\Pi_1 = p_{11} + p_{12}$  as nonidentifiable, and would have little else to say on the matter. None of the many non-Bayesian methods of nuisance parameter elimination listed in Basu (1977) apply to the present case. On the other hand, a Bayesian regards a parameter as an unknown entity that exists in its own right. It enters into the sampling distribution of a properly planned experiment but is not defined by the experiment. Nonidentifiability is, therefore, a non-problem from the Bayesian viewpoint.

With a suitable representation  $\xi$  of his/her opinion about  $\underline{p} = (p_{11}, p_{21}, p_{12}, p_{22})$ , the Bayesian will proceed to derive the posterior distribution by matching  $\xi$  with the likelihood function (1.3). The posterior marginal distribution of the parameter of interest  $\Pi_1$  will be obtained by integration.

In Section 2 we demonstrate how the choice of a Dirichlet prior for  $\underline{p}$  simplifies the Bayesian operation. The more general case where the respondents are classified into  $k$  (instead of 2) categories,  $A_1, \dots, A_k$ , is analyzed in a similar fashion. Since the inference is based on the data, it is of interest to study the distribution of the data under the considered prior. Section 3

introduces the Dirichlet-Multinomial distribution and some of its properties. This distribution, besides being the marginal distribution of the data, plays an important role in the rest of the paper.

Sections 4 and 5 deal with the case of sampling from a finite population; that is, the case where the statistical model is Hypergeometric or, more generally, Multivariate Hypergeometric.

For the case where  $k = 2$ , instead of  $p_{11}$ ,  $p_{21}$ ,  $p_{12}$ , and  $p_{22}$ , the unknown frequency counts  $\theta_{11}$ ,  $\theta_{21}$ ,  $\theta_{12}$ , and  $\theta_{22}$  must be considered. As in (1.2), the population parameters may be displayed as:

(1.4)

	R	R'	
$A_1$	$\theta_{11}$	$\theta_{12}$	$\theta_1$
$A_2$	$\theta_{21}$	$\theta_{22}$	$\theta_2$
	$\psi$	$N - \psi$	$N$

with  $\theta_2 = N - \theta_1$ , and the parameter of interest being  $\theta_1 = \theta_{11} + \theta_{12}$ . A Dirichlet-Multinomial prior for  $\underline{\theta} = (\theta_{11}, \theta_{21}, \theta_{12}, \theta_{22})$  greatly simplifies the analysis of the data (1.1) vis-à-vis the parameter of interest  $\theta_1$ .

NOTATION: Let  $x$ ,  $y$ , and  $z$  be either random variables or random vectors. When  $x$ ,  $y$ , and  $z$  are mutually independent we write  $x \perp\!\!\!\perp y \perp\!\!\!\perp z$ . By  $x \perp\!\!\!\perp y | z$  it is meant that  $x$  and  $y$  are conditionally independent given  $z$ , and if  $x$  and  $y$  have the same distribution we write  $x \sim y$ .

Let  $\underline{p} = (p_1, \dots, p_k)$  be a  $k$ -dimensional positive random vector such that  $\sum_1^k p_i = 1$ . We write  $\underline{p} \sim D(\alpha_1, \dots, \alpha_k)$  to indicate that the distribution of  $\underline{p}$  is a Dirichlet with nonnegative real parameters  $\alpha_1, \alpha_2, \dots, \alpha_k$ . For  $k = 2$ , instead of  $(p_1, p_2) \sim D(\alpha_1, \alpha_2)$ , we use the conventional Beta distribution notation,  $p_1 \sim B(\alpha_1, \alpha_2)$ .

Let  $\underline{x} = (x_1, \dots, x_k)$  be a  $k$ -dimensional nonnegative integer random vector with fixed  $n = \sum_1^k x_i$ . We write  $\underline{x}|\underline{p} \sim M(n; \underline{p})$ , where  $\underline{p}$  is defined as above, to indicate that the conditional distribution of  $\underline{x}$  given  $\underline{p}$  is Multinomial with parameters  $n$  and  $\underline{p}$ . For  $k = 2$ , instead of  $(x_1, x_2)|(p_1, p_2) \sim M(n; (p_1, p_2))$ , we use the conventional Binomial distribution notation,  $x_1|p_1 \sim \text{Bi}(n; p_1)$ .

When  $\underline{\theta} = (\theta_1, \dots, \theta_k)$  is a nonnegative integer random vector with  $\sum_1^k \theta_i = N$  fixed, we write  $\underline{x}|\underline{\theta} \sim H(N, n, \underline{\theta})$  to indicate that the conditional distribution of  $\underline{x}$  given  $\underline{\theta}$  is Multivariate Hypergeometric with parameter  $(N, n, \underline{\theta})$ . For  $k = 2$ , instead of

$(x_1, x_2)|(\theta_1, \theta_2) \sim H(N, n, (\theta_1, \theta_2))$ , we use the conventional notation for Hypergeometric distributions,  $x_1|\theta_1 \sim h(N, n, \theta_1)$ .

The probability function corresponding to  $H(N, n, \underline{\theta})$  may be expressed in the following two ways:

$$f(\underline{x}|\underline{\theta}) = \frac{\binom{\theta_1}{x_1} \binom{\theta_2}{x_2} \cdots \binom{\theta_k}{x_k}}{\binom{N}{n}}$$

$$= \frac{\binom{n}{x_1 \cdots x_k} \binom{N-n}{\theta_1-x_1 \cdots \theta_k-x_k}}{\binom{N}{\theta_1 \cdots \theta_k}}$$

2 - NONRESPONSE: THE MULTINOMIAL MODEL

First we consider the case of  $k = 2$ , where the data, the population parameters, and the likelihood are described by (1.1), (1.2), and (1.3) respectively.

In the full response model, it is well known that the family of Dirichlet distributions of the correct dimension is the natural conjugate family for the Bayesian analysis. That is, if  $y_1$  and  $y_2$  were the observations in  $R'$ , and

$$(2.1) \quad \underline{p} = (p_{11}, p_{21}, p_{12}, p_{22}) \sim D(\alpha_{11}, \alpha_{21}, \alpha_{12}, \alpha_{22})$$

a priori, then the posterior distribution would be

$$D(\alpha_{11} + x_1, \alpha_{21} + x_2, \alpha_{12} + y_1, \alpha_{22} + y_2).$$

To introduce a Bayesian solution to the nonresponse case, it is useful to consider the following reparametrization:

$$(2.2) \quad q = p_{11} + p_{21}, \quad q_{11} = \frac{p_{11}}{q}, \quad \text{and} \quad q_{12} = \frac{p_{12}}{1 - q}$$

with the reverse transformation being

$$(2.3) \quad \begin{aligned} p_{11} &= qq_{11}, \quad p_{12} = (1 - q)q_{12} \\ p_{21} &= q(1 - q_{11}), \quad \text{and} \quad p_{22} = (1 - q)(1 - q_{12}) \end{aligned}$$

The following general result for Dirichlet distributions is a key to the solution. Let  $m \in \{2, \dots, k - 1\}$  be fixed.

LEMMA 1

The following set of conditions is necessary and sufficient to have  $(p_1, \dots, p_k) \sim D(\alpha_1, \dots, \alpha_k)$ :

$$(i) \quad y = \sum_1^m p_i \sim B(\sum_1^m \alpha_i, \sum_{(m+1)}^k \alpha_i)$$

$$(ii) \quad \frac{1}{y}(p_1, \dots, p_m) \sim D(\alpha_1, \dots, \alpha_m)$$

$$\frac{1}{1-y}(p_{m+1}, \dots, p_k) \sim D(\alpha_{m+1}, \dots, \alpha_k),$$

and (iii)  $y \perp\!\!\!\perp \frac{1}{y}(p_1, \dots, p_m) \perp\!\!\!\perp \frac{1}{1-y}(p_{m+1}, \dots, p_k)$ .

The proof of this result is straightforward and therefore is omitted.

Suppose that, a priori, (2.1) is considered. By Lemma 1, this is equivalent to

$$(2.4) \quad q \sim B(\alpha_{.1}, \alpha_{.2}), \quad q_{11} \sim B(\alpha_{11}, \alpha_{21}),$$

$$q_{12} \sim B(\alpha_{12}, \alpha_{22}), \quad \text{and } q \perp\!\!\!\perp q_{11} \perp\!\!\!\perp q_{12}$$

where  $\alpha_{.j} = \alpha_{1j} + \alpha_{2j}$ , ( $j = 1, 2$ ).

The reparametrization (2.2) changes the likelihood (1.3) to

$$(2.5) \quad L = q^{n_1} (1-q)^{n_2} q_{11}^{x_1} (1-q_{11})^{x_2}.$$

By matching the prior (2.4) with (2.5), we derive the posterior distribution of  $(q, q_{11}, q_{12})$ :



$$(i) \quad q \prod q_{11} \prod q_{12} | X,$$

$$(ii) \quad q_{11} | X \sim B(\alpha_{11} + x_1, \alpha_{21} + x_2),$$

$$q_{12} | X \sim q_{12} \sim B(\alpha_{12}, \alpha_{22}),$$

$$\text{and (iii) } q | X \sim q | n_1 \sim B(\alpha_{.1} + n_1, \alpha_{.2} + n_2).$$

As expected,  $n_1$  is sufficient to predict  $q$ , and  $q_{12}$  is independent of the data. Since  $\alpha_{.2} = \alpha_{12} + \alpha_{22} \leq \alpha_{.2} + n_2$ , the posterior distribution of the original parameter  $\underline{p}$  is again Dirichlet if and only if  $n_2 = 0$ . It is, however, a mixture of Dirichlet distributions, and  $(p_{11}, p_{21}, (1 - q)) | X \sim D(\alpha_{11} + x_1, \alpha_{21} + x_2, \alpha_{.2} + n_2)$ . Note that these properties of the posterior allow one to define a "nice" conjugate family of distributions for the nonresponse case. That is, the prior given by (2.4) would be conjugate if, instead of  $q \sim B(\alpha_{.1}, \alpha_{.2})$ , we had  $q \sim B(\alpha_{.1}, \beta)$ , where  $\beta \geq \alpha_{.2}$ .

To proceed with the estimation of  $\Pi_1$ , the parameter of interest, we recall (2.3) to write  $\Pi_1 = q q_{11} + (1 - q)q_{12}$ , and consider  $\alpha = \alpha_{11} + \alpha_{21} + \alpha_{12} + \alpha_{22}$ , and  $\alpha_{i.} = \alpha_{i1} + \alpha_{i2}$  ( $i = 1, 2$ ). Under the squared error loss function, the Bayes estimator of  $\Pi_1$  is given by:

$$\hat{\Pi}_1 = E\{\Pi_1 | X\} = E\{q q_{11} + (1 - q)q_{12} | X\}.$$

In view of the posterior distribution (2.6), we finally have:

$$\begin{aligned} \hat{\pi}_1 &= E\{q|X\}E\{q_{11}|X\} + E\{(1-q)|X\}E\{q_{12}|X\} \\ (2.7) \quad &= \frac{1}{\alpha + n}(\alpha_{1.} + x_1 + \frac{\alpha_{12}}{\alpha_{.2}} n_2) \end{aligned}$$

We notice that (see Example in Section 3)  $\frac{\alpha_{12}}{\alpha_{.2}} n_2$  is the conditional expectation of  $y_1$  - the sample frequency of nonrespondents that belong to  $A_1$  - given the data. Therefore,  $\hat{\pi}_1$  is an intuitive estimator since in the case of full response we would have  $y_1$  in place of  $\frac{\alpha_{12}}{\alpha_{.2}} n_2$ .

The generalization of the above analysis to the case of  $k$  categories,  $A_1, \dots, A_k$  ( $k \geq 2$ ), is straightforward. Tables (1.1) and (1.2) are replaced respectively by:

(2.8)

	R	R'	
$A_1$	$x_1$	$n_2$	
$\cdot$	$\cdot$		
$\cdot$	$\cdot$		
$A_k$	$x_k$		
	$n_1$	$n_2$	$n$

(2.9)

	R	R'	
$A_1$	$p_{11}$	$p_{12}$	$\pi_1$
$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$
$A_k$	$p_{k1}$	$p_{k2}$	$\pi_k$
	$q$	$1 - q$	$1$

The parameter of interest is now  $\underline{\Pi} = (\pi_1, \dots, \pi_k)$ , and the data is  $X = (x_1, \dots, x_k, n_2)$ . In place of (2.1), a priori, we consider that

$$(2.10) \quad \underline{p} = (p_{11}, \dots, p_{k1}, p_{12}, \dots, p_{k2}) \sim D(\alpha_{11}, \dots, \alpha_{k1}, \alpha_{12}, \dots, \alpha_{k2}).$$

Analogous to (2.2) and (2.3) the following reparametrization is considered:

$$(2.11) \quad q = \sum_{i=1}^k p_{i1}, \quad q_{i1} = \frac{p_{i1}}{q}, \quad q_{i2} = \frac{p_{i2}}{1-q} \quad (i = 1, \dots, k)$$

$$Q_1 = (q_{11}, \dots, q_{k1}), \quad \text{and} \quad Q_2 = (q_{12}, \dots, q_{k2}).$$

Conversely,

$$(2.12) \quad p_{i1} = qq_{i1}, \quad p_{i2} = (1-q)q_{i2} \quad (i = 1, \dots, k)$$

$$\text{and} \quad \underline{\Pi} = qQ_1 + (1-q)Q_2.$$

With the reparametrization (2.11) the likelihood is given by:

$$(2.13) \quad L = q^{n_1} (1-q)^{n_2} \prod_{i=1}^k q_{i1}^{x_i}.$$

Again, by Lemma 1, to consider (2.10) a priori is equivalent to considering the following set of conditions:

$$(2.14) \quad q \perp\!\!\!\perp Q_1 \perp\!\!\!\perp Q_2, \quad q \sim B(\alpha_{.1}, \alpha_{.2}),$$

$$Q_1 \sim D(\alpha_{11}, \dots, \alpha_{k1}), \quad \text{and} \quad Q_2 \sim D(\alpha_{12}, \dots, \alpha_{k2}),$$

where  $\alpha_{.j} = \sum_i \alpha_{ij}$  ( $j = 1, 2$ ).

