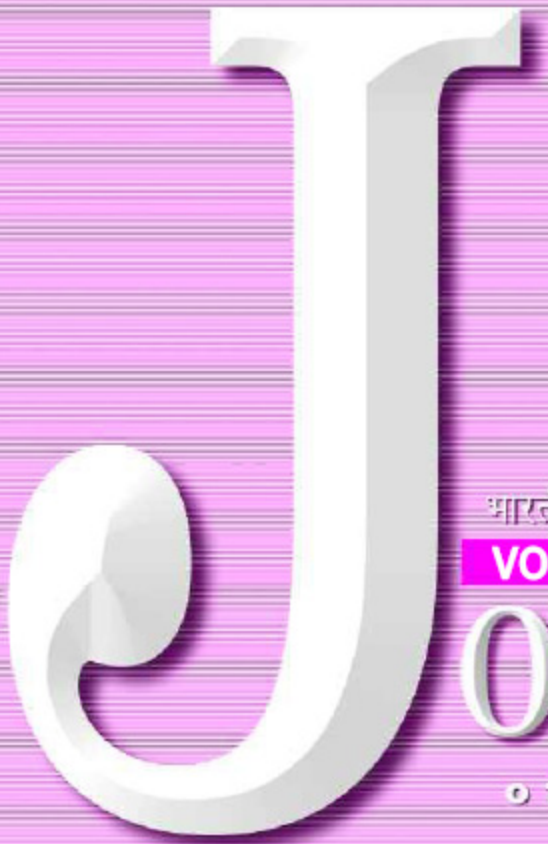

ISAS

भारतीय कृषि सांख्यिकी संस्था की पत्रिका

VOL. 63 NO. 3 Dec. 2009

JOURNAL

of the

INDIAN SOCIETY OF AGRICULTURAL STATISTICS

(Registered under the Societies Registration Act XXI of 1860)

Office Bearers of the Indian Society of Agricultural Statistics 2009

President

MANGALA RAI, ICAR, Krishi Bhavan, New Delhi – 110 114. E-mail : mrai@icar.org.in

Executive President

PREM NARAIN, B-3/27A, Lawrence Road, Delhi – 110 035. E-mail : narainprem@hotmail.com

Sessional President

VK BHATIA, IASRI, New Delhi – 110 012. E-mail: vkbhatia@iasri.res.in

Patron

BAL BPS GOEL, B-77, Naraina Vihar, New Delhi – 110 028. E-mail : alkag@nda.vsnl.net.in

KC RAUT, 41B, Pocket B2, Lawrence Road, Delhi – 110 035. E-mail : kahnucharanraut@yahoo.ca

Vice Presidents

ALOKE DEY, Indian Statistical Institute, New Delhi – 110 016. E-mail : adey@isid.ac.in

VK GUPTA, IASRI, New Delhi – 110 012. E-mail : vkgupta@iasri.res.in

SM JHARWAL, Ministry of Agriculture, Krishi Bhavan, New Delhi – 110 114. E-mail: paa-moa@nic.in

SD SHARMA, ICAR, Krishi Anusandhan Bhavan-II, New Delhi – 110 012. E-mail : adghrd@icar.org.in

Members

HVL BATHLA, 17, Pocket B-7, Sector 4, Rohini, Delhi – 110 085. E-mail : hvl_bathla@yahoo.com

SN MEGERI, Department of Agricultural Statistics, University of Agricultural Sciences, Dharwad – 580 005
E-mail : megerisn@rediffmail.com

S RAVICHANDRAN, Directorate of Rice Research, Rajendranagar, Hyderabad – 500 030
E-mail : srgmravi@yahoo.com

SC RAI, KB-58, Kavi Nagar, Ghaziabad. E-mail : naivedya.kashyap@gmail.com

VK SINGH, Directorate of Agriculture, U.P., Lucknow. E-mail : agristat@sify.com

BS KULKARNI, Department of Statistics & Mathematics, Acharya NG Ranga Agricultural University, Rajendranagar, Hyderabad – 500 030. E-mail: bskstat@rediffmail.com

BVS SISODIA, Department of Agricultural Statistics, ND University of Agriculture & Technology, Narendra Nagar, Kumarganj, Faizabad – 224 229. E-mail : bvs@india.com

DK JAIN, DESM Division, National Dairy Research Institute, Karnal – 132 001. E-mail : dkjn@rediffmail.com

VK MAHAJAN, IASRI, New Delhi – 110 012. E-mail: vkm@iasri.res.in

RC AGRAWAL, NBPGR, New Delhi – 110 012. E-mail : rakesh@nbpgr@ernet.in

HUKUM CHANDRA, IASRI, New Delhi – 110 012. E-mail: hchandra@iasri.res.in

KISHORE SINHA, Department of Agricultural Statistics, Birsa Agricultural University, Kanke, Ranchi – 834 006.
E-mail: skish55@yahoo.co.in

Secretary

VK BHATIA, IASRI, New Delhi – 110 012. E-mail: vkbhatia@iasri.res.in

Joint Secretaries

RS KHATRI, IASRI, New Delhi – 110 012. E-mail : rskhatri.rs@gmail.com

RAJENDER PARSAD, Division of Design of Experiments, IASRI, New Delhi – 110 012.
E-mail : rajender@iasri.res.in, rajender1066@yahoo.co.in

PK MALHOTRA, Division of Computer Application, IASRI, New Delhi – 110 012. E-mail : pkm@iasri.res.in

Treasurer

AK VISHANDASS, Department of Chemicals & Petrochemicals, Ministry of Chemicals & Fertilizers, Janpath Bhavan, Janpath, New Delhi – 110 001. E-mail : vishandass@nic.in

Ex-Officio Members (By Designation)

Director, Indian Agricultural Statistics Research Institute, New Delhi – 110 012.

Economic & Statistical Adviser, Ministry of Agriculture, Govt. of India, New Delhi.

Director General, Central Statistical Organization, Govt. of India, New Delhi.

Director General, National Sample Survey Organization, Govt. of India, New Delhi.

Chair Editor, Journal of the Indian Society of Agricultural Statistics, New Delhi.

JOURNAL
OF THE
INDIAN SOCIETY
OF
AGRICULTURAL STATISTICS

भारतीय कृषि सांख्यिकी संस्था की पत्रिका

December 2009 **VOL. 63** No. 3

Indian Society of Agricultural Statistics

Founded: January 03, 1947

The Indian Society of Agricultural Statistics is a scientific society. Its aims are to promote the study of and research in Statistical Theory in the widest sense of the term and its applications to Agriculture, Animal Husbandry, Agricultural Economics and allied fields. The Society has broadened its coverage by including research in Computer Applications also into its gamete. Its membership is open to all persons and institutions interested in the aims of the Society. The membership consists of Honorary Members, Patrons, Life Members, Annual Members and Institutional Members. The membership fee is Rs. 210 per annum (foreign, US $ 100) for Annual Members and Rs. 2000 (foreign, US $ 300) for Life Members. Membership fee includes subscription to the Journal. The Permanent Institutional Membership fee is Rs. 10,000. Additional information about the Society and application forms for membership may be obtained from the Secretary, Indian Society of Agricultural Statistics, IASRI Campus, Library Avenue, Pusa, New Delhi-110012 or may be downloaded from the website of the Society at www.isas.org.in.

Journal of Indian Society of Agricultural Statistics

The Society also publishes an International Peer Reviewed Journal called "Journal of the Indian Society of Agricultural Statistics" with ISSN 0019-6363. Three issues of the Journal (April, August and December) are published annually. The first volume of the Journal was released in 1948. The Journal devoted to the publication of original research papers on all aspects of Statistics and Computer Applications preferably with innovative applications in Agricultural Sciences or that have a potential application in Agricultural Sciences. The review articles of the topics of current interest are welcome. The Journal also accepts books, monographs and periodicals for review. Special issues on thematic areas of both national and international importance are also brought out. The Journal has a very strong Editorial Board comprising of Chair Editor and Associate Editors.

Abstracted/Indexed in: *Mathematical Reviews, Zentralblatt MATH, Statistical Theory & Method Abstracts.*

Manuscript Submission: Manuscripts should be prepared as per style of the latest issue of the Journal available at www.isas.org.in. The manuscript may be submitted for a possible publication through e-mail at isas.jisas@gmail.com.

Review Process: Journal of Indian Society of Agricultural Statistics is peer reviewed journal. The paper is assigned a manuscript number immediately on manuscript submission. An associate editor is then assigned for handling the review process of the paper. Normally the first review of the paper is sent to the authors within 18-20 weeks of its submission. The editorial process is expected to complete within 36 weeks of its submission.

Note: The contents of a paper published in this Journal are the sole responsibility of the author or authors, and its publication does not imply the concurrence of the Executive Council of Indian Society of Agricultural Statistics or its Editors.

Laser typeset at COMPUTER CORNER, A-65, GF, Shanker Garden, Vikas Puri, New Delhi-110018, and printed on behalf of Computer Corner at RASHTRIYA PRINTERS, M-135, Panchsheel Garden, Naveen Shahdara, Delhi-110032

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009

CONTENTS

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 209-218

Food and Nutrition Security in India: Some Contemporary Issues*

H.S. Gupta**

Indian Agricultural Research Institute, New Delhi

SUMMARY

This paper examines the main elements of Indian food policy in the context of recent economic and agricultural developments. Specifically, implications of recent global food crisis, climate change and shortfall in the agricultural growth are discussed. Given the rapid increase in demand for food grains globally and likely adverse impact of climate change on yield; the policy of self-sufficiency and government interventions to ensure physical and economic access should be strengthened. The country need to pay more attention to nutritional issues as 1/3rd of our people suffer from malnutrition. Concerted efforts should also be made to raise food grain yields not only to meet food requirement but also release some area for cultivation of high value crops like fruits and vegetables. Application of available stock of knowledge and technology for bridging the yield gap, particularly in the eastern region, can have immediate impacts. A long term strategy should also target developing plant varieties and crop management practices for adaptation and mitigation of climate change.

Key words : Trends in foodgrain production, Temporal and spatial trends, Demand projections, Food and nutritional security.

1. INTRODUCTION

India has successfully followed the policy of food security for all sections of the society. The key elements of this policy are self-sufficient in food production and improving physical and economic access of poor to food through appropriate government interventions. This policy has worked both for consumers and producers and also insulated Indian food economy from volatility of the world market. However, some recent developments, global and national, have forced to re-assess the current scenario and strategy to ensure food security. First and foremost is slow down of agricultural growth for a number of years. Likely shortfall in kharif production because of deficit rainfall has raised further concerns. Second important issue is implications of global food crisis seen due to shortfall in foodgarin production, mainly wheat, in some major producing countries. The possibility of such a crisis in future can't be ruled out, and reinforces the need for self-reliance. Third major issue is fluctuating incentives for diversification of agricultural production, mainly because of increasing demand for high value commodities. Farmers often find remunerative to grow high value crops but in some years high foodgrain prices shift farmers back to rice and wheat crops which also have low market risk. This paper revisits the food security scenario and related policy issues in the present context. Specifically, both supply and demand side factors are taken into consideration. The main premise of the strategy is to ensure required growth in foodgrain production through technological options.

2. TRENDS IN FOODGRAIN PRODUCTION AND AVAILABILITY

Foodgrain production in the country maintained a sharp uptrend since the green revolution era. It increased from 108.4 million tonnes (MT) in 1970-71 to 176.4 MT in 1990-91 which further rose to current

Dr. Rajendra Prasad Memorial Lecture delivered at 63rd Annual Conference of the Indian Society of Agricultural Statistics held at Pusa, Samastipur, Bihar during Dec. 3-5, 2009

**E-mail address* : director@iari.res.in

highest level of 230.8 MT in 2007-08. Area under foodgrains stabilized since 1980s but yield maintained an uptrend and contributed to the growth in production since then (Fig. 1). However, rate of productivity growth and hence production decelerated during the last decade or so. This deceleration in growth could not keep pace with population growth and as a result, per capita production and availability declined in the recent past (Fig. 2). This in some years was lower than what was during the green revolution period. This is most worrying aspect of food security. It is only recently in last two years that food grain production went up sharply and per capita production was restored to a reasonably high level. Although decline in per capita production is a disturbing trend but this should be adjusted with the fact that there is decline in per capita consumption of cereals. The effect of low production is felt sharply when there is a need for import of food grains and there is not adequate stock available in the international market.

Temporal and Spatial Trends in Foodgrain Production

Agricultural production in India has kept pace with the food needs of the growing population owing to

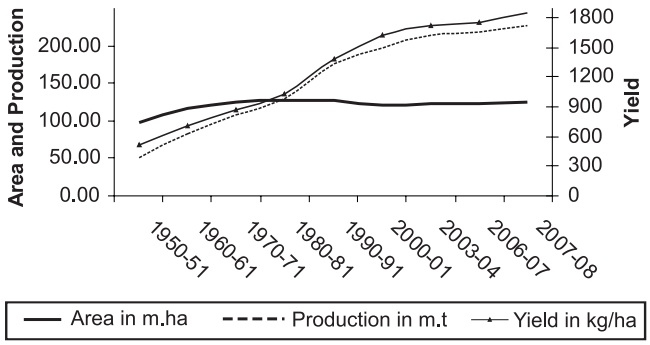

Fig. 1. Area, Production and Yield of Foodgrains in India

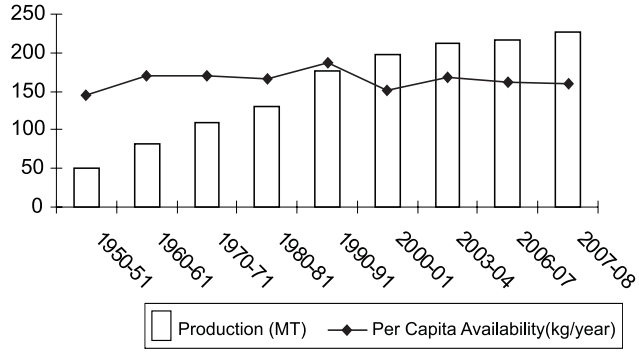

Fig. 2. Foodgrain Production and Per Capita Availability in India

increased yield of almost all crops, especially cereals. Rice production has increased from 74 million tonnes in 1990-91 to 97 million tonnes in 2007-08 showing around 30 per cent increase in production during the period. Wheat, another important staple food of the country showed 44 per cent increase, from 55 million tonnes in 1990-91 to 79 million tonnes in 2007-08. Cereals and pulses together constitute a little more than one third of total value of crops in the country. The acreage under foodgrains to cropped area declined since the early eighties, but there was hardly any decline in the share of area under rice and wheat, against a sharp decline in coarse cereals. Hence, the growth in cereals has been largely accounted for yield growth during the period. The average yield of foodgrains at national level has risen from 1.4 t/ha to 1.8 t/ha during the last two decades showing a consistent increase over the period. Among cereals, rice and wheat showed impressive yield rise in contrast to the stagnancy in pulses productivity (Table 1). There has been an adverse trend in area growth for foodgrains, mainly resulting from diversion of area from coarse cereals to oilseeds, in spite of the fact that there has been significant yield growth in coarse cereals as compared to the superior grains like rice and wheat especially during recent times. This is mainly because of increasing adoption of modern varieties in coarse cereals in rainfed areas of the country.

Among foodgrains, the growth scenario is completely different for pulses as is visible by its declining growth in area in the country. The dismal performance in pulse production, especially during the nineties, has been mainly accounted for by stagnancy in yield growth, which has been hovering around 0.5 to 0.6 t/ha since last two decades. However, during the last decade, there was appreciable acceleration in the growth of pulse production, owing to significant reduction in crop duration, particularly in pigeon pea and chick pea, now grown in the non-traditional areas.

This impressive national growth scenario masks large variation in the crop productivity per unit of cropped area in different states of India, though yields have tended to increase overtime in most of the states. The foodgrains especially cereals productivity was the highest in high growth states like Punjab (4 t/ha) and Haryana (3.4 t/ha) during 2006-07. West Bengal needs special mention as the state showed an impressive rise

Table 1. State-wise trend in foodgrain yields, tonne/ha

Sl. No.	State	Year	Rice	Wheat	Pulses	Total foodgrains
1	Andhra Pradesh	1990-91	2.4	0.9	0.4	1.6
		2000-01	2.9	0.6	0.6	2.1
		2006-07	3.0	0.9	0.7	2.2
2	Assam	1990-91	1.3	1.3	0.4	1.3
		2000-01	1.5	1.2	0.6	1.4
		2006-07	1.3	1.1	0.6	1.3
3.	Bihar	1990-91	1.2	1.8	0.8	1.3
		2000-01	1.5	2.1	0.9	1.7
		2006-07	1.5	1.9	0.7	1.7
4	Gujarat	1990-91	1.5	2.0	0.7	1.1
		2000-01	1.5	2.3	0.3	1.2
		2006-07	1.9	2.5	0.6	1.4
5	Haryana	1990-91	2.8	3.5	0.7	2.4
		2000-01	2.6	4.1	0.6	3.1
		2006-07	3.2	4.2	0.8	3.4
6	Karnataka	1990-91	2.1	0.6	0.3	0.9
		2000-01	2.6	0.9	0.5	1.4
		2006-07	2.5	0.8	0.4	1.3
7	Madhya Pradesh	1990-91	1.1	1.5	0.6	1.0
		2000-01	0.6	1.4	0.6	0.9
		2006-07	0.8	1.8	0.8	1.2
8	Maharashtra	1990-91	1.5	1.1	0.4	0.9
		2000-01	1.3	1.3	0.5	0.8
		2006-07	1.7	1.3	0.6	0.9
9	Orissa	1990-91	1.2	1.6	0.6	1.0
		2000-01	1.0	1.5	0.4	1.0
		2006-07	1.5	1.5	0.4	1.4
10	Punjab	1990-91	3.2	3.7	0.7	3.5
		2000-01	3.5	4.6	0.7	4.0
		2006-07	3.9	4.2	0.9	4.0
11	Rajasthan	1990-91	1.2	2.4	0.5	0.9
		2000-01	0.9	2.4	0.3	0.9
		2006-07	1.6	2.8	0.5	1.1
12	Tamil Nadu	1990-91	3.1		0.4	1.9
		2000-01	3.4		0.5	2.5
		2006-07	3.4		0.5	2.6
13	Uttar Pradesh	1990-91	1.8	2.2	0.9	1.7
		2000-01	2.0	2.7	0.8	2.1
		2006-07	1.9	2.7	0.7	2.1
14	West Bengal	1990-91	1.8	2.0	0.6	1.7
		2000-01	2.3	2.5	0.9	2.2
		2006-07	2.6	2.3	0.7	2.5
15	All India	1990-91	1.7	2.3	0.6	1.4
		2000-01	1.9	2.7	0.5	1.6
		2006-07	2.1	2.7	0.6	1.8

Source: Ministry of Agriculture data

in rice yields from 1.8 t/ha in 1990-91 to 2.6 t/ha in 2006-07. This was mainly due to late spread of modern varieties and increase in area under summer (*boro*) paddy, which was taken under better irrigation and input management conditions. Maize is one of the crops showing exceptional productivity growth because of rapid spread of modern hybrids. The area under this crop continued to expand even in non-traditional maize growing states of southern India. Much of the produce was for poultry feed coming up in a major way in the southern states.

Although there is growth in yield and production of foodgrains in most of the states, there are some worrisome trends. First is low food grain yield in all the eastern states, except West Bengal which has shown significant growth. These states experience all production constraints, including low use of inputs, low seed replacement rate etc. These states also witness frequent weather fluctuations. The second disturbing factor is recent decline or no increase in wheat yield in several states, including north-west region. This is mainly because of higher temperature but lack of superior varieties could not be rules out. Third most disquieting feature is continued low productivity of pulses in all the states, pushing total foodgrain yield downward. The reasons for the low productivity are well documented-high yield loss to biotic stress, low input use, high risk, low seed replacement rate etc. Increasing pulse productivity is essential for nutritional security.

3. ELEMENTS AND EFFECTIVENESS OF FOOD POLICY

Government Interventions in Foodgrain Markets

Government intervention in foodgrain markets has been an integral part of Indian food policy. Since the green revolution era, the government procures food grains at the predetermined price, now minimum support price (MSP). This is mainly done for rice and wheat in surplus states of Punjab, Haryana and Uttar Pradesh. Some rice is now also procured from Andhra Pradesh and eastern states. This procurement is to meet the requirement of public distribution at a price lower than the ruling market price and also to meet requirement of special welfare or employment schemes. Part of the stocks is also used to maintain buffer stock to moderate year-to-year fluctuations in food grain production. The procurement of food grains is open ended depending upon level of production and market prices, while public distribution is governed by the scale of allocation and its offtake by the beneficiaries. Excess stock if any is disposed through open market sale or exports.

Since procurement is ruled by market conditions, there are considerable year-to-year variations in the stocks held by the Food Corporation of India. The stocks varied between 45 to 58 million tonnes during 2001-2003 when rice and wheat production was comfortable. The stocks reduced to 17-19 million tonnes during 2006-2008, especially due to low stocks of wheat (Fig. 3). In fact, procurement of wheat during this period reduced by 4-5 million tones, owing to better open market prices. Under such condition, offtake of rice and wheat for public distribution also increases - it increased from 24.2 million tonnes in 2003-04 to 33.5 million tonnes in 2007-08 and most of this increase was due to higher distribution to the beneficiaries which are Above the Poverty Line. The allocation of Below the Poverty Line households has been around 15 million tonnes during the last few years (GoI 2009).

The government operations have been successful in improving access of the poor to food, reducing year-to-year fluctuations in the availability and providing remunerative price to farmers. However, there are few issues which need some discussion. First, now size of foodgrain economy is too large and government alone can't manage the size of operation needed. It is suggested that private trade should be encouraged,

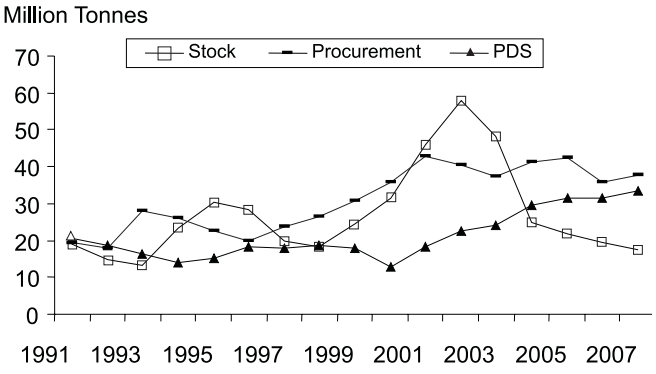

Fig. 3. Procurement, PDS and Stock of Cereals

especially in the surplus states, and if necessary, regulations on procurement and storage of food grains should be relaxed for participation of private sector. These regulations should be invoked or applicable during the year scarcity. The government should focus on those areas where surplus is emerging now so that farmers are able to get benefit of MSP. The second issue relates to procurement through levy on rice mills. This is sometimes taxing for the mills when there is shortage of production and does not benefit farmers. This, as suggested by the Sen Committee, should be replaced by custom milling of paddy procured under MSP.

4. CONSUMPTION AND FUTURE REQUIREMENT OF FOODGRAINS

Trends in Foodgrain Consumption

Although foodgrains continue to be staple food in India, there are some important changes in the consumption pattern which have significant implications for the national food and nutrition security. There has been a persistence decline in per capita consumption of cereals since the green revolution period. Annual per capita consumption of total cereals decreased from 173.8 kg in 1973-74 to 139.8 kg in 2004-05. This trend was more drastic for coarse cereals due to increase in income level. Per capita consumption of rice also decreased but this was noticed during the last two decades or so. On the other hand, per capita consumption of wheat rose significantly since the 1980s. These trends in cereal consumption are seen in both rural and urban areas. Currently in 2004-05, per capita annual consumption of food grains is 148.8 kg, of this 139.9 kg is cereals and 9 kg is pulses. The consumption of rice and wheat is 73.8 kg and 53.8 kg/ year, respectively (Fig. 4).

In addition to direct demand of foodgrains as human food, there is indirect food demand as animal feed. Some amount of production is also used as seed and wasted during transportation, storage etc. There are no reliable estimates for these demand components. Conventionally, official estimates used for this purpose is 12.5 per cent of the total production. This estimate is used since 1950s and there could be some under estimation with respect to feed demand in view of rapid growth in livestock industry. Therefore, an objective assessment about feed demand could help better manage the food economy.

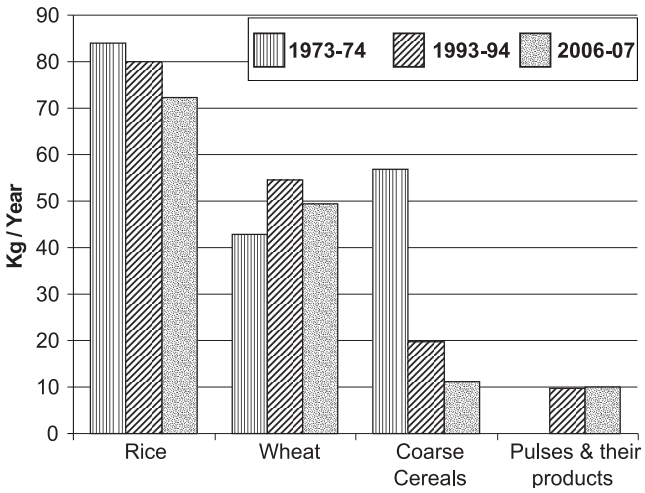

Fig. 4. Trend in per capita consumption of foodgrains in India

Demand Projections for Foodgrains

Demand for food grains comprises of two parts. First is direct food demand. This demand is influenced by trends in population growth, per capita income and change in food taste and preferences due to urbanization and other demographic changes. The second component is 'other demand' mainly for feed, seed and industrial and other uses. Several projections are made using different assumptions about population and income growth. Chand (2009) assumes an annual growth in per capita income more than seven per cent and population growth 1.2 per cent. Total demand for food grain as household food is project as 179.1 million tonnes (MT) in 2020-21. Adding to this indirect demand for food (i.e. feed) and other uses, total demand for food grains will be 280.6 million tonnes. This will comprise 261.5 MT cereals and 19.1 MT pulses. This means that the country need to produce another 50 MT of food grains in 11 years, needing annual increase of 5 MT which is rather challenging given the trend during the last decade. On demand side also these estimates looks on higher side. In another study, Kumar *et al*. (2009) projected total demand for food grains as 253 MT in 2021. The Planning Commission (2002) maintains a projected demand of 247 MT in 2020. Of this, 129 MT is for rice, 92 MT for wheat, 16 MT for coarse cereals and 20 MT for pulses. These projections look within the reach and food grain production should increase about 2 MT every year, which is a reasonable target.

Diversification of Production

Agricultural diversification is driven by rising income levels and urbanization process. There is

increasing consumption of high value products like fruits, vegetables and livestock products. This trend is reflected in markets and farmers have responded to higher prices of these commodities. As a result, growth in these commodities during tenth Plan has been three per cent or more against 1.3 per cent in cereals. This growth rate will be higher if we take recent three years or so.