By matching (2.14) with (2.13), we obtain the posterior distribution which is defined by the conditions below.

$$(2.15) \quad q \prod_1 Q_1 \prod_2 Q_2 | X, q | X \sim q | n_1 \sim B(\alpha_{.1} + n_1, \alpha_{.2} + n_2),$$

$$Q_1 | X \sim D(\alpha_{11} + x_1, \dots, \alpha_{k1} + x_1), \text{ and}$$

$$Q_2 | X \sim Q_2 \sim D(\alpha_{12}, \dots, \alpha_{k2}).$$

Again,  $p | X$  is distributed as Dirichlet if and only if  $n_2 = 0$ .

It is, however, a mixture of Dirichlet distributions and

$$(p_{11}), \dots, p_{k1}, (1 - q) | X \sim D(\alpha_{11} + x_1, \dots, \alpha_{k1} + x_1, \alpha_{.2} + n_2).$$

As before, we might consider a conjugate family of distributions by taking  $\beta \geq \alpha_{.2}$  for  $\alpha_{.2}$  in (2.14).

The Bayes estimator for the parameter of interest

$\underline{\Pi} = (\Pi_1, \dots, \Pi_k)$ , analogous to (2.7), has the following form:

$$(2.16) \quad \hat{\underline{\Pi}} = E\{\underline{\Pi} | X\} = \frac{1}{\alpha + n} [(\alpha_{1.}, \dots, \alpha_{k.}) + XM]$$

where  $M$  is a  $(k + 1) \times k$ -matrix with the  $(k + 1)$ th row being

$(\alpha_{12}, \dots, \alpha_{k2}) \frac{1}{\alpha_{.2}}$ , the diagonal elements being the unity, and the remaining elements being zero.

The next section deals with the study of the distribution of the data  $X$ . The covariance matrix of  $\hat{\underline{\Pi}}$  is presented at the end of the section.

### 3 - THE DIRICHLET - MULTINOMIAL DISTRIBUTION: PROPERTIES

When the discrete data follow the Multinomial model, the family of Dirichlet distributions has been widely used by Bayesians since it is a conjugate family large enough to accommodate various shades of prior opinion. The study of the mixture of Multinomial distributions by a Dirichlet distribution therefore becomes relevant because the (marginal) distribution of the data is then a mixture of this kind. Generalizing the definition of the Beta-Binomial (Ferguson [1967]) this mixture is called here the Dirichlet-Multinomial distribution. More specifically, for  $k \geq 2$ , let  $\underline{x} = (x_1, \dots, x_k)$  be a nonnegative integer random vector such that  $\sum_1^n x_i = n$  is fixed, and let  $\underline{p} = (p_1, \dots, p_k)$  be a nonnegative real random vector with  $\sum_1^k p_i = 1$ .

#### DEFINITION

If  $\underline{p} \sim D(\alpha_1, \dots, \alpha_k)$  and  $\underline{x}|\underline{p} \sim M(n; \underline{p})$ , then the distribution of  $\underline{x}$  is called Dirichlet-Multinomial (DM) with parameter  $(n; \alpha_1, \dots, \alpha_k)$ , and we write  $\underline{x} \sim DM(n; \alpha_1, \dots, \alpha_k)$ . When  $k = 2$ , in place of  $(x_1, x_2) \sim DM(n; \alpha_1, \alpha_2)$ , we write  $x_1 \sim BB(n; \alpha_1, \alpha_2)$  to indicate that  $x_1$  is distributed as Beta-Binomial.

It is easy to check that the probability function (p.f.) associated with the DM distribution is given by:

$$(3.1) \quad f(\underline{x}) = \frac{n! \Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{i=1}^k \frac{\Gamma(\alpha_i + x_i)}{x_i! \Gamma(\alpha_i)},$$

where  $\alpha = \sum_1^k \alpha_i$ .

Some of the important properties of the DM distributions are given below. Let  $\underline{x} = (x_1, \dots, x_k) \sim \text{DM}(n; \alpha_1, \dots, \alpha_k)$ .

PROPOSITION 1

If  $(i_1, \dots, i_k)$  is a permutation of  $(1, \dots, k)$ , then  $(x_{i_1}, \dots, x_{i_k}) \sim \text{DM}(n; \alpha_{i_1}, \dots, \alpha_{i_k})$ .

PROPOSITION 2

If  $m \in \{1, 2, \dots, k\}$  is fixed, then for  $\beta = \sum_{i=1}^m \alpha_i$   
 $(x_1, \dots, x_m, n - \sum_{i=1}^m x_i) \sim \text{DM}(n; \alpha_1, \dots, \alpha_m, \alpha - \beta)$ , and  
 $n_1 = \sum_{i=1}^m x_i \sim \text{BB}(n; \beta, \alpha - \beta)$ .

These two results are immediate consequences of analogous properties of the Multinomial and the Dirichlet distributions.

PROPOSITION 3

For  $m$  and  $n_1$  defined as above, we have that

$$(x_1, \dots, x_m) | n_1 \sim \text{DM}(n_1; \alpha_1, \dots, \alpha_k).$$

Proof

Note that the conditional probability function of  $(x_1, \dots, x_m) | n_1$  is obtained by dividing the p.f. of  $(x_1, \dots, x_m, n - n_1)$  by the p.f. of  $n_1$ , which is the p.f. of a  $\text{DM}(n; \alpha_1, \dots, \alpha_k)$ .  $\square$

The result we present next is an important characterization of the DM distribution which will be used in the sequel.

Let  $(x_1, \dots, x_k)$  be a nonnegative integer random vector with  $\sum_{i=1}^k x_i = n$  fixed. Choose an integer  $m \in \{2, \dots, k-1\}$ , and

denote  $n_1 = \sum_1^m x_i$  with  $n_2 = n - n_1$ . Consider now the following set of conditions:

$$(3.2) \quad \begin{aligned} & \text{(i)} \quad (x_1, \dots, x_m) \perp\!\!\!\perp (x_{m+1}, \dots, x_k) | n_1 \\ & \text{(ii)} \quad (x_1, \dots, x_m) | n_1 \sim \text{DM}(n_1; \alpha_1, \dots, \alpha_m), \\ & \quad (x_{m+1}, \dots, x_k) | n_1 \sim \text{DM}(n_2; \alpha_{m+1}, \dots, \alpha_k), \\ & \text{and (iii)} \quad n_1 \sim \text{BB}(n; \sum_1^m \alpha_i, \alpha - \sum_1^m \alpha_i). \end{aligned}$$

### THEOREM 1

The above set of conditions (3.2) are necessary and sufficient to have:

$$\text{(iv)} \quad (x_1, \dots, x_k) \sim \text{DM}(n; \alpha_1, \dots, \alpha_k).$$

### PROOF

By Propositions 1, 2, and 3 (iv)  $\Rightarrow$  (ii) and (iii). To prove the remaining implications we need only note that (3.1) may be factored as:

$$\begin{aligned} f(\underline{x}) = & \left[ \frac{n! \Gamma(\alpha)}{\Gamma(\alpha + n)} \frac{\Gamma(\beta + n_1) \Gamma(\alpha - \beta + n_2)}{n_1! n_2! (\beta) \Gamma(\alpha - \beta)} \right] \\ & \times \left[ \frac{n_1! \Gamma(\beta)}{\Gamma(\beta + n_1)} \prod_{i=1}^m \frac{\Gamma(\alpha_i + x_i)}{x_i! \Gamma(\alpha_i)} \right] \times \left[ \frac{n_2! \Gamma(\alpha - \beta)}{\Gamma(\alpha - \beta + n_2)} \prod_{i=m+1}^k \frac{\Gamma(\alpha_i + x_i)}{x_i! \Gamma(\alpha_i)} \right] \end{aligned}$$

where, as before,  $\alpha = \sum_1^k \alpha_i$ , and  $\beta = \sum_1^m \alpha_i$ . The first factor is the p.f. of a  $\text{BB}(n; \beta, \alpha - \beta)$ , the second is the p.f. of a  $\text{DM}(n_1; \alpha_{m+1}, \dots, \alpha_k)$ , and the third is the p.f. of a  $\text{DM}(n_2; \alpha_{m+1}, \dots, \alpha_k)$ .  $\square$

EXAMPLE

Recalling the Bayes estimator  $\hat{\pi}_1$  presented in (2.7), we notice that  $(x_1, x_2) \perp\!\!\!\perp (y_1, y_2) | n_1$ , and then  $y_1 | X \sim y_1 | n_2 \sim \text{BB}(n_2; \alpha_{12}, \alpha_{22})$  which implies (see (3.3) below) that  $E\{y_1 | X\} = E\{y_1 | n_2\} = n_2 \frac{\alpha_{12}}{\alpha_{.2}}$ .  $\square$

An interesting property of the DM distribution is given below where we consider the finite sequence  $(z_1, \dots, z_k)$  with  $z_j = \sum_{i=1}^j x_i$  ( $j = 1, \dots, k$ ). Clearly,  $z_1 = x_1$ ,  $z_m = n_1$ , and  $z_k = n$ .

COROLLARY

If  $(x_1, \dots, x_k) \sim \text{DM}(n; \alpha_1, \dots, \alpha_k)$ , then  $(z_1, \dots, z_k)$  forms a Markov Chain.

It is intuitive that we might give a characterization of the DM distribution in terms of  $(z_1, \dots, z_k)$ . This, however, would go beyond our needs.

To present the mean vector and the covariance matrix of the DM distribution we introduce the vector  $a = (\alpha_1, \dots, \alpha_k)$ , and the matrix

$$A = \begin{pmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_k \end{pmatrix}$$

From Propositions 1 and 2, we notice that  $x_i \sim \text{BB}(n; \alpha_i, \alpha - \alpha_i)$ , and  $x_i + x_j \sim \text{BB}(n; \alpha_i + \alpha_j, \alpha - \alpha_i - \alpha_j)$  for  $i, j = 1, \dots, k$  with



$i \neq j$ . From easy computations when using the definition of BB we have that

$$(3.3) \quad E\{x_i\} = n \frac{\alpha_i}{\alpha}$$

$$\text{Var}\{x_i\} = \alpha_i - \frac{\alpha_i^2}{\alpha} \frac{\alpha + n}{\alpha(\alpha + 1)} n,$$

$$\begin{aligned} \text{and } \text{Var}\{x_i + x_j\} &= [\alpha_i + \alpha_j - \frac{(\alpha_i + \alpha_j)^2}{\alpha}] \frac{\alpha + n}{\alpha(\alpha + 1)} n = \\ &= \text{Var}\{x_i\} + \text{Var}\{x_j\} + 2 \text{cov}\{x_i, x_j\}. \end{aligned}$$

From this last equation, it follows that

$$\text{cov}\{x_i, x_j\} = -\frac{\alpha_i \alpha_j}{\alpha} \frac{\alpha + n}{\alpha(\alpha + 1)} n.$$

Finally, the mean vector and the covariance matrix are given by:

$$E\{\underline{x}\} = \frac{n}{\alpha} \underline{a}$$

$$\text{Cov}\{\underline{x}\} = [A - \frac{1}{\alpha} \underline{a}'\underline{a}] \frac{\alpha + n}{\alpha(\alpha + 1)} n$$

where  $\underline{a}'$  is the transpose of  $\underline{a}$ .

The data vector  $X = (x_1, \dots, x_k, n_2)$ , for the nonresponse data presented in Section 2, follows the DM model; that is,

$X \sim \text{DM}(n; \alpha_{11}, \dots, \alpha_{k1}, \alpha_{.2})$ . In this case

$$(3.4) \quad a = (\alpha_{11}, \dots, \alpha_{k1}, \alpha_{.2}), \text{ and} \quad A = \begin{pmatrix} \alpha_{11} & 0 & \dots & 0 & 0 \\ 0 & \alpha_{21} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \alpha_{k1} & 0 \\ 0 & 0 & & 0 & \alpha_{.2} \end{pmatrix}.$$

The mean vector and the covariance matrix for  $\hat{\Pi}$ , the Bayes estimator given by (2.16), are:

$$(3.5) \quad E\{\hat{\Pi}\} = \frac{1}{\alpha + n} [(\alpha_{1.}, \dots, \alpha_{k.}) + E\{X\}M], \text{ and}$$

$$\text{Cov}\{\hat{\Pi}\} = \left(\frac{1}{\alpha + n}\right)^2 M' \text{Cov}\{X\}M.$$

Using (3.4), we have that

$$E\{\hat{\Pi}\} = \frac{1}{\alpha} (\alpha_{1.}, \dots, \alpha_{k.}),$$

$$M'AM = \begin{pmatrix} \alpha_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \alpha_{k1} \end{pmatrix} + \frac{1}{\alpha_{.2}} (\alpha_{12}, \dots, \alpha_{k2})' (\alpha_{12}, \dots, \alpha_{k2}),$$

$$\text{and } \frac{1}{\alpha} M'a'aM = \frac{1}{\alpha} (\alpha_{1.}, \dots, \alpha_{k.})' (\alpha_{1.}, \dots, \alpha_{k.}),$$

which imply

$$(3.6) \quad \text{Var}\{\hat{\Pi}_i\} = \frac{n}{\alpha(\alpha + 1)(\alpha + n)} \left[ \alpha_{i1} + \frac{\alpha_{i2}^2}{\alpha_{.2}} - \frac{1}{\alpha} \alpha_{i.}^2 \right]$$

$$\text{Cov}\{\hat{\Pi}_i, \hat{\Pi}_j\} = \frac{n}{\alpha(\alpha + 1)(\alpha + n)} \left[ \frac{\alpha_{i2} \alpha_{j2}}{\alpha_{.2}} - \frac{1}{\alpha} \alpha_{i.} \alpha_{j.} \right]$$

for  $i, j = 1, \dots, k$  and  $i \neq j$ .