Besides demand side consideration, there has been significant development on the supply side. Growth of vegetables could also be attributed to adoption of improved hybrids through the seed policy reforms in the late 1980s. Seed of foreign hybrids and planting material were permitted resulting in high productivity growth. In livestock, much of higher growth could be attributed to poultry sector and development of dairy industry, owing to improved breed, feeding and heath practices. Also, there has been extension of market infrastructure and institutional development under cooperative sector, which really contributed to increased milk production and now the country is the largest producer of milk in the world. Livestock now contributes to 27 per cent of total value of agricultural output in 2006-07, as against 17 per cent by fruits and vegetables (Fig. 5). Foodgrains contributes 25 per cent to the total value and in absolute terms (at 1999-2000 prices), it increased from Rs 152 thousand crore in 1999-2000 to Rs 157 thousand crore in 2006-07 (CSO 2008).

Malnutrition, Poverty and Gender

Nearly half of the population still suffers from chronic under-nutrition. The most vulnerable sections of the society are children, women and the elderly, especially among the lower income groups. While the number of children suffering from severe malnutrition declined significantly in the 1990s, the prevalence of mild and moderate under-nutrition, especially among the low income group is still high. Estimates show that about 40 per cent of the undernourished children in the world are in India although India accounts for less than 20 per cent of the children in the world. Within the country, nearly half of the children are below three years age are undernourished and about one-fifth of them are severely affected. There are substantial inter-state variations in the malnutrition levels of children. The per centage of moderately and severely malnourished children in 1998-99 varied between 27.4 per cent in Kerala and 55.7 per cent in Madhya Pradesh among the major states. In terms of nutritional status of children, middle income states such as Kerala (27.4 per cent), Tamil Nadu (37.7 per cent) and Andhra Pradesh (38.7 per cent) performed better than states with higher per capita SDP, such as Maharashtra (50.7 per cent) and West Bengal (49.7 per cent). Malnutrition in Kerala and Tamil Nadu declined by about half in the past 25 years while in other states, the decline was much less. Not surprisingly, poorer states such as Madhya Pradesh, Bihar and Orissa showed the worst performance. North-eastern states showed better performance in terms of nutritional status (Radhakrishna and Ravi 2004).

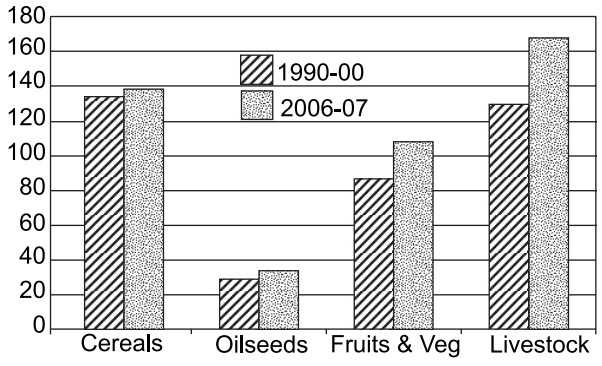

Fig. 5. Trends in value of agriculture output, 1999-2000 ('000 crore Rs)

The extent of malnutrition among the adults was high though lower than that of children. Malnutrition among women of developed states like Maharashtra (40 per cent), and West Bengal (44 per cent) were of a similar order of magnitude as the less developed states like Bihar (40 per cent), Madhya Pradesh (39 per cent) and Orissa (48 per cent). About 37.4 per cent of adult males and 39.4 per cent of adult females in 2000-01 suffered from Chronic Energy Deficiency (CED) in rural areas of India [Arnold *et al*. 2003]. Gender differences seem to exist in some states, particularly in Tamil Nadu where it was extreme and was comparatively higher in West Bengal and Kerala. North eastern states other than Assam had lower CED among females. Malnutrition was high among women in households with low standard of living and child malnutrition was associated with mother's CED. Furthermore, intra-family distribution of food is inequitable in the poor households and the pre-school children get much less than their physiological needs as compared to adult males and females (Radhakrishna 2006).

Thus there is unacceptable level of malnutrition among children and women which cannot simply be addressed by increase in food production. Targeted food for work programmes and nutrition programmes have addressed the problem temporarily. A long term solution could be to ensure employment opportunities for increasing the purchasing power of the poor to meet their nutritional requirements. Thus, employment or livelihood security becomes an essential and inseparable component of a comprehensive strategy for national food security and therefore should be accorded high priority.

5. EMERGING ISSUES IN FOOD AND NUTRITIONAL SECURITY

Continued growth of the agriculture sector is particularly important not only for ensuring the national food and nutritional security but also because of its vital role in enhancing purchasing power of the rural population. Spread of the green-revolution technologies in new regions and continued growth of productivity in the north-western states, enabled India to achieve a 3.8 per cent annual growth rate in agricultural production during the 1980s. The overall growth slowed to 2.7 per cent in the 1990s, which was associated with a reduction in public investment in agriculture, slower growth and imbalanced use of fertilizer and depletion of soil fertility. The Planning Commission has set a growth target of 4 per cent per annum for agriculture sector which has been rather an elusive goal so far. For maintaining self-sufficiency in foodgrain production, the country needs to produce at least another 20 MT of additional foodgrains and for this target, foodgrain yield should increase by one-third of the current level. This requires reassessment of our development priorities and evolving a strategy to meet this growth target. Here some of important elements of the strategy are spelled out. The main argument is that it is technology-centric approach supported with adequate rural infrastructure will help unleash growth potential of Indian agriculture.

Implications of Global Food Scenario

Food deficit countries faced food crisis in the recent past and food prices increased sharply in the world market. But with improved production performance because of higher price incentives, global food prices started to decline. A long-term trend will depend upon demand and production scenario in developing countries where most of the increase in population will take place. IFPRI (1999) study has shown that the world needs to produce 40 per cent more cereals to feed the growing population in 2020. Also, there will be deficit in many developing countries and therefore food important will double by 2020. This implies that any shortfall in production may deplete food stocks and prices may again rise. In case this shortfall occurs in large country like China and India, there may not be adequate stock to meet the import demand. Therefore, it would be advisable to follow the policy of self-reliance in food production. If necessary, storage of food grains may be increased and there should not be problem of disposing the stocks in international market, if required.

Government Support for Food Production

There are a number of areas which continue to demand government support in a substantial manner. First is the public investment for expansion of agriculture and other rural infrastructure. There was some complacency in terms of public investment during the 1990s and as a result, there was significant slowdown of agricultural growth since the mid-1990s. This was corrected to some extent recently and real public investment maintained uptrend (Fig. 6), resulting into positive agricultural growth, including in foodgrain production. However, there is a need for sustaining this uptrend in public investment and also take appropriate measures to increase efficiency of public investment

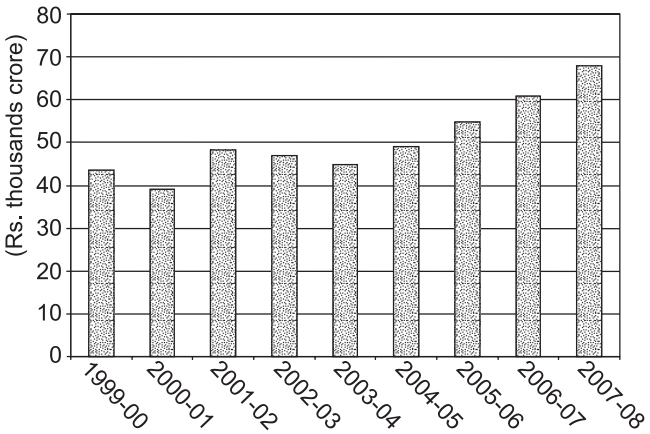

Fig. 6. Trend in Gross Capital Formation in Agriculture in India, 1999-00 prices

through targeting in high-payoff areas and promoting institutional innovations to manage the resources and infrastructure so created (Roy and Pal 2002). Surface irrigation is one area where most of public investment has been made and institutional reforms can help modernize this and other sources of surface irrigation, e.g. irrigation tanks. It is also suggested that part of input subsidies could also be diverted to long-term investment which will also encourage private investment and thereby contribute to higher food grain production.

Delivery of farm inputs and services is another area which needs government attention. There is increasing participation of private sector in delivery of inputs like fertilizer, seeds, pesticides etc which needs to be encouraged. However, there is lot of information asymmetry and quality problems in input markets which should be addressed through appropriate regulatory measures. Improved seed needs especial attention because of its immediate impact on crop productivity. There is a need to address quality issues, increased information flow and now encourage early flow of protected varieties into farmers' seed system. Planning material for horticultural crops needs adequate attention of the government as not many players are action in this area and farmers need not only planting material but also reliable information to make varietal choice involving long-term investment decisions (Pal and Tripp 2002).

Bridging yield gap which varies between 60 to 100 per cent depending upon crop and region, is a high payoff option. The gap is especially high in the eastern region, both for rice and wheat, and bridging this yield gap will contribute to substantial increase in foodgrain production. For dryland area, yield gap is an issue but more importantly raising yield level through water conservation measures can contribute to higher productivity of cereals and pulses. For this watershed development approach is followed but this needs effective participation of farming community for ensuring its success. Of course, revitalization of state extension system with more resources, accountability to stakeholders and linkages with NARS needs immediate attention. Since this revitalization process may take some time and there frontline extension system (KVKs) and state agricultural universities (SAUs) should take proactive steps to fill the gap. Some of the SAUs need funding and other support which should be accorded high priority (NAAS 2009).

The government has also taken some important steps to accelerate agricultural growth and meet the target of foodgrain production. The National Food Security Mission is being implemented by the Government in XI plan in 312 districts of 17 states with a budget of Rs 4882 crore. The target is to produce additional 20 MT of food grains during XI Plan. Efforts made by the government have successfully raised food grain production during 2006-07 and 2007-08. The government is also encouraging state governments to step investment in agriculture through Rashtriya Krishi Yojana. This scheme essentially focuses on decentralized planning and implementation of agricultural development in the country.

Climate Change and Food Security

Climate change is a global phenomenon and needs a global response for addressing this challenge. Agriculture is contributing to this climate change, mainly through changes in land use pattern, which in turn will have far reaching impacts on agriculture, especially in developing countries. Although the specific impacts and their magnitude of climate change will be known during its tipping point, but some broad directions are well understood. These include water shortages due to shrinking of glaciers, increased risk due to erratic weather, increased pest problems, and impact on crop yields and crop distribution. Indian agriculture is already prone to weather events and further intensification of erratic events will make it more vulnerable. IPCC has projected 1-3°C increase in temperature which may reduce yield up to 10 per cent in 2020 for Asia (IPCC 2007). Recent Indian study indicates 5-7 per cent decline in wheat yield for every degree Celsius increase in temperature given that current level of irrigation does not erode (Aggarwal 2008). Some of these adverse impacts are already seen on wheat yield due to changes in temperature. These yield impacts, coupled with reduced availability of

water will have tremendous pressure on productivity of irrigated system like rice-wheat system. Obviously, this will have significant implications for the national food security. A long term adaptation and mitigation strategy has to be evolved for Indian agriculture. This should entail development of heat tolerance varieties, change in land use and management practices, risk management, efficient use of water resources, and carbon sequestration. For ensuring food security there are some other implications which shall need urgent attention. This will include storage of more food grains which have cost implications for the government.

Agricultural Diversification

The trend of agricultural diversification towards high value commodities will further intensify after revival of the economy from current slowdown. Demand for Indian agricultural products in world market will further intensify this trend. This raises the question of food grain security versus diversification. But a rational approach could help realize both these objectives. Self-sufficiency in foodgrain production is a must given the global food scenario and therefore foodgrain production should be increased. This will be mainly yield driven. Increase in yield of food grains is also essential to release some area for high value commodities which are in high demand. In addition, there is considerable area under rice-fellow in eastern region which could be brought under cultivation by adopting better moisture management practices. Technological options like single cross hybrids in maize, hybrid rice, system of rice intensification, water saving technologies (drip and sprinkler irrigation), IPM in pulses and site-specific nutrient management can contribute significantly to yield improvement of food grain crops.

Management of Food Economy

Current policy of government interventions in foodgrain markets should continue; in fact, its need will be felt more in the years to come. This is because private trade is not well developed in large part of the country and farmers should be ensured remunerative price of their products. This coverage, as mentioned earlier should be extended to other food surplus areas. Secondly, variability in foodgrain production may increase because of erratic weather events and year-to-year changes in food grain production in absolute quantity may be quite high. This implies more storage of food grains which will need more resources for creating storage capacity and carrying stocks. The need for such an intervention is underscored by likely shortfall in rice production in 2009-10. Official statistics on first advance estimate just released pointed a shortfall to the order to 15 MT of rice. In order to minimize the cost, the feasibility of community food storage should be explored.

6. CONCLUDING REMARKS

We may conclude from the foregoing discussion that India has successfully achieved self-sufficiency in food grain production and ensured physical and economic access of the poor to food. All sign points to continuation of this policy and ability to meet the growing demand for food. However, there are some concerns which need attention for avoiding any adverse scenario. First is paying more attention to pulse crop for raising their productivity to a reasonable level. Second issue is maintaining uptrend in wheat productivity which has slipped for few years because of adverse weather conditions, particularly rise in temperature. This also raises the issue of breeding and crop management strategy for adaptations and mitigation of climate change. But immediate impact in terms of higher yield will be seen through application of available stock of knowledge and technology for bridging the yield gap and wherever possible raising yield potential. Finally, increasing food production and making it available to people may not solve the problem of hunger and malnutrition, which will require increasing income of poor through creation of employment opportunities and promoting agricultural diversification towards high value commodities like fruits and vegetables.

REFERENCES

Aggarwal, P.K. (2008). Global climate change and food security in South Asia: An adaptation and mitigation framework. Lead paper for the international symposium on “Climate Change and Food Security in South Asia,” Dhaka, 25-30 Aug 2008.

Chand (2009). *Demand for foodgrains during 11th Plan and towards 2020*. Brief 28, NCAP, New Delhi.

CSO (2008). *National Accounts Statistics*. New Delhi.

IFPRI (1999). World Food Prospects: Critical issues for the early 21st century. Washington, DC.

IPCC (2007). *Climate Change 2007: Impacts, Adaptations and Vulnerability*. Fourth Assessment Report of IPCC. Cambridge University Press.

Kumar, P., Joshi, P.K. and Birthal, P.S. (2009). Demand projects for foodgrains in India. *Agril. Eco. Res. Rev.,* **22**, 237-243.

NAAS (2009). *State of Indian Agriculture*. New Delhi.

Pal, Suresh and Tripp, R. (2002). India's seed industry reforms: Prospectus and issues. *Ind. J. Agric. Eco.,* **57(3)**, 443-458.

Planning Commission (2002). *Indian Vision 2020*. New Delhi.

Roy, B.C. and Pal, Suresh (2002). Investment, agricultural productivity and rural poverty in India: A state-level analysis. *Ind. J. Agric. Eco.,* **57(4)**, 653-678.

Arnold, F., Nangia, P. and Kapila, U. (2003). Indicators of nutrition for women and children in India: Current status and programme recommendations. Presented at workshop on *National Family Health Survey*, Centre for Economic and Social Studies, Hyderabad.

Radhakrishna, R. and Ravi, C. (2004). Malnutrition in India: Trends and determinants. *Eco. Pol. Weekly*, February 14.

Radhakrishna, R. (2006). Food Consumption and nutritional status in India: Emerging trends and perspectives. Keynote Paper for presentation at the *66th Annual Conference of the Indian Society of Agricultural Economics,* Umiam (Barapani), Meghalaya, November 8-10.

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 219-225

A Reflection on the Choice of Covariates in the Planning of Experimental Designs*

Bikas K. Sinha**

Faculty, Bayesian and Interdisciplinary Research Unit [BIRU], Applied Statistics Division, Indian Statistical Institute, Kolkata

SUMMARY

In the context of design of experiments, reference is drawn to well-known examples involving the use of covariates. Recent advances in the Theory of Optimal Covariates' Designs suggest possibility of significant improvements through 'optimal' choice of the covariate values. A blend of theory and applications is discussed.

Key words : Optimal designs of covariates models, A-optimality, Efficiency.

1. INTRODUCTION

Optimal Designs for Covariates Models is of relatively recent research interest and preliminary results were discussed in the papers by Lopes Troya (1982a, 1982b). After a considerable gap, renewed interest in this fascinating topic was found in the work of Liski *et al.* (2002), Das *et al.* (2002) and Rao *et al.* (2003). Since then this topic has picked up a momentum and has grown quite rapidly. See a few further references, listed at the end, on theory and applications in various design settings.

The purpose of this article is to popularize this area of research in the statistical community of teachers and researchers by narrating some of the results with examples of experiments taken from standard text books.

The settings are those of ANCOVA Models in the standard design layouts such as CRDs, RBDs, LSDs, BIBDs etc. The focus is on optimal / most efficient estimation of covariates' parameters incorporated in the model(s). The combinatorially challenging problem is to accommodate maximum number of covariates in an optimal manner in different design settings. Considerable effort and attention have been paid on these issues. We do not intend to touch upon these issues here. See the references cited above. Instead, we enter into a detailed discussion of results in some standard set-ups, following examples from standard text books such as Montgomery, Cochran and Cox, Federer, Hinkelman and Kempthorne:

Example 1. Comparison of Three Machines

Study variable: breaking strength of material in lbs;
Covariate: diameter of the material in 10^{-3} inches

Example 2. Comparison of Three Hand-trucks

Study variable: delivery time in minutes;
Covariate: volume delivered [in litres]

Example 3. Comparison of Four Glue Formulations

Study variable: strength of raw material in lbs;
Covariate: thickness of raw material in 0.01 inches

Example 4. Comparison of Three Cutting Speeds
Study variable: amount of metal removed;
Covariate: hardness of the specimen

**Dr. V.G. Panse Memorial Lecture delivered at 63rd Annual Conference of the Indian Society of Agricultural Statistics held at Pusa, Samastipur, Bihar during Dec. 3-5, 2009*

***E-mail address* : sinhabikas@yahoo.com

Example 5. Comparison of Three Chemical Processes
Study variable: yield or final product;
Covariate: impurity of the raw material

Example 6. Comparison of Six Varieties of Corn
Study variable: yield (lbs/plot) of ear corn;
Covariate: number of plants in each plot

2. COMPLETELY RANDOMIZED DESIGNS

Consider the simplest set-up of varietal designs viz., Completely Randomized Designs [henceforth abbreviated as CRDs] involving v treatments and n experimental units, with the treatment allocation numbers $n_1, n_2, ..., n_v$ so that $n_1 + n_2 + ... + n_v = n$. Suppose there is also available a controllable covariate X, assuming values in a finite closed interval $[a, b]$ on every experimental unit in the design layout.

Our purpose in this section is to make a comparison of the designs already used [as illustrated through the examples in text books cited above] and the alternative designs that could be suggested involving one or two covariates.

3. CRD: THEORETICAL CONSIDERATIONS

In a CRD, given n and v, various optimality considerations suggest $n_1 = n_2 = ... = n_v = n/v$ whenever n is divisible by v. We assume this divisibility condition to be satisfied and set $n_0 = n/v$. Further, by a location and scale change applied to each available covariate, it is assumed that for each one, the range of covariate values is the closed interval $[-1, 1]$. Set Z as the single covariate, which possesses values $z_{i_1}, z_{i_2}, ..., z_{i_{n0}}$ on the experimental units underlying the i-th treatment. Optimality considerations suggest $z_{ij} = +1/-1$ for each (i, j) combination and at the same time, $\Sigma_j z_{ij} = 0$ for each $i = 1, 2, ..., v$. This is possible only when there are plenty of experimental units available and one can identify such units with specified values of the covariate. We again assume this is feasible in an experimental CRD set-up. Whenever n_0 is an even number, routine split-half rule can be implemented to select the experimental units underlying each treatment, possessing exactly 50-50 per cent $+1/-1$ as the associated Z-values. This ensures most efficient CRD, as regards simultaneous inference on the treatment contrasts [of the form $(\tau_i - \tau_j)$] as also on the 'regression coefficient', say γ, involved in the model.

We now refer to Example 3 dealing with four glue formulations. The data refer to the chosen experimental units having the covariate values [thickness of the glue in suitable unit] as given below:

Formulation 1 : 12, 12, 13, 14, 14

Formulation 2 : 10, 11, 12, 12, 14

Formulation 3 : 10, 11, 11, 14, 15

Formulation 4 : 10, 11, 12, 15, 16

We notice that the minimum and maximum values of the covariate values covered in the study are respectively given by 10 and 16.

Had there been only 4 experimental units under each treatment, an optimal choice of the experimental units would correspond to those having the covariate values as 10, 10, 16 and 16. However, in this case there are 5 units under each treatment but we also have 4 treatments to be compared.

We propose to develop some comparison results in this section, based on the above features of the data.

Assume n_0 to be odd and v to be an even number. We work with the location-scale changed Z-values in the closed interval $[-1, 1]$ so that '10' corresponds to '-1' and '16' corresponds to '$+1$'. The model underlying a CRD is of the form

$$y_{ij} = \mu + \tau_i + \gamma z_{ij} + e_{ij}$$

Since n_0 is odd, we set the value '$+1$' for $(n_o - 1)/2$ units and the value '-1' for other $(n_0 - 1)/2$ units, leaving exactly one unit with the Z-value unspecified at z_i for treatment i.

We propose to develop results pertaining to 'optimal' choice of these $z_1, z_2, ..., z_v$ values. In order that the γ parameter is estimated 'orthogonally' [i.e., orthogonal to μ), we must have $\Sigma_i z_i = 0$ and we impose this condition to start with. Further to this, information on γ is maximized whenever the z_i's assume values $+1/-1$. Since v is even, this is also possible to achieve. On the other hand, let us check the nature of the C-matrix for the treatment parameters in the general set-up with the $z_1, z_2, ..., z_v$ values. It is shown in the Appendix that

$$C_z = n_0 I_v - n_0^2 J/n - zz'/(n - v + T(z)_2)$$
$$= n_0(I_v - J_v/v) - zz'/(n - v + T(z)_2)$$

where $z = (z_1, z_2, ..., z_v)'$, $T(z)_2 = \Sigma z_i^2$

Therefore,

$$\text{trace } (C_z) = n_0(v - 1) - \frac{T(z)_2}{n - v + T(z)_2}$$

It turns out that the trace is maximum when $T(z)_2 = 0$ i.e., when $z_i = 0$; $i = 1, 2, ..., v$. In that case, C-matrix is also completely symmetric (c.s.) and hence this choice leads to Universally Optimal [U.O.] design for treatment parameters. But on the other hand, an expression for the information on γ is given by $n - v + T(z)_2(n_0 - 1)/n_0$ which attains its maximum when $T(z)_2 = v$ i.e., when $z_i = +1/-1$ subject to $\Sigma z_i = 0$. This is a contrasting scenario. To arrive at a compromise solution, we consider the full parameter vector of the form

$$\eta = [\gamma \; P\tau]$$

where $P\tau$ refers to a full set of orthonormal treatment parameter contrasts.

For average variance as also for generalized variance, we need to work with the positive eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{v-1}$ of C_z-matrix which are given by n_0 with multiplicity $(v - 2)$ and $n_0 - \dfrac{T(z)_2}{n - v + T(z)_2}$.

For D-optimality, we need to maximize $\lambda_1.\lambda_2...\lambda_{v-1}.(n - v + T(z)_2(1 - n_0^{-1}))$ which is equivalent to maximizing

$$\left[n_0 - \frac{T(z)_2}{n - v + T(z)_2}\right](n - v + T(z)_2 (1 - n_0^{-1}))$$

or, equivalently,

$$[(n - v + T(z)_2(1 - n_0^{-1}))]^2/[(n - v + T(z)_2)]$$

It turns out that this last expression attains its maximum when $T(z)_2 = v$. Hence, over-all D-optimal design corresponds to $z_i = +1/-1$-values subject to $\Sigma z_i = 0$.

For A-optimality, it trivially follows that we need the same set of solutions for the z-values.

Thus, when the treatment parameter contrasts are to be estimated without any reference to the covariate parameter γ, a characterization of UO design obtains whenever $T(z)_2 = 0$ i.e., z_i's are all 0's. This is based on the properties of (i) maximum trace and (ii) cs of the C-matrix. On the other hand, taking into consideration also the γ parameter, specific optimality criteria lead to a different characterization viz., one with z_i's equal to $+1/-1$. We would tend to recommend the latter design in practice since consideration of the covariate parameter is also of importance in such models.

Turning back to the illustrative Example 3, our recommendation [Recommendation I] would have been

Formulation 1 : 10, 10, 16, 16, 16

Formulation 2 : 10, 10, 16, 16, 16

Formulation 3 : 10, 10, 10, 16, 16

Formulation 4 : 10, 10, 10, 16, 16

and it would lead to the C-matrix given by

$$C_0 = [(3.7, -1.3, -1.2, -1.2); (3.7, -1.2, -1.2); (3.7, -1.3); (3.7)]$$

whereas the original design yields the C-matrix given by

$$C_1 = [(3.63, -1.1080, -1.1954, -1.3264); (3.5822, -1.3187, -1.1597); (3.7252, -1.2152); (3.7013)]$$

The positive eigenvalues of the two C-matrices are given by [5.00, 5.00, 4.80] and [5.00, 5.00, 4.63] respectively. Hence, trivially, a comparison between the two C-matrices results in the superiority of C_0 over the other in respect of (i) the trace, (ii) the product of the eigenvalues, (iii) the sum of reciprocals of the eigenvalues and (iv) the smallest eigenvalue.

As regards estimation of the γ-parameter, it is known that information on γ in a CRD model is proportional to $SSW(z) = \Sigma\Sigma(z_{ij} - \bar{z}_i)^2$. Computations yield :

$SSW(z) = 69.20$ for the given CRD whereas this is 172.80 for the modified design.

Further to this, from theoretical considerations, we could as well recommend [Recommendation II] the design

Formulation 1 : 10, 10, 13, 16, 16

Formulation 2 : 10, 10, 13, 16, 16

Formulation 3 : 10, 10, 13, 16, 16

Formulation 4 : 10, 10, 13, 16, 16

which would be best [UO] for inference on the treatment contrasts, without any reference to the γ parameter. The C-matrix for this design is c.s. with the scalar multiplier being 5.00 and hence it has maximum trace. Another alternative to this design [Recommendation III] would be

Formulation 1 : 10, 13, 13, 13, 16

Formulation 2 : 10, 13, 13, 13, 16

Formulation 3 : 10, 13, 13, 13, 16

Formulation 4 : 10, 13, 13, 13, 16

which would have the same form of the C-matrix for treatment comparisons !

As to the status of these two latter designs for inference on γ-parameter, it is evident that these two alternatives should perform in between the earlier two ! Computations yield the $SSW(z)$ values for these designs respectively as 144.00 and 72.00. Therefore, gradual improvements are seen in the order $69.20 < 72.00 < 144.00 < 172.80$.

Before concluding this section, we may as well mention the context of a CRD when two covariates are involved in the experiment. This time, we want the two covariate parameters to be orthogonally estimated between them as also to be so with respect to each of the treatment parameters. This can be achieved by adopting the experimental units with the following pairs of values of the covariates, based on our earlier Recommendation I:

Formulation 1 : *Lm LM Hm HM HM*

Formulation 2 : *Lm LM Hm HM Hm*

Formulation 3 : *Lm LM Hm HM LM*

Formulation 4 : *Lm LM Hm HM Lm*

In the above, '*L*' and '*H*' denote the Lowest and Highest Values of one covariate and '*m*' and '*M*' represent those of another covariate. It turns out that both the covariate parameters are orthogonally and most efficiently estimated, each with information given by 19.20, in the standardized scale wherein the range of each covariate is taken to be [–1, 1]. For the treatment parameters, we have similar results as in Recommendation I. This time there are two covariates and so the C-matrix will be given by

C_0 = [(3.65, –1.25, –1.25, –1.15); (3.65, –1.15, –1.25); (3.65, –1.25); (3.65)]

whose eigenvalues would be 5.00, 4.80, 4.80.

With our Recommendation II - extended to the case of two covariates, the experimental units should be selected by adhering to the pairs of covariate values as given below :

Formulation 1 : *Lm LM Hm HM NN*

Formulation 2 : *Lm LM Hm HM NN*

Formulation 3 : *Lm LM Hm HM NN*

Formulation 4 : *Lm LM Hm HM NN*

where '*N*' refers to the mid-value of the covariate range for each covariate. This time, as before, the treatment comparisons would achieve UO status while the covariate parameters would be suboptimally estimated with $SSZ = 16.00$ for each one measured in the standardized scale of [–1, 1], as before.

4. RANDOMIZED BLOCK DESIGNS [RBDs]

The study of optimal use of covariates in the set-up of an RBD was initiated in Das *et al.* (2002) and continued further in subsequent work by Rao *et al.* (2003). Here we will take up an illustrative example and discuss some results.

We refer to Example 6 which deals with a Latin Square Design [LSD] of order 6 involving one covariate. Here are the covariate (X) values across different rows :

(18, 16, 18, 14, 15, 17)

(16, 15, 16, 19, 21, 14)

(15, 16, 16, 19, 15, 17)

(18, 18, 20, 18, 22, 17)

(15, 18, 19, 16, 16, 15)

(18, 20, 19, 17, 17, 15)

We use this example for illustrative purpose and assume that instead of a Latin Square Design, it may be treated as a Randomized Block Design [RBD] involving 6 blocks and 6 treatments with the rows representing the blocks and the columns representing the treatments. (In Section 6, we will discuss results related to the Latin Square Design. Also in Section 5, we will discuss further results, suitably modifying this example into that of a Balanced Incomplete Block Design [BIBD]).

Note that the minimum and maximum values of the covariate covered in the given design are respectively given by 14 and 22. Had there been available suffcient number of 'experimental units' with only these covariate values, we would have 'profitted' substantially by using 18 of each kind in the manner explained below.

(14, 14, 14, 22, 22, 22)

(14, 14, 14, 22, 22, 22)

(14, 14, 14, 22, 22, 22)

(22, 22, 22, 14, 14, 14)

(22, 22, 22, 14, 14, 14)

(22, 22, 22, 14, 14, 14)

In this format, in terms of the 'transformed' values [14 converted to '–1' and 22 converted to '+1'], it turns out that the 36 × 1 vector of the covariate values [written in a lexicographic order, starting with the elements of the first row, for example], is 'orthogonal' to the vectors representing the block effects as also the treatment effects. Hence the presence of the covariate values does not influence the analysis of the RBD nor is the information on the covariate parameter reduced because of the block and treatment effects. Thus we have an optimal display of the covariate values for most effcient estimation of the covariate parameter. It may be mentioned in passing that the above specification is not unique for optimal estimation of the covariate parameter. It may also be noted that most efficient estimation [in terms of orthogonality with block and treatment effects parameter-vectors] may not possible in other set-ups viz., BIBDs or LSDs. These will be studied in the next two sections.

5. BALANCED INCOMPLETE BLOCK DESIGNS [BIBDs]

This time, we slightly modify the layout to serve as a Symmetric BIBD [SBIBD] with parameters $b = v = 6$, $r = k = 5$ and $\lambda = 4$. We reproduce the design layout for covariate values, considering the rows representing the blocks and columns representing the treatments and using the incidence matrix as $\mathbf{N} = \mathbf{J} - \mathbf{I}$.

(–, 16, 18, 14, 15, 17)
(16, –, 16, 19, 21, 14)
(15, 16, –, 19, 15, 17)
(18, 18, 20, –, 22, 17)
(15, 18, 19, 16, –, 15)
(18, 20, 19, 17, 17, –)

Optimality theory has been developed for situations dealing with SBIBDs having $k = r = 0$ (*mod*4). Here we have $r = k = 5$ and we, therefore, propose to introduce a highly efficient design. We denote the least value of the covariate by '*L*' and the maximum value of the covariate by '*M*'. Here we propose use of an SBIBD with the following distribution of the covariates :

(–, *L*, *H*, *L*, *H*, *L*)
(*L*, –, *H*, *L*, *L*, *H*)
(*L*, *H*, –, *H*, *L*, *L*)
(*H*, *L*, *H*, –, *L*, *H*)
(*H*, *H*, *L*, *H*, –, *L*)
(*H*, *L*, *L*, *H*, *H*, –)

As mentioned before, we will take $L = -1 < 1 = M$ as the two extreme values of the covariate. The row totals of the covariate values against the blocks are $\delta_{bl} = (-1, -1, -1, 1, 1, 1)$ and those against the column totals are $\delta_{tr} = (1, -1, 1, 1, -1, -1)$. It turns out that the incidence matrix for the treatment parameters and the covariate parameter is a $(v + 1)\ v + 1$ matrix as outlined below. It is a partitioned matrix in the form of $[P\ Q]$; $[Q'\ R]$ where

$P = C$-matrix for the BIBD
$Q' = \delta_{tr} - k^{(-1)}\, \delta'_{bl}\, N'$
$R = 30 - 6/k$, a scalar

Hence, the information matrix for the treatment parameters, after eliminating the effect of the covariate parameter, is given by $C - D$ where $D = QR^{(-1)}Q'$. It follows that D has rank 2 and its positive eigenvalues δ_1 and δ_2 satisfy the relations : (i) sum = (6 + 6/25)/R; (ii) product = $32/25R^2$ whence the eigenvalues are given by 0.2093 and 0.0074.

In the absence of the covariate parameter in the model, the SBIBD would have constant non-zero eigenvalue of the C-matrix given by $\lambda v/k = 4.8$. Hence the eigen-values of the C-matrix in the presence of the covariates are given by 4.8, 4.8, 4.8, 4.8 – 0.2093 = 4.5907 and 4.8 – 0.0074 = 4.7926.

Therefore, efficiency of the recommended design is computed as

Efficiency = [5/4.8] / [3/4.8 + 1/4.5907 + 1/4.7926]
= 99.07 per cent

As regards the covariate parameter, in the absence of any impact of block effects and treatment effects, the information would have been 30. However, because of non-orthogonality, it gets reduced to $R - Q'P^{(-)}Q$ which simplifies to 27.712. Hence the efficiency of the design is computed as 92.37 per cent.

6. LATIN SQUARE DESIGN

Now we turn to the Example 6 [without any modification of the layout] which describes a Latin Square Design [LSD] of order 6 involving one covariate was utilized. Here are the covariate (X) values across the different rows :

(18, 16, 18, 14, 15, 17)
(16, 15, 16, 19, 21, 14)
(15, 16, 16, 19, 15, 17)
(18, 18, 20, 18, 22, 17)

(15, 18, 19, 16, 16, 15)

(18, 20, 19, 17, 17, 15)

The original LSD is slightly modified here to demonstrate the strength of the theory for optimal choice of covariate values. Here is the ORIGINAL LSD, with treatment allocations shown row-wise :

(A, B, C, D, E, F)

(B, C, D, E, F, A)

(C, E, A, F, B, D)

(D, F, B, A, C, E)

(E, D, F, B, A, C)

(F, A, E, C, D, B)

Below we display the MODIFIED LSD :

(A, B, C, D, E, F)

(B, A, D, C, F, E)

(F, E, A, B, C, D)

(E, F, B, A, D, C)

(D, C, F, E, B, A)

(C, D, E, F, A, B)

We shall work with the modified LSD. Computations yield as follows [ignoring the variance σ^2 in the expressions below] :

(i) C-matrix for the treatment parameters, after eliminating row, column and covariate effects, is given by

(4.9929, –0.9955, –0.9839, –1.0264, –1.0110, –0.9761)
(4.9971, –1.0102, –0.9832, –0.9930, –1.0152)
(4.9633, –0.9399, –0.9751, –1.0542)
(4.9014, –1.0409, –0.9110)
(4.9832, –0.9632)
(4.9197)

The eigenvalues are 6.00020, 5.99999, 5.99995, 5.99988, 5.75759.

(ii) The information on the covariate parameter is given by $I(\gamma) = 487.8333$.

Next note that the minimum and maximum values of the covariate covered in the given design are respectively given by 14 and 22. Had there been available sufficient number of 'experimental units' with only these covariate values, we would have 'profited' substantially by using 18 of each kind in the manner explained below.

(22, 22, 22, 14, 14, 14)

(22, 22, 14, 14, 14, 22)

(22, 14, 14, 14, 22, 22)

(14, 14, 14, 22, 22, 22)

(14, 14, 22, 22, 22, 14)

(14, 22, 22, 22, 14, 14)

For the modified LSD, it turns out that with the above recommendation on the choice of experimental units with the stated values of the covariates, the C-matrix is completely symmetric with eigenvalues each equal to 6.0000 and it has maximum trace. Hence for treatment comparisons, the LSD is UO !

Further to this, the information on the covariate parameter is given by 576.00 as against 487.8333 computed above.

So, strictly speaking, the given design is quite efficient with respect to treatment comparisons but has scope for improvement in terms of precision on the covariate parameter.

Consider again the modified LSD but this time, with deletion of the last column. Then the Row Design reduces to a BIBD ($b = v = 6$, $r = k = 5$, $\lambda = 4$) while the Column Design is still an RBD. So it is a Youden Design. In the presence of a covariate (X), we can still obtain UO Youden Design for treatment comparisons provided the covariate values are chosen as follows ['H' for highest value i.e., 22 and 'L' for lowest value i.e., 14 and '0' for mid-value i.e., 18] :

(A-H, B-L, C-H, D-L, E-0)

(B-L, A-H, D-L, C-H, F-0)

(F-H, E-L, A-L, B-H, C-0)

(E-L, F-H, B-H, A-L, D-0)

(D-H, C-L, F-L, E-H, B-0)

(C-L, D-H, E-H, F-L, A-0)

The explanation is quite straightforward. The Youden Design has all its treatment parameter vectors orthogonal to the covariate parameter vector in the model description.

7. APPENDIX

We set $\theta=(\mu, \gamma, \tau_1, \tau_2, ..., \tau_v)$ and note that for the CRD with z-values [subject to $\Sigma z_i = 0$] underlying the experimental units, as indicated, the information matrix for θ is given by

$$\mathbf{I}(\theta; z) = [(n, 0, n_0, n_0, ..., n_0); (n - v + T(z)_2, z_1, z_2, ..., z_v); (n_0, 0, 0, ..., 0); ...., (n_0)]$$

From the above, we deduce that for the vector parameter τ, the C-matrix is given by

$$\begin{aligned} C_z &= n_0.\mathbf{I} - [(n_0, z_1); (n_0, z_2); ...; (n_0, z_v)] \\ &\quad [(n, 0); (n - v + T(z)_2)]^{-1} \\ &\quad [(n_0, n_0, ..., n_0); (z_1, z_2, ..., z_v)] \\ &= n_0.\mathbf{I} - n_0^2.\mathbf{J}/n - zz'/(n - v + T(z)_2) \\ &= n_0(\mathbf{I} - \mathbf{J}/v) - zz'/(n - v + T(z)_2) \end{aligned}$$

Next, let us again set θ as before but this time, we transform the vector parameter τ to $\overline{0}\tau$ where $\overline{0}$ is an orthogonal matrix with the first row vector proportional to the vector of 1's and the submatrix of order $(v - 1) \times v$ is denoted by $\mathbf{P}$ so that $\mathbf{P}$ satisfies (i) $\mathbf{P}\ 1 = 0$, (ii) $\mathbf{P}\ \mathbf{P}' = \mathbf{I}$, (iii) $\mathbf{P}'\mathbf{P} = \mathbf{I} - \mathbf{J}/v$. Clearly, $\mathbf{P}\tau$ describes a full set of orthonormal treatment effect contrasts. We intend to derive the joint information matrix of $\mathbf{P}\tau$ and γ. Set $\Gamma = [(\mu, \gamma, \overline{0}\tau]$ so that $\Gamma = \mathbf{M}\theta$ where $\mathbf{M}$ = partitioned matrix with the first part an identity matrix of order 2 and the second part the matrix $\overline{0}$.

For this, we start with the C-matrix for θ as a whole and make a transformation from τ to $\overline{0}\tau$ and then proceed to derive the desired information matrix. We set, for brevity, T_2 for $T(z)_2$.

(i) $\mathbf{I}(\theta) = \mathbf{X}'\mathbf{X}$

(ii) $\mathbf{I}(\Gamma) = \mathbf{M}\mathbf{X}'\ \mathbf{X}\ \mathbf{M}' = [(n, T_1, n_0\ \sqrt{(v)}, 0, 0, .., 0); (n - v + T_2(1 - n_0^{-1}), 0, z'\mathbf{P}'), (n_0, 0, 0, ..., 0); (n_0, 0, 0, ... 0); ...; (n_0)]$

(iii) $\mathbf{I}(\gamma, \mathbf{P}\tau) = [(n - v + T_2, z'\mathbf{P}'); (n_0, 0, 0, .., 0); ..., (n_0)]$

(iv) $\mathbf{I}(\gamma) = n - v + T_2 - z'\mathbf{P}'\ \mathbf{P}z/n_0$
$= n - v + T_2\,(1 - n_0^{-1})$, as expected.

(v) $\mathbf{I}(\mathbf{P}\tau) = n_0\mathbf{I} - \mathbf{P}zz'\mathbf{P}'/(n - v + T_2)$

It turns out that the eigenvalues of $\mathbf{I}(\mathbf{P}\tau)$ are n_0 with multiplicity $(v\text{-}2)$ and $n_0 - T_2(n - v + T_2)^{-1}$.

We also note that the above expression for $\mathbf{I}(\mathbf{P}\tau)$ follows from the expression for C_z given earlier in view of

$$\mathbf{I}(\mathbf{P}\tau) = \mathbf{P}\mathbf{I}(C_z)\mathbf{P}'$$

since $\mathbf{P}\mathbf{P}' = \mathbf{I} - \mathbf{J}/v$ and $\mathbf{P}1 = 0$.

ACKNOWLEDGEMENTS

I sincerely thank Professor G.M. Saha for helping me with the computations and Professors N.K. Mandal, Premadhis Das, Manisha Pal and Dr. Ganesh Dutta for giving a patient hearing while I was working on the problem.

REFERENCES

Das, K., Mandal, N.K. and Sinha, Bikas K. (2002). Optimal experimental designs with covariates. *J. Statist. Plann. Inf.*, **115**, 273-285.

Dutta, G. (2004). Optimum choice of covariates in BIBD set-up. *Cal. Stat. Assoc. Bull.*, **55**, 39-55.

Das, Premadhis, Dutta, G. and Mandal, N.K. (2007). Optimum choice of covariates for a series of SBIBDs obtained through projective geometry. *Jour. Modern Appld. Stat. Methods*, **6**, 649-656.

Das, Premadhis, Dutta, G. and Mandal, N.K. (2009). Optimum covariate designs in split-plot and strip-plot design set-ups. *Jour. Appld. Stat.*, **36**.

Das, Premadhis, Dutta, G. and Mandal, N.K. (2009). Optimum covariate designs in partially balanced incomplete block (PBIB) design set-ups. *J. Statist. Plann. Inf.*

Das, Premadhis, Dutta, G. and Mandal, N.K. (2009). Optimum covariate designs in binary proper equi-replicate block design set-up. To appear in *Discrete Mathematics*.

Das, Premadhis, Dutta, G. and Mandal, N.K. (2010). D-optimal designs for co-variate parameters in block design set-up. To appear in *Comm. Statist.–Theory Methods.*

Liski, E., Mandal, N.K., Shah, K.R. and Sinha, Bikas K. (2002). *Topics in Optimal Design*. Lecture Notes Series in Statistics, No. 163, Springer.

Lopes Troya, J. (1982a). Optimal designs for covariates models. *J. Statist. Plann. Inf.*, **6**, 373-419.

Lopes Troya, J. (1982b). Cyclic designs for a covariate model. *J. Statist. Plann. Inf.,* **7**, 49-75.

Rao, PSSNVP, Rao, S.B., Saha, G.M. and Sinha, Bikas K. (2003). Optimal designs for covariates' models and mixed orthogonal arrays. *Electronic Notes in Discrete Mathematics*, **15**, 157-160, Elsevier.

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 227-235

Evaluation of Variation in Socio-Economic Development in the States of Eastern Region

Prem Narain, V.K. Bhatia and S.C. Rai*

Indian Society of Agricultural Statistics, New Delhi

SUMMARY

The status of development of different states of eastern region of the country was estimated with the help of composite index based on optimum combination of a number of socio-economic indicators. Five major states and seven smaller states of the region are included in the study. The data on various indicators for the year 2001-02 are used in the analysis. The level of development is examined separately for agricultural sector, infrastructural facilities and overall socio-economic field. West Bengal is ranked first among the major states and Mizoram obtains the first position among smaller states in socio-economic development. Wide disparities in the level of development are found among different states. Infrastructural facilities are found influencing the socio-economic development in the positive direction both for major and smaller states. For bringing out uniform regional development, potential targets of different indicators are estimated for low developed states.

Key words : Developmental indicators, Composite index, Model states, Potential targets.

1. INTRODUCTION

Socio-economic development is a process which improves the quality of life. Developmental programmes have been initiated in the country in a planned way through various Five Year Plans with the main objective of enhancing the quality of life of people by providing the basic necessities as well as effective improvement in their social and economic well being. The green revolution in agricultural sector and commendable progress in industrial front have certainly increased the total crop production and manufactured goods but there is no indication that these achievements have been able to reduce substantially the level of regional disparities in terms of development. For focusing the attention of scientists, planners, policy makers and administrators on the problems of estimation of level of development, a seminar was organized jointly by Planning Commission, Government of India and State Planning Institute, Government of Uttar Pradesh during April 1982. Realizing the seriousness and importance of the problems of estimation of level of development, the Indian Society of Agricultural Statistics conducted a series of research studies in this direction. The level of socio-economic development was estimated for different states for the year 1971-72 and 1981-82 by Narain *et al*. (1991). The study revealed that there were wide disparities in the level of development among different states. For making deeper analysis, the data mostly pertaining for the year 1991-92 were utilized for estimating the status of development at district level. Studies for estimating the level of development at district level have been completed by Narain *et al.* for the states of Orissa (1992, 1993); Andhra Pradesh (1994, 2009); Kerala (1994, 2004); Uttar Pradesh (1995, 2001); Maharashtra (1996); Karnataka (1997, 2003); Tamil Nadu (2000); States of Southern Region (1999); Madhya Pradesh (2002); and Jammu & Kashmir (2005) and Rai *et al.* for Assam (2004). It was found that entire parts of the low developed districts are not backward but some parts are also better developed.

* *Corresponding author* : S.C. Rai
E-mail address : naivedya.kashyap@gmail.com

The present study relates to the estimation of variation in socio-economic development in the states of eastern region of the country. The states of Arunachal Pradesh, Assam, Bihar, Jharkhand, Manipur, Meghalaya, Mizoram, Nagaland, Orissa, Sikkim, Tripura and West Bengal are situated in the eastern part of the country. Out of these twelve states, seven states namely Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura are smaller states in area and population. These states cover about 5.6 per cent area and 1.2 per cent population of the country. The states of Assam, Bihar, Jharkhand, Orissa and West Bengal are comparatively major states and they cover about 15.3 per cent area and 24.7 per cent population of the country. The level of development has been estimated separately for major and small states. These states are primarily agricultural states and this sector alone contributes more than 50 per cent of state domestic product and provides employment to about 70 per cent of the total working force. The data on socio-economic variables for the year 2001-02 have been utilized to estimate the level of development. The level of development for agricultural sector, infrastructural facilities and socio-economic field was estimated for different states. The study also throws light on the association of development in different sectors of economy. The improvements required in various indicators for enhancing the level of development have also been suggested.

2. DEVELOPMENTAL INDICATORS

Socio-economic development is a continuous process of improvement in the level of living. The level of development cannot be fully estimated by single indicator. Moreover, a number of indicators when analyzed individually, do not provide an integrated and easily comprehensible picture of reality. For this study, states have been taken as the unit of analysis. Each state faces situational factors of development unique to it as well as common administrative and financial factors. The developmental indicators common to all the states have been included in the study. The composite indices of development have been obtained for different states on the basis of following developmental indicators:

01. Yield rate of total cereals (kg/ha)
02. Yield rate of pulses (kg/ha)
03. Yield rate of total foodgrains (kg/ha)
04. Yield rate of sugarcane (kg/ha)*
05. Per capita foodgrains production (kg.)
06. Fertilizer consumption (kg/ha)
07. Percentage of gross irrigated area to gross cropped area
08. Net area sown per cultivator (ha)
09. Per capita domestic consumption of electricity (KWH)
10. Percentage of agricultural workers to total workers
11. Population density
12. Decennial growth rate of population (1991-92 to 2001-02)
13. Sex ratio
14. Percentage of SC and ST population
15. Percentage of main workers to total population
16. Literacy percentage (male)
17. Literacy percentage (female)
18. Life expectancy at birth (male)*
19. Life expectancy at birth (female)*
20. Annual birth rate
21. Annual death rate
22. Infant mortality rate
23. No. of students in primary and secondary schools per ’000 population
24. Per capita gross output in industries*
25. No. of motor vehicles per lakh population
26. Total road length per ’00 square km. of area
27. No. of fair price ration shops per lakh population
28. No. of banks per lakh population
29. Per capita bank deposit
30. Per capita bank credits

*These indicators are not included for the estimation of development in smaller states.

These indicators may not form an all inclusive list but these are the major interacting components of socio-economic development. Out of these 30 developmental indicators, 10 indicators are directly connected with the development in agricultural sector. Remaining 20 indicators demonstrate the level of infrastructural facilities. For smaller states 9 indicators are from agriculture sector and 17 indicators are from infrastructural facilities.

3. METHOD OF ANALYSIS

There are several statistical methods which are used for estimation of level of development but most of these methods are having their own limitations. The major limitation arises from the assumptions made about the developmental indicators themselves and their weightage in aggregate index. Keeping in view the limitations of different methods, the following statistical procedures are used in the study.

The variables for different indicators are taken from different population distributions and they might be recorded in different units of measurement. The values of the variables are not quite suitable for combined analysis. Hence, for combined analysis the values of the variables are transformed as given below.

Let $[\mathbf{X}_{ij}]$ be the values of the variables of ith state and jth indicator, where (i = 1, 2,….., n) and j = (1, 2,… k).

$[\mathbf{X}_{ij}]$ is transformed to $[\mathbf{Z}_{ij}]$ as follows

$$[\mathbf{X}_{ij}] = \frac{X_{ij} - \bar{X}_j}{S_j}$$

where $\bar{X}_j$ is the mean of jth indicator

and S_j is the standard deviation of jth indicator.

From $[\mathbf{Z}_{ij}]$, identify the best value of each indicator. The best value will be either the maximum value or minimum value of the indicator depending upon the direction of the impact of the indicator on the level of development. For obtaining the composite index of development, the statistical procedures given by Narain *et al*. (1991) are applied. The value of composite index is non-negative. Smaller values of composite indices indicate high level of development and higher values of composite indices indicate low level of development. Based on developmental distances between different states and composite indices of development, model states are identified and potential targets of various indicators are estimated for low developed states.

For classificatory purposes, a simple ranking of the states on the basis of composite index of development is sufficient. However, a more meaningful characterization of different stages of development can be obtained on the basis of Mean and S.D. of composite indices as given below:

It appears quite valid to assume that the states having the composite indices ≤ (Mean – SD) are in high developed category, the states having composite indices in between (Mean – SD) to (Mean) are in high middle level category; the states having composite indices in between (Mean) to (Mean + SD) are in low middle level category and the states having composite indices ≥ (Mean + SD) are in low developed category.

4. RESULTS AND DISCUSSIONS

4.1 The Level of Development

The composite indices of development have been worked out for different states of eastern region, in respect of agricultural sector, infrastructural facilities and socio-economic sector. The levels of development have been estimated separately for major states and smaller states. The states have been ranked on the basis of composite index of development. The composite indices and rank of the major states are presented in Table 1.

Table 1. The Composite Indices of Development and Ranks of Major States

S.No.	States	Agricultural Development		Infrastructural Facilities		Socio-economic Development	
		C.I.	Rank	C.I.	Rank	C.I.	Rank
1.	Assam	0.47	3	0.59	3	0.58	3
2.	Bihar	0.50	4	0.65	4	0.63	4
3.	Jharkhand	0.59	5	0.68	5	0.68	5
4.	Orissa	0.47	2	0.49	2	0.49	2
5.	West Bengal	0.09	1	0.31	1	0.28	1

It may be seen from the above table that the State of West Bengal has been ranked first and the State of Jharkhand has been placed on the last position among the eastern states in case of agricultural development. The composite indices varied from 0.09 to 0.59. In case of infrastructural facilities, West Bengal is found to be on the first place and Jharkhand is placed on the last position. The composite indices varied from 0.31 to 0.68. Regarding overall socio-economic development, West Bengal again found to occupy the first position whereas the State of Jharkhand is on the last place. The composite indices varied from 0.28 to 0.68. The State of West Bengal is observed to occupy the first place in respect of agricultural development, infrastructural facilities and overall socio-economic development among eastern states. The State of Jharkhand is on the last position in the above three sectors. The composite index of development for smaller states of eastern region is presented in Table 2. The states have been ranked on the basis of composite index.

In case of smaller states of eastern region, for agricultural sector, Tripura is found to be the best developed state whereas Meghalaya is on the last place. The composite indices varied from 0.57 to 0.86. With respect to infrastructural facilities, the State of Mizoram is on the first place and Nagaland is on the last position. The composite indices varied from 0.44 to 0.78. As regards socio-economic development, the State of Mizoram is found to be on the first position whereas the State of Arunachal Pradesh is on the last place. The composite indices varied from 0.55 to 0.81. When the status of development of these states were compared with the development of newly formed states of Jharkhand, Chhattisgarh and Uttarakhand, the State of Mizoram was found to occupy the first position in socio-economic development.