To conclude this section we notice that some of the important discrete distributions are particular cases of the DM distribution. For example, for  $\alpha_1 = \dots = \alpha_k = 1$ , the probability function is  $f(\underline{x}) = \binom{n+k-1}{n}^{-1}$  which goes by the name Bose-Einstein statistic in Statistic Mechanics. See Feller (1968) for additional discussion.

#### 4 - THE DM DISTRIBUTION: A NATURAL FAMILY OF PRIORS FOR FINITE POPULATION STUDIES

A sample of fixed size  $n$  is taken from a population of finite size  $N$  which is partitioned in  $k \leq N$  categories. The category frequency counts are represented by  $\theta_1, \dots, \theta_k$  with  $\sum_1^k \theta_i = N$ . From the sample, an inference about  $(\theta_1, \dots, \theta_k)$  is required. Corresponding to each  $\theta_i$  ( $i = 1, \dots, k$ ),  $x_i$  is the sample frequency count of the  $i$ -th category where  $\sum_1^k x_i = n$ .

The above problem may be viewed in a simple way by considering a bag with  $N$  balls of  $k \leq N$  different colors that are identified by  $c_1, c_2, \dots, c_k$ . The number of balls with the  $i$ -th color ( $i = 1, \dots, k$ ) is represented by  $\theta_i$  where  $\sum_1^k \theta_i = N$ . Suppose that the  $N$  balls are separated into two bags in such a way that one bag (bag number 1) contains  $n$  balls, and the other (bag number 2)  $N-n$  balls. The statistician is allowed to look at the composition of bag 1 and record the numbers  $x_1, \dots, x_k$ , the frequency counts of the  $k$  colors. Now, the unknown quantities for the statistician are  $\theta_1 - x_1, \dots, \theta_k - x_k$ ; that is, the composition of bag 2.

We restrict the choice of the prior distribution for  $\underline{\theta}$  to the family of DM distributions. Given the sample  $\underline{x} = (x_1, \dots, x_k)$ , the composition of the first bag, we want to derive the posterior distribution of  $\underline{\theta} - \underline{x} = (\theta_1 - x_1, \dots, \theta_k - x_k)$ , the composition of the second bag. In order to reach this goal, we use only intuitive arguments since an algebraic analysis, besides being tedious (albeit easy), would bury the beauty of the argument.

Let  $e_1 = (1, 0, \dots, 0), \dots, e_k = (0, \dots, 0, 1)$  be the standard orthonormal basis for  $R^k$ . To each unit  $j$  ( $j = 1, \dots, N$ ) of the population  $P$ , we associate an incidence vector  $y_j$  which is equal to  $e_i$  if the color of  $j$  is  $c_i$ . More specifically, let  $P = \{1, \dots, N\}$  be an enumeration of the population of balls. Associated with  $P$  are the incidence vectors  $y_1, \dots, y_N$  described above. The unknown vector is  $\underline{\theta} = (\theta_1, \dots, \theta_k) = \sum_1^N y_j$ . We are considering the case where the sampling selection is noninformative. That is, the selection of the  $n$  balls (sample) from  $P$ , is based only on the labels  $1, \dots, N$ , which are themselves uninformative about the incidence vectors  $y_1, \dots, y_N$ .

A natural way to introduce the prior model  $\underline{\theta} \sim DM(N; \alpha_1, \dots, \alpha_k)$  is to consider a random vector  $\underline{p} = (p_1, \dots, p_k) \sim D(\alpha_1, \dots, \alpha_k)$ , and to stipulate that for  $j = 1, \dots, N$

$$y_j | \underline{p} \sim M(1; \underline{p}), \text{ and } y_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp y_N | \underline{p}.$$

In other words, given  $\underline{p}$  the  $y_j$ 's are i.i.d. with common distribution  $M(1; \underline{p})$ . Since  $(y_1, \dots, y_N)$  is an exchangeable finite sequence, without loss of generality, we can consider our sampled items as being the first  $n$  population items, say  $\{1, 2, \dots, n\}$ . That is, the two bags are  $\{1, \dots, n\}$  and  $\{(n+1), \dots, N\}$ . Now, the sample is represented by the vector  $\underline{x} = \sum_{i=1}^n y_i$ , and the unknown quantity of interest is the vector  $\underline{\theta} - \underline{x} = \sum_{i=n+1}^N y_i$ .

In terms of the pseudo parameter  $\underline{p}$  we then have, a priori, the following:

$$\begin{aligned}
 & \text{(i)} \quad \underline{p} \sim D(\alpha_1, \dots, \alpha_k), \\
 & \text{(ii)} \quad \underline{x} | \underline{p} \sim M(n; \underline{p}), \\
 (4.1) \quad & \text{(iii)} \quad \underline{x} \sim DM(n; \alpha_1, \dots, \alpha_k), \\
 & \text{(iv)} \quad (\underline{\theta} - \underline{x}) | \underline{p} \sim M(N - n; \underline{p}), \\
 & \text{(v)} \quad (\underline{\theta} - \underline{x}) \sim DM(N - n; \alpha_1, \dots, \alpha_k), \\
 & \text{(vi)} \quad (\underline{\theta} - \underline{x}) \perp\!\!\!\perp \underline{x} | \underline{p}. \\
 & \text{and (vii)} \quad \underline{\theta} | \underline{p} \sim M(N; \underline{p}).
 \end{aligned}$$

The result below is useful to our discussion.

#### LEMMA 2

If two independent random vectors  $X$ , and  $Y$  are such that  $X \sim M(n_1; \underline{p})$  and  $Y \sim M(n_2; \underline{p})$ , then the conditional distribution of  $X | X + Y$  is the Multivariate Hypergeometric with parameter  $(n_1 + n_2, n_1, X + Y)$ ; that is,  $X | X + Y \sim H(n_1 + n_2, n_1, X + Y)$ . (Note that this distribution does not depend on the value of  $\underline{p}$ .)

The following conclusion based on (4.1), and Lemma 2 is the first important result since defines the likelihood function:

$$(4.2) \quad \underline{x} | (\underline{\theta}, \underline{p}) \sim \underline{x} | \underline{\theta} \sim H(N, n, \underline{\theta})$$

From the Bayesian analysis of the multinomial case, we recall that  $\underline{p} | \underline{x} \sim D(\alpha_1 + x_1, \dots, \alpha_k + x_k)$ . On the other hand we notice that the conditional distribution of  $(\underline{\theta} - \underline{x}) | \underline{x}$  may be viewed as a composition of the distribution of  $(\underline{\theta} - \underline{x}) | \underline{p} \sim (\underline{\theta} - \underline{x}) | (\underline{p}, \underline{x})$  (see (4.1)) by the distribution of  $\underline{p} | \underline{x}$ . Now, by the definition of the DM distribution, we have that  $(\underline{\theta} - \underline{x}) | \underline{x} \sim DM(N - n; \alpha_1 + x_1, \dots, \alpha_k + x_k)$ . This is the main result of this section and may be summarized as:

#### THEOREM 2

For the finite population sampling described, if  $\underline{\theta} \sim DM(N; \alpha_1, \dots, \alpha_k)$  a priori, then  $(\underline{\theta} - \underline{x}) | \underline{x} \sim DM(N - n; \alpha_1 + x_1, \dots, \alpha_k + x_k)$  a posteriori.

The next section is devoted to the nonresponse problem in finite populations.

#### 5 - NONRESPONSE: THE MULTIVARIATE HYPERGEOMETRIC MODEL

The data for the nonresponse problem is presented in the  $k \times 2$ -tabular form as in (2.7). Instead of (1.4), the population parameters have the following representation:

(5.1)

	R	R'	
A <sub>1</sub>	θ <sub>11</sub>	θ <sub>12</sub>	θ <sub>1</sub>
A <sub>2</sub>	θ <sub>21</sub>	θ <sub>22</sub>	θ <sub>2</sub>
⋮	⋮	⋮	⋮
A <sub>k</sub>	θ <sub>k1</sub>	θ <sub>k2</sub>	θ <sub>k</sub>
	ψ	N - ψ	N

Now, the parameter of interest is  $\underline{\theta} = (\theta_1, \dots, \theta_k)$ , and the likelihood may be written as:

(5.2)

$$L = \frac{\prod_{i=1}^k \binom{\theta_i}{x_i}}{\binom{\psi}{n_1}} \frac{\binom{\psi}{n_1} \binom{N-\psi}{n_2}}{\binom{N}{n}}$$

Suppose that, a priori, a DM distribution for  $\theta = (\theta_{11}, \dots, \theta_{k1}, \theta_{12}, \dots, \theta_{k2})$  is considered; that is, a priori

(5.3)  $\theta \sim \text{DM}(N; \alpha_{11}, \dots, \alpha_{k1}, \alpha_{12}, \dots, \alpha_{k2})$ .

An additional notation is introduced below:

$$\theta_1 = (\theta_{11}, \dots, \theta_{k1}), \text{ and } \theta_2 = (\theta_{12}, \dots, \theta_{k2}).$$

Recall that,  $\underline{\theta} = \theta_1 + \theta_2$  is the parameter of interest, and that  $X = (x_1, \dots, x_k, n_2)$  is the data vector which may also be represented in the slightly abbreviated form  $\underline{x} = (x_1, \dots, x_k)$ .

$$\alpha_{.j} = \sum_{i=1}^k \alpha_{ij} \quad (j = 1, 2), \quad \alpha_{i.} = \alpha_{i1} + \alpha_{i2}, \quad (i = 1, \dots, k), \text{ and}$$

$$\alpha = \sum \sum \alpha_{ij}.$$

Writing the parameters in the extended form  $(\psi, \theta_1, \theta_2)$ , it is convenient to describe the prior (5.3) in the following equivalent form (see Theorem 1):

$$(5.4) \quad \begin{aligned} & \text{(i)} \quad \psi \sim \text{BB}(N; \alpha_{.1}, \alpha_{.2}), \\ & \text{(ii)} \quad \theta_1 | \psi \sim \text{DM}(\psi; \alpha_{11}, \dots, \alpha_{k1}), \\ & \quad \quad \theta_2 | \psi \sim \text{DM}(N - \psi; \alpha_{12}, \dots, \alpha_{k2}), \\ & \text{and (iii)} \quad \theta_1 \perp\!\!\!\perp \theta_2 | \psi \end{aligned}$$

The theorem presented below is the main result of this section. It allows a simple derivation of the Bayes estimator.

### THEOREM 3

The posterior distribution derived from the Bayes operation, when (5.2) is the likelihood and (5.4) defines the prior, is given by the following set of conditions:

$$(5.5) \quad \begin{aligned} & \text{(i)'} \quad (\psi - n_1) | X \sim (\psi - n_1) | n_1 \sim \text{BB}(N - n; \alpha_{.1} + n_1, \alpha_{.2} + n_2) \\ & \text{(ii)'} \quad (\theta_1 - \underline{x}) | (\psi, X) \sim \text{DM}(\psi - n_1; \alpha_{11} + x_1, \dots, \alpha_{k1} + x_k), \\ & \quad \quad \theta_2 | (\psi, X) \sim \theta_2 | \psi, \\ & \text{and (iii)'} \quad \theta_1 \perp\!\!\!\perp \theta_2 | (\psi, X) \end{aligned}$$

### PROOF

The second condition of (ii)' follows from the fact that the likelihood does not depend on  $\theta_2$  when  $\psi$  is known. This fact together with the prior condition (iii), implies (iii)'.

To prove (i)' and the first condition of (ii)' we consider (as in Section 2) the invisible nonresponse sample frequency counts,



say  $\underline{y} = (y_1, \dots, y_k)$ . If we had full response, the data would have been represented by  $(\underline{x}, \underline{y})$ . From Theorems 1 and 2 we have that

$$a) \quad \psi - n_1 | (\underline{x}, \underline{y}) \sim \text{BB}(N - n; \alpha_{.1} + n_1, \alpha_{.2} + n_2)$$

$$b) \quad \theta_1 - \underline{x} | (\psi, \underline{x}, \underline{y}) \sim \text{DM}(\psi - n_1; \alpha_{11} + x_1, \dots, \alpha_{k1} + x_k)$$

From a) and b) it follows that  $(\psi - n_1) | (\underline{x}, \underline{y}) \sim (\psi - n_1) | n_1$ , and that  $(\theta_1 - \underline{x}) | (\psi, \underline{x}, \underline{y}) \sim (\theta_1 - \underline{x}) | (\psi, X)$  which imply (i)' and the first condition of (ii)' respectively.  $\square$

Note that we showed above that  $\psi \perp\!\!\!\perp X | n_1$ ; that is,  $n_1$  is partially Bayes sufficient to predict  $\psi$ . See Basu (1977) for a more complete discussion of this concept.