4.2 Different Stages of Development

On the basis of system of classification mentioned in section 3, the states are put in four stages of development as high, high middle, low middle and low. Table 3 presents the number of states along with the percentage area and population lying in different stages of development.

There are five major states in the eastern region. These states cover about 5 lakh sq. kilometer of area and about 2535.92 lakh population. Out of these five states, one state is found to be in the high developed category in agricultural sector. About 18 per cent area and 32 per cent population are covered by this state. Out of these five states, four states are found in low middle category of development. The area and population covered by these states are 82 per cent and 68 per cent respectively. None of the states is found in high middle category and low category of development.

In case of infrastructural facilities, one state having about 18 per cent area and 32 per cent population is in high category of development. One state is also found in high middle level developed category. This state is having about 31 per cent area and 14 per cent population. The remaining three states having about 51

Table 2. The Composite Indices of Development and Ranks of Smaller States

S.No.	States	Agricultural Development		Infrastructural Facilities		Socio-economic Development	
		C.I.	Rank	C.I.	Rank	C.I.	Rank
1.	Arunachal Pradesh	0.78	6	0.76	6	0.81	7
2.	Manipur	0.60	2	0.56	2	0.60	3
3.	Meghalaya	0.86	7	0.61	4	0.72	4
4.	Mizoram	0.70	4	0.44	1	0.55	1
5.	Nagaland	0.67	3	0.78	7	0.80	6
6.	Sikkim	0.72	5	0.70	5	0.75	5
7.	Tripura	0.57	1	0.56	3	0.59	2

Table 3. Number of States, Percentage Area and Population lying under Different Stages of Development

Stage of Development	Number of States		Area (%)		Population (%)	
Agricultural Development						
High	1	(1)	18	(5)	32	(26)
High Middle	–	(2)	–	(21)	–	(34)
Low Middle	4	(3)	82	(62)	68	(21)
Low	–	(1)	–	(12)	–	(19)
Infrastructural Facilities						
High	1	(1)	18	(11)	32	(7)
High Middle	1	(3)	31	(29)	14	(64)
Low Middle	3	(1)	51	(4)	54	(4)
Low	–	(2)	–	(56)	–	(25)
Socio-economic Development						
High	1	(2)	18	(18)	32	(33)
High Middle	1	(1)	31	(12)	14	(18)
Low Middle	3	(2)	51	(16)	54	(23)
Low	–	(2)	–	(54)	–	(26)

Note. Details regarding number of states, area and population percentages are given in brackets for smaller states.

per cent area and 54 per cent population are found in low middle category of development. None of these states is found in low developed category.

In case of socio-economic development, one state is in high developed category. The area and population covered by this state is 18 per cent and 32 per cent respectively. One state having about 31 per cent area and 14 per cent population is found in high middle level developed category. The remaining three states having 51 per cent area and 54 per cent population are found in low middle level developed category. None of the states is in low developed category.

West Bengal is the only major state in the eastern region which is found to be in high developed category in agricultural sector, infrastructural facilities and socio-economic sector.

There are seven smaller states in the eastern region. These states cover about 1.83 lakh square kilometer of area and about 122 lakh population. In agricultural sector, one state having 5 per cent area and 26 per cent population is found in high developed category. Two states are in high middle level developed category. These states are having about 21 per cent area and 34 per cent population. Three states are found to be in low middle level developed category. These states are having about 62 per cent area and 21 per cent population. One state with 12 per cent area and 19 per cent population is found in low developed category.

As regards infrastructural facilities, one state with 11 per cent area and 7 per cent population is found to be in high level category. Three states are found to be in high middle level category. These states cover about 29 per cent area and 64 per cent population. One state with 4 per cent area and 4 per cent population is found in low middle level developed category. Two states having 56 per cent area and 25 per cent population are found in low developed category.

With respect to socio-economic development, two states having about 18 per cent area and 33 per cent population are found to be in high developed category. One state is in high middle level developed category. This state is having about 12 per cent area and 18 per cent population. Two states are found in low middle level developed category. These states cover about 16 per cent area and 23 per cent population. Two states with 54 per cent area and 26 per cent population are found in low level developed category.

4.3 Inter-relationship among Different Sectors of Economy and Literacy Rate

For proper development, it is essential that different sectors of economy should flourish together. Similarly, system of education and literacy rate envisages all round development of manpower and human resources required for socio-economic activities. A large population below an acceptable economic level poses serious problems. Massive poverty in the country characterizes its economy. The correlation coefficients between the development of different sectors of economy and literacy level are given in Table 4.

For major states, agricultural development is found to be positively influenced by infrastructural facilities. Socio-economic development is positively affected by agricultural development. Infrastructural facilities also influenced the socio-economic development in the positive direction. Literacy rates for male and female do not influence the agricultural development and socio-economic development. However, literacy rates for male and female are positively associated among themselves. Infrastructural facilities do not affect the literacy rate.

In case of smaller states, agricultural development was not influenced by literacy rates for male and female and other infrastructural facilities. It did not influence the socio-economic development also. Infrastructural facilities influenced the socio-economic development and literacy rate for male and female in the positive direction. Literacy rates for male and female were positively associated among themselves.

4.4 Improvement required in Low Developed States

It is quite important and useful to examine the extent of improvement needed in various developmental indicators for the low developed states. This will help the administrators and planners to readjust the resources for bringing about uniform regional development. For estimation of potential targets of developmental indicators of low developed states, model states have been identified on the basis of composite index of development and developmental distances between different states. In case of major states, none of the states is found in low developed category. Three states namely Assam, Bihar and Jharkhand are found in low middle level developed category for overall socio-economic development. In case of smaller states, Arunachal Pradesh and Nagaland are found to be in low category of socio-economic development. List of model states for these low middle level and low level developed states is presented in Table 5.

Table 4. Correlation Coefficients

Factors	Agricultural Development	Infrastructural Facilities	Socio-economic Development	Literacy Rate (Male)	Literacy Rate (Female)
Agricultural Development	1(1)	0.94* (0.27)	0.96** (0.53)	–0.64 (–0.61)	–0.72 (–0.25)
Infrastructural Facilities		1 (1)	0.99** (0.96**)	–0.82 (.0.78*)	–0.81 (–0.79*)
Socio-economic Development			1 (1)	– 0.79 (–0.86*)	– 0.79 (–0.75)
Literacy Rate (Male)				1 (1)	0.93* (0.86*)
Literacy Rate (Female)					1 (1)

Note : Correlation coefficients for smaller states are given in brackets.

*significant at 0.05 probability level.

** significant at 0.01 probability level.

Table 5. Model States for Low Developed States

S.No.	Low Developed States	Model States
1.	Assam	West Bengal
2.	Bihar	West Bengal
3.	Jharkhand	West Bengal and Orissa
4.	Arunachal Pradesh	Manipur, Mizoram and Tripura
5.	Nagaland	Manipur, Mizoram and Tripura

Model states are better developed. West Bengal is found to be model state for middle level developed states of Assam, Bihar and Jharkhand. In case of smaller states, Manipur, Mizoram and Tripura are model states for low developed states of Arunachal Pradesh and Nagaland. The best value of the developmental indicators of model state is taken as potential target of low developed state. The present value of developmental indicators along with the potential target of low developed states is presented in Table 6.

It may be seen that values of potential targets for some of the indicators are quite high. Suitable action should be taken to achieve the potential target of developmental indicators for enhancing the level of development. For suggesting specific action at various locations in the state, study at district or tehsil/block levels might be conducted. However, broad suggestions for improving the level of development of low developed states are given below:

1. Assam

This State is low middle level developed in agricultural sector, infrastructural facilities and socio-economic sector. Yield rates of major crops in the State are quite low. Yield rates should be enhanced by creating irrigation facilities and using high doses of fertilizer. Educational, medical, banking and road transport facilities should be improved in the State.

Table 6. Present Value of Developmental Indicators and Potential Target

S.No.	Developmental Indicators	Assam	Bihar	Jharkhand	Arunachal Pradesh	Nagaland
01.	Yield rate of total cereals (kg/ha)	1469 (2488)	1537 (2488)	1359 (2488)	1238 (2398)	1669 (2398)
02.	Per capita foodgrains production (kg.)	144 (192)	128 (192)	103 (192)	215 (215)	196 (215)
03.	Fertilizer consumption (kg/ha)	48 (115)	88 (115)	51 (115)	3 (38)	2 (38)
04.	Gross irrigated area (%)	5 (51)	58 (58)	10 (51)	16 (28)	28 (28)
05.	Net area sown per cultivator (ha)	0.7 (1.0)	0.7 (1.0)	0.5 (1.0)	0.6 (0.9)	0.6 (0.9)
06.	Decennial growth rate of population (1991-92 to 2001-02)	19 (18)	29 (18)	23 (18)	27 (16)	65 (16)
07.	SC and ST population (%)	19 (19)	17 (17)	38 (29)	65 (37)	89 (37)
08.	Percentage of main workers	27 (29)	25 (29)	24 (29)	38 (41)	35 (41)
09.	Literacy rate (Male)	71 (77)	60 (77)	67 (77)	64 (80)	71 (80)
10.	Literacy rate (Female)	35 (60)	33 (60)	39 (60)	44 (87)	61 (87)
11.	Birth rate	25 (17)	30 (17)	27 (17)	23 (15)	16 (15)
12.	Death rate	8.7 (6.4)	8.1 (6.4)	7.9 (6.4)	5.9 (4.1)	3.8 (3.8)
13.	Infant mortality rate	68 (38)	61 (38)	50 (38)	37 (13)	18 (13)
14.	No. of motor vehicles (per lakh population)	2588 (3045)	854 (3048)	4295 (4295)	1842 (4238)	8226 (8226)
15.	No. of fair price ration shops (per lakh population)	123 (123)	57 (103)	31 (103)	91 (104)	21 (104)
16.	No. of banks (per lakh population)	4.2 (5.3)	3.9 (5.3)	5.1 (5.3)	5.9 (8.3)	3.4 (8.3)

Note : Potential target for different indicators is given in brackets.

2. Bihar

This State is low middle level developed in agricultural and socio-economic sectors. Infrastructural facilities are also not very good. Crop yields are generally poor. Steps should be taken to enhance the productivity of various crops by applying suitable doses of fertilizers and using irrigation facilities. Literacy rate among male and female is very low. Action is needed to enhance the literacy rate in the State. Educational, medical, transport and banking facilities should be improved in the State.

3. Jharkhand

The State is found to be low middle level developed in agricultural sector and socio-economic field. Major improvements are required to improve the crop yield by providing irrigation facilities and suitable doses of fertilizer. Improved technique of dry land farming might also be adopted in the area where irrigation facilities are not sufficient. Infrastructural facilities for transport, medical and banking may be improved. The level of literacy for female is very low. Steps should be taken to enhance the literacy rate among women.

4. Arunachal Pradesh

This State is low developed in agricultural sector and infrastructural facilities. Productivity of crops is very low. Steps should be taken to enhance the crop productivity by increasing irrigation facilities and using fertilizers. Literacy rate among female is low which requires improvement. Medical and transport facilities should be improved.

5. Nagaland

The State is low developed in socio-economic field and infrastructural facilities. The crop productivity should be improved by irrigation and use of fertilizers. Infrastructural facilities are very poor. Medical, transport and banking facilities should be improved. Literacy rate among male and female population is quite good but steps are needed to make further improvement in it.

5. CONCLUSIONS

The broad conclusions emerging from the study are as follows :

1. Among the major states of eastern region, the state of West Bengal is found to be better developed as compared to the remaining four states of Assam, Bihar, Jharkhand and Orissa in socio-economic field. In case of smaller states, Mizoram and Tripua are found to be better developed as compared to the remaining five states. The state of Arunachal Pradesh and Nagaland are found to be low developed.

2. As regard agricultural development, the state of West Bengal is better developed in comparison to other major states of eastern region. In case of smaller states, Tripura is found to be better developed in comparison to other states. The state of Meghalaya is low developed.

3. Infrastructural facilities regarding education, medical, banking and road transport are found in high category in the state of West Bengal. In case of smaller states, these facilities are better in Mizoram as compared to other smaller states of the region.

4. In case of major states, overall socio-economic development is found to be positively associated with the development in agricultural sector. Infrastructural facilities also influence the socio-economic development in positive direction. For smaller states, infrastructural facilities influence the socio-economic development in the positive direction. These facilities are also found to be positively associated with literacy rates both for male and female. Literacy rates for male and female are found to be positively associated with each other.

5. Wide disparities in the level of development among different states have been observed both among major states and smaller states.

6. For enhancing the level of development of low developed states, model states have been identified and potential targets of important developmental indicators have been estimated.

7. It would be useful to examine and evaluate the level of development at micro level (say district, tehsil or block) for making location specific recommendations.

ACKNOWLEDGEMENT

Authors are thankful to Ms. Vijay Bindal, Technical Officer, Indian Agricultural Statistics Research Institute, New Delhi for helping in scrutiny and basic analysis of data for this study.

REFERENCES

Narain, P., Rai, S.C. and Shanti Sarup (1991). Statistical evaluation of development on socio-economic front. *J. Ind. Soc. Agril. Statist.*, **43**, 329-345.

Narain, P., Rai, S.C. and Shanti Sarup (1992). Evaluation of economic development in India. Souvenir of 11th Economic Development Conference in "Complementarily of Agriculture and Industry in Development". Instt. Trade & Industrial Development, New Delhi, 67-77.

Narain, P., Rai, S.C. and Shanti Sarup (1992). Classification of districts based on socio-economic development in Orissa. *Yojana*, **36**, No. 23, 9-12.

Narain, P., Rai, S.C. and Shanti Sarup (1993). Evaluation of economic development in Orissa. *J. Ind. Soc. Agril. Statist.*, **45**, 249-278.

Narain, P., Rai, S.C. and Shanti Sarup (1994). Regional dimensions of socio-economic development in Andhra Pradesh. *J. Ind. Soc. Agril. Statist.*, **46**, 156-165.

Narain, P., Rai, S.C. and Shanti Sarup (1994). Inter-districts disparities in socio-economic development in Kerala. *J. Ind. Soc. Agril. Statist.*, **46**, 362-377.

Narain, P., Rai, S.C. and Shanti Sarup (1995). Regional disparities in the levels of development in Uttar Pradesh. *J. Ind. Soc. Agril. Statist.*, **47**, 288-304.

Narain, P., Rai, S.C. and Shanti Sarup (1996). Dynamics of socio-economic development in Maharashtra. *J. Ind. Soc. Agril. Statist.*, **48**, 360-372.

Narain, P., Rai, S.C. and Bhatia, V.K. (1997). Regional pattern of socio-economic development in Karnataka. *J. Ind. Soc. Agril. Statist.*, **50**, 380-391.

Narain, P., Rai, S.C. and Bhatia, V.K. (1999). Inter district variation of development in southern region. *J. Ind. Soc. Agril. Statist.*, **52**, 106-120.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2000). Regional disparities in socio-economic development in Tamil Nadu. *J. Ind. Soc. Agril. Statist.*, **53**, 35-46.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2001). Regional dimensions of disparities in crop productivity in Uttar Pradesh. *J. Ind. Soc. Agril. Statist.*, **54**, 62-79.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2002). Dimensions of regional disparities in socio-economic development in Madhya Pradesh. *J. Ind. Soc. Agril. Statist.*, **55**, 88-107.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2003). Evaluation of economic development at micro level in Karnataka. *J. Ind. Soc. Agril. Statist.*, **56**, 52-63.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2004). Estimation of socio-economic development in hilly states. *J. Ind. Soc. Agril. Statist.*, **58**, 126-135.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2005). Estimation of socio-economic development of different districts in Kerala. *J. Ind. Soc. Agril. Statist.*, **59**, 48-55.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2005). Dimensions of socio-economic development in Jammu & Kashmir. *J. Ind. Soc. Agril. Statist.*, **59**, 243-250.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2005). Evaluation of socio-economic development and identification of backward areas in Orissa. Research work carried out in Research Unit of ISAS during 2005.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2007). Statistical evaluation of socio-economic development of different states in India. *J. Ind. Soc. Agril. Statist.*, **61**, 320-335.

Narian, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2009). Inter-district variation of socio-economic development in Andhra Pradesh. *J. Ind. Soc. Agril. Statist.*, **63**, 35-42.

Rai, S.C. and Bhatia, V.K. (2004). Dimensions of regional disparities in socio-economic development of Assam. *J. Ind. Soc. Agril. Statist.*, **57** (Special Volume), 178-190.

Regional Dimensions of India's Economic Development. Proceedings of Seminar held on April 22-24, 1982 sponsored by Planning Commission, Govt. of India and State Planning Institute, Govt. of U.P.

Economic Survey of Maharashtra (2006-07). Directorate of Economics & Statistics, Govt. of Maharashtra.

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 237-238

Indian Society of Agricultural Statistics :
Review of Activities for the Year 2009

The Indian Society of Agricultural Statistics is a scientific body which was founded on January 03, 1947 with the main objective of promoting and undertaking research in Statistics and its application to Agriculture, Animal Husbandry, Fishery, Agricultural Economics, Computer Applications and allied fields. The Society was fortunate to have Late Dr. Rajendra Prasad, the then Union Minister of Agriculture, Government of India as its Founder President. He guided the Society for the first sixteen years of its inception even after becoming the President of the Republic of India. The Society had the privilege of receiving patronage and guidance from several eminent personalities from time to time as its Presidents in the past who took keen interest and were a source of great inspiration. In fact, the Society could attain its present status due to the untiring efforts of its Presidents in the past and the continued patronage and guidance from the current President Dr. Mangala Rai, Secretary, Department of Agricultural Research and Education, Ministry of Agriculture, Government of India and Director General, Indian Council of Agricultural Research, New Delhi. The farsightedness, overall guidance and unstinting support from the eminent statisticians and founder members, Late Prof. P.V. Sukhatme and Late Dr. V.G. Panse have been fundamental to the growth of the Society.

The Society organizes annually a conference in different parts of the country which provides a wide platform for exchange of ideas on various issues of national as well as regional importance through symposia besides holding invited lectures by eminent scientists, paper presentations, and awards and incentives in various forms. Last year, the Society held its 62nd Annual Conference at S.V. Agricultural College, Acharya N.G. Ranga Agricultural University, Tirupati, Andhra Pradesh from 24 to 26 November 2008. The Society conveyed its grateful thanks to the authorities of the Acharya N.G. Ranga Agricultural University, Hyderabad for organizing the Conference. This year, the Society is grateful to the Vice Chancellor, Rajendra Agricultural University, Pusa, Samastipur, Bihar for inviting the Society to hold its 63rd Annual Conference at Rajendra Agricultural University, Pusa, Samastipur during December 03–05, 2009. Symposia on (i) **Statistical and Computational Genomics**, and (ii) **Statistical and Informatics Perspectives of Climate Change** are being organized during the Conference.

The Society also publishes an International Peer Reviewed Journal called "Journal of the Indian Society of Agricultural Statistics" with ISSN 0019-6363. Three issues of the Journal (April, August and December) are published in a volume in a year. The first Volume of the Journal was released in 1948. The Journal devoted to the publication of original research papers on all aspects of Statistics and Computer Applications preferably with innovative applications in Agricultural Sciences or that have a potential application in Agricultural Sciences. The review articles of the topics of current interest are welcome. Journal also accepts books, monographs and periodicals for review. Special issues on thematic areas of both national and international importance are also brought out. The Hindi Supplement continues to be a special feature of the Journal. The high standard of the Journal has been maintained due to the sincere efforts of the Chair Editor, Associate Editors and Reviewers. The Society is thankful to them for their keen interest in its activities related to the publication of the Journal. The Journal is reviewed by Mathematical Reviews and indexed in Zentralblatt MATH, Statistical Theory & Methods Abstracts. With a view to promoting research in Statistics and improving the standard of its Journal, the Society has been awarding prizes for the best papers published in the Journal for every biennium from 1987 in the fields of Design of Experiments, Sampling Theory, Statistical Genetics, Statistical Methodology, Applied Statistics and Computer Applications.

The membership of the Society which is drawn from all parts of India as well as from abroad during the year was

Permanent Institutional Members	29
Life Members	602
	631

During the year under report, 07 new life members, 02 annual members and 02 institutional members were enrolled. In addition to its regular members, the Society has a number of institutional subscribers to its Journal in India and abroad. The number of Indian subscribers during the year was 185. Thus, the total number of life members, annual members, institutional members and subscribers during the year was 827.

In order to perpetuate the memory of its Founder President, Late Dr. Rajendra Prasad, a memorial lecture is being organized during the Conference since 1965. The Society has organized 45 lectures so far and the memorial lecture organized during this Conference was 46th in the series and was delivered by Dr. H.S. Gupta, Director, IARI, New Delhi. Also since 1973, the Society has been organizing a lecture in the memory of Late Dr. V.G. Panse, who had been the guiding spirit behind the Society and its activities. The Society has so far organized 30 lectures in his memory and the current memorial lecture was 31st in the series and was delivered by Dr. Bikas Sinha, Ex-Member, National Statistical Commission, Government of India and Professor, Indian Statistical Institute, Kolkata.

The Society has a Research Unit to undertake research on specific problems of current interest under the guidance of a Research Direction Committee.

This year, a study relating to the estimation of variation of socio-economic development in the states of eastern region of the country was conducted and its details were presented separately in a different session during this Conference by Shri S.C. Rai.

The problem of finance for scientific activities, printing of the Journal and other ad-hoc publications could be overcome to a certain extent through grant-in-aid received from the Ministry of Agriculture, Government of India and Indian Council of Agricultural Research. The Society wishes to acknowledge very gratefully the financial assistance received from them during the year under report.

The Society continues to be a member of the International Statistical Institute, Netherlands, Indian Association of Social Science Institutions, New Delhi and Federation of Indian Societies of Agricultural Sciences and Technology, New Delhi.

The Society organized a five day training programme under the technical supervision of FAO Statistics Division, United Nations in close collaboration with Ministry of Statistics & Programme Implementation, Government of India at IASRI, New Delhi from 25 to 29 May, 2009 on "National Demonstration Centre" on Food Security Analysis of National Household Surveys Food Consumption Data. Twenty two participants from Bangladesh, China, India, Indonesia, Myanmar, Nepal, Pakistan, Philippines, Thailand and Vietnam attended the training programme. The NDC was successfully conducted and a large number of guests from the Food & Agriculture Organization of the United Nations, Ministry of Statistics & Programme Implementation, Ministry of Agriculture, Central Statistical Organization, National Sample Survey Organization, World Food Programme, Indian Council of Agricultural Research and other national institutions participated in either the opening or closing ceremonies. FAO Representative in India has been very supportive in the success of the NDC.

The accounts of the Society for the year ending 31 March 2009 were audited by M/s K.L. Sehgal & Co., Chartered Accountants, Professional Auditors and were presented in the General Body Meeting.

The work of the Society during the year was made possible through the advice and help of the members of the Executive Council, Editorial Board and the Research Direction Committee. The burden of the entire Secretariat of the Society has been willingly borne by my colleagues, Dr. V.K. Gupta, Shri R.S. Khatri, Dr. Rajender Parsad, Dr. P.K. Malhotra, Dr. A.K. Vishandass and Shri S.C. Rai. In the end, I wish to thank the staff of the Society for their devoted work.

V.K. BHATIA
Secretary

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 239-242

Symposium on Statistical and Computational Genomics

Chairman : Prof. Prem Narain, Executive President, ISAS
Conveners : 1. Dr. Rajender Parsad, IASRI, New Delhi
2. Dr. B.M. Prasanna, IARI, New Delhi
3. Dr. V.K. Shahi, RAU, Pusa, Samastipur

Three papers covering various aspects related with the theme of the symposium were presented by the following speakers:

1. Statistical Genomics for Crop Improvement: Opportunities and Challenges – B.M. Prasanna
2. Design and Analysis for 2-colour Microarray Experiments – Rajender Parsad
3. Bioinformatics in Agriculture – V.K. Bhatia

After long deliberations, following recommendations were emerged out:

1. Computing environment (both in terms of software and hardware) should be strengthened at par with International level so as to enable the researchers to perform voluminous data analysis of available in Genomics.
2. Identify the faculty members with aptitude in statistical genomics and provide advanced training to them in leading Institutions.
3. To introduce M.Sc. programme in Bioinformatics with special reference to Mathematics, Statistics and Computational Biology.
4. To create infrastructure for National Repository of Agricultural Genomic databases.
5. Concerted efforts need to be made for development of efficient and robust experimental designs and efficient analytical technique for single and multi-factor microarray experiments.
6. Basic and Applied research in the areas of proteomics, functional genomics also needed to be undertaken through a strong network of inter-institutional programme.

ABSTRACTS OF THE PAPERS PRESENTED

1. Statistical Genomics for Crop Improvement: Opportunities and Challenges

B.M. Prasanna

In the last one decade, exciting advances have been made in statistical genomics, besides development of high throughout genotyping. Powerful statistical methods and tools are immensely aiding research progress in molecular plant breeding and in bridging the genotype-phenotype divide. This is largely due to the evolution in statistical and computational means for analyzing the patterns of molecular diversity, formulating core collections, understanding genotype × environment interactions, identifying QTLs, detecting epistatic interactions, undertaking association mapping, and analyzing microarray data. The presentation will review the national and international developments in these important areas, with particular focus on crop plants, and shall identify critical gaps for strengthening statistical genomics in the Indian context.

The new focus on genomics has also highlighted a particular challenge: how to integrate the different views of the genome that are provided by various types of experimental data and provide a proper biological

perspective that can lead to crop improvement. Mapping and studying the genetic architecture of complex traits, and understanding the dynamic network of gene interactions that determine the physiology of an individual organism over time is another major challenge that requires novel, quantitative and testable statistical solutions.

Besides development of comprehensive computational tools to integrate information regarding genotypic performance, pedigree relationships, germplasm diversity and genomic data, there is also an immense need to develop a new "breed" of geneticists and statisticians in developing and implementing novel, more effective and efficient translational bioinformatic systems that the scientists can routinely use in breeding strategies. The availability of efficient computational algorithms/softwares is essential to the scientific community. However, it is equally important that these tools are applied with thorough understanding of the genetic data and the tools themselves.

Indian Agricultural Research Institute, New Delhi

2. Design and Analysis for 2-colour Microarray Experiments

Rajender Parsad and V.K. Gupta

Microarrays are microscopic arrays of single-stranded DNA molecules immobilized on a solid surface by biochemical synthesis. **DNA microarrays** are created in two basic forms viz. (i) by DNA depositions and (ii) by in situ synthesis of oligonucleotide arrays. These are also known as *DNA chips*, *gene chips*, *biochips*, *DNA microarrays* or simply the *arrays*. Microarray is an important genomics tool that can identify the expression of several thousand genes at a time.