As in the multinomial case, the posterior (5.5) does not define a distribution in the same class as the prior was chosen from; that is, (5.5) does not define a DM distribution. It is easy to check, however, that  $(\theta_{11} - x_1, \dots, \theta_{k1} - x_k, N - \psi - n_2) \sim \text{DM}(N - n; \alpha_{11} + x_1, \dots, \alpha_{k1} + x_k, \alpha_{.2} + n_2)$ . A more complete class might be considered by taking in (5.4) a  $\beta \geq \alpha_{.2}$  for  $\alpha_{.2}$  in (i).

From the posterior (5.5) we obtain the following results:

$$E\{\psi - n_1 | X\} = (N - n) \frac{\alpha_{.1} + n_1}{\alpha + n}$$

$$E\{N - \psi | X\} = n_2 + (N - n) \frac{\alpha_{.2} + n_2}{\alpha + n}$$

$$E\{\theta_{i1} - x_i | (\psi, X)\} = (\psi - n_1) \frac{\alpha_{i1} + x_i}{\alpha_{.1} + n_1}$$

$$E\{\theta_{i2} | (\psi, X)\} = E\{\theta_{i2} | \psi\} = (N - \psi) \frac{\alpha_{i2}}{\alpha_{.2}}$$

Using now the properties of conditional expectation we have the Bayes estimators,

$$\begin{aligned} \hat{\theta}_i &= E\{\theta_i | X\} = E\{\theta_{i1} + \theta_{i2} | X\} \\ &= E\{\theta_{i1} | X\} + E\{\theta_{i2} | X\} \\ &= x_i + \frac{\alpha_{i1} + x_i}{\alpha_{.1} + n_1} E\{\psi - n_1 | X\} + \frac{\alpha_{i2}}{\alpha_{.2}} E\{N - \psi | X\} \\ &= \frac{\alpha + N}{\alpha + n} \left( x_i + n_2 \frac{\alpha_{i2}}{\alpha_{.2}} \right) + (N - n) \frac{\alpha_{i.}}{\alpha + n} \end{aligned}$$

Similarly to (2.15), the Bayes estimator of the parameter of interest  $\underline{\theta} = \theta_1 + \theta_2$  is given by:

$$\underline{\hat{\theta}} = E\{\underline{\theta} | X\} = \frac{\alpha + N}{\alpha + n} XM + \frac{N - n}{\alpha + n} (\alpha_{1.}, \dots, \alpha_{k.})$$

Using the results (3.4), and (3.5) we finally have:

$$E\{\underline{\hat{\theta}}\} = \frac{N}{\alpha} (\alpha_{1.}, \dots, \alpha_{k.})$$

$$\text{Cov}\{\underline{\hat{\theta}}\} = \left( \frac{\alpha + N}{\alpha + n} \right)^2 M' \text{Cov}\{X\}M$$

which implies that

$$(5.6) \quad \text{Cov}\{\hat{\theta}_i, \hat{\theta}_j\} = \frac{n(\alpha + N)^2}{(\alpha + n)(\alpha + 1)\alpha} \left( \delta_{ij} \alpha_{i1} + \frac{\alpha_{i2} \alpha_{j2}}{\alpha_{.2}} - \frac{\alpha_{i.} \alpha_{j.}}{\alpha} \right),$$

where  $\delta_{ij}$  is the Kronecker delta, and  $\text{Var}\{\theta_i\} = \text{Cov}\{\theta_i, \theta_i\}$ .

#### 6 - FINAL REMARKS

(i) There are many follow-up techniques used to obtain response among some of the  $n_2$  units that have not responded initially. For example, from the  $n_2$  nonrespondents in our sample, we select a subsample of size  $n'_2 \leq n_2$  and offer an incentive to those who now would respond. In that way, information about  $Q_2$  in (2.14) or about  $\theta_2|\psi$  in (5.4) might be improved. See Kaufman and King (1973), and Singh and Sedrausk (1978) for a more specific discussion on this two stage sampling.

(ii) Although we have restricted ourselves to the nonresponse problem, it should be understood that our method applies equally well to the general problem of categorical data with missing entries. Consider, for instance, the categorical data where all but the first  $n$  cell entry data are missing. By using Lemma 1 for the multinomial case or Theorem 1 for the hypergeometric case, we would, analogously to (2.6) or (5.4), obtain the posterior distribution for the cell parameters.

(iii) One word about the relevance of the variance of Bayes estimators as presented in (3.6), and (5.6). Note that we are not talking about conditional variances (with the parameter fixed) but the variance of the marginal distribution of the estimator. Consider the  $k = 2$  multinomial case for instance. It is clear that  $\text{Var}\{\hat{\pi}_1\} = \text{Var}\{\pi_1\} - E\{\text{Var}\{\pi_1|X\}\}$ ; that is, the variance of  $\hat{\pi}_1$  may be

regarded as the expected amount of uncertainty removed, when uncertainty (De Groot (1962)) about the parameter is measured by its variance. Thus, the variance of the Bayes estimator is a kind of a measure of the amount of information in the experiment. The larger the variance of  $\hat{\Pi}_1$  is, the better off we are!

The variance of the Bayes estimator may be used (see Appendix) to study the amount of information lost when the nonresponse portion of the sample is neglected as in many classical procedures.

## REFERENCES

1. BASU, D. (1977). On the elimination of nuisance parameters. JASA, 72, 355-66.
2. BASU, D. (1979). A discussion on survey theory. Symposium on Incomplete Data, August 10-11, Washington, D.C..
3. DEGROOT, M. H. (1962). Uncertainty, information, and sequential experiments. Ann. Math. Statistics., 33, 404-19.
4. FELLER, W. (1968). An Introduction to Probability Theory and Its Applications. Vol. I, 3rd ed. Wiley, N.Y..
5. FERGUSON, T. S. (1967). Mathematical Statistics: A Decision Theoretic Approach. Academic Press Inc. N.Y..
6. KAUFMAN, G. M., and KING, B. (1973). A Bayesian analysis of nonresponse in dichotomous processes. JASA, 68, 670-678.
7. SINGH, B., and SEDRANSK, J. (1978). A two-phase sample design for estimating the finite population mean when there is nonresponse. In N.K. Namboodiri, Ed., Survey Sampling and Measurement. Academic Press Inc. N.Y..

APPENDIX

The usual non-Bayesian methods for analyzing data with nonresponse do not use the nonresponse portion of the sample. The likelihood in this case is defined by the conditional probability of  $\underline{x} = (x_1, \dots, x_k)$ , the nonresponse vector, given  $n_1 = \sum_{i=1}^k x_i$ . For instance, in the multinomial case this "conditional" likelihood is  $L_c = \prod_{i=1}^k q_{i1}^{x_i}$ . It is intuitive that, by considering this reduction, one is not using the complete information (about the parameter of interest) contained in the data. In order to clarify this point we define a reasonable measure of information and compute, in a particular case, its values for both the original and the conditional model.

Consider the Multinomial model for the case of two categories ( $k = 2$ ). Let the prior be the uniform distribution; that is,  $\alpha_{11} = \alpha_{21} = \alpha_{12} = \alpha_{22} = 1$ . By using the variance (see Section 6, iii) as the uncertainty function, we define the measure of information as:

$$I(\text{data}) = (\text{Var}\{\Pi_1\})^{-1} \text{Var}\{E\{\Pi_1 | \text{data}\}\}.$$

Considering the original likelihood, the information measure is given by  $I(X) = I = \frac{n}{2(4+n)}$  since  $\text{Var}\{\Pi_1\} = (20)^{-1}$  and  $\text{Var}\{\hat{\Pi}_1\} = \frac{n}{40(4+n)}$ .

The posterior distribution under the conditional model (and some prior) is defined by the following conditions:

$$q' \sim B(2, 2), \quad q'_{11} \sim B(1 + x_1, 1 + x_2)$$

$$q'_{12} \sim B(1, 1), \quad \text{and } q' \perp\!\!\!\perp q'_{11} \perp\!\!\!\perp q'_{12}.$$

The Bayesian estimator in this case is given by

$$\frac{1}{2} \frac{1 + x_1}{2 + n_1} + \frac{1}{4},$$

and the respective measure of information is

$$I(\underline{x}|n_1) = I_c = 5 \text{ Var}\left\{\frac{1 + x_1}{2 + n_1}\right\}.$$

Relative to the uniform prior, the distribution of  $(x_1, x_2, n_2)$  is  $DM(n; 1, 1, 2)$ . Considering the particular case of  $n = 4$  we obtain the following results:

$$I = \frac{1}{4}, \quad I_c = \frac{51}{280}, \quad \text{and } \frac{I - I_c}{I} \doteq .27.$$

Here we might say that if an inference about  $\Pi_1$  is required, then 27% of the information is expected to be lost (relatively) when the nonresponse portion is neglected.

Note that it is possible to have an analogous analysis for the Hypergeometric model. However, in addition to the value of  $n$ , we would have to fix a value for  $N$ , the population size. Here, the conditional model is given by  $L_c = \binom{\psi}{n_1}^{-1} \prod_{i=1}^k \binom{\theta_i}{x_i}$ . For particular choices of  $N$ , the relative loss of information would appear to be more extreme.

CHAPTER III. THE INFLUENCE OF THE SAMPLE ON THE  
POSTERIOR DISTRIBUTION

1 - INTRODUCTION

As before,  $\underline{\theta}$  is the parameter of interest and  $\underline{x}$  is the data on which the inference about  $\underline{\theta}$  is based. The Bayesian operation (prior to posterior) supplies the answer to the question of how to use the information (about  $\underline{\theta}$ ) provided by the data. In particular, in Chapter 2 we showed how the nonresponse sample portion (being a source of information) is incorporated into the analysis. Now, we shift our attention to another general question: What kind of information about  $\underline{\theta}$  does the sample possess?

Consider again the population of colored balls separated randomly into two urns, as described in Section 4 of Chapter 2. The composition  $\underline{x}$  (the data) of urn 1 is expected to reflect the composition  $\underline{\theta}$  of the population in the sense that the larger  $x_i$  is, the larger stochastically will  $\theta_i$  be a posteriori. We also usually expect this relationship (a posteriori) to apply to the composition  $\underline{\theta} - \underline{x}$  of urn 2. This is clearly not true. For example, if we know in advance the exact value of  $\theta_1$ , then  $\theta_1 - x_1$  decreases as  $x_1$  increases. Whitt (1979) considers the case of two colors ( $k = 2$ ), and restricts himself to the class of exchangeable priors for  $y_1, \dots, y_N$ , the incidence vectors described in Chapter 2.



He shows, by using the Monotone Likelihood Ratio (MLR) ordering, that  $\theta_1$  always increases when  $x_1$  increases, and presents conditions on the prior under which  $\theta_1 - x_1$  increases (or decreases) when  $x_1$  increases. Note that by considering only exchangeable priors, and by using (4.1) of Chapter 2 we may conclude that: The posterior distribution of  $\underline{\theta} - \underline{x}$  (the unobserved part of the population) given  $\underline{x}$  (the data) is independent of the value of  $\underline{x}$  for every sample size  $n$  if and only if the prior distribution of  $\underline{\theta}$  is Multinomial with fixed parameter  $(N; \underline{p})$ . This is the multivariate generalization of the Corollary presented by Whitt (1979).

In the present chapter, we show that Whitt's key ideas may be extended to the case of multivariate absolutely continuous distributions. Our basic notions are multivariate total positivity of order 2, and multivariate reverse rule of order 2, introduced and explained by Karlin and Rinott (1980a, b). These concepts are briefly described in Section 2.

In Section 3, we present sufficient conditions on the likelihood function and on the prior distribution under which the posterior distribution insures  $\theta_1$  be increasing in  $x_1$  and decreasing in  $x_j$  for  $j \neq 1$ . In Section 4, we apply these results to some well known distributions.

## 2 - PRELIMINARIES

In this section we present definitions, notation, and basic facts used throughout this chapter.

DEFINITION 1

A random vector  $\underline{x}$  is said to be stochastically increasing in a random vector  $\underline{y}$  if  $E\{\phi(\underline{x})|\underline{y}\}$  is increasing in  $\underline{y}$  for every increasing bounded real function  $\phi$ . (A function  $\phi: \mathcal{R}^k \rightarrow \mathcal{R}$  is said to be increasing if it is increasing in each of its arguments.)

The following concepts of total positivity of order 2 ( $TP_2$ ), reverse rule of order 2 ( $RR_2$ ), and Pólya frequency function of order 2 ( $PF_2$ ) may be found in Karlin (1968).

DEFINITION 2

(i) A nonnegative real function  $f: \mathcal{R}^2 \rightarrow \mathcal{R}$  is  $TP_2$  ( $RR_2$ ) if

$$f(x_1, x_2)f(x'_1, x'_2) \geq (\leq) f(x_1, x'_2)f(x'_1, x_2),$$

whenever  $x'_1 \geq x_1$ , and  $x'_2 \geq x_2$ .

(ii) A nonnegative real function  $\zeta: \mathcal{R} \rightarrow \mathcal{R}$  is  $PF_2$  if  $f(x_1, x_2) = \zeta(x_1 - x_2)$  is  $TP_2$ .

The definitions below appear in Karlin and Rinott (1980a, b). For every  $\underline{x}, \underline{y} \in \mathcal{R}^k$ , denote:

$$\underline{x} \geq \underline{y} \text{ if } x_i \geq y_i \quad \forall i = 1, \dots, k,$$

$$\underline{x} \vee \underline{y} = (\max(x_1, y_1), \dots, \max(x_k, y_k)),$$

$$\text{and } \underline{x} \wedge \underline{y} = (\min(x_1, y_1), \dots, \min(x_k, y_k)).$$

The following is the natural generalization of Definition 2(i):

DEFINITION 3

Consider a nonnegative real function  $f: \mathbb{R}^k \rightarrow \mathbb{R}$ . We say that  $f(x)$  is multivariate totally positive of order 2 or  $MTP_2$  (multivariate reverse rule of order 2 or  $MRR_2$ ) if:

$$f(\underline{x} \vee \underline{y}) f(\underline{x} \wedge \underline{y}) \geq (\leq) f(\underline{x})f(\underline{y})$$

for every  $\underline{x}, \underline{y} \in \mathbb{R}^k$ .

Karlin and Rinott (1980a) show that  $MTP_2$  is a concept of strong positive dependence. They show, however, in their second report (1980b) that the  $MRR_2$  property fails to be a good concept of negative dependence. In the same report they seem to solve the problem by introducing the following definition. For additional illustration of its usefulness we refer to Block, Savits, and Shaked (1980).

Let  $(i_1, \dots, i_k)$  be any permutation of  $(1, 2, \dots, k)$ .

DEFINITION 4

An  $MRR_2$  function  $f: \mathbb{R}^k \rightarrow \mathbb{R}$  is said to be strongly- $MRR_2$  (S- $MRR_2$ ) if for any set of  $k$   $PF_2$  functions  $\{\zeta_1, \dots, \zeta_k\}$ , and for each  $j \leq k$ , the function

$$g_{k-j}(x_{i_{j+1}}, \dots, x_{i_k}) = \int \cdots \int f(\underline{x}) \prod_{m=1}^j \zeta_m(x_{i_m}) dx_{i_m}$$

is  $MRR_2$  whenever the integral exists.

The multivariate generalization of the familiar Monotone Likelihood Ratio property is given by:

DEFINITION 5

Let  $f_1$  and  $f_2$  be two probability density or mass functions (p.d.f.) on  $\mathbb{R}^k$ . If for every  $\underline{x}, \underline{y} \in \mathbb{R}^k$ ,

$$f_2(\underline{x} \vee \underline{y}) f_1(\underline{x} \wedge \underline{y}) \geq f_1(\underline{x}) f_2(\underline{y}),$$

then we say that  $f_2$  is "larger than"  $f_1$  in the  $TP_2$  sense, and write

$$f_2 >_{TP_2} f_1.$$

It is well known that the monotone likelihood ratio ordering implies stochastic ordering. The following result is a generalization of this fact.

THEOREM 1

Let  $f_1$  and  $f_2$  be two p.d.f.'s on  $\mathbb{R}^k$  such that  $f_2 >_{TP_2} f_1$ .

If  $\phi: \mathbb{R}^k \rightarrow \mathbb{R}$  is an increasing function, then

$$\int \cdots \int \phi(\underline{x}) f_1(\underline{x}) d\underline{x} \leq \int \cdots \int \phi(\underline{x}) f_2(\underline{x}) d\underline{x}.$$

For a proof of this result, see Karlin and Rinott (1980a).

3 - THEORETICAL RESULTS

In the sequel, let  $\underline{\theta}$  be a parameter taking values in a subset  $\Theta$  (the parameter space) of  $\mathbb{R}^k$ , and  $\underline{x} \in \mathbb{R}^n$  ( $n \geq k$ ) be the data vector.

The likelihood function or the sample p.d.f. is represented by

$f(\underline{x}|\underline{\theta})$ . The prior and the posterior p.d.f.'s are denoted

respectively by  $\xi(\underline{\theta})$  and  $\xi^*(\underline{\theta}|\underline{x})$ .

The following result is the  $TP_2$  version of Theorem 4 of Whitt (1979). In the above notation suppose that  $k = n = 1$ ,  $\theta$  denotes the parameter, and  $x$  denotes the data.

THEOREM 2

Consider  $f(x|\theta)$  and  $\xi^*(\theta|x)$  as bivariate real functions of  $\theta$  and  $x$ . Then  $f(x|\theta)$  is  $TP_2(RR_2)$  if and only if  $\xi^*(\theta|x)$  is  $TP_2(RR_2)$ .

PROOF

For  $[H(x)]^{-1} = \int f(x|\theta)\xi(\theta)d\theta$  we notice that  $\xi^*(\theta|x) = H(x)\xi(\theta)f(x|\theta)$ . Then,  $\frac{\xi^*(\theta|x)}{\xi^*(\theta'|x)} \geq(\leq) \frac{\xi^*(\theta|x')}{\xi^*(\theta'|x')}$  holds if and only if  $\frac{f(x|\theta)}{f(x|\theta')} \geq(\leq) \frac{f(x'|\theta)}{f(x'|\theta')}$ .  $\square$

Theorem 2 motivates the results of this section.

Let  $h: R^n \rightarrow R$  and  $g: R^{n+k} \rightarrow R$  be two nonnegative real functions.

THEOREM 3

Suppose that

- (a)  $f(\underline{x}|\underline{\theta}) = h(\underline{x}) g(\underline{x}, \underline{\theta})$ , where  $g$  is  $MTP_2$ ,  
and (b)  $\xi(\underline{\theta})$  is  $MTP_2$ .

Then

$$\xi^*(\underline{\theta}|\underline{x}') >_{TP_2} \xi^*(\underline{\theta}|\underline{x})$$

for every  $\underline{x}, \underline{x}' \in R^n$  such that  $\underline{x}' \geq \underline{x}$ .

PROOF

The posterior p.d.f. may be factored as:

$$\xi^*(\underline{\theta}|\underline{x}) = H(\underline{x}) g(\underline{x}, \underline{\theta}) \xi(\underline{\theta}),$$

where  $[H(\underline{x})]^{-1} = \int g(\underline{x}, \underline{\theta}) \xi(\underline{\theta}) d\underline{\theta}.$

Consider two sample points  $\underline{x}'$  and  $\underline{x}$  such that  $\underline{x}' \geq \underline{x}$ , and denote

$$\xi_0(\underline{\theta}) = \xi^*(\underline{\theta}|\underline{x}) \text{ and } \xi_1(\underline{\theta}) = \xi^*(\underline{\theta}|\underline{x}').$$

Let  $\underline{\theta}$  and  $\underline{\theta}'$  be two points in the parameter space  $\Theta$ . Now, we can write

$$\begin{aligned} \xi_0(\underline{\theta}) \xi_1(\underline{\theta}') &= H(\underline{x}) H(\underline{x}') g(\underline{x}, \underline{\theta}) g(\underline{x}', \underline{\theta}') \xi(\underline{\theta}) \xi(\underline{\theta}') \\ &\leq H(\underline{x}) H(\underline{x}') g(\underline{x} \wedge \underline{x}', \underline{\theta} \wedge \underline{\theta}') g(\underline{x} \vee \underline{x}', \underline{\theta} \vee \underline{\theta}') \xi(\underline{\theta} \wedge \underline{\theta}') \xi(\underline{\theta} \vee \underline{\theta}') \end{aligned}$$

since both  $g$  and  $\xi$  are  $MTP_2$ . Since  $\underline{x}' \geq \underline{x}$ , we finally have

$$\begin{aligned} \xi_0(\underline{\theta}) \xi_1(\underline{\theta}') &\leq [H(\underline{x}) g(\underline{x}, \underline{\theta} \wedge \underline{\theta}') \xi(\underline{\theta} \wedge \underline{\theta}')] \times \\ &\quad [H(\underline{x}') g(\underline{x}', \underline{\theta} \vee \underline{\theta}') \xi(\underline{\theta} \vee \underline{\theta}')] \\ &= \xi_0(\underline{\theta} \wedge \underline{\theta}') \xi_1(\underline{\theta} \vee \underline{\theta}'). \end{aligned}$$

Thus, it is equivalent to say that  $\xi_1 >_{TP_2} \xi_0$ .  $\square$

COROLLARY

If conditions (a) and (b) of Theorem 3 hold, then  $\underline{\theta}$  is stochastically increasing in  $\underline{x}$ .

PROOF

The result follows immediately from Theorems 1 and 3.  $\square$

In many cases, conditions (a) and (b) of Theorem 3 are too strong. A more realistic result is presented below where  $\underline{x}$  is considered to be the data reduced by sufficiency, and to have the same dimension of  $\underline{\theta}$ ; that is,  $n = k$ .

Let  $c: \mathbb{R}^k \rightarrow \mathbb{R}$ , and  $g_i: \mathbb{R}^2 \rightarrow \mathbb{R}$  ( $i = 1, \dots, k$ ) be nonnegative functions and represent the posterior marginal p.d.f. of  $\theta_i$  ( $i = 1, \dots, k$ ) by  $\xi_i^*(\theta_i | \underline{x})$ .

THEOREM 4

Suppose that

$$f(\underline{x} | \underline{\theta}) = h(\underline{x}) \prod_{i=1}^k g_i(x_i, \theta_i) c(\underline{\theta})$$

where, for  $i = 1, \dots, k$ ,  $g_i$  is  $TP_2$ . Then, for every  $i = 1, \dots, k$ , the following condition (for the posterior marginal density of  $\theta_i$ ) holds:

$$\xi_i^*(\theta_i | \underline{x}') >_{TP_2} \xi_i^*(\theta_i | \underline{x})$$

for  $\underline{x}'$  equal to  $\underline{x}$  except for the  $i$ -th coordinate, where  $x'_i (\geq x_i)$  replaces  $x_i$ .

PROOF

Without loss of generality we assume  $i = 1$ . The posterior marginal p.d.f. of  $\theta_1$  is

$$\xi_1^*(\theta_1 | \underline{x}) = H(\underline{x}) g_1(x_1, \theta_1) \int \cdots \int C(\underline{\theta}) \prod_{i=2}^k g_i(x_i, \theta_i) d\theta_i,$$

where  $C(\underline{\theta}) = c(\underline{\theta}) \xi(\underline{\theta})$ , and  $H(\underline{x})$  is defined as above. Now, define

$$G_1(\underline{x}, \theta_1) = \int \cdots \int C(\underline{\theta}) \prod_{i=2}^k g_i(x_i, \theta_i) d\theta_i,$$

which is constant in  $x_1$ . Thus,

$$\xi^*(\theta_1 | \underline{x}) = H(\underline{x}) g_1(x_1, \theta) G_1(\underline{x}, \theta_1)$$

Consider two sample points that differ only in the first coordinate, say

$$\underline{x} = (x_1, x_2, \dots, x_k), \text{ and } \underline{x}' = (x'_1, x_2, \dots, x_k),$$

where  $x'_1 \geq x_1$ .

Define

$$\xi_{10}(\theta_1) = \xi_1^*(\theta_1 | \underline{x})$$

$$\text{and } \xi_{11}(\theta_1) = \xi_1^*(\theta_1 | \underline{x}').$$

Since by definition

$$G_1(\underline{x}, \theta_1) = G_1(\underline{x}', \theta_1),$$

we thus have

$$\frac{\xi_{11}(\theta_1)}{\xi_{10}(\theta_1)} = \frac{H(\underline{x}') g_1(x'_1, \theta_1)}{H(\underline{x}) g_1(x_1, \theta_1)},$$

which is increasing in  $\theta_1$  by the  $TP_2$  property of  $g_1$ . Hence,

$$\xi_{11} > TP_2 \xi_{10}. \quad \square$$



REMARKS

1 - Note that the result in Theorem 4 pertains to the posterior marginal p.d.f. of  $\theta_i$ . Also it holds irrespective of the choice of the prior distribution.

2 - It follows from Theorems 1 and 4 that  $E\{\phi(\theta_i)|\underline{x}\}$  is increasing in  $x_i$  for every increasing real function  $\phi: R \rightarrow R$ .

In many applications, we noted that the posterior distribution of  $\theta_i$  stochastically decreases in  $x_j$  for every  $j \neq i$ . This fact is included in the following result.

THEOREM 5

Suppose that

$$\xi^*(\theta|\underline{x}) = H(\underline{x}) \prod_{i=1}^k g_i(x_i, \theta_i) C(\theta),$$

where

- (i)  $g_i(\theta_i, x_i)$  is  $TP_2 \forall i = 1, \dots, k$ ,
- (ii) for fixed  $x_i (i = 1, \dots, k)$ ,  $g_i(x_i, \theta_i)$  is  $PF_2$  (in  $\theta_i$ ),
- and (iii)  $C(\theta)$  is  $S\text{-}MRR_2$ .

Then, for every  $i = 1, \dots, k$ ,

$$\xi_i^*(\theta_i|\underline{x}) >_{TP_2} \xi_i^*(\theta_i|\underline{x}')$$

whenever  $\underline{x}' \geq \underline{x}$  and the  $i$ -th coordinates of  $\underline{x}'$  and  $\underline{x}$  are equal;

that is,  $x_i = x_i'$  and  $x_j \leq x_j' \forall j \neq i$ .