In 2-colour cDNA microarray experiments, four basic experimental factors viz., array (A), dye (D), variety (V) and gene (G) are studied. These four factors give rise to 15 effects that include 4 main effects, 6 two-factor interactions, 4 three-factor interactions and one four factor interaction. But all the main effects and selected two-factor interactions viz. array-gene interaction (AG), dye-gene interaction (DG), variety-gene interaction (VG) are the seven effects of interest to the experimenter. An important and common question in DNA microarray experiments is the identification of differentially expressed genes, that is, genes whose expression levels are associated with a response or covariate of interest. Designing of microarray experiments is an important issue to get precise comparisons of variety ´ gene interactions. Reference designs, alternating loop designs and dye-swap designs are the most commonly used designs for these experiments.

The investigations carried out so far deal with the situations where same set of genes is spotted on each array in microarray experiments. Therefore, genes/ gene specific effects (G, AG, DG, VG) are orthogonal to global effects (A, D, V). Therefore, optimality aspects of designs for microarray experiments have been studied by taking only array, dye and variety effects and leaving gene specific effects from the model. Designs that are efficient under the model containing only global effects are also efficient under the model containing both global and gene specific effects. Efficient designs have been obtained for these experiments under the assumption that the dye effects are orthogonal with respect to variety effects under a fixed effects model. Catalogue of efficient block designs for microarray experiments are now available in the literature. Further, array effects may be random, and as a consequence, model becomes a linear mixed effects model. The robustness aspects of efficient designs obtained under a fixed effects model have been investigated under mixed effects model. There is need to obtain designs by considering the dye effects also in the model.

In the research investigations carried under a mixed effects model, the ratio of inter and intra block variances has been assumed as constant. One may consider putting a beta prior on the ratio of the variance components and then derive efficient designs after taking the expectation with respect to beta distribution. This may be achieved through Bayesian inference.

Further, in all the investigations carried so far, it has tacitly been assumed that the variability in gene expression is constant across all the genes. Depending upon the underlying biology, the gene expressions may be heteroscedastic and depend upon the gene of interest. This is another important issue which needs attention. This amounts to obtaining efficient block designs under a heteroscedatic set up.

In most of the investigations carried out so far it has been assumed that all the genes are spotted on all the arrays. There is a need to obtain efficient designs when genes spotted on each array differ, i.e. array and gene effects are non-orthogonal.

The above discussion is for the experimental situations where there is only one factor that is causing variation in gene expression levels, i.e. variety (different types of tissues, drug treatments or time points of a biological process). There, however, do occur experimental situations, where it is desired to compare more than one factor for studying the gene expression levels. For example, one may be interested to study and compare the two mutants at times zero hour and 24 hours. The interest is in measuring the changes over time. Therefore, there are two factors viz. varieties (two mutants) and time (0 hour and 24 hours) and there is a need to obtain efficient designs for factorial and time-course microarray experiments.

For factorial microarray experiments, balanced factorial designs may be useful. The application of designs for balanced factorial experiments in 2-colour microarray experiments require that these designs should be constructed with block size two since only two varieties/ treatment combinations can be accommodated on one array. Some methods of construction of designs for balanced factorial experiments with block size two are available in the literature. These methods are heuristic in nature and give only designs for very few parametric combinations. It is, therefore, required to obtain efficient designs for balanced factorial experiments with block size two that can have an application in factorial or time-course microarray experiments. The efficient designs for balanced factorial experiments are only useful for the situations where a natural baseline does not exist for at least one of the factors, like gender lacks a natural baseline. This is known as orthogonal parameterization problem. For orthogonal parameterization, EGD designs may be useful and catalogues of EGD designs with block size two need to be prepared.

There do exists situations, where a natural baseline or null state exists such as the situation involving two mutants where one proliferates a particular disease and other does not. The mutant that does not proliferate into disease is baseline. In toxicological study with binary factors, each representing presence or absence of a toxin factor, absence can be regarded as a natural baseline level of each factor. Null state or baseline level of a factor need not strictly mean zero level on some scale, but may as well refer to a standard or control level like the one currently being used in practice. This is known as baseline parameterization. Definitions of main effects under the two parameterizations (orthogonal and baseline) are entirely different. Therefore, the designs that are efficient for one parameterization may not be optimal/ efficient under other parameterization. Problem of obtaining efficient designs under baseline parameterization has received a little attention. There is a need to put research emphasis on obtaining efficient designs for factorial microarray experiments for baseline parameterization.

It is also required to prepare a catalogue of such efficient designs to serve as a ready recknoer for the experimenters. Further, it is also required to arrange these designs in row-column structure so as to display both the variability factors viz. arrays and dyes.

Several other statistical issues involved in the analysis of gene expression data include data quality, data analysis and validation. The biological question of differential expression can be stated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels and the responses or covariates. Researchers are interested in determining the direction of rejection of the null hypothesis, which is, in determining whether genes are over or under expressed. Generalized p-values are being used for these situations. One may also be interested in testing several hypotheses simultaneously for each gene. Another aspect in which one may be interested is the two-dimensional testing problem where several hypotheses are tested simultaneously for each of thousands of genes. This problem of multiple hypothesis testing requires attention of the statisticians. The genes are spotted on each array one along the other. Therefore, the fluorescence intensities from the nearby points may be correlated. Therefore, developing analytical techniques of data in the presence of spatially correlated observations is an unprecedented challenge. Another important problem in microarray experiments is classification of genes using the data from microarray experiments. Keeping in view the above, efforts will be made to obtain efficient designs for microarray experiments and to develop statistical analysis procedures for identification of differentially expressed genes.

Some other issues are how many biological samples should be taken and to what extent does biological averaging holds? Further, with the availability of multiple platforms employing cDNA or oligonucleotides, which may also differ in probe preparation methods and array surface chemistry, raises the question of cross-platform agreement in gene

expression measurements. Several studies have been carried out for comparing different microarray platforms.

Minimum Information about a Microarray Experiment (MIAME) is needed to enable interpretation of results of experiment unambiguously and to reproduce experiment. Six most critical elements contributing towards MIAME are: (i) Raw data for each hybridization, (ii) Final processed (normalised) data for the set of hybridisations in the experiment, (iii) Essential sample annotation including experimental factors and their values, (iv) Essential sample annotation including experimental factors and their values, (v) Sufficient annotation of the array and (vi) Essential laboratory and data processing protocols. The public repositories Array_Express at the EBI (UK), GEO at NCBI (US) and CIBEX at DDBJ (Japan) are designed to accept, hold and distribute MIAME compliant microarray data.

Several statistical packages are now available that deals with microarray experiments. To name a few: JMP Genomics, SAS Genomics, SAS Macros AnovArray, GenStat 12, S+ArrayAnalyzer , TM4 suite, Bioconductor, etc. For E-leraning on *Statistical Genomics*, a new link has been created on Design Resources Server (ww.iasri.res.in/design) that aims to provide support towards statistical and computational aspects of plant genomics. The main purpose of initiating a link on Statistical Genomics is to provide an E-learning platform to the experimenters in their analysis of data.

Indian Agricultural Statistics Research Institute, New Delhi

3. Bioinformatics in Agriculture

V.K. Bhatia

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an exponential growth in the biological data generated by the scientific community. The appropriate analysis and management of genomic data has resulted very useful information and knowledge about the underlying phenomenon. With the advent of newer informatics techniques this needs to be further looked into for extracting additional deeper knowledge. Thus, the information generated from biological experiments is essentially required to be stored, organized, understood and analyzed for furthering the knowledge and insight of the genomic data. This search of newer knowledge has led to the emergence of an integrated discipline, called Bioinformatics. This rapid technological advancement has impacted agriculture significantly and thus, the research community involved in agricultural sciences need to participate and contribute to the ensuing global bioinformatics revolution.

The application of bioinformatics has been carried out in various ICAR labs for agricultural/biological sciences. Looking at the current status of bioinformatics research at International and National levels some research gaps pertaining to agriculture have been identified. The gaps basically identified are related to computational biology, statistical genomics and development of newer algorithms for handling voluminous genomic and proteomic data. In order to fill up these gaps ICAR initiated a process of creating a Centre of Bioinformatics for processing genomic data collected at ICAR institutes across all species. The present article deals with the importance of bioinformatics in agriculture and the initiatives taken along with identifying the future research strategies. The emphasis for development of a National Agricultural Bioinformatics Facility under ICAR to meet the challenges ahead in the field of bioinformatics in agriculture has been highlighted.

The bioinformatics initiative will enable the flow of information from the partners in terms of database and problems to be core group for eliciting meaningful and useful inferences by leveraging the tools of information technology for arriving at solutions to their problems. This will not only expedite layout of experiments and generating more such data on a continuous and ongoing basis but also serve as a veritable knowledge base. The partners would be in a position to interact with one another directly or indirectly for solving problems of mutual and common nature. The information so gathered will be warehoused in the centre. The centre will also take care of the needs for collaboration with other Indian and international organizations for enhancing the database, by way of human resource development and also widening the perspectives of bioinformatics research by deeper insights into various aspects. Its collaboration is also expected to enable the country to emerge as a global player for genomic research in agriculture.

Indian Agricultural Statistics Research Institute, New Delhi

Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 243-246

Symposium on Statistical and Informatics Perspective of Climate Change

Chairman : Prof. Bikas Sinha, Visiting Prof. ISI, Kolkata
Conveners : 1. Dr. P.K. Malhotra, IASRI, New Delhi
2. Dr. R.C. Agrawal, NBPGR, New Delhi

Three papers covering various aspects related with the theme of the symposium were presented by the following speakers:

1. Spatio-temporal data mining for monitoring of climate change – Anil Rai
2. Experimental designs for mitigation and adaption of climate change – Rajender Parsad
3. Modeling climate effects – V.K. Bhatia

After detailed discussions, the following recommendations emerged out:

(i) Multivariate analysis and data reduction techniques need to be used for monitoring climate change.

(ii) Data sets of spatio-temporal variables on climate as well as derived parameters based on spectral data from MODIS must be used in climate change studies.

(iii) For development of cultivars/agricultural practices for mitigation and adaption of climate change, study on designing of experiments and analysis of experimental data should be taken up vigorously.

(iv) Data modeling should be given emphasis for identifying the variables for climate change studies.

(v) Development of knowledgebase required for climate modeling and climate impact assessment studies.

ABSTRACTS OF THE PAPERS PRESENTED

1. Spatio-Temporal Mining for Monitoring Climate Change

Anil Rai, K.K. Chaturvedi and P.K. Malhotra

Climate is usually defined as the "average weather", or more rigorously, as the statistical description of the weather in terms of mean and variability of relevant parameters over periods of several decades (typically three decades as defined by WMO). These parameters are most often surface variables such as temperature, precipitation and wind, but in a wider sense the "climate" is the description of the state of the climate system. The climate system consists of five major components as (a) the atmosphere, (b) the oceans, (c) the terrestrial and marine biospheres, (d) the cryosphere (sea ice, seasonal snow cover, mountain glaciers and continental scale ice sheets) and (e) the land surface. These components interact with each other, and through this collective interaction, determine the earth's surface climate. It has been observed that due to global warming, global average temperature is continuously increasing. There

is increase in global average temperature by one degree from 1961 to 1990. As a consequence of this sea level is increasing and overall area under snow cover is decreasing. Climate change scenarios for future can be best project using magic quadrant of Special Report on Emission Scenario (SRES). Agriculture is one of the most venerable sectors, which will have high impact of this global warming. Increase in CO_2 in the atmosphere will increase photosynthesis capacity of the plants, decreases stomatal conductance, enhanced water use efficiency, altered photosynthesis partitioning and finally its impact on plant depends on elevation and C3/ C4 plants type. Increase in temperature will have negative impact on crop yield especially during winter season. In this process, there may be increase in yield, which will have positive effect on paddy crop but at the same time there will be increase in extreme events which will have negative impact on agricultural production. Further, aerosols in the atmosphere will increase and as a consequence of this there will be reduced solar radiations on the earth, which will have negative impact on agricultural production. Therefore, the combined effect of climate change will be very complex and difficult to quantify in relation to agriculture. However, monitoring of climate and its effect can help in better planning and minimizing the damage to global agriculture. Construction of Ocean Climatic Index (OCI) is one of the important techniques to monitor the climate and its impact at global level. OCI captures ocean and land relationship. It is based on time series data and it summarises behaviour of selected area on land surface based on climatic parameters of ocean. El Nino effect is one of the well known pattern of changing direction of trade winds in such a way that it brings drought in Australia, warmer winters in North America, flooding in coastal Peru and increased rainfall in east Africa. Important climatic variables are Sea Level Pressure (SLP), Sea Surface Temperature (SST) and precipitation. One of the most important dependent variable, which indirectly measures the plant growth is Net Primary Production (NPP). NPP measures net assimilation of atmospheric Carbon Dioxide in to organic matter by plant and helps in understanding global carbon cycle at regional and global level. It is driven by solar radiation and constraint by precipitation and temperature. These variables are measured at spherical grid level. During pre-processing of this data, monthly Z-score are calculated to remove seasonality and spatial as well as temporal correlations. There are other numbers of known climatic indices such as Southern Oscillation Index (SOI), North Atlantic Oscillation (NAO) and Dipole Mode Index (DMI) etc. DMI is very important for prediction of rainfall during monsoon in India. Singular Value Decomposition (SVD) analysis is one of the important techniques for development of OCI. It is similar to Principle Component Analysis (PCA) and captures the spatial and temporal pattern in the climatic data. This is also known as Empirical Orthogonal Function (EOF). But, it also has number drawbacks. SVD is able to detect only few strongest patterns as strong patterns mask the detection of weaker patterns. Interpretability is poor due to orthogonal vectors. Further, analysis cannot take into accounts of different lags in spatial patterns. Recently, graph based clustering has been used to avoid these drawbacks. Most popular techniques are Shared Nearest Neighbour (SNN), Jarvis-Patrick Clustering etc. These cluster based OCI's are found to be very effective in monitoring of climate at global level. Now Moderate Resolution Imagine Spectro-radiometer (MODIS) provides wealth of data on various climatic parameters along with derived variables which can be used in development of OCI's for monitoring the global climate more effectively.

Indian Agricultural Statistics Research Institute, New Delhi

2. Experimental Designs for Mitigation and Adaptation Strategies of Climate Change

Rajender Parsad[1], V.K. Gupta[1] and
Raj. S. Malhotra[2]

Climate change refers to the changes in the mean and /or variance of the parameters such as temperature, rainfall, presence of greenhouse gases, etc. over a long period of time. Obviously, this change should be identifiable using statistical tests. Therefore, the first and foremost requirement of studying climate change is the data availability of given parameters at a given location over an extended period say at least 3 to 4 decades. To see the spatial pattern over geographic locations, the data availability should be at different locations geographically. In the study of climate changes, there are two approaches viz.

(i) to detect the changes over time and space and

(ii) to study the causes of those changes.

The first one can be achieved by studying the trends through saptio-temporal models. In this case, no reasons are to be assigned. According to Levine and Berliner (1999, *Journal of Climate*, 12(2), 564-574), the main difficulty in early detection of changes resulting from human induced/ anthropogenic forcing is that the natural variability overwhelms the climate change signal in the observed data. To overcome this problem, fingerprint procedures have been developed to express the climate data in terms of low-dimensional signal patterns. Fingerprint approach is a mathematical procedure for optimally detecting a climate change signal above the background natural climate variability noise. For detection, in statistical perspective, it is testing of hypothesis that the change is not due to natural variability. Some people also profess Bayesian approach to multi-model information processing for developing climate forecasts.

To study the causes of changes, one requires the data on the variables and one has to develop models for cause and effect relationships. For establishing these relationships controlled comparative experiments may be helpful.

The effect of global warming and climate change on agriculture would be in terms of, rise in global temperature, elevated levels of carbon dioxide in the atmosphere, altered rainfall patterns, greater severity and frequency of extreme weather events, including droughts and floods, etc. This will result in decrease in availability of water, may be more severe droughts, sudden weather fluctuations, increased vulnerability to diseases and insect pests, etc. There may be swings in temperature and rainfall. There may be heavy summer and winter rainfalls but with only small changes in annual rainfall in totality.

Therefore, the need is to strengthen the adaptive capacity of the communities, and resilience of farming systems, in short, mitigation and adaptation.

For deciding upon mitigation and adaptation measures is to know the effect of increased levels of CO_2, rise in temperatures, extended drought periods, we need to study the effects of these effects of climate change including swings on agriculture and food security. A lot of studies have been conducting for studying the effect of elevation in level of CO_2, rise in temperature, drought periods etc. taken one factor at a time. The interactions among CO_2, temperature and water can be substantial and the combined effect on the biological systems of several factors may not be predicted from experiments with one or a few factors. Therefore, it is required to conduct multi-factor experiments involving a large set of factors such as taking together different levels of elevation of CO_2, rise in temperature, drought periods etc. One such experiment was conducted with two levels of CO_2 (ambient and elevated level 510 ppm) two levels of temperature (ambient and elevated temperature of 2^0C) and two levels of droughts (present condition and induced drought in late spring summer) at Brandbjerg, Copenhagen, Denmark during 2005-2007 using split plot designs in 6 replications on a grassland eco-system. For details on this experiment a reference may be made to Mikkelsen *et al*. (2008, *Function and Ecology*, 22, 185-195). In this experiment only the extreme levels of different factors have been used and it does not provide any information on the intervening levels of these factors. Further, the effects of factors of climate change may vary from region to region and crop to crop. Therefore, region specific experiments should be conducted using varying levels of different factors of climate change. This requires efficient designing of experiments under scarce resources.

Once these effects are known, the mitigation strategy is to identify/ develop varieties which are drought tolerance, cold tolerant, resistant against diseases and pests, varieties which can sustain the high concentrations of CO_2, etc. In other words, we have to develop/ identify varieties that are resistant to biotic and abiotic stresses. We have also to think of identifying improved cropping systems and practices that make efficient use of natural resources. In all these efforts, statistical designing and analysis of experimental data may be of help. In the present talk, we shall concentrate on statistical issues related to design and analysis of experiments for identification of cultivars that are tolerant to biotic stress (living organisms which can

harm plants, such as viruses, fungi, and bacteria, and harmful insects, etc.) and/or abiotic stress {negative impact of non-living factors on the living organisms in a specific environment, for example, extreme temperatures (high and low) during whole or part of crop growth, drought, flood, and other natural disasters, soil conditions, etc.). In other words, crops/ cultivars containing genes that will enable them to withstand biotic and abiotic stresses need to be developed. Here, care should be taken that not only the cultivars are high yielding but also retain the nutrition quality in terms of protein, fat, carbohydrates, minerals contents. In this talk, we shall explain it through some situations that we encountered during our advisory services with the subject matter specialists.

[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*ICARDA, Aleppo, Syria*

3. Modeling Climate Effects

V.K. Bhatia

Field crops are very much influenced by climate effects. It is more for perennial crops because they remain in the field for many years and are subjected to climate changes over a longer period of time. For a researcher, the primary objective is to assess the potential impacts of climate change and identify possible adaptation strategies. Thus, the main challenges in front of researchers are, how to study the effects of climate particularly for the situations of large number of variables related to climate parameters in comparison to very small number of observations of main character of interest of crop production or productivity. This problem, therefore, finally leads to the very important statistical problem of handling the situation of large p (number of independent factors) with small n (number of observations of the response variable). The solution of this problem has been studied by number of research workers. The contribution of these researchers has been reviewed in brief.

Mainly there are two issues, firstly that of model selection which deals with estimating the performance of different models in order to choose the best model and secondly having chosen a final model, estimating its prediction error on the new data. With these issues in mind, the applicability of different methodologies of Least Square Techniques (Regression), Lasso (Least Absolute Shinkage and Selection Operator) as given by Tibshirani (1996); Lars (Least Angle Regression) as given by Efron *et al*. (2004); Dantzig Selector by Efron (2007) and Dasso (Connections between the Dantzig selector and Lasso) by James *et al*. (2009) have been highlighted. The concept of the regression tree modeling has also been highlighted. In predictive modeling having large number of predictors, problem is to find "what subset of the effects which provides the best model for the data". In this area, the problems of identifying the best model are also brought to the notice of researchers. Finally, it is concluded that model selection is very important for building a high performance model and there is a need to develop methods for special data sets.

REFERENCES

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, (with discussion). *Ann. Statist*., **32,** 407-451.

Efron, Bradley, Hastie, T and Tibshirani, R. (2007). Discussion of "the Dantzig selector". *Ann. Statist*., **35**, 2358-2364.

James, G.M., Radchenko, P. and Jinchi, Lv. (2009). DASSO: connections between the Dantzig selector and Lasso. *J. Roy. Statist. Soc*., **B71**, 127-142.

Tibshirani, R. (1996). Regression shrinkage and selection via. the lasso. *J. Roy. Statist. Soc*., **B58**, 267-288.

Indian Agricultural Statistics Research Institute, New Delhi

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 247-257

ABSTRACTS OF PAPERS

1. Unbiased Regression Estimators in Finite Population Sampling

B.V.S. Sisodia and K.K. Mourya

Micky (1959), William (1962), and Singh and Srivastava (1980) provided some sampling schemes, which yielded regression estimators of population mean in finite population sampling. The variances of the estimators were, however, quite complicated and faced computational hard-work. A new sampling scheme is proposed in the present paper, which is quite simple and provides an unbiased regression estimator. Its variance is derived which is simple one and is easily comparable with the usual biased regression estimator. An empirical study is also carried out to illustrate the precision of the unbiased regression estimators.

N.D. University of Agriculture & Technology, Faizabad

2. Estimation of Population Mean under Non-Response in Two-Occasion Rotation Patterns

G.N. Singh and Jaishree Prabha Karna

The present work is an attempt to study the effect of non-response at both occasions in search of good rotation patterns over two occasions. Ratio type estimators have been proposed for estimating the population mean at current occasion in presence of non-response at both the occasions in two-occasion successive (rotation) sampling. Detailed behaviors of proposed estimators have been studied. Proposed estimators are compared with the estimator using no information from previous (first) occasion. Performances of the proposed estimators have been demonstrated via empirical studies.

Indian School of Mines, Dhanbad

3. On the Precision of Ratio Estimators using Linear Transformation

Sunil Kumar, Amar Singh and B.V.S. Sisodia

Various authors have proposed different version of linear transformation of the auxiliary variable (x) in sample surveys to reduce the bias and mean square error (MSE) of usual ratio estimator. Notable among them are Mohanty and Das (1971), Reddy (1974), Srivenkatramana (1978), Das and Tripathi (1980), Sisodia and Dwivedi (1981), Singh and Kakran (1993), Mohanty and Sahoo (1995), Upadhyay and Singh (1999), Swain (2009) etc. A comprehensive review of transformed ratio estimators is presented in this paper. Moreover, when a prior value of *y*–intercept in the simple linear regression of *y* on *x* is available, a linear transformation of study variable *y* is suggested to find out another modified transformed ratio estimator. Following Swain (2009), a linear transformation of *x* based on minimum and maximum values of *x* is also proposed and accordingly another new transformed ratio estimator is developed. Properties of proposed modified transformed ratio estimators are studied. An empirical study with some real populations is also carried out to highlight the precision of transformed ratio estimators.

N.D. University of Agriculture & Technology, Faizabad

4. Comparison of Ratio Estimators with One Auxiliary Variable using Monte Carlo Simulation

M. Krishna Reddy[1] and K. Ranga Rao[2]

Ratio estimators are often employed in sample surveys for estimating the population mean $\bar{Y}$ of a characteristic of interest *Y* or the population ratio $\frac{\bar{Y}}{\bar{X}}$ utilizing a supplementary variate *X* that is positively correlated with *Y*. It is well known that the classical ratio estimator $\frac{\bar{y}}{\bar{x}}$ is biased and often, in practice, the bias may be negligible compared to standard error and can be neglected. In recent years, considerable attention has been given to the development of unbiased and almost unbiased ratio estimators. The classical ratio estimator, two unbiased and two almost unbiased

estimators available in literature in this paper and are compared with respect to relative bias, mean square error, skewness and kurtosis using Monte Carlo simulation by generating random samples from different bivariate population with known correlation coefficients. The Monte Carlo simulation is adopted for comparison of these ratio estimators because analytical comparisons are not possible.

[1] *Osmania University, Hyderabad*
[2] *Sri Venkateswara Veterinary University, Kamareddy*

5. Allocation in Stratified Sampling using R-Software

S. Maqbool[1], Mahesh Kumar[2] and S.P. Singh[2]

Stratified sampling techniques have been widely used for surveys because of their efficiency. The purpose of stratification is to partition the population into disjoint sub-populations so that the power consumption characteristics within each sub-population are more homogenous than in the original population. In this paper, we have developed computer programs in R-Software which can be implemented for allocation problems using equal and proportional allocation methods and also for drawing inferences of related parameters.

[1] *Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar*
[2] *RAU, Pusa, Samastipur*

6. Robust Estimation in Finite Population Sampling

R.P. Kaushal and B.V.S. Sisodia

Following Scott *et al*. (1978), a BLU predictor of population total under the model $\xi(0, 1 : x_{hk}^2)$ in stratified sampling when the slope of the model is common across to strata is constructed. Its robustness and optimality is studied when some general polynomial model of degree *J*, i.e. $x(d_0, d_1, ..., d_J : x_{hk}^2)$ is true in real practice. It has been found that the proposed predictor is robust and optimal for stratified balanced sample and is also more efficient than that of due to Scott *et al*. (1978) when slopes are varying from stratum to stratum.

N.D. University of Agriculture & Technology, Faizabad

7. Secondary Stage Education in India: An Analytical Overview

Virendra P. Singh[1] and Sandeep K. Sharma[2]

The present study has analytically overviewed the secondary stage education based on the recent past educational surveys. It provides tangible comparisons pertaining to secondary stage education on the schooling facilities in rural areas, schooling facilities in habitations predominantly populated by scheduled castes and scheduled tribes, schooling facilities in villages, secondary schools, secondary sections in the schools, enrolment in classes IX and X, science laboratories and computer education, guidance services, pre-vocational courses, secondary stage in oriental schools following general system of education, schools admitting children with disabilities and enrolment taking place in India. The study utilizes secondary data collected during recent successive two surveys, namely, sixth and seventh on school education conducted by the National Council of Educational Research and Training under administrative and financial control of the Government of India.

[1] *National Council of Educational Research & Training, New Delhi*
[2] *Indian Agricultural Research Institute, New Delhi*

8. Some Statistical Aspects on Major Operational Incentive Schemes in Indian Schools

Virendra P. Singh

This paper considers some statistical aspects on major operational incentive schemes prevailing to attract children in the Indian schools. Attempts have been made to analyze the availability of major incentive schemes in the schools covering incentive schemes on free uniforms, free textbooks, supply of free textbooks, attendance scholarship for girls and beneficiaries thereof by social groups, namely, scheduled casts, scheduled tribes, educationally backward minority community, respectively including midday meals and types thereof at primary stage in the country. The present paper is based mainly on secondary data on school education collected during seventh school education survey conducted by the National Council of Educational Research and Training under administrative and financial control of Government of India.