(For  $k = 2$ , condition (ii) is not required.)

PROOF

Without loss of generality we assume  $i = 1$  and

$$\underline{x} = (x_1, x_2, x_3, \dots, x_k) \text{ and}$$

$$\underline{x}' = (x_1, x_2', x_3, \dots, x_k),$$

where  $x_2' > x_2$ ; that is,  $\underline{x}$  and  $\underline{x}'$  differ only in the second coordinate.

(A) We consider first the case of  $k > 2$ .

The posterior marginal p.d.f. of  $\theta_1$  is

$$\xi_1^*(\theta_1 | \underline{x}) = H(\underline{x}) g_1(x_1, \theta_1) \int g_2(x_2, \theta_2) G(\theta_1, \theta_2, x_3, \dots, x_k) d\theta_2$$

where

$$G(\theta_1, \theta_2, x_3, \dots, x_k) = \int \dots \int C(\underline{\theta}) \prod_{i=1}^k g_i(x_i, \theta_i) d\theta_i.$$

Since  $C(\underline{\theta})$  is S-MRR<sub>2</sub> and  $g_i(\theta_i, x_i)$  is PF<sub>2</sub> in  $\theta_i$ , it follows from Definition 4 that  $G$  is RR<sub>2</sub> in  $(\theta_1, \theta_2)$  for every fixed  $(x_3, \dots, x_k)$ .

By the basic composition formula (Karlin [1968]) and the fact that  $g_2(x_2, \theta_2)$  is TP<sub>2</sub>, it follows that

$$G_1(\theta_1, \underline{x}) = \int g_2(x_2, \theta_2) G(\theta_1, \theta_2, x_3, \dots, x_k) d\theta_2$$

is RR<sub>2</sub> in  $(\theta_1, x_2)$  for every fixed  $(x_3, \dots, x_k)$ . (Note that  $G_1$  is constant in  $x_1$ .)

As before, let

$$\xi_{10}(\theta_1) = \xi_1^*(\theta_1 | \underline{x}), \text{ and } \xi_{11}(\theta_1) = \xi_1^*(\theta_1 | \underline{x}')$$

and note that

$$\frac{\xi_{11}(\theta_1)}{\xi_{10}(\theta_1)} = \frac{H(\underline{x}') G_1(\theta_1, \underline{x}')}{H(\underline{x}) G_1(\theta_1, \underline{x})}$$

is decreasing in  $\theta_1$  since  $G_1$  is  $RR_2$  in  $(\theta_1, x_2)$ . Hence,

$$\xi_{10} > TP_2 \xi_{11}.$$

(B) When  $k = 2$ , from (iii),  $C(\theta_1, \theta_2)$  is  $RR_2$  and by the basic composition formula,

$$G_1(\theta_1, x_2) = \int g_2(\theta_2, x_2) C(\theta_1, \theta_2) d\theta_2$$

is  $RR_2$ . Thus, if  $x_2' > x_2$

$$\frac{\xi_1^*(\theta_1 | (x_1, x_2'))}{\xi_1^*(\theta_1 | (x_1, x_2))} = \frac{H(x_1, x_2') G_1(\theta_1, x_2')}{H(x_1, x_2) G_1(\theta_1, x_2)}$$

is decreasing in  $\theta_1$  and the result follows.  $\square$

#### REMARKS

3 - Note that Theorem 5 involves conditions on the prior distribution.

4 - It follows from Theorems 1 and 5 that  $E\{\phi(\theta_i) | \underline{x}\}$  is decreasing in  $x_j$  whenever  $j \neq i$  and  $\phi: \mathcal{R} \rightarrow \mathcal{R}$  is increasing.

#### 4 - APPLICATIONS

In this section, we show how the results of Section 3 apply to some important probability distributions.

EXAMPLE 1 - MULTIVARIATE NORMAL DISTRIBUTION

Let  $\underline{x}$  be a k-dimensional random vector whose coordinates,  $x_i$  ( $i = 1, \dots, k$ ), are independent. Suppose that for  $i = 1, \dots, k$ ,  $\theta_i = E\{x_i\}$  is unknown and  $\sigma_i^2 = \text{Var}\{x_i\}$  is known. Then, whatever appropriate prior we choose, Theorem 4 applies, and by Remark 2,  $E\{\phi(\theta_i) | \underline{x}\}$  is increasing in  $x_i$  for every increasing real function  $\phi$ .

Suppose that our prior opinion about  $\underline{\theta}$  is represented by a nonsingular k-dimensional normal distribution, with mean vector  $\mu$  and covariance matrix  $V^{-1}$ , which is a conjugate prior. Let  $v_{ij}$  ( $i, j = 1, \dots, k$ ) be the  $(i, j)$ -th element of  $V$ . If  $v_{ij} \leq 0$  for every  $i \neq j$ , then  $\xi(\underline{\theta})$ , the prior p.d.f., is  $\text{MTP}_2$  (Barlow and Proschan [1975]). Thus, Theorem 3 and its corollary apply yielding the conclusion that  $\underline{\theta}$  is stochastically increasing in  $\underline{x}$ .

If the prior  $\xi(\underline{\theta})$  is negatively dependent in the  $\text{S-MRR}_2$  sense, then Theorem 5 applies, and by Remark 4, for every  $i = 1, \dots, k$ ,  $E\{\phi(\theta_i) | \underline{x}\}$  is decreasing in  $x_j$  ( $\forall j \neq i$ ) for every increasing real function  $\phi$ . A normal prior is  $\text{S-MRR}_2$  if

$$V^{-1} = D - \alpha'\alpha$$

where  $D$  is a positive definite diagonal matrix, say

$D = \text{diag}(d_1, \dots, d_k)$  with  $d_i > 0$ , and  $\alpha = (\alpha_1, \dots, \alpha_k)$  with  $\alpha_i \geq 0$  and  $\sum_{i=1}^k \alpha_i^2 d_i^{-1} < 1$  (Karlin and Rinott [1980b]). In particular, if the correlation matrix for the normal prior distribution is

$$\begin{pmatrix} 1 & \rho & \dots & \rho \\ \vdots & & & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix},$$

where  $\rho \leq 0$ , then the prior p.d.f. is S-MRR<sub>2</sub> (Block, Savits, and Shaked [1980]).  $\square$

#### EXAMPLE 2 - MULTIVARIATE BERNOULLI TRIALS

Let  $y_1, y_2, \dots$  be a sequence of i.i.d. k-dimensional vectors with common multinomial distribution with parameters  $n = 1$  and  $\underline{p} = (p_1, \dots, p_k)$ , where  $\sum_1^k p_i = 1$ . That is,  $y_1 \perp\!\!\!\perp y_2 \perp\!\!\!\perp \dots$  and  $y_i \sim M(1, \underline{p}) \forall i = 1, 2, \dots$ . Note that for any finite sequence  $y_1, \dots, y_m$ , the vector  $\underline{x} = \sum_1^m y_i = (x_1, \dots, x_k)$  is a sufficient statistic since the probability mass function of  $y_1, \dots, y_m$  is

$$L = \prod_{i=1}^k p_i^{x_i} I(\underline{p})$$

where  $I(\underline{p})$  is the indicator function of  $\sum_1^k p_i = 1$ . Clearly,  $f(\underline{x}|\underline{p}) = h(\underline{x})L$ .

We notice now that Theorem 4 applies since  $p_i^{x_i}$  is TP<sub>2</sub> in  $(x_i, p_i)$ . Hence, by Remark 2, for  $i = 1, \dots, k$ ,  $E\{\phi(p_i)|\underline{x}\}$  is increasing in  $x_i$  for every increasing real function  $\phi$ .

Suppose that a Dirichlet distribution with parameters each no smaller than unity is chosen to represent our prior opinion about  $\underline{p}$ . With this choice, the prior p.d.f. for  $\underline{p}$  is S-MRR<sub>2</sub> (see Karlin and Rinott [1980b] or Block, Savits, and Shaked [1980]). Thus, since  $p_i^{x_i}$  is PF<sub>2</sub> in  $p_i$  ( $i = 1, \dots, k$ ), Theorem 5 applies and by

Remark 4, for  $i = 1, \dots, k$ ,  $E\{\phi(p_i) | \underline{x}\}$  is decreasing in  $x_j$  whenever  $j \neq i$  and  $\phi$  increasing.  $\square$

REMARK

5 - Note that Example 2 includes both the multinomial and the negative multinomial models.

EXAMPLE 3 - MULTIVARIATE HYPERGEOMETRIC DISTRIBUTION

The probability mass function in this case may be expressed as

$$f(\underline{x} | \underline{\theta}) = h(\underline{x}) \prod_{i=1}^k \binom{\theta_i}{x_i} I(\underline{\theta}),$$

where  $I(\underline{\theta})$  is the indicator function of  $\sum_{i=1}^k \theta_i = N$ . Again, Theorem 4 applies since  $\binom{\theta_i}{x_j}$  is  $TP_2$ . Thus, for  $i = 1, 2, \dots, k$ ,  $E\{\phi(\theta_i) | \underline{x}\}$  is increasing in  $x_i$  whenever  $\phi$  is an increasing function.

Suppose that a Dirichlet-Multinomial distribution (see Chapter 2) with shape parameters,  $\alpha_i$ 's, each no smaller than unity is chosen to represent our opinion about  $\underline{\theta}$ . In addition, note that  $\binom{\theta_i}{x_i}$  is  $PF_2$  in  $\theta_i$  for every fixed  $x_i$ . Thus, Theorem 5 and Remark 4 apply. To conclude this example, recall that the posterior distribution of  $\underline{\theta} - \underline{x}$  (the unsampled population) is  $DM(N; \alpha_1 + x_1, \dots, \alpha_k + x_k)$  whose component means are given by  $(\sum_{j=1}^k (\alpha_j + x_j))^{-1} (\alpha_i + x_i)$  for  $i = 1, \dots, k$ . Thus, for every  $i = 1, \dots, k$ ,  $E\{\theta_i - x_i | \underline{x}\}$  is decreasing in  $x_j \forall j \neq i$ .  $\square$

REMARK

6 - The above Example may be viewed as a natural generalization of Theorems 2 and 3 of Whitt (1979).

EXAMPLE 4 - UNIFORM DISTRIBUTION

Suppose that  $t_1, \dots, t_n$  is a random sample from a uniform distribution on the real interval  $(\theta_1, \theta_2)$ . Let  $(x_1, x_2)$  be the usual sufficient statistic; that is,  $x_1 = \min(t_1, \dots, t_k)$  and  $x_2 = \max(t_1, \dots, t_k)$ . Suppose that a bilateral Pareto distribution with parameter  $(r_1, r_2, \alpha)$  represents our prior opinion. This distribution is a conjugate prior for the uniform distribution case (De Groot [1970], pp. 62-63 and pp. 172-174). The prior p.d.f. is given by

$$\xi(\theta_1, \theta_2) = \alpha(\alpha + 1)(r_2 - r_1)^\alpha (\theta_2 - \theta_1)^{-(\alpha+2)} I(\theta_1, \theta_2),$$

where  $\alpha > 0$ ,  $r_1 < r_2$ , and  $I(\theta_1, \theta_2)$  is the indicator function of  $\theta_1 < r_1$  and  $\theta_2 > r_2$ . The likelihood function may be expressed as:

$$L = (\theta_2 - \theta_1)^{-n} I_1(x_1, \theta_1) I_2(x_2, \theta_2),$$

where  $I_1$  and  $I_2$  are the indicator functions of  $x_1 > \theta_1$  and  $x_2 < \theta_2$  respectively. Since  $I_1$  and  $I_2$  are  $TP_2$  functions and  $(\theta_2 - \theta_1)^{-b}$  is  $RR_2$  for  $b > 0$ , Theorems 4 and 5 ( $k = 2$ ) apply. Then by Remarks 2 and 4, we conclude that  $E\{\phi(\theta_1) | (x_1, x_2)\}$  increases in  $x_1$  and decreases in  $x_2$  for every increasing function  $\phi$ . The analogous result for  $\theta_2$  is obvious.  $\square$

EXAMPLE 5 - THE EXPONENTIAL FAMILY

This application is of a general nature. Consider that the distribution of  $\underline{x}$  belongs to the exponential family. With a proper reparametrization we may consider  $\underline{\theta}$  such that

$$f(\underline{x}|\underline{\theta}) = h(\underline{x}) \exp\{\underline{\theta}'\underline{x}\}c(\underline{\theta}).$$

Since  $e^{\theta_i x_i}$  is  $TP_2$ , Theorem 4 applies and for any increasing  $\phi$ ,  $E\{\phi(\theta_i)|\underline{x}\}$  is increasing in  $x_i$ . With a suitable choice for the prior, Theorem 5 and Remark 4 apply.

#### 5 - ACKNOWLEDGEMENTS

This Chapter is based on a joint paper of Fahmy, Pereira, and Proschan (1980). Originally, we stated Theorem 5 for  $k = 2$ . This fact greatly restricts our applications. The present general version of Theorem 5 was suggested by Professor Moshe Shaked of Indiana University.