National Council of Educational Research and Training, New Delhi.

9. Estimation from Independent Sub-samples of a Population

Jagbir Singh

In this paper Minimum Variance Linear Unbiased Estimators for population mean (i.e. milk yield/day) in the first year, second year and third year; change and average thereof over years thereof have been developed by making use of Projective Geometry approach and adopting the following sampling design: In a district, two independent sub-samples of villages are drawn with simple random sampling without replacement of size n_1 and n_2 respectively. Next year m_1 and m_2 sample villages are retained in sub-sample 1 and 2 respectively and u_1 and u_2 sample villages are drawn using SRS without replacement. Here $n_1 = m_1 + u_1$ and $n_2 = m_2 + u_2$. In the third year m_1 and m_2 sample villages are replaced by another set of villages using SRS without replacement, while u_1 and u_2 sample villages remain unchanged. In each selected village out of H, h households are selected and in each selected household two cows in milk are selected using SRS without replacement for collecting information on milk yield/day.

Indian Agricultural Statistics Research Institute, New Delhi

10. A Note on Variance Estimation of Ratio Estimators in Two Phase Sampling

R. Arnab[1] and U.C. Sud[2]

An alternative estimator of the variance of the ratio estimator under two phase sampling scheme has been proposed. The proposed variance estimator is approximately unbiased and more efficient than the existing classical estimator due to Cochran (1977) and Rao-Sitter (1995) under a linear superpopulation model with intercept.

[1]*University of Botswana, Botswana*

[2]*Indian Agricultural Statistics Research Institute, New Delhi*

11. Some Investigations on Sampling Variance of Genetic Correlation

S.D. Wahi and A.R. Rao

The present investigation is an attempt to compare the estimated, predicted, empirical and bootstrap standard errors for different combinations of population heritability, genetic and phenotypic correlation for different family sizes and structures under half-sib mating design. The data under half-sib model are simulated by taking sire effects following normal as well as gamma distribution. It is observed that the empirical standard error of genetic correlation when sire effects are from gamma distribution are invariably higher as compared to the data with sire effects following normal distribution irrespective of the sample size heritability and genetic correlation of the traits. The empirical standard error of genetic correlation estimates are very high for lowly heritable traits for whole range of positive and negative genetic correlation. The large sample approximation of standard error given by Tallis is always underestimating the standard error even for large family size of 30 to 50 and should not be used in practice.

Baring small sample size the bootstrap estimates of SE are very close to predicted SE and can be used as an estimate of SE of genetic correlation. The bootstrap estimates of standard error of genetic correlation are found to be very close to the predicted standard error for sample size 500 and above in case of lowly heritable traits for both positive and negative genetic correlation values. In case of moderately and highly heritable traits the bootstrap estimates of standard error are found very close to predicted standard error for all values of genetic correlation and for all the sample sizes and family structures expect for sample size 200 (10, 20) in case of moderately heritable traits. Hence, it can be said that the bootstrap estimates of standard error which are very close to predicted values can be used to estimate the standard error instead of approximate formulae given in literature. It is also found that in case of non-normal data sets with sire effects following gamma distribution the bootstrap estimates of standard error of genetic correlation are always underestimated.

Indian Agricultural Statistics Research Institute, New Delhi

12. Growth Pattern and Current Scenario of Pulse Production in Uttar Pradesh

M.K. Sharma and B.V.S. Sisodia

The Uttar Pradesh is most populous state of the country. It accounts for about 12 per cent (2003-04) of total area of pulses while 16 per cent (2003-04) of total pulse production in the country. The average yield of

pulses has been between 7.00 to 9.29 quintal per hectare since 1987-88. However, the area under pulse crops has declined substantially over years to the extent of 2.7 million hectare in 2003-04 from 3.0 million tonnes in 1987-88. Similarly, the scenario of pulse production has also not been satisfactory since it has witnessed tremendous fluctuations ranging between 2.36 to 2.62 million tonnes since 1987-88.

Uttar Pradesh has four regions i.e. Western, Central, Bundelkhand and Eastern regions. This study is based on the different regions of Uttar Pradesh having important role in term of pulse production. Various statistical tools used to determine the growth rates, instability of variation of area, production and productivity of different pulse crops and total pulse crop for two different periods before and after the launch of technological mission on pulse production in the country.

The growth rate, sustainability etc. on area, production and productivity of pulse crops of U.P. have been worked out. The time series data pertaining to 1960-61 to 2005-06 have been considered, which has been divided into two periods, before (1960-61 to 1989-90) and after (1990-91 to 2005-06) launch of Technology Mission on pulse production. The pulse crops i.e. arhar, pea, gram and lentil alongwith total pulses were taken for the study. It has been found that the area of pea of Bundelkhand region have been more stable in pre technology mission. The same trend follows in the case of production. Increase or decrease in area and production of different regions have been carried out over different decades. The overall results indicate that the Bundelkhand region has been found more stable in respect of all the pulse crops considered under study.

N.D. University of Agriculture & Technology, Faizabad

13. Measurement of Risk in Yield of Cotton in Amravati Division

P.D. Deshmukh, R.K. Kolhe and A.S. Tingre

The study was conducted to measure the risk in yield of cotton in Amravati division of Vidarbha region over a period of 30 years. In general, for overall period (1975-76 to 2004-05) growth rates of production and yield of cotton were positive and highly significant. Constant variability in area of cotton were observed for all the districts under study as well as for Amravati division as a whole. In overall period lowest probability of crop failure were observed in Yavatmal district (0.56) with crop loss ratio of 22.20 per cent. The yield uncertainty arising due to vagaries of nature vitiates farmers production programme and causes instability in production and income of the farmers.

The measurement of risk involved in crop production is of paramount importance to suggest remedial measure of technical and social nature. Keeping in view the above aspects it is proposed to measure risk involved in the production of cotton crop of Amravati division of Vidarbha, which is the major Cotton growing area of the Vidarbha region of Maharashtra. Presently cotton is a backbone of Maharashtra's economy Maharashtra state ranks first in the country as regards to the area of 27.60 lakh ha. under cotton with production of 271 kg. lint / ha. Vidarbha region has about 65 percent of the total area under cotton in Maharashtra i.e. 12.50 lakh ha. with the production of 24 lakh bales and productivity of 326 kg lint / ha. The study was undertaken with the following objectives:

(1) To examine the growth and variability of cotton.
(2) To measure the risk in yield of cotton.

Dr. Panjabrao Deshmukh Krishi Vidyapeeth, Akola

14. Statistical Evaluation of Socio-economic Development of WSHGs through Aquaculture Activities in Keonjhar and Koraput Districts of Orissa

Nirupama Panda[1] and K.B. Dutta[2]

Aquaculture is an economic, employment generating and developmental activity for women self-help groups. Koraput and Keonjhar are two tribal dominated districts of Orissa and are inhabited with more than 76% of households living below poverty line. The present study was conducted to statistically evaluate the socio-economic development of the members of the women self-help groups undertaking aquaculture activities in these two districts. Seventeen developmental indicators common to all the WSHGs were considered for analysis. The composite indices of development for the WSHGs in three dimensions e.g.

economic, social and empowerment along with the overall development index have been estimated. On the basis of the level of development, the WSHGs were classified into four categories of development e.g. high level, high middle, low middle and low level.

It is revealed from the study that most of the WSHGs have developed to the stages of high middle and low middle level of development through aquaculture activities. The members of the WSHGs were empowered to a higher level (Composite Index = 0.37) compared to economic development (Composite Index = 0.56) and social development (Composite Index = 0.64) during the year 2008-09. The correlation coefficient between the composite index of overall development and composite index of social development was highest (r = 0.89) and the correlation coefficient between economic development and empowerment was the lowest (r = 0.04).

[1] *Central Institute of Freshwater Aquaculture, Bhubaneswar*
[2] *Sambalpur University, Sambalpur*

15. Poultry Egg Production in India

Shiv Prasad[1], Rajendra Singh[1] and D.P. Singh[2]

Poultry is a part of our livestock and provides protein in the form of meat and eggs. The fowls and ducks contributed 93.54% and 6.21% to total poultry population (489.01 million) in the year 2003. About 1832 million eggs were produced in 1950-51 and increased to 53532 million eggs in 2007-08 with 6.10% annual growth rate so that the availability of eggs per head per year increased from 5 eggs in 1950-51 to 51 eggs in 2007-08. The fowls and ducks contributed 94.69% and 2.88% respectively to total egg production during 2006-07. The average annual egg production of improved fowls (250) was higher than Desi fowls (90) and ducks. Andhra Pradesh and Tamil Nadu ranked first and second with respect to annual egg production. Different growth functions viz. linear, quadratic, non-polynomial, exponential, logistic and Gompertz were fitted to total egg production from 1950-51 to 2007-08. On the basis of different measures of goodness of fit, exponential model was found the best to describe the egg production curve.

[1]*Indian Veterinary Research Institute, Izatnagar*
[2]*Central Avian Research Institute, Izatnagar*

16. Some Investigation on the Statistical Properties of Goodness of Fit Criteria of Non-linear Growth Curves through Bootstrap Technique

A.K. Paul, Mohan Das Singh and S.D. Wahi

The three non-linear growth models that are Logistic, Gompertz, and Von Bertalanffy have been considered to investigate the statistical properties of goodness of fit criteria. The monthly body weight data from birth to 12 months age on 110 animals obtained from CIRG goat farms, Makhdoom, Uttar Pradesh for the year 2005 has been used for study. The bootstrap samples are drawn with replacement from this data to obtain statistical measures of the goodness of fit criteria like R^2, RMSE and ARR of all the three growth models. The statistical analysis based on the bootstrap samples shows the non normal distributions of these criteria. The overall mean value of RMSE is 0.7876 with standard deviation 0.0267 and overall mean value of ARR is 32.2067 with standard deviation 3.1833 are found least in case of Von-bertallanffy model. So, among the three models studied the Von bertalanffy model comes out to be the best model to describe the growth pattern goats.

Indian Agricultural Statistics Research Institute, New Delhi

17. Market Integration in Coarse Cereals in India: A Case of Maize and Jowar

Prawin Arya, A.K. Vasisht, Sivaramane N. and D.R. Singh

Coarse cereals, being the staple diet of millions of poor people of India play a very important role in maintaining their nutritional status. However, due to change in per capita income and subsequently the dietary pattern, there is a gradual shift in consumption from coarse cereals to major cereals such as rice and wheat. Due to the imbalances in the production performance of different crops and relatively poor performance or stagnation of the important coarse cereals along with wide fluctuations in prices, the coarse cereals are decelerating at a very fast pace. Several studies have elucidated factors such as poor demand, rise in the per capita income, increase in irrigated area, and increase in the relative output prices in favour of major cereals, and market imperfections as the major reasons for this shift. In this backdrop, this study has been conducted to investigate spatial integration of Maize (*Zea mays*) and Jowar (*Sorghum bicolor*) markets in India. The multivariate

cointegration methodology using Johansen's procedure was used to study the extent and nature of cointegration among the maize and jowar markets. The time series data pertaining to the period Jan. 2001 to Dec. 2008 on the wholesale prices of sixteen maize markets and eleven jowar markets were used for the analysis. The data were screened to impute the missing values and to remove the outliers. Further, the data was transformed by detrending and deseasonalizing. The results showed that the nine jowar price series and twelve maize price series were non-stationary and were integrated of the first order. The maximum eigen value and trace tests showed that there were six and five cointegrating vectors for jowar and maize markets respectively revealing reasonably good degree of integration among markets. The series though integrated in the longrun had shown disturbances in the shortrun. The ECM showed the pace of adjustment of the shortrun disturbances towards the longrun equilibrium.

Indian Agricultural Statistics Research Institute, New Delhi

18. On Methods of Estimation of Generalized Negative Binomial Distribution

Abhay Kumar[1], R.C. Bharati[1], S.K. Singh[2], A. Mishra[2], and K.M. Singh[1]

The negative binomial distribution was perhaps the first probability distribution, considered in statistics, whose variance is larger than its mean. On account of wide variety of available discrete distributions, the research workers in applied fields have begun to wonder which distribution would be most suitable one in a particular case and how to choose it. With the aim of reducing this problem, Jain and Consul (1971) gave a Generalized Negative Binomial Distribution (GNBD) by compounding the negative binomial distribution with another parameter which takes into account the variations in the mean and variance. This GNBD reduces to the binomial or the negative binomial distribution as particular cases and converges to a Poisson-type distribution in which the variance may be more than, equal to or less than the mean, depending upon the value of the parameter. A number of methods for estimation of parameters of GNBD, like weighted discrepancies method, minimum chi-square method etc. are available but these methods produce such equations which are not simple to be solved directly and hence some iterations has to be applied to find the solution. An alternative estimator has been suggested here, which is capable of giving more or less as good results as given by the moment estimators. Although, the probability of the observed value of χ^2 to be exceeded, are slightly higher in case of the suggested method that in case of method of moments, these differences do not seem to be much significant and can be considered due to sample fluctuation. Further, the suggested method has one definite advantage over the other methods in certain situations. It can be applied when the method of moments fails to give estimates of the parameters. Moreover, it is relatively very quick to be obtained and so it may be preferred to others where very quick results are required.

[1] *ICAR Research Complex for ER, Patna*

[2] *Patna University, Patna*

19. Growth Pattern and Technological Impact of Oilseeds Production in Uttar Pradesh

A.K. Bharti, L.K. Dube, B.V.S Sisodia and M.K. Sharma

An attempt was made in present investigation to find the trend and growth rates and the impact of Technology Mission on Total oilseeds including the oilseeds crops i.e. groundnut, rapeseed-mustard and linseed. The time-series data on area, production and productivity of oilseeds crops and total oilseeds in Uttar Pradesh pertaining to the period 1970-71 to 2005-06 were used for the investigation of trend and growth of oilseeds and also impact of technological changes on oilseeds production in the state of Uttar Pradesh. The relevant statistical tools & technique like regression analysis etc. have been used for the purpose of investigation.

An attempt was also made in present investigation to fit the parametric & nonparametric regression models to arrive at a methodology that can precisely estimate the trend and growth rates in area, production & productivity of rapeseed-mustard crops grown in different agro-climatic zones of U.P. For a period of 36 years time-series data from 1970-71 to 2005-06 on area, production and productivity of rapeseed-mustard crop of the U.P. were collected. In the case of parametric models, first, second, third degree polynomials, exponential and Gompertz models were considered. The statistically most suited parametric models were

selected on the basis of adjusted R^2, significant partial regression coefficients and significant coefficient of determination (R^2), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values and assumptions of residuals. The statistically sound model was selected on the basis of various goodness of fit criteria viz. RMSE, MAE and assumptions of residuals (Shaprio-Wilk test for normality and Lijung & Box test for randomness). Relative growth rates of area, production and productivity of the rapeseed-mustard crop were calculated for the successive years starting from 1970-71 to 2005-06.

The results indicated that none of the parametric models were found suitable to fit the trend in area, production and productivity of the rapeseed-mustard crop under investigation due to either non-significant partial regression coefficients lack of assumptions of the residuals. Non-parametric regression model was selected as the best fitted function for the area, production and productivity of the rapeseed-mustard crop.

On an overall it can be concluded that increase in area in case of rapeseed-mustard & groundnut has made some breakthrough in their production. Increase in productivity of groundnut, rapeseed-mustard and linseed has also contributed to their production to some extent. The production of oilseed has however not been satisfactory in the State in terms of its productivity as well as production because there is still shortfall of our desired level of oilseed production. It has been also found that technological changes have made significant structural technological changes have made significant structural difference in the production process of rapeseed-mustard between two periods. More efforts should be made to concentrate on these crops besides rapeseed-mustard by the policy makers in the State.

N.D. University of Agriculture & Technology, Faizabad

20. Family Size Distribution and Correlation Between the Numbers of Two Types of Children in Family: Evidence from Six African Countries

H.L. Sharma[1], R.N. Singh[2] and Roshni Tiwari[1]

This paper is concerned with the investigation of the family size distribution as Polya-Aeppli distribution, truncated below one and computation of correlation between the numbers of male and female children in a family. The distribution involves two parameters θ and q and these are estimated by methods of moments. The suitability of the distribution and correlation is tested using 10% sample of new DHS data gathered recently on six countries namely *Lesotho, Namibia, Kenya, Swaziland, Zambia,* and *Zimbabwe* related to Eastern-Southern sub-Saharan Africa. The observed and expected correlation coefficients are found to be significant and the same in each country. There is more variation in the estimated values of θ rather than that of q. The results reveal that the average family size mainly depends on θ, average size of groups per couple in a family, not on q. Through the values of χ^2 at 5% level, we deem the fit of the distribution to be very good. The values of the parameters of this distribution may be used to generate the underlying sibship sizes for the simulation study.

[1]*J.N. Agricultural University, Jabalpur*

[2]*Agriculture Research Institute, RAU, Patna*

21. Scenario of Fertilizer Consumption in Eastern Uttar Pradesh

Annu, L.K. Dubey and M.K. Sharma

Eastern Uttar Pradesh is the most populous and occupies an important place in fertilizer consumption in the state. It covers 25 districts. Spread over in three –climatic zones viz. North Eastern plain zone (NEPZ), Eastern plain Zone (EPZ) and Vindhyan Zone (VZ). The study is based on secondary data which was procured from Directorate of Agricultural Statistics and Crop Insurance covering the period 1970-71 to 2004-05. This paper has been prepared to study the aspect of fertilizer consumption during pre and post green revolution period along with trend of fertilizer consumption. The three yearly moving average methods used for smoothing of data and also finding the trend of fertilizer consumption. The different growth curves were used for growth rate of fertilizer consumption. The trend of fertilizer is very aggressive due to awareness of farmers. The year 1985-86 has been found to consume highest fertilizers consumption for all the zones. The eastern plain zone recorded highest growth rate in term of fertilizer consumption during green revolution whereas, North Eastern Plain Zone showed highest growth rate in post green revolution period. The study shows that there has been an abrupt increase in

growth rate in fertilizer consumption in Eastern Uttar Pradesh, which might be because of certain favourable conditions.

N.D. University of Agriculture & Technology, Faizabad

22. Balanced Block Designs for All Order Neighbour Correlations

Anurup Majumder and Aatish Sahu

The work of Morgan and Chakraborti (1988) for optimality results for block designs under first and second order (NN1 and NN2, respectively) neighbour correlations has been extended foe all order (NNt, where t = 1, 2,..., k – 1 and k is the block size) neighbour correlations. Conditions for optimality and minimality are presented for NNt model.

Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia

23. Forecasting of Arrivals and Prices in Ramnagar and Siddlaghatta Cocoon Markets

R. Bharathi, Y.N. Havaldar, S.N. Megeri and G.M. Patil

The cocoon study was carried out to forecast the arrivals and prices of cocoon of Ramnagar and Siddlaghatta market. The main aim was to forecast the arrivals and prices of cocoon in both the markets. The data on price and arrivals of cocoon was collected for those markets from 1998-99 to 2007-08. ARIMA (Auto Regressive Integrated Moving Average) model was used for forecasting of arrivals and prices in both the markets. Suitable model was identified based on the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function). The adequacy of the model was judged based on the values of Box-Pierce Q statistics and AIC (Akike Information Co-efficient) and the accuracy of forecasts for both Ex-ante and Ex-post were tested by using MSE (Mean Square Error) and MAPE (Mean Average Percentage Error). The model (1,1,3) (1,1,1) was tentatively identified for arrivals and model (0,1,0) (1,1,1) was identified for prices of cocoon in Ramnagar market and model (2,1,1) (1,1,1) was identified for arrivals and (0,1,0) (1,1,1) model for prices in Siddlaghatta market. Forecasted values of arrivals showed increasing trend in both the markets and price showed decreasing trend in Siddlaghatta market.

University of Agricultural Sciences, Dharwad

24. Prediction Models for Consumption of Pesticide

S.N. Megeri and R. Veena

A study was conducted based on the secondary data procured from the Agricultural Research, Data Book 2004 and Compendium of Environment Statistics, 2001, Central Statistical Organization, Ministry of Statistics and Programme Implementation & Ministry of Chemicals and Fertilizers, Govt. of India & Ministry of Agriculture, Govt. of India. Here we tried to find out the suitable models to predict the pesticide consumption in India and some of the southern states of India. For the present study 15 years data was collected on pesticide consumption in India and for southern states viz. Karnataka, Andhra Pradesh, Tamil Nadu and Goa. The result reveals that for pesticide consumption the best fitted models were linear and quadratic regression model with R^2 values 0.932 and 0.943 respectively for India. Similarly for Karnataka, Andhra Pradesh, Tamil Nadu and Goa also linear and quadratic regression models fits well. We noticed decreasing trend in the consumption of pesticides. This indicates that instead of using excessive chemical pesticides in various crops to control insect pests of the farmers are using alternative techniques of pest control like cultural, mechanical, biological methods and bio pesticides. Using the best fitted models we forecasted requirement of pesticide for next five years.

University of Agricultural Sciences, Dharwad

25. Forecasting of Rice (*Oryza sativa* L.) Production in Tamil Nadu by using Statistical Model

Y.N. Havaldar, S.N. Megeri and K. Padmanaban

In this paper, an attempt has been made to forecast rice production using statistical time-series modelling technique– simple exponential smoothing. In Tamil Nadu the rice is the important crop among all crops. The study indicated that there is declining trend in production due to low rainfall and also the some of the rice growing sea shore districts are affected by natural calamities in 2003-04.Then introduction of some new rice growing technology and water management system, the production of rice growing increasing trend.

University of Agricultural Sciences, Dharwad

26. Forecasting Models for Fertilizer Consumption in India

S.N. Megeri and H.S. Anil Kumar

Long term fertilizer requirement and dependency are key to the success of long term plans for global food security and the profitability of the fertilizer industry. The production of nitrogen (*N*) and phosphorus (*P*) fertilizer together has increased from mere 0.3 lakh MT in 1950-51 to about 147 lakh MT in nutrients terms in 2001-02. The overall consumption of fertilizers in nutrient terms (*N*, *P* & *K*) currently is about 175 lakh MT per annum. Hence in the present study investigation is carried out, on *NPK* fertilizer consumption. It explains fertilizer demand in relation to net irrigated area and dry land area in India. The multiple regression was used to estimate the requirement of the fertilizer, it was noticed $R^2 = 0.97$ and is significant. With regard to fertilizer consumption over years different models were tried and cubic model was found best based on R^2 value followed by logistic model, power, linear and exponential model. Using the selected models the fertilizer consumption was estimated and also forecasted for next ten years.

University of Agricultural Sciences, Dharwad

27. Use of Discriminant Function Analysis for Forecasting Wheat Yield

Ranjana Agrawal, Chandrahas and Kaustav Aditya

Weather based modeling is one of the major approaches for forecasting crop yields. The approach utilizes time series data on weather variables and trend as explanatory variables. Mostly multiple regression technique is used taking trend along with weather variables as such or in some derived form or indices as regressors. The present paper deals with use of discriminant function analysis for developing wheat yield forecast model for Kanpur. Discriminant function analysis is a technique of obtaining linear / quadratic function which discriminates the best among populations and as such, provides qualitative assessment of the probable yield. In the present paper, quantitative forecasts of yield have been obtained using multiple regression technique taking regressors as weather scores obtained through discriminant function analysis. The time series data of 30 years (1971-2000) have been divided into three categories: congenial, normal and adverse based on yield distribution. Taking these three groups as three populations, discriminant function analysis has been carried out. These discriminant functions have been used to obtain weather scores which have been used as regressors in the modeling. Various strategies of using weekly weather data have been adopted. The models have been used to forecast yield in the subsequent three years 2000-01, 2001-02, 2002-03 (which were not included in model development). The approach provided reliable yield forecast about two months before harvest.

Indian Agricultural Statistics Research Institute, New Delhi

28. Comparison of Statistical Models for Rice Yield Forecasting

Y.A. Garde[1], A.K. Shukla[1], R.K. Sharma[1], S.K. Tewari[1] and S. Singh[2]

A study was undertaken for forecasting yield of rice crop based on time series data for 27 years (1981-82 to 2007-08) of yield and weather parameters obtained from G. B. Pant University of Agriculture and Technology, Pantnagar, District Udham Singh Nagar, Uttarakhand. To study the association between yearly crop yields and different weekly weather parameters, Karl-Pearson's correlation techniques were applied. For forecasting the yield of rice two statistical models were applied. It was found that proposed Modified Model-II based on technical and statistical indicators for forecasting the rice yield was better than Model-I of Agrawal *et al*. (2001). It was concluded that Model II may be effectively used for early pre-harvest forecasting of crop yield particularly up to two and half month before harvesting.

[1]*Govind Ballabh Pant University of Agriculture and Technology, Pantnagar*

[2]*Banaras Hindu University, Varanasi*

29. Trends and Decomposition Analysis of Lentil in India

Hemant Kumar, Devraj and Purushottam

The growth in lentil area and production in the country has shown increasing trends. The study shows that the best trend in area and production is quadratic in nature. The significantly positive square term in the quadratic equation indicates acceleration in lentil area and production in the country during the study period.

Instability analysis shows that area and production of lentil remains almost stagnant during the classified periods. This variation indicates that stability in lentil area and production has not been obtained in the country. The present study highlights that the country as a whole showed significant growth rate in area and production of lentil during the period under study (1970-71 to 2006-07). Thus, further change in the lentil production in the country might be attributed to all the three effects i.e., area effect, yield effect and interaction effect. During the overall period, the change in the total production of lentil was completely due to the change in area under the crop as the yield and interaction effects were very small. Therefore, it is concluded that production growth in lentil over the past forty years has been slow and unstable with substantial temporal variation in the country.

Indian Institute of Pulses Research, Kanpur

30. Forecasting Technological Needs in Genetics and Plant Breeding for Sustainable Agriculture

Ramasubramanian V., V.K. Bhatia, Amrender Kumar, Satya Pal and Sarvesh Kumar Premi

Technology Forecasting (TF) is the process of prediction of feasible or desirable characteristics of performance parameters in future technologies such as machines or techniques. TF methodologies range from intuitive (e.g. Delphi, Brainstorming) to statistical (e.g. trend extrapolation, growth models) to normative methods (e.g. relevance trees) and also include the scientometric methods. The domain of genetics and plant breeding has continued to evolve with a much broader scope and potential than in the past more so with incorporation of new knowledge from other fields of science. The conventional plant breeding programs can reduce the time frame for evolving new technologies if it takes full advantage of emerging fields like biotechnology. As a TF exercise, a Brainstorming session was organized at IARI, New Delhi which provided a platform to plant breeders, geneticists etc. in scripting the future technological needs of agriculture pertaining to the domain of genetics and plant breeding for converting the crop varieties/commodities into viable products for productivity improvement and effective utilization of modern tools for value addition and genetic enhancement. Information from experts was obtained through questionnaires for identification of specific technologies with greater utility which was then statistically analyzed for prioritizing future technological needs. Attempts were also made to analyze the available information using multi-dimensional scaling approach.

Indian Agricultural Statistics Research Institute, New Delhi

31. Monsoon and India's Rice Production - Rice Forecasting through ANN Modelling Methodology

S. Ravichandran

Artificial Neural Network (ANN) is an information processing system that roughly replicates the behaviour of a human brain by emulating the operations and connectivity of biological neurons. ANN modelling methodology is widely utilized for modelling time-series data and subsequent forecasting. Monsoon is the lifeline of Indian agriculture. Due to global climate change, monsoon also fluctuates. Indian sub-continent receives good rains in some years and fails to receive sufficient rainfall in remaining years due mainly to climate change. Modelling and forecasting all-India rice production is carried out by utilising data on all-India rice area, production and yield for the period 1950-51 to 2008-09 along with all-India rainfall data from June to September for the corresponding period. India, a monsoon dependent country, received one of the lowest rainfall (only 353 mm) during the peak monsoon months of June and July in the last 50 years. The previous lowest rainfall in June and July was during 2002-03 (306 mm), which happened to a severe drought year. Previous lowest June and July rainfall in some other years in the last 59 years are: 334 mm in 1972-73, 337 mm in 1987-88, 377 mm in 1982-83, 391 mm in 1962-63, 393 mm in 1965-66, 397 mm in 1979-80. Rainfall received from 1st June to 30th September, 2009 was 689.8 mm, which is 23% below the average rainfall. Based on the data for the period 1950-51 to 2008-09, several ANN models were developed using ANN models by making use of 70% of the data for training the model, 20% for testing and remaining 10% of data is utilised for validating the model. Forecasting was carried out by making use of the most efficient model. Based on the model, forecasting rice production for 2009-10 would be 82.19 million tons. This would be lesser by 16.9% compared to rice production during 2008–09.

Directorate of Rice Research, Rajendranagar, Hyderabad

32. Implications of Global Warming on Agricultural Production in India

Sushila Kaul

Agriculture is vulnerable to global warming and the world's most widely eaten grains i.e. rice, wheat and corn are exquisitely sensitive to higher temperature. The impact of global warming on agriculture in developing countries, and particularly in countries like India, that depend on rain for irrigation, are likely to be devastating. Rice crop, in much of India will be affected by the global warming. Most of the hunger, resulting from global warming, is likely to be felt by those who are not responsible for contributing to the cause of the problem i.e. the people of developing countries. With climate change, the agricultural areas in the tropics will decline, causing a situation that those who are well off now will be better off in the future, and those who are in problems will be prone to greater problems. A rough rule of thumb developed by crop scientists is that, for every 1-degree Celsius increase in temperature, above the mid-30s, during key stages in the growing season, such as pollination, yields fall about 10 per cent. Optimum growing conditions for most of the crops, generally range from about 20 to 35 degrees, and then diminish sharply. At 40 degrees, heat stress causes photosynthesis to shut down. The climate change may erase all the gains that accrued, as a result of the technological advancements. The present study has been undertaken to assess the impact of climate related variables and agricultural production in India. For this purpose, two crops viz. rice and jowar have been selected. These two crops are predominantly grown in monsoon season and any change in climate, particularly rainfall and temperature would affect the productivity of these crops significantly. In the present paper an econometric model of crop production has been attempted. The analysis has been undertaken for the country as a whole, using state wise data for both these crops. For rice crop, the state of Orissa has been selected, because this is the main crop of the state. For jowar, the state of Karnataka has been selected, because of its significance in the region. Data on different variables for various states has been collected from the publications of Directorate of Economics and Statistics, Ministry of Agriculture, as well as from the Economics and Statistics Directorates of various states. Similarly district-wise data has been primarily collected from the publications of these two state governments. The study reveals that excessive rains and extreme variation in temperature would affect the productivity of these crops adversely, thereby affecting the incomes of farming families in a negative manner. Thus suitable strategies pertaining to resource use, planting flood and drought resistant varieties of crops, better irrigation networks, and crop mix are to be adopted for mitigating the harmful effects of climatic changes.

Indian Agricultural Statistics Research Institute, New Delhi

33. Growth Trends of Area, Production and Productivity of Rice (*Oryza sativa L.*) in Bihar

Mahesh Kumar[1], Manish Sharma[2], N.K. Azad[1], Nidhi[1] and S.K. Sinha[1]

As rice is important staple food crop, its demand increases with increasing production. Looking to this important of rice crop in Bihar, an attempt has been made to study growth trend to rice regarding area, production and productivity data in Bihar. The 60 years data (1947-2007) were collected from Directorate of Statistics & Evaluation, Govt. of Bihar, Patna.

For this purpose data of 60 years is divided into two phases each with 30 years i.e. in first phase 1946-47 to 1976-77 and second phase 1977-78 to 2006-07. Further, phase first was divided into base period as 1946-47 to 1949-50 and current period as 1974-75 to 1976-77 phase second base period as 1977-78 to 1980-81 and current period as 2003-04 to 2006-07 by taking triennium average of each period in order to avoid fluctuations of year to year data. Further, the instability in area, production and productivity were calculated by using the coefficient of variation (C.V.).

[1] *RAU, Pusa, Samastipur*

[2] *Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, Jammu*

Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 259-265

Methodology for Estimation of Production of Flowers on the Basis of Market Arrivals

A.K. Gupta*, H.V.L. Bathla, U.C. Sud and K.K. Tyagi
Indian Agricultural Statistics Research Institute, New Delhi

(Received: January 2009, Revised: June 2009, Accepted: August 2009)

SUMMARY

Appropriate methodology has been developed for estimation of production of important flowers in Delhi on the basis of market arrivals data in flower mandis. The estimated market arrivals figures were in agreement with the estimated figures obtained through village survey.

Key words : Flower mandi, Market arrivals, Village survey approach.

1. INTRODUCTION

India has a vide range of climate and soil conditions which enable cultivation of an array of horticulture crops such as fruits, vegetables, floriculture plants, plantation crops etc. Among these, floriculture is a fast expanding dynamic industry which has gained momentum with the liberization of economic and industrial trade policies. The Government of India has identified floriculture as a major thrust area for export because floriculture industry in India has made significant progress in the recent years. Flowers, of all kind, besides being a source of essential oils are in great demand for decoration and various other purposes. Floriculture is now a remunerable venture for unemployed youth and women, yielding higher returns and generating employment opportunities for rural people in the villages. With varied agro-climatic conditions available in the country, it is possible to grow almost all the major flower crops of the world, either from tropical, sub-tropical or temperate region.

Following the Green Revolution in Agriculture, flowers are considered to be one of the best alternatives for diversification.

Currently, no scientific methodology is available for reporting area and production of flowers in the country. The existing crop-cutting experiments approach being followed for estimation of production of food crops may not be appropriate in case of flowers when seen from the point of view of cost and time involved. This is due to the fact that the production of flowers involves multiple pickings. The enumerator is thus required to make multiple visits for recording the production of flowers. Recording the produce of every picking may turn out to be very cumbersome.

With the specific objective of strengthening the existing database pertaining to flowers, the National Statistical Commission recommended that a cost-effective suitable sampling methodology be developed for estimation of production of important flowers on the basis of market arrivals. Accordingly, a pilot study funded by Central Statistical Organization, Ministry of Statistics and Programme Implementation, Government of India, entitled "Pilot sample survey to develop sampling methodology for estimation of area, production and productivity of important flowers on the basis of market arrivals", was planned and conducted in Delhi State during September 2003 to August 2004 in which estimates of production of flowers were developed by considering two approaches namely, (i) on the basis of market arrivals of flowers and (ii) on the basis of village survey.

* *Corresponding author* : A.K. Gupta
E-mail address : akgupta@iasri.res.in

2. MATERIALS AND METHODS

(i) Market Survey Approach

There are three flower mandis in Delhi namely, Hanuman Mandir Mandi, Khari Baoli Mandi and Mehrauli Mandi. Cut flowers of Rose, Gladiolus, Chrysanthemum, Tube-rose and Carnation etc. are mainly traded in the Hanuman Mandir Mandi while trading of loose flowers of Marigold, Rose, Margaret and Jaffrey etc. is carried out in Khari Baoli and Mehrauli Mandi by the commission agents and self-selling farmers. The commission agents and self-selling farmers were selected as per the following sampling design adopted for collection of data on varieties of flowers sold in the three mandis.

Sampling Design: A stratified random sampling design was followed in each flower mandi. Commission agents comprise the first stratum while the self-selling farmers the second stratum. Within the first stratum, seven random groups of commission agents were formed to cover all the commission agents trading in both Hanuman Mandir Mandi and Khari Baoli Mandi. The survey work was carried out from September 2003 - August 2004. The entire one year survey work was divided into three periods viz. Period-1: September - December 2003, Period-2: January - April 2004 and Period-3: May - August 2004. One random group of commission agents was randomly selected and observed for a fortnight in one period for data collection purpose. The remaining six random groups were observed in a similar manner. All the seven groups were observed in seven fortnights in one period. This process was repeated in the other two periods also. All the commission agents of Mehrauli mandi were observed for 8 days in each period. A suitable number of self-selling farmers were selected to collect inquiry based data on flowers sold. A self-selling farmer once chosen was not repeated in the particular period of inquiry.

(ii) Village Survey Approach

Sampling Design: The sampling design adopted for estimation of area under floriculture in flower growing villages of Delhi was one of stratified uni-stage random sampling with villages as the sampling units. For estimating production of important flowers on the basis of village survey, the sampling design was stratified two stage random sampling with villages as first stage sampling units and flower growing farmers as the second stage sampling units. All the flower growing villages of Delhi were divided, in each period, into three strata as follows; Stratum I: villages having area up to 5 ha under flower, Stratum II: villages having area more than 5 ha and less than 10 ha, and Stratum III: villages having area more than 10 ha under floriculture. Out of 92 flower growing villages in Delhi, a simple random sample of 15 flower growing villages was selected. This sample of 15 flower growing villages was allocated among the three strata using proportional allocation according to area under flowers in each village in each of the three Periods. Accordingly, among each of the three strata, 3, 7 and 5 villages were selected in Period-1; 5, 4 and 6 villages in Period-2; and 5, 3 and 7 villages in Period-3 respectively. The area estimates were obtained by complete enumeration of all the flower growing farmers in each of the selected villages. Villages having 15 or less than 15 flower growing farmers were completely enumerated for compilation of production figures. The production estimates for villages having more than 15 flower growing farmers were made on the basis of a random sample of 15 flower growing farmers selected in such a way that each kind of flower grown by the farmers got properly represented. The sampling units at both the stages were selected by simple random sampling without replacement.

3. ESTIMATION PROCEDURE

(i) Market Survey Approach

Estimation of total market arrivals in Delhi for a particular kind of flower in the i^{th} period

The entire survey period of one year (366 days) was divided into 3 periods viz. Period-1 of 122 days, Period-2 of 121 days and Period-3 of 123 days. Two strata viz. stratum-1 of Commission Agents/Mashakhors and stratum-2 of self-selling farmers and seven groups (each group comprising of suitable number of commission agents/mashakhors from first stratum and self-selling farmers from the second stratum) were formed.

Let y_{ighad} be the quantity of market arrivals of the flower for i^{th} period, g^{th} group, h^{th} stratum, a^{th} commission agent/ self-selling farmer on d^{th} day. Mean market arrivals per day for i^{th} period, g^{th} group, h^{th} stratum, a^{th} commission agent/self-selling farmer is given by

$$\bar{y}_{igha.} = \frac{1}{m_{igha}} \sum_{d}^{m_{igha}} y_{ighad}$$

where m_{igha} is the number of days for which a^{th} commission agent/self-selling farmer was observed for i^{th} period, g^{th} group, h^{th} stratum.

Estimated mean market arrivals per day per commission agent/self-selling farmer in the i^{th} period, g^{th} group, h^{th} stratum is given by

$$\bar{y}_{igh..} = \frac{1}{n_{igh}} \sum_{a}^{n_{igh}} \bar{y}_{igha.}$$

where n_{igh} is the number of commission agents observed in the i^{th} period, g^{th} group, h^{th} stratum.

The variance of $\bar{y}_{igh..}$ is given by

$$V(\bar{y}_{igh..}) = \left(\frac{1}{n_{igh}} - \frac{1}{N_{igh}} \right) \frac{1}{(N_{igh} - 1)} \sum_{a}^{N_{igh}} \left(\bar{y}_{igha.} - \bar{Y}_{igh..} \right)^2$$

where $\bar{Y}_{igh..} = \frac{1}{N_{igh}} \sum_{a}^{N_{igh}} \bar{y}_{igha.}$

The estimator of $V(\bar{y}_{igh..})$ is given by

$$\hat{V}(\bar{y}_{igh..}) = \left(\frac{1}{n_{igh}} - \frac{1}{N_{igh}} \right) \frac{1}{(n_{igh} - 1)} \sum_{a}^{n_{igh}} \left(\bar{y}_{igha.} - \bar{y}_{igh..} \right)^2$$

where N_{igh} is the total number of commission agents pertaining to the i^{th} period, g^{th} group, h^{th} stratum.

Estimated total market arrivals per day for the mandi on the basis of g^{th} group is given by

$$\hat{Y}_{ig...} = \sum_{h=1}^{2} N_{igh} \bar{y}_{igh..}$$

The variance of $\hat{Y}_{ig...}$ is given by

$$V(\hat{Y}_{ig...}) = \sum_{h=1}^{2} N_{igh}^2 V(\bar{y}_{igh..})$$

and the corresponding variance estimator by

$$\hat{V}(\hat{Y}_{ig...}) = \sum_{h=1}^{2} N_{igh}^2 \hat{V}(\bar{y}_{igh..})$$

Estimated total market arrivals per day for the mandi averaged over all the groups for the i^{th} period is

$$\hat{Y}_{i....} = \frac{1}{g} \sum_{g=1}^{7} \hat{Y}_{ig...}$$

The variance of $\hat{Y}_{i....}$ is given by

$$V\left(\hat{Y}_{i....}\right) = \frac{1}{g^2} \sum_{g=1}^{7} V\left(\hat{Y}_{ig...}\right)$$

and the estimator of $V\left(\hat{Y}_{i....}\right)$ is given by

$$\hat{V}\left(\hat{Y}_{i....}\right) = \frac{1}{g^2} \sum_{g=1}^{7} \hat{V}\left(\hat{Y}_{ig...}\right)$$

Estimator of total market arrivals for the entire year (366 days) for a specified kind of flower is given by

$$\hat{Y}..... = 122\ \hat{Y}_{1....} + 121\ \hat{Y}_{2....} + 123\ \hat{Y}_{3....}$$

The variance of $\hat{Y}.....$ is given by

$$V(\hat{Y}_{.....}) = (122)^2 V(\hat{Y}_{1....}) + (121)^2\ V(\hat{Y}_{2....}) + (123)^2 V(\hat{Y}_{3....})$$

and the estimator of variance of the estimate of total market arrivals for the entire year is given by

$$\hat{V}(\hat{Y}_{.....}) = (122)^2 \hat{V}(\hat{Y}_{1....}) + (121)^2\ \hat{V}(\hat{Y}_{2....}) + (123)^2 \hat{V}(\hat{Y}_{3....})$$

Estimates of all kind of flowers grown in Delhi on the basis of market arrivals have been obtained on the similar lines.

(ii) Village Survey Approach

Estimation of total production for a particular kind of flower for a period in the villages of Delhi

Let y_{hij} be the production of a particular kind of flower for the j^{th} flower growing farmer of the i^{th} flower growing village in the h^{th} stratum (j = 1, 2, …, M_{hi} ; i = 1, 2, …, N_h ; h = 1, 2, 3), M_{hi} being the number of

flower growing farmers in the i^{th} village of the h^{th} stratum and N_h, the total number of flower growing villages in the h^{th} stratum. The average production of a particular kind of flower in the i^{th} village of the h^{th} stratum is given by

$$\bar{y}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$$

where m_{hi} is the number of selected flower growing farmers in the i^{th} village of h^{th} stratum.

Total production per village of a particular kind of flower in the h^{th} stratum is given by

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} M_{hi} \bar{y}_{hi}$$

where n_h denotes the number of selected villages in the h^{th} stratum.

Let y'_{1i} is the production of a particular kind of flower reported from i^{th} village in the 1^{st} stratum. Hence

$\bar{y}'_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y'_{1i}$ is the estimate of average production of a particular kind of flower in 1^{st} stratum. Again

$\bar{y}'_h = \frac{1}{n_h} \sum_{i}^{n_h} M_{hi} \bar{y}_{hi}$ is the estimate of average production of a particular kind of flower in h^{th} stratum (h = 2, 3).

Accordingly, an estimator of the total production of a particular kind of flower in Delhi is given by

$$\hat{Y} = \frac{N_1}{n_1} \sum_{i=1}^{n_1} y'_{1i} + \sum_{h=2}^{3} N_h \bar{y}'_h$$
$$= \hat{Y}_1 + \hat{Y}_2$$

where $\hat{Y}_1 = \frac{N_1}{n_1} \sum_{i=1}^{n_1} y'_{1i}$ and $\hat{Y}_2 = \sum_{h=2}^{3} N_h \bar{y}'_h$

The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = N_1^2 \left(\frac{1}{n_1} - \frac{1}{N_1}\right) S_{b1}^2 + \sum_{h=2}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{bh}^2 + \sum_{h=2}^{3} \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi}^2 \left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}}\right) S_{hi}^2$$

where

$$S_{b1}^2 = \frac{1}{(N_1 - 1)} \sum_{i=1}^{N_1} (y'_{1i} - \bar{Y}_1)^2$$

$$S_{bh}^2 = \frac{1}{(N_h - 1)} \sum_{i=1}^{N_h} \left(M_{hi} \bar{Y}_{hi} - \bar{Y}_h\right)^2$$

$$S_{hi}^2 = \frac{1}{(M_{hi} - 1)} \sum_{j=1}^{M_{hi}} (y_{hij} - \bar{Y}_{hi})^2$$

$$\bar{Y}_1 = \frac{1}{N_1} \sum_{i}^{N_1} y'_{1i}$$

$$\bar{Y}_{hi} = \frac{1}{M_{hi}} \sum_{j}^{M_{hi}} y_{hij}$$

and $\bar{Y}_h = \frac{1}{N_h} \sum_{i}^{N_h} M_{hi} \bar{Y}_{hi}$

An estimator of $V(\hat{Y})$ is given by

$$\hat{V}(\hat{Y}) = N_1^2 \left(\frac{1}{n_1} - \frac{1}{N_1}\right) s_{b1}^2 + \sum_{h=2}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{bh}^2 + \sum_{h=2}^{3} \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}^2 \left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}}\right) s_{hi}^2$$

where

$$s_{b1}^2 = \frac{1}{(n_1 - 1)} \sum_{i=1}^{n_1} (y'_{1i} - \bar{y}'_1)^2$$

$$s_{bh}^2 = \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (M_{hi} \bar{y}_{hi} - \bar{y}'_h)^2$$

and $s_{hi}^2 = \frac{1}{(m_{hi} - 1)} \sum_{j=1}^{m_{hi}} (y_{hij} - \bar{y}_{hi})^2$

Combined estimate for the three periods has been obtained by adding the estimates of different kind of flowers to get an estimate for the entire duration of one year.

Estimation of area for a particular kind of flower for *i*th period on the basis of village survey

Let $a_{h'j}$ be the area under a particular kind of flower in the j^{th} village of h'^{th} stratum of the i^{th} period and $\bar{a}_{h'}$ be the corresponding average area per village in the h'^{th} stratum.

We define $\bar{a}_{h'} = \sum_{j}^{n_{h'}} \frac{a_{h'j}}{n_h}$, where $n_{h'}$ is the number of sampled flower growing villages in the h'^{th} stratum.

Let $\hat{A}_{h'}$ be the estimated total area under flower cultivation in the h'^{th} stratum.

Therefore,

$$\hat{A}_{h'} = N_{h'} \times \bar{a}_{h'}$$

An estimator of variance for $\hat{A}_{h'}$ is given by

$$\hat{V}(\hat{A}_{h'}) = N_{h'}^2 \left(\frac{1}{n_{h'}} - \frac{1}{N_{h'}} \right) s_{ah'}^2$$

where $$s_{ah'}^2 = \frac{\sum_{j=1}^{n_{h'}} (a_{h'j} - \bar{a}_{h'})^2}{n_{h'} - 1}$$

Estimated total area in Delhi State for a particular kind of flower in the i^{th} period is given by

$$\hat{A} = \sum_{h'=1}^{3} N_{h'} \bar{a}_{h'}$$

and the variance estimator is given by

$$\hat{V}(\hat{A}) = \sum_{h'=1}^{3} N_{h'}^2 \left(\frac{1}{n_{h'}} - \frac{1}{N_{h'}} \right) s_{ah'}^2$$

Estimates of area for all kind of flowers grown in Delhi for the other periods have been obtained on similar lines. The estimates for the three periods are added to get an estimate for the entire period.

4. RESULTS AND DISCUSSION

(i) Market Survey Approach

Estimates of the total market arrivals of loose flowers in Metric Tonnes (MT) as well as of cut flowers in lakh numbers along with their percentage standard errors in the three flower mandis of Delhi are presented in Table 1. A close perusal of Table 1 reveals that the estimate of the total market arrivals of loose flowers from the villages of Delhi in the flower mandis was 14570.91 MT with 2.51% standard error (SE). The corresponding figure for cut flowers was 670.69 lakhs with 1.53% SE.

Table 1. Estimate of total market arrivals of different kind of flowers in Delhi on the basis of market arrivals

Flowers	Loose (MT) Estimate		Cut (Lakh Nos.) Estimate	
Rose	1896.55	(1.02)	571.46	(1.74)
Marigold	1727.17	(3.62)		–
Guldawari	33.65	(*)	26.94	(7.75)
Rajnigandha	13.52	(*)	10.82	(6.11)
Jaffrey	8897.82	(3.85)	–	
Margaret (White/Yellow)	1899.03	(5.69)	–	
Gladiolus	–		14.05	(3.77)
Gerbera	–		–	
Orchid	–		–	
Carnation	–		–	
Tube Rose (Double)	–		2.76	(*)
Others	103.16	(2.61)	44.65	(2.63)
Total production	14570.91	(2.51)	670.69	(1.53)

Note: Figures within parentheses indicate corresponding percent standard errors.

* Estimates are based on small number of observations.

(ii) Village Survey Approach

Table 2 provides period-wise as well as stratum-wise estimated area (ha) of loose and cut flowers separately in the villages of Delhi. The area under loose flowers was estimated to be 2583.28 ha and that under cut flowers was 442.59 ha. Thus, during the survey period, 3025.87 ha area was estimated to be under floriculture in the flower growing villages of Delhi. Out of this, 85.37% area was under loose flowers while 14.63% was under cut flowers. However, the percentage standard errors of the estimates were on the higher side as these were based on small number of observations.

Period-wise and stratum-wise estimates of production of loose and cut flowers are presented in Table 3. A close perusal of Table 3 reveals that estimated production of loose flowers was to the tune of 1359.10 MT with 7.03% SE in Period-2 of stratum I; 668.50 MT with 7.30% SE in Period-1 of stratum II and 8277.30 MT with 4.31% SE in Period-2 of stratum III. Pooled over all the periods, these figures were

Table 2. Area (ha) under important flowers in the villages of Delhi

Period	Stratum I			Stratum II			Stratum III			Total		
	Loose	Cut	Total	Loose	Cut	Total	Loose	Cut	Total	Loose	Cut	Total
1	130.08	0	130.08	61.72	23.60	85.32	627.90	98.95	726.85	819.70	122.55	942.25
	(32.42)		(32.42)	(31.66)	(*)	(26.21)	(*)	(*)	(*)	(35.16)	(*)	(32.15)
2	123.05	0	123.05	34.16	75.43	109.59	796.74	85.26	882.00	953.95	160.69	1114.64
	(30.30)		(30.30)	(*)	(30.32)	(31.03)	(33.82)	(*)	(31.83)	(28.64)	(*)	(25.59)
3	136.01	0	136.01	0	84.47	84.47	673.62	74.88	748.50	809.63	159.35	968.98
	(31.98)		(31.98)		(8.94)	(8.94)	(34.60)	(*)	(32.45)	(29.28)	(*)	(25.48)
Total	389.14	0	389.14	95.88	183.50	279.38	2098.26	259.09	2357.35	2583.28	442.59	3025.87
	(18.28)		(18.28)	(33.24)	(14.40)	(14.82)	(21.72)	(*)	(20.23)	(17.90)	(32.13)	(15.99)

Table 3. Estimates of production of important flowers in the villages of Delhi

Period	Stratum I		Stratum II		Stratum III		Total	
	Loose (MT)	Cut (Lakh)	Loose (MT)	Cut (Lakh)	Loose (MT)	Cut (Lakh)	Loose (MT)	Cut (Lakh)
1	1325.38	0	668.55	36.82	4278.51	322.29	6272.447	359.12
	(13.74)		(7.30)	(6.10)	(17.28)	(*)	(7.47)	(1.98)
2	1359.09	0	527.33	91.42	8277.33	165.96	10163.747	257.38
	(7.03)		(*)	(15.22)	(4.31)	(*)	(3.73)	(5.84)
3	307.65	0	0	72.16	984.83	44.76	1292.49	116.92
	(17.32)			(30.54)	(3.79)	(*)	(10.62)	(17.97)
Total	2992.126	0	1159.88	200.39	13540.67	533.02	17728.68	733.41
	(6.80)		(4.95)	(9.19)	(8.18)	(*)	(4.09)	(3.54)

Note: Figures within parentheses indicate the corresponding percent standard errors (Table 2 & Table 3).

* Estimates are based on very few observations (Table 2 & Table 3).

2992.10 MT with 6.80% SE, 1159.80 MT with 4.95% SE and the highest 13540.70 MT with 8.18% SE for the three strata respectively. The pooled estimate of production of loose flowers was significantly higher in Period-2 i.e. of the order of 10163.70 MT with 3.73% SE followed by 6272.40 MT with 7.47% SE in Period-1 and 1292.50 MT with 10.62% SE in Period-3. The overall estimated production of loose flowers in the villages of Delhi was to the tune of 17728.70 MT with 4.09% SE. The period-wise estimated production of cut flowers were 359.11 lakhs with 1.98% SE, 257.38 lakhs with 5.84% SE and 116.92 lakhs with 17.97% SE respectively. The overall estimated production of cut flowers was 733.41 lakhs with 3.54% SE.

(iii) Comparative Study of both the approaches

A comparative study of the estimates of production of loose and cut flowers from the market arrivals survey approach and village survey approach is presented in Table 4. The results reveal that a maximum 91.4% of loose flowers produced in the flower growing villages of Delhi arrived for trading in the flower mandis of Delhi in Period-2 (peak period of flowers production) while 98.7% cut flowers produced in the villages of Delhi arrived for trading in the flower mandis of Delhi in Period-1. When pooled over the three periods, the percentage arrivals were to the tune of 82.2% and 91.5% respectively.