REFERENCES

1. BARLOW, R. E. and PROSCHAN, F. (1975). Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston, New York.
2. BLOCK, H. D., SAVITS, T. H., and SHAKED, M. (1980). Some Concepts of Negative Dependence. Unpublished report.
3. DE GROOT, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill, New York.
4. FAHMY, S., PEREIRA, C. A. B. and PROSCHAN, F. (1980). The Influence of the Sample on the Posterior Distribution. Unpublished report.
5. KARLIN, S. (1968). Total Positivity. Stanford University Press.
6. KARLIN, S. and RINOTT, Y. (1980a). Classes of Orderings of Measures and Related Correlation Inequalities. I. Multivariate Totally Positive Distributions. Unpublished report.
7. KARLIN, S. and RINOTT, Y. (1980b). Classes of Orderings of Measures and Related Correlation Inequalities. II. Multivariate Reverse Rule Distributions. Unpublished report.
8. WHITT, W. (1979). A note on the Influence of the Sample on the Posterior Distribution. JASA, 74, 424-426.

CHAPTER IV. ON THE CHARACTERIZATION OF DISTRIBUTIONS IN  
TERMS OF SUFFICIENCY AND COMPLETENESS

1 - INTRODUCTION

The discussions in this chapter are motivated by the Bayesian analysis of samples from finite populations discussed in Section 4 of Chapter 2.

As before, we have a finite population  $P = \{1, \dots, N\}$  of  $N$  units which is partitioned into  $k$  categories  $c_1, \dots, c_k$ . The incidence ( $k$ -dimensional) vectors,  $y_1, \dots, y_N$ , associated with the population units, indicate the category allocation of each unit. In order to select a sample of size  $n \leq N$ ,  $P$  is divided into two disjoint subsets  $S$  and  $\bar{S}$  where  $S \cup \bar{S} = P$ , and the number of units of  $S$  is  $n$ . The parameter of interest and the data are respectively  $\underline{\theta} = \sum_1^N y_i$  and  $\underline{x} = \sum_{i \in S} y_i$ . When the choice of the prior distribution of  $(y_1, \dots, y_N)$  is restricted to the class of exchangeable distributions, by considering a pseudo parameter  $\underline{p} = (p_1, \dots, p_k)$  with  $p_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\sum_1^k p_i = 1$ , if a priori  $\underline{\theta} | \underline{p} \sim M(N; \underline{p})$ , then

$$(i) \underline{x} | \underline{p} \sim M(n; \underline{p}),$$

$$\text{and (ii) } \underline{x} | (\underline{\theta}, \underline{p}) \sim \underline{x} | \underline{\theta} \sim H(N, n, \underline{\theta}).$$

First we notice that the conditional distribution of  $\underline{x} | \underline{\theta}$  (the sample model) is independent of the method used to partition  $P$

into  $S$  and  $\bar{S}$ ; that is, the sample model is independent of the sample design. This fact supports an often repeated slogan of D. Basu (1969, 1970) that the nature of sampling randomization has no place in data analysis. The basic argument is that two samples containing the same units must give exactly the same information about  $\underline{\theta}$ , no matter how different the selection methods were.

Finally, we notice that the distribution of  $\underline{x} | (\underline{\theta}, \underline{p})$  is the same for every value of  $\underline{p}$  and it must be a Multivariate Hypergeometric distribution. The likelihood then is completely defined by our choice of the prior distribution. This fact shows the intuition behind a characterization of the Hypergeometric distribution discussed by M. Skibinsky (1970). The following statement is the original version of Skibinsky's characterization:

"A family of  $N + 1$  probability distributions (indexed, say, by  $j = 0, \dots, N$ ), each supported on a subset of  $\{0, 1, \dots, n\}$  is the Hypergeometric family having population and sample size parameters  $N, n$ , respectively (the remaining parameter of the  $j$ -th member being  $j$ ), if and only if for each number  $p$ ,  $0 \leq p \leq 1$ , the mixture of the family with Binomial  $(N; p)$  mixing distribution is the Binomial  $(n; p)$  distribution."

In Section 2 we introduce some notation and briefly review the notions of transition functions and sufficient experiments (Blackwell and Girshick [1954]). By applying these concepts to our sampling problem, we obtain an elegant proof of Skibinsky's characterization, and introduce its multivariate version. In

Section 3 we use the same technique to obtain analogous characterizations of the Binomial, the Multinomial the Beta-Binomial, and the Dirichlet-Multinomial distributions.

## 2 - CHARACTERIZATION OF THE HYPERGEOMETRIC MODELS

In this section we present definitions, notation, and basic facts used throughout this chapter. Although we restrict ourselves to the discrete case, the discussion can be extended to other cases.

Let  $\lambda$  be an unknown parameter taking values in the parameter space  $\Lambda$ . A random variable (or random vector)  $Y$ , taking values  $y$  in a countable space  $Y$ , and having probability functions  $g(\cdot|\lambda)$  for every  $\lambda \in \Lambda$ , characterizes an (statistical) experiment. Consider a second experiment  $X$  taking values  $x$  in the set  $X$  with probability functions  $f(\cdot|\lambda)$  for every  $\lambda \in \Lambda$ .

### DEFINITION 1

A real function  $q: X \times Y \rightarrow [0, 1]$  is said to be a transition function from  $Y$  to  $X$  if, for every  $y \in Y$ ,  $q(\cdot, y)$  is a probability function on  $X$ .

### REMARK

1 - Note that if  $X$  and  $Y$  are defined as random variables (random vectors) on a common sample space where the conditional distribution of  $X|Y$  does not involve  $\lambda$ , then the conditional probability function of  $X|Y$  is a transition function from  $Y$  to  $X$ .

DEFINITION 2

The experiment  $Y$  is Blackwell sufficient for the experiment  $X$  (in symbols  $Y \succ X$ ) if there is a transition function from  $Y$  to  $X$  such that

$$f(\cdot|\lambda) = \sum_y q(\cdot, y)g(y|\lambda) \quad \forall \lambda \in \Lambda.$$

EXAMPLE 1

In the sample problem described in Section 1, let  $k = 2$ . Then for  $0 < p < 1$ ,  $\theta_1 \sim B(N; p)$ ,  $x_1 \sim B(n; p)$  and  $x_1|\theta_1 \sim h(N, n, \theta_1)$ . Here  $x_2 = n - x_1$  and  $\theta_2 = N - \theta_1$ . By Remark 1, we have that

$$(2.1) \quad q(x, \theta) = \frac{\binom{n}{x} \binom{N-n}{\theta-x}}{\binom{N}{\theta}}$$

is a transition function from  $\{0, 1, \dots, N\}$  to  $\{0, 1, \dots, n\}$ .

If  $f(\cdot|p)$  and  $g(\cdot|p)$  are respectively the probability functions of  $B(n; p)$  and  $B(N; p)$ , then

$$(2.2) \quad f(\cdot|p) = \sum_{\theta} q(\cdot|\theta)g(\theta|p), \quad \forall 0 < p < 1,$$

since  $q(\cdot, \cdot)$  is the conditional probability function of  $x_1|\theta_1$ .

That is,  $\theta_1 \succ x_1$ .  $\square$

REMARK

2 - Note that (2.2) is a relation between the functions  $f$ ,  $q$ , and  $g$ . It does not depend on the stochastic relationship between  $x_1$  and  $\theta_1$ . This shows that if any two experiments  $X$  and  $Y$  are such

that for  $n \leq N$ ,  $X \sim B(n; p)$  and  $Y \sim B(N; p)$ ,  $0 < p < 1$ , then  $Y \succ X$ .

This example and remark are generalized as follows:

### EXAMPLE 2

Consider now the general case of our sample problem; that is,  $k \geq 2$ . Recall that for every  $\underline{p} = (p_1, \dots, p_k)$ , where  $p_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\sum p_i = 1$ , we had  $\underline{\theta} \sim M(N; \underline{p})$ ,  $\underline{x} \sim M(n; \underline{p})$ , and  $\underline{x} | \underline{\theta} \sim H(N, n, \underline{\theta})$ . Then by Remark 1,

$$(2.3) \quad q(\underline{x}, \underline{\theta}) = \frac{\binom{n}{x_1, \dots, x_k} \binom{N-n}{\theta_1 - x_1, \dots, \theta_k - x_k}}{\binom{N}{n}}$$

is a transition function from  $\Theta$  (the domain of  $\underline{\theta}$ ) to  $X$  (the domain of  $\underline{x}$ ). If  $f(\cdot | \underline{p})$  and  $g(\cdot | \underline{p})$  are respectively the probability functions of  $M(n; \underline{p})$  and  $M(N; \underline{p})$ , then

$$(2.4) \quad f(\cdot | \underline{p}) = \sum_{\underline{\theta}} q(\cdot, \underline{\theta}) g(\underline{\theta} | \underline{p})$$

for every  $\underline{p} = (p_1, \dots, p_k)$ , where  $p_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\sum p_i = 1$ . That is,  $\underline{\theta} \succ \underline{x}$ .  $\square$

### REMARK

3 - Again note that (2.4) is a relation between the functions  $f$ ,  $q$ , and  $g$ . It does not depend on the nature of  $\underline{\theta}$  and  $\underline{x}$ . This shows that if any two experiments  $X$  and  $Y$  are such that for  $n \leq N$ ,  $X \sim M(n; \underline{p})$  and  $Y \sim M(N; \underline{p})$ , where  $\underline{p} = (p_1, \dots, p_k)$ ,  $p_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\sum p_i = 1$ , then  $Y \succ X$ .

Let  $\phi: X \rightarrow \mathbb{R}$  be a bounded function.

DEFINITION 3

The experiment  $X$  is said to be boundedly complete if for every  $\lambda \in \Lambda$ ,

$$\sum_x \phi(x) f(x|\lambda) = 0$$

implies  $\phi \equiv 0$ .

The following remarks are well known results:

REMARKS

4 - If  $Y \sim B(N; p)$ ,  $0 < p < 1$ , then  $Y$  is boundedly complete.

5 - If  $Y \sim M(N; \underline{p})$ , where  $\underline{p} = (p_1, \dots, p_k)$ ,  $p_i \geq 0$  ( $i = 1, \dots, k$ ), and  $\sum p_i = 1$ , then  $Y$  is boundedly complete.

The following result is the statistical version of Skibinsky's characterization. Let  $f(\cdot|p)$  and  $g(\cdot|p)$  be respectively the probability functions of the experiments ( $n \leq N$ )  $X \sim B(n; p)$  and  $Y \sim B(N; p)$ ,  $0 < p < 1$ .

THEOREM

A transition function  $q(\cdot, \cdot)$  from  $\{0, 1, \dots, N\}$  to  $\{0, 1, \dots, n\}$  satisfies the relation

$$f(\cdot|p) = \sum_y q(\cdot, y) g(y|p) \quad \forall 0 < p < 1$$

if and only if for each  $y \in \{0, 1, \dots, N\}$ ,  $q(\cdot, y)$  is the Hypergeometric probability function (2.1) with parameters  $N$ ,  $n$  and  $y$ .

PROOF

← This follows from relation (2.2) and Remark 2.

→ Suppose that  $q^*(\cdot, \cdot)$  satisfies the relation

$$f(\cdot | p) = \sum_y q^*(\cdot, y) g(y | p).$$

Since  $Y$  is complete,  $q^*(\cdot, y) = q(\cdot, y)$  for every  $y \in \{0, 1, \dots, N\}$ .  $\square$

To conclude this section we state a generalization of the above theorem. It follows directly from Example 2 and Remarks 3 and 5. Here  $X$  and  $Y$  are two sets of integer vectors

$\underline{x} = (x_1, \dots, x_k)$  and  $\underline{y} = (y_1, \dots, y_k)$ , respectively, such that  $\sum_1^k x_i = n \leq \sum_1^k y_i = N$ . Consider two multinomial experiments

$\underline{X} \sim M(n; \underline{p})$  and  $\underline{Y} \sim M(N; \underline{p})$ , where  $\underline{p} = (p_1, \dots, p_k)$ ,  $p_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\sum p_i = 1$ . Let  $f(\cdot | \underline{p})$  and  $g(\cdot | \underline{p})$  be respectively the probability functions of  $X$  and  $Y$ .

PROPOSITION 1

A transition function  $q(\cdot, \cdot)$  from  $Y$  to  $X$  satisfies the relation

$$f(\cdot | \underline{p}) = \sum q(\cdot, \underline{y}) g(\underline{y} | \underline{p})$$

for every  $\underline{p} = (p_1, \dots, p_k)$ ,  $p_i \geq 0$  ( $i = 1, \dots, k$ ) and  $\sum p_i = 1$ , if and only if for each  $\underline{y} \in Y$ ,  $q(\cdot, \underline{y})$  is the Multivariate Hypergeometric probability function (2.3) with parameters  $N$ ,  $n$ , and  $\underline{y}$ .

3 - CHARACTERIZATION OF OTHER DISTRIBUTIONS

In this section, the technique illustrated in Section 2 is applied to some important discrete distributions.



We write  $Y \sim \text{Poi}(\lambda)$ ,  $\lambda > 0$ , to indicate that the experiment  $Y$  has Poisson distribution with parameter  $\lambda > 0$ . That is, if  $g(\cdot|\lambda)$  is the probability function of  $Y$ , then for  $y \in N = \{0, 1, \dots\}$

$$g(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad \forall \lambda > 0.$$

In this case it is well known that  $Y$  is complete. Consider now two experiments  $X$  and  $Y$  such that for a fixed number  $r \in [0, 1]$ ,  $X \sim \text{Poi}(r\lambda)$  and  $Y \sim \text{Poi}(\lambda)$ ,  $\lambda > 0$ . Then the following result shows that  $Y \succ X$  and characterizes the Binomial distributions.