5. CONCLUSION

The study demonstrates the feasibility of estimating the production of flowers with a reasonable degree of precision. The methodology developed for estimating production of important flowers on the basis of market arrivals needs to be tested in some other representative areas before it can be recommended for large scale adoption.

Table 4. Estimates of production of important flowers based on Market Arrivals Survey Approach and Village Survey Approach

Period	Loose (MT)		Cut (Lakh Numbers)	
	Market Arrivals Survey Approach	Village Survey Approach	Market Arrivals Survey Approach	Village Survey Approach
1	4393.15 (70.0%)	6272.45	354.35 (98.7%)	359.11
2	9290.06 (91.4%)	10163.75	231.60 (90.0%)	257.38
3	887.67 (68.7%)	1292.49	84.74 (72.5%)	116.91
Total	14570.91 (82.2%)	17728.68	670.69 (91.5%)	733.41

Note: Figures within parentheses indicate the percentage of estimated flower production of flowers based on market arrivals approach to the estimated production of flowers based on village survey approach.

REFERENCES

Gupta, A.K., Jain, V.K., Narang, M.S., Tyagi, K.K. and Sud, U.C. (2004). Pilot sample survey to develop sampling methodology for estimation of area, production and productivity of important flowers on the basis of market arrivals. Project Report published by IASRI, New Delhi and funded by CSO, Ministry of Statistics & Programme Implementation, Government of India.

Murthy, M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Report of the National Statistical Commission (2001). National Statistical Commission, Government of India, Volume 1, 66-67.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press, AMES, Iowa, U.S.A. and Indian Society of Agricultural Statistics, New Delhi, India.

Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 267-276

Estimation of Small Area Proportions Under Unit Level Spatial Models

Hukum Chandra*
Indian Agricultural Statistics Research Institute, New Delhi

(Received: March 2009, Revised: September 2009, Accepted: October 2009)

SUMMARY

Generalized linear mixed models (GLMMs) containing fixed and random area-specific effects are often used for small area estimation (SAE) of discrete variables (McGilchrst 1994 and Rao 2003). In GLMM, the random area effects take account for between areas variation beyond that is explained by auxiliary variables included in the model. These area effects are generally assumed to be independent in SAE. However, in practice area effects are correlated with neighbouring areas and the correlation decays to zero as distance increases. In this paper we investigate SAE based on GLMM with spatially correlated random area effects where the neighbourhood structure is described by a contiguity matrix. We use simulation studies to compare the performances of empirical best predictor for small area proportions under such models with and without spatially correlated area effects. The simulation studies are based on two real data sets. Our empirical results show only marginal gains when spatial dependence between small areas is incorporated into the SAE model.

Key words : Cost function, Domain estimation, Optimal sample design, Probability of item response.

1. INTRODUCTION

The demand of reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of statistical methods from both the theoretical and empirical point of view. The traditional estimators (i.e. design-based direct estimators) for small area quantities based on survey data alone are often unstable because of sample size limitations. In this perspective the model-based methodologies allow for the construction of efficient estimators by borrowing the strength through use of a suitable small model. Such estimators are often referred as the indirect estimators, see Rao (2003).

Commonly used model for small area estimation (SAE) of discrete or non-normal data (e.g., binary or count data) is a generalized linear mixed model (GLMM) containing fixed and random effects, see Rao (2003) and McGilchrst (1994). The indirect estimators for small areas under GLMM are the EBLUP type estimators, often known as the empirical best predictors (EBP) for small area quantities, see Saei and Chambers (2003) and Manteiga *et al.* (2007). The mean squared error (MSE) estimation of the EBP is also described in Manteiga *et al.* (2007). These authors have also shown the performance of MSE estimator for the EBP. The area-specific random effects in GLMM take account for the between area dissimilarities beyond that is explained by auxiliary variables included in the fixed part of the model. Although it is customary to assume that these random area effects are independent, in practice most small area boundaries are arbitrary and there appears to be no good reason why population units just one side of such a boundary should not generally be correlated with population units just on the other side. In particular, it is often reasonable to assume that the effects of neighbouring areas (defined, for example, by a contiguity criterion) are correlated, with the correlation decaying to zero as the distance between these areas increases (Pratesi and Salvati 2008, 2009, and Petrucci and Salvati 2006). That is, small area models should allow for spatial correlation of area

* *Corresponding author* : Hukum Chandra
E-mail address : hchandra@iasri.res.in

random effects, See Cressie (1991). Such models allow efficient use of spatial auxiliary information (Chandra *et al.* 2007; Pratesi and Salvati 2008, 2009; Petrucci and Salvati 2006; and Singh *et al.* 2005).

In this paper we consider unit level generalized linear mixed models (Rao 2003, chapter 5 and Manteiga *et al.* 2007) and we extend the EBP for SAE (Saei and Chambers 2003 and Manteiga *et al.* 2007) to account for spatial correlation between the small areas where the neighbourhood structure is described by a contiguity matrix. We then use simulation studies to compare the performances of EBP under such models with and without spatially correlated area effects to examine the gains by incorporating the spatial dependence between the areas. The rest of the paper is organised as follows. In section 2 we review the EBP for SAE under a GLMM with spatially independent small area effects (Saei and Chambers 2003, and Manteiga *et al.* 2007) and discuss the extension of EBP for SAE to account for spatial dependence between the areas. We define the resulting estimator for the small area proportions and their mean squared error estimator. In section 3 we describe the design of our simulation studies and present empirical results and their discussion. In simulation studies we use two real data sets. The first data comes from consumer expenditure survey of the National Sample Survey Organisation (NSSO) for rural areas of state of the Uttar Pradesh in India and the second data from the Environmental Monitoring and Assessment Program (EMAP) survey of lakes in the north-east of the USA. It is noteworthy that two data are from two different real life surveys (i.e., social survey and environmental survey) and very different from each other. This clearly gives us an opportunity to examine the performance of proposed approach of SAE in two different life situations. Finally, in section 4 we provide some concluding remarks and identify further research prospects.

2. THE EMPIRICAL BEST PREDICTOR FOR THE SMALL AREAS

2.1 Models with Spatially Independent Random Area Effects

To start, let us consider a finite population U of size N and assumed to partitioned into D non-overlapping sub-groups (or small areas or small domains) U_i each of sizes N_i with $i = 1, ..., D$ such that $N = \sum_{i=1}^{D} N_i$. Let j and i respectively index the unit j within small area i, y_{ij} is the survey variable of interest and known for sampled units, $\mathbf{x}_{ij}$ is the vector of auxiliary variables (including the intercept), known for the whole population. Let s_i and r_i respectively denotes the sample (of size n_i) and non-sample (of size $N_i - n_i$) in small area i. We assume that y_{ij} is typically a binary variable. Let π_{ij} be the probability that a unit j in area i assumes value 1. Let u_i denotes the random area effect for the small area i and assumed to be normally distributed with mean zero and variance ϕ. We assume that u_i's are independent and $y_{ij}|u_i \sim Bin(1, \pi_{ij})$ with $E(y_{ij}|u_i) = \mu_{ij} = \pi_{ij}$ and $Var(y_{ij} \mid u_i) = \sigma_{ij} = \pi_{ij}(1 - \pi_{ij})$. A popular model for this type of data is the logistic linear mixed model of the form

$$\log it\,(\pi_{ij}) = \log\{\pi_{ij}/(1 - \pi_{ij})\} = \eta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i, \quad j = 1, ..., N_i;\ i = 1, ..., D \qquad (1)$$

where $\boldsymbol{\beta}$ $(p \times 1)$ is the vector of regression parameters. For estimation of unknown model parameters, it is common practice to express model (1) at the population level (Rao 2003, chapter 6). What follows next, we aggregate model (1) and write a population level version of this model as below.

Let $\mathbf{y}_U$ be the $N \times 1$ vector of response variable with elements $y_{ij}(j = 1, ..., N_i;\ i = 1, ..., D)$, $\mathbf{X}_U$ be the $N \times p$ known design matrix with rows $\mathbf{x}_{ij}$, $\mathbf{G}_U = diag(\mathbf{1}_{N_i};\ 1 \le i \le D)$ is the known matrix of order $N \times D$, $\mathbf{1}_k$ is a column vector of ones of size k, $\mathbf{u} = (u_1, ..., u_D)'$ and $\boldsymbol{\eta}_U$ denotes the $N \times 1$ vector of linear predictors η_{ij} given by (1). We define $\boldsymbol{\mu} = E(\mathbf{y}_U \mid \mathbf{u})$ the conditional mean function of $\mathbf{y}_U$ given $\mathbf{u}$ with elements μ_{ij} and $Var(\mathbf{y}_U \mid \mathbf{u}) = \text{diag}\ \{\sigma_{ij}\}$ the conditional covariance matrix. Let $g(\cdot)$ be a monotonic function, the link function (McCullagh and Nelder 1989, page 27), such that $g(\boldsymbol{\mu})$ can be expressed as the linear model of form

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta}_u = \mathbf{X}_U\boldsymbol{\beta} + \mathbf{G}_U\mathbf{u} \qquad (2)$$

The model (2) defines a GLMM, if $\mathbf{y}_U$ given $\boldsymbol{\mu}$ are independent and belong to the exponential family of distribution. Evidently, the vector of random area effects $\mathbf{u}$ has mean $\mathbf{0}$ and variance $\Omega(\delta) = \phi\mathbf{I}_D$, where $\mathbf{I}_D$ is the identity matrix of order D. For binomial response variable the link function $g(\cdot)$ is a logit function, see equation (1). We note that the logistic linear mixed model (1) is a special case of GLMM for logit link. The relationship among $\mathbf{y}_U$ and $\boldsymbol{\eta}_U$ is represented through a known function $h()$, defined by $E(\mathbf{y}_U \mid \mathbf{u}) = h(\boldsymbol{\eta}_U)$. Suppose that our interest is to predict the vector of linear parameters for small areas $\theta = \mathbf{a}_U\mathbf{y}_U$, where

$\mathbf{a}_U = diag\{\mathbf{a}_i', i = 1,..,D\}$ is a $D \times N$ matrix and $\mathbf{a}_i' = (a_{i1},...,a_{iN_i})$ is a vector of known elements. For example, when y_{ij} is a binary variable and our aim is to estimate proportion for small area i,

$$p_i = N_i^{-1}\sum_{j\in U_i} y_j = N_i^{-1}\left\{\sum_{j\in s_i} y_j + \sum_{j\in r_i} y_j\right\}$$

then $\mathbf{a}_i'$ (i = 1, ..., D) denote the population vector with value N_i^{-1} for each population unit in area i. The estimation of parameter of interest θ is carried out as follows.

Without loss of generality, we arrange the vector $\mathbf{y}_U$ so that its first n elements correspond to the sample units, and then partition $\mathbf{a}_U$, $\mathbf{y}_U$, $\boldsymbol{\eta}_U$, $\mathbf{X}_U$ and $\mathbf{G}_U$ according to sample and non-sample units as

$$\mathbf{a}_U, = \begin{bmatrix}\mathbf{a}_s\\ \mathbf{a}_r\end{bmatrix}, y_U = \begin{bmatrix}\mathbf{y}_s\\ \mathbf{y}_r\end{bmatrix}, \boldsymbol{\eta}_U = \begin{bmatrix}\boldsymbol{\eta}_s\\ \boldsymbol{\eta}_r\end{bmatrix}$$

$$\mathbf{X}_U = \begin{bmatrix}\boldsymbol{X}_s\\ \boldsymbol{X}_r\end{bmatrix} \text{ and } \boldsymbol{G}_U = \begin{bmatrix}\boldsymbol{G}_s\\ \boldsymbol{G}_r\end{bmatrix}$$

Here a subscript s denotes components defined by the n sample units while a subscript r is used to denote corresponding components defined by the remaining $N - n$ non-sample units. We then write $E(\mathbf{y}_s \mid \mathbf{u}) = h(\boldsymbol{\eta}_s)$ and $E(\mathbf{y}_r \mid \mathbf{u}) = h(\boldsymbol{\eta}_r)$. Typically, $h()$ is obtained as $g^{-1}()$. Using sample and non-sample deposition of various quantities, parameter of interest $\theta = \mathbf{a}_U\mathbf{y}_U$ can be expressed as

$$\theta = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r\mathbf{y}_r = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{G}_r\mathbf{u}) \qquad (3)$$

Here $\mathbf{y}_s$ the vector of sample values is known, whereas the second term of (3), which depends on the non-samples values $\mathbf{y}_r = h(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{G}_r\mathbf{u})$ is unknown and can be predicted by fitting model (3) for sample data. In our case $\mathbf{y}_s = \{y_{sij}\}$ denotes the vector of sample values of the binary survey variable which takes value 1 or 0. Similarly, $\mathbf{y}_r = \{y_{rij}\}$ represents the vector of non-samples values of the survey variable. It is obvious that the parameter of interest p_i for each small area can be obtained by using as prediction of each element $\{y_{rij}\}$. The problem then reduced to prediction of y_{rij} under model (2) which has two unknown components $\boldsymbol{\beta}$ and $\mathbf{u}$. A major difficulty in use of GLMM for SAE is the estimation of unknown model parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ since the likelihood function for GLMM often involves high dimensional integrals (computed by integrating a product of discrete and normal densities, which has no analytical solution) which are difficult to evaluate numerically. Although computationally attractive alternatives to the likelihood method are available, they can suffer of inconsistency (Jiang 1998).

For known $\boldsymbol{\Omega}(\boldsymbol{\delta})$, the values of $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated by Penalized Quasi Likelihood (PQL) under model (3) fitted for sample data (Breslow and Clayton 1993). The PQL approach is most popular estimation procedure for the GLMM and it constructs a linear approximation of the distribution of non-normal response variable and assumes the linearised dependent variable is approximately normal. This approach is reliably convergent but it has been noticed that the PQL tends to underestimate variance components as well as fixed effect coefficients (Breslow and Clayton 1993). McGilchrst (1994) introduced the idea to use BLUP to obtain approximate restricted maximum likelihood (REML) estimates for GLMMs. This link between BLUP and REML is illustrated in Harville (1977) for the normal case. For given $\boldsymbol{\Omega}(\boldsymbol{\delta})$, an iterative procedure to obtain maximum Penalized Quasi Likelihood (MPQL) estimate of $\boldsymbol{\beta}$ and $\mathbf{u}$ is described in Saei and Chambers (2003) as below.

1. Assign initial values to $\boldsymbol{\beta}$ and $\mathbf{u}$.
2. Update these values via

$$\begin{bmatrix}\boldsymbol{\beta}_{new}\\ \mathbf{u}_{new}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}_{old}\\ \mathbf{u}_{old}\end{bmatrix} + \mathbf{V}_s^{-1}\begin{bmatrix}\mathbf{X}_s'\\ \mathbf{G}_s'\end{bmatrix}\left(\frac{\partial l_1}{\partial \eta_s}\bigg|_{\boldsymbol{\beta}_{old},\mathbf{u}_{old}}\right) - \mathbf{V}_s^{-1}\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\Omega}^{-1}\mathbf{u}_{old}\end{bmatrix}$$

where $\mathbf{V}_s = \begin{bmatrix}\mathbf{X}_s'\\ \mathbf{G}_s'\end{bmatrix}\left(\frac{\partial^2 l_1}{\partial\eta_s\partial\eta_s'}\bigg|_{\boldsymbol{\beta}_{old},\mathbf{u}_{old}}\right)\begin{bmatrix}\mathbf{X}_s & \mathbf{G}_s\end{bmatrix} + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Omega}^{-1}\end{pmatrix}$

and $\frac{\partial l_1}{\partial \eta_s}$, $\frac{\partial^2 l_1}{\partial\eta_s\partial\eta_s'}$ are first and second derivatives of l_i with respect to η_s.

3. Return to step 2.

At convergence, this gives the best linear unbiased estimate (BLUE) for $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) for $\mathbf{u}$. Hence, using (3) we obtain the BLUP type estimator of θ (i.e, the MPQL estimate of θ).

In practice the variance components parameters defining the matrix $\boldsymbol{\Omega}(\delta)$ are unknown and have to be estimated from sample data. Following Saei and Chambers (2003) an iterative procedure that combines the MPQL estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with ML estimation of $\boldsymbol{\Omega}$ is:

1. Assign initial values to $\boldsymbol{\beta}$, $\mathbf{u}$ and δ
2. Update $\boldsymbol{\Omega}$
3. Update $\boldsymbol{\beta}$ and $\mathbf{u}$ using the iterative PL estimation procedure described in above para
4. Update $\boldsymbol{\eta}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{G}_s\mathbf{u}$
5. Update $\mathbf{B}_s = -(\partial^2 l_1 / \partial\boldsymbol{\eta}_s\partial\boldsymbol{\eta}_s')$
6. Update $\mathbf{T}_s = (\boldsymbol{\Omega}^{-1} + \mathbf{G}_s'\mathbf{B}_s\mathbf{G}_s)^{-1}$
7. Update $\boldsymbol{\delta}$
8. Return to step 2 and repeat the procedure until the values of the different parameters converges.

The corresponding iterative procedure used to obtain the REML estimators is exactly the same except that the role of $\mathbf{T}_s$ in step 6 of ML algorithm is replaced by the $\mathbf{T}_{22}$ submatrix of $\mathbf{T}$ defined in section 6.1 in Saei and Chambers (2003). In our empirical results reported in section 3, we adopted the REML algorithm for parameters estimation.

Using estimated value $\hat{\boldsymbol{\delta}}$ of the $\boldsymbol{\delta}$ leads to the empirical BLUE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and the empirical BLUP (EBLUP) $\hat{\mathbf{u}}$ for $\mathbf{u}$ and the EBLUP type estimator (i.e., empirical best predictor (EBP)) of θ is

$$\hat{\boldsymbol{\theta}} = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{G}_r\hat{\mathbf{u}}) \qquad (4)$$

Turning now to estimation of mean squared error of the EBLUP type predictor or EBP (4) we define

$$\mathbf{H}_r = \mathbf{H}(\hat{\eta}_r) = \partial h(\eta_r)/\partial\eta_r\big|_{\eta_r=\hat{\eta}_r}$$

and

$$\hat{\mathbf{B}}_s = \partial^2 l_1/\partial\eta_s\partial\eta_s'\big|_{\eta_s=\hat{\eta}_s}$$

the matrix of second derivatives of l_1 (the log-likelihood function l_1 defined by the vector $\mathbf{y}_s$ given $\mathbf{u}$) with respect to η_s at $\eta_s = \hat{\eta}_s$. Similarly, $\hat{\mathbf{B}}_r = \partial^2 l_1/\partial\eta_r\partial\eta_r'\big|_{\eta_r=\hat{\eta}_r}$.

We put $\mathbf{X}_r^* = \mathbf{a}_r\mathbf{H}_r\mathbf{X}_r$ and $\mathbf{G}_r^* = \mathbf{a}_r\mathbf{H}_r\mathbf{G}_r$. Then an approximate estimate of the mean squared error for the EBP (4) (see Saei and Chambers 2003; Manteiga *et al.* 2007) is

$$mse(\hat{\boldsymbol{\theta}}) = m_1(\hat{\boldsymbol{\delta}}) + m_2(\hat{\boldsymbol{\delta}}) + 2m_3(\hat{\boldsymbol{\delta}}) + m_4(\hat{\boldsymbol{\delta}}) \qquad (5)$$

where

$$m_1(\hat{\boldsymbol{\delta}}) = \mathbf{G}_r^*\hat{\mathbf{T}}_s\mathbf{G}_r^{*\prime} \text{ with } \hat{\mathbf{T}}_s = (\hat{\boldsymbol{\Omega}}^{-1} + \mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s)^{-1}$$

$$m_2(\hat{\boldsymbol{\delta}}) = \mathbf{C}_r\left(\mathbf{X}_s'\hat{\mathbf{B}}_s\mathbf{X}_s - \mathbf{X}_s'\hat{\mathbf{B}}_s\mathbf{G}_s\hat{\mathbf{T}}_s\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{X}_s\right)^{-1}\mathbf{C}_r'$$

with $\mathbf{C}_r = \left\{\mathbf{X}_r^* - \mathbf{G}_r^*\hat{\mathbf{T}}_s\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{X}_s\right\}$

$$m_3(\hat{\boldsymbol{\delta}}) = \left\{tr\left((\hat{\nabla}_t\hat{\Sigma}_s\hat{\nabla}_k')v(\hat{\boldsymbol{\delta}})\right)\right\}$$

with $\hat{\Sigma}_s = \mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s + \phi\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s\mathbf{G}_s'\hat{\mathbf{B}}_s\mathbf{G}_s$

and $m_4(\hat{\boldsymbol{\delta}}) = \mathbf{a}_r\hat{\mathbf{B}}_r\mathbf{a}_r'$

Let $\varsigma = \mathbf{G}_r^*\hat{\mathbf{T}}_s$ and $\mathbf{G}_{rt}^*$ be the t^{th} row of the matrix $\mathbf{G}_r^*$, then $\hat{\nabla}_t = \partial(\varsigma_t)/\partial\delta\big|_{\delta=\hat{\delta}} = \hat{\phi}^{-2}\mathbf{G}_{rt}^*\hat{\mathbf{T}}_s\hat{\mathbf{T}}_s$. Here $v(\hat{\boldsymbol{\delta}})$ is the asymptotic covariance matrix of estimates of variance components $\hat{\boldsymbol{\delta}}$ which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\boldsymbol{\delta}}$, see Saei and Chambers (2003). Manteiga *et al.* (2007) described the EBP (4) for small area proportion estimation and estimates of their mean squared error. They have studied the empirical performance of MSE estimator (5). However, they have not taken account of spatial dependence between the small areas, which is main objective of this article.

2.2 Models with Spatial Dependence Random Area Effects

In many situations the physical location of the small areas is so relevant that the assumption of spatial independence of the small area models becomes questionable. That is, small area data exhibit a spatial structure and therefore use of spatial models becomes essential. Spatial dependency is the extent to which the value of an attribute in one location depends on the value of the attribute in nearby locations or small areas. Recently the problem has been addressed by introducing a common autocorrelation parameter among small areas extending the linear mixed model through

the Simultaneously Autoregressive (SAR) process (Pratesi and Salvati 2008, 2009; Singh *et al.* 2005; Petrucci and Salvati 2006, Chandra *et al.* 2007). These extensions of small area models (e.g., area level models described in Pratesi and Salvati 2008, 2009; Singh *et al.* 2005 and unit level models discussed in Chandra *et al.* 2007) to spatial small area models are the special case of linear mixed models. The focus here is on the introduction of the SAR process in the generalised linear mixed models (GLMMs) where the vector of random area effects $\mathbf{v} = (v_i)$ satisfies

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u} \qquad (6)$$

where ρ is spatial autoregressive coefficient which determines the degree of spatial dependency of the model, $\mathbf{W}$ is proximity or contiguous matrix of order D. This matrix is symmetric and encapsulates the relative spatial arrangement (i.e. neighbourhood structure) of the small areas whereas ρ defines the strength of the spatial relationship among the random effects associated with neighbouring areas. The simplest way to define such a matrix is as simple contiguity: the elements of $\mathbf{W} = \{w_{jk}\}$ take non-zero values only for those pairs of areas that are contiguous to each other. Generally, for ease interpretation, the general spatial weight matrix is defined in row-standardized form; in this case ρ is called spatial autocorrelation parameter (Banerjee *et al.* 2004). In row-standardised form this becomes

$$w_{jk} = \begin{cases} d_j^{-1} & \text{if } j \text{ and } k \text{ are contiguous} \\ 0 & \text{otherwise} \end{cases}$$

where d_j is the total number of areas that share an edge with area j (including area j itself). Contiguity is the simplest but not necessarily the best specification of a spatial interaction matrix. It may be more informative to express this interaction in a more detailed way, e.g. as some function of the length of shared border between neighbouring areas or as a function of the distance between certain locations in each area. Furthermore, the concept of neighbours of a particular area can be defined not just in terms of contiguous areas, but also in terms of all areas within a certain radius of the area of interest. In the empirical evaluations reported later in this paper, however, we used simple contiguity (row-standardized) to define the spatial interaction between different areas. Here

$E(\mathbf{u}) = \mathbf{0}$ and $\text{Var}(\mathbf{u}) = \phi\mathbf{I}_D$

$E(\mathbf{v}) = \mathbf{0}$ and $Var(\mathbf{v}) = \mathbf{\Omega}(\phi\ \rho)$

$= \phi[(\mathbf{I}_D - \rho\mathbf{W})\ (\mathbf{I}_D - \rho\mathbf{W}^T)^{-1}$

where $\mathbf{\Omega}(\phi, \rho) = \mathbf{\Omega}(\delta)$ is the SAR dispersion matrix. To define the EBP under spatially correlated area effects or spatial-EBP (denoted by SEBP), the linear predictor η_U is expressed as

$$\eta_U = \mathbf{X}_U\beta + \mathbf{G}_U\mathbf{v} \qquad (7)$$

where the vector $\mathbf{v}$ is an D-vector of spatially correlated area effects that satisfies SAR model (7). For estimation of unknown model parameters we adopt an iterative procedure similar to one described earlier in this section. However, variance components are now $\delta = (\phi, \rho)$ and $\hat{\mathbf{u}}$ is replaced by $\hat{\mathbf{v}}$. This leads to the spatial EBP of θ (i.e., SEBP) as

$$\hat{\theta} = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r\hat{\beta} + \mathbf{G}_r\hat{\mathbf{v}}). \qquad (8)$$

The MSE of the SEBP (8) are followed from (5) using the variance components $\delta = (\phi, \rho)$ and $\hat{\mathbf{v}}$ in place of $\hat{\mathbf{u}}$.

3. EMPIRICAL EVALUATIONS

In this section we present simulation studies to contrast the performance of the two SAE methods: (i) the empirical best predictor (4) under GLMM with spatially independent area effects, denoted by EBP (see Saei and Chambers 2003 and Manteiga *et al.* 2007) and (ii) the proposed empirical best predictor (8) under GLMM with spatially dependent area effects, denoted by SEBP. The empirical evaluations are based on design-based simulation studies using two real data sets. This evaluates the performance of these methods in the context of real population and realistic sampling methods. The two data sets used in the design-based simulations are from two different types of surveys and are very different to each other. They are

i) The National Sample Survey Organisation (NSSO) Consumer Expenditure Survey: The basis data comes from the survey that underpins the empirical results reported in Sud *et al.* (2008). I used the 61st round survey of NSSO (July 2004-June 2005), the quinquennial series of consumer expenditure survey for rural areas of the state of Uttar Pradesh in India. From this survey, I consider a sample of 307 household from $D = 10$ selected districts (districts are the small area of interest) of state of Uttar Pradesh