PROPOSITION 2

A transition function  $q(\cdot, \cdot)$  from  $N$  to  $N$  satisfies the relation

$$(3.1) \quad \frac{e^{-\lambda r} (\lambda r)^x}{x!} = \sum_{y=0}^{\infty} q(x, y) \frac{e^{-\lambda} \lambda^y}{y!}$$

for every  $x \in N$  and  $\lambda > 0$ , if and only if for every  $y \in N$ ,  $q(\cdot, y)$  is the Binomial probability function with parameters  $y$  and  $r$ .

PROOF

$$\begin{aligned} \text{If } q(x, y) &= \binom{y}{x} r^x (1-r)^{y-x} \text{ for } x \leq y \\ &= 0 \text{ otherwise,} \end{aligned}$$

then for every  $x \in N$ ,

$$\begin{aligned} \sum_{y=0}^{\infty} q(x, y) \frac{e^{-\lambda} \lambda^y}{y!} &= \frac{(\lambda r)^x}{x!} e^{-\lambda r} \sum_{y=x}^{\infty} \frac{[(1-r)\lambda]^{y-x}}{(y-x)!} e^{-\lambda(1-r)} \\ &= \frac{(\lambda r)^x}{x!} e^{-\lambda r} \end{aligned}$$

Since the Poisson experiment  $Y$  is complete, for  $x \in N$ ,  $q(x, \cdot)$  is the only solution for (3.1).  $\square$

To generalize this result to the multinomial case we consider the experiment  $\underline{X} = (X_1, \dots, X_k)$  with independent components where, for a fixed nonnegative real vector  $\underline{r} = (r_1, \dots, r_k)$  with  $\sum r_i = 1$  and for each  $i = 1, \dots, k$ ,  $X_i \sim \text{Poi}(\lambda r_i)$ ,  $\lambda > 0$ . The following result shows that if  $Y \sim \text{Poi}(\lambda)$ ,  $\lambda > 0$ , then  $Y \succ \underline{X}$ , and it characterizes the Multinomial distribution. Here  $N^k$  is the Cartesian product  $N \times N \times \dots \times N$  ( $k$ -times).

### PROPOSITION 3

A transition function  $q(\cdot, \cdot)$  from  $N$  to  $N^k$  satisfies the relation

$$\prod_{i=1}^k \frac{(\lambda r_i)^{x_i}}{x_i!} e^{-\lambda r_i} = \sum_{y=0}^{\infty} q(\underline{x}, y) \frac{\lambda^y}{y!} e^{-\lambda},$$

for every  $\underline{x} = (x_1, \dots, x_k) \in N^k$  and  $\lambda > 0$ , if and only if for every  $y \in N$ ,  $q(\cdot, y)$  is the Multinomial probability function with parameter  $(y; \underline{r})$ .

The proof follows the steps of the previous one and therefore is omitted.

Consider now a sequence of Bernoulli trials with probability of success  $p$  ( $0 < p < 1$ ). If  $Y$  is the number of failures needed to obtain a fixed number  $\alpha$  of successes, then  $Y$  is said to be a Negative Binomial experiment with parameter  $(\alpha; p)$  and we write  $Y \sim \text{NB}(\alpha; p)$ ,  $0 < p < 1$ . The probability function of  $Y$  is

$$(3.2) \quad g(y|p) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)y!} p^\alpha (1 - p)^y$$

for every  $y \in N$  and  $0 < p < 1$ . Note that  $\sum_{y=0}^{\infty} \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)y!} (1 - p)^y = p^{-\alpha}$

for every  $\alpha > 0$ . Then, the following results hold not only for  $\alpha \in N$  but in general for any  $\alpha > 0$ . In this case, we still write  $Y \sim \text{NB}(\alpha; p)$  to indicate that the probability function of  $Y$  is (3.2). It is easy to check that  $Y$  is complete.

For  $\alpha \geq \alpha_1 > 0$ , let  $X$  and  $Y$  be two experiments such that  $X \sim \text{NB}(\alpha_1, p)$  and  $Y \sim \text{NB}(\alpha, p)$ ,  $0 < p < 1$ . The following result shows that  $Y \succ X$  and characterizes the Beta-Binomial distribution.

#### PROPOSITION 4

A transition function  $q(\cdot, \cdot)$  from  $N$  to  $N$  satisfies the relation

$$(3.2) \quad \frac{\Gamma(\alpha_1 + x)}{\Gamma(\alpha_1)x!} p^{\alpha_1} (1 - p)^x = \sum_{y=0}^{\infty} q(x, y) \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)y!} p^\alpha (1 - p)^y,$$

for every  $x \in N$  and  $0 < p < 1$ , if and only if for every  $y \in N$ ,  $q(\cdot, y)$  is the Beta-Binomial probability function (see Chapter 2) with parameter  $(y; \alpha_1, \alpha_2)$ , where  $\alpha_2 = \alpha - \alpha_1$ .

PROOF

$$\text{Let } g(x, y) = \frac{y! \Gamma(\alpha)}{\Gamma(\alpha + y)} \frac{\Gamma(\alpha_1 + x)}{x! \Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + y - x)}{(y - x)! \Gamma(\alpha_2)} \text{ for } x \leq y$$

$$= 0 \text{ otherwise.}$$

Then,

$$\sum_{y=0}^{\infty} q(x, y) g(y|p) = \frac{\Gamma(\alpha_1 + x)}{x! \Gamma(\alpha_1)} p^{\alpha_1} (1 - p)^x \sum_{y=x}^{\infty} \frac{\Gamma(\alpha_2 + y - x)}{(y - x)! \Gamma(\alpha_2)} p^{\alpha_2} (1 - p)^{y-x}$$

$$= \frac{\Gamma(\alpha_1 + x)}{x! \Gamma(\alpha_1)} p^{\alpha_1} (1 - p)^x.$$

Since the Negative Binomial experiment  $Y$  is complete, for every  $x \in \mathbb{N}$ ,  $q(x, \cdot)$  is the only solution of (3.2).  $\square$

To generalize this result to the Dirichlet-Multinomial distribution case, we consider the experiment  $\underline{X} = (X_1, \dots, X_k)$  with independent components, where for a fixed nonnegative real vector  $(\alpha_1, \dots, \alpha_k)$  with  $\sum_{i=1}^k \alpha_i = \alpha$  and for each  $i = 1, \dots, k$ ,  $X_i \sim \text{NB}(\alpha_i; p)$ ,  $0 < p < 1$ . The following result characterizes the Dirichlet-Multinomial distribution and shows that  $Y \succ \underline{X}$ , where  $Y \sim \text{NB}(\alpha; p)$ ,  $0 < p < 1$ .

PROPOSITION 5

A transition function  $q(\cdot, \cdot)$  from  $\mathbb{N}$  to  $\mathbb{N}^k$  satisfies the relation

$$\prod_{i=1}^k \frac{\Gamma(\alpha_i + x_i)}{\Gamma(\alpha_i) x_i!} p^{\alpha_i} (1 - p)^{x_i} = \sum_{y=0}^{\infty} q(\underline{x}, y) \frac{\Gamma(\alpha + y)}{\Gamma(\alpha) y!} p^{\alpha} (1 - p)^y,$$

for every  $\underline{x} = (x_1, \dots, x_k) \in N^k$  and  $0 < p < 1$ , if and only if for every  $y \in N$ ,  $q(\cdot, y)$  is the Dirichlet-Multinomial probability function (see (3.1) Chapter 2) with parameter  $(y; \alpha_1, \dots, \alpha_k)$ .

The proof is omitted since it follows the steps of the previous one.

REMARK

As in the Hypergeometric case, these results are intuitive. For instance recall that if  $X_1, \dots, X_k$  are  $k$  independent Poisson random variables with parameters  $\lambda_1, \dots, \lambda_k$  respectively, then the conditional joint distribution of  $(X_1, \dots, X_k)$  given the sum  $\sum_1^k X_i = N$  is Multinomial with parameter  $(N; r_1, \dots, r_k)$ , where  $r_i = \lambda_i (\sum_1^k \lambda_j)^{-1}$ . On the other hand, if  $X_1, \dots, X_k$  are  $k$  independent Negative-Binomial random variables with parameters  $(\alpha_1; p), \dots, (\alpha_k; p)$  respectively, then the conditional joint distribution of  $(X_1, \dots, X_k)$  given the sum  $\sum_1^k X_i = N$  is Dirichlet-Multinomial with parameter  $(N; \alpha_1, \dots, \alpha_k)$ .

To conclude this chapter we present another interesting application of the concepts of transition function and Blackwell sufficiency.

Let  $\lambda \in \Lambda$  be the unknown parameter and consider two parametric functions  $a: \Lambda \rightarrow [0, 1]$  and  $b: \Lambda \rightarrow [0, 1]$ . Suppose that  $X$  and  $Y$  are two Bernoulli experiments with success probabilities  $a(\lambda)$  and  $b(\lambda)$  respectively. The objective is to find a necessary and sufficient condition to have  $X \succ Y$ .

First note that a transition function  $q(\cdot, \cdot)$  from  $\{0, 1\}$  to  $\{0, 1\}$  can be expressed in a tabular form as

	0	1	
0	1 - p	p	$0 < p < 1$
1	1 - q	q	$0 < q < 1$

Now, if  $X \succ Y$  we are able to express  $b(\lambda)$  in terms of  $a(\lambda)$  and  $q(\cdot, \cdot)$  as follows:

$$\begin{aligned} b(\lambda) &= a(\lambda)q + [1 - a(\lambda)]p \\ &= p + (q - p)a(\lambda). \end{aligned}$$

This clearly shows that:  $X \succ Y$  if and only if there are two numbers  $\alpha$  and  $\beta$ , where  $0 < \alpha < 1$  and  $0 < \alpha + \beta < 1$  such that  $b(\lambda) = \alpha + \beta a(\lambda)$  for every  $\lambda \in \Lambda$ . To illustrate this fact, consider the plane  $\mathbb{R}^2$ . Then  $X \succ Y$  if and only if there is a line  $L$  (see figures below) intersecting the line  $x = 0$  between the points  $(0, 0)$  and  $(0, 1)$ , and intersecting the line  $x = 1$  between the points  $(1, 0)$  and  $(1, 1)$  such that the set

$$\{(a(\lambda), b(\lambda)); \lambda \in \Lambda\}$$

lies on  $L$ . The following figures show two possibilities for  $L$ .

Note that there are many ways a line may intersect the square inside the points  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 1)$  and  $(1, 0)$ . However, to have  $X \succ Y$  the line  $L$  must intersect (through the square) both side lines  $x = 0$  and  $x = 1$ .

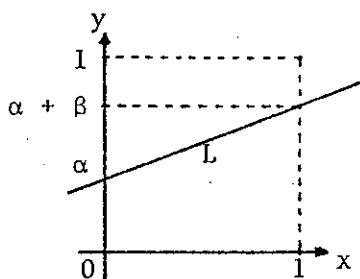


Figure 1

A case of  $X \succ Y$   
with  $q > p$ . ( $\beta > 0$ )

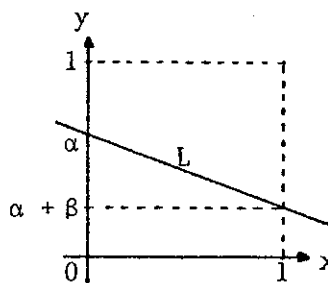


Figure 2

A case of  $X \succ Y$   
with  $q < p$ . ( $\beta < 0$ )

For an application of this result (see Lehmann [1958] pp. 75-77) take  $\Lambda = [0, 1]$ ,  $a(\lambda) = \lambda$  and  $b(\lambda) = r\lambda$  for a fixed  $r \in (0, 1)$ . This example shows that the relation  $X \succ Y$  is not related to the variances of  $X$  and  $Y$  as one might think.

REFERENCES

1. BASU, D. (1969). Role of the Sufficiency and the Likelihood Principles in Sample Survey Theory. Sankhya 31, 441-454.
2. BASU, D. (1978). Relevance of Randomization in Data Analysis. Survey Sampling and Measurement. (N. K. Namboodiri, editor), Academic Press, N.Y.
3. BLACKWELL, D. and GIRSHICK, M. A. (1954). Theory of Games and Statistical Experiments. John Wiley, N.Y.
4. LEHMANN, E. L., (1959). Testing Statistical Hypothesis. John Wiley, N.Y.
5. SKIBINSKY, M. (1970). A Characterization of Hypergeometric Distributions. JASA 65, 926-929.



## VITA

Carlos Alberto de Bragança Pereira

### BIRTH

July 1, 1946; Rio de Janeiro, R. J. Brasil

### ACADEMIC TRAINING

B.S. - Statistics (1969), ENCE, Rio de Janeiro, R. J. Brasil.

M.S. - Statistics (1971), Universidade de Sao Paulo, SP, Brasil.

M.S. - Statistics (1978), The Florida State University, Tallahassee, Florida.

Ph.D. - Statistics (1980), The Florida State University, Tallahassee, Florida.

### PROFESSIONAL EXPERIENCE

Statistics Instructor, Universidade de Sao Paulo, Sao Paulo, SP, Brasil, 1969-.

### SOCIETIES

American Statistical Association  
Institute of Mathematical Statistics  
Sociedade Brasileira dos Estatisticos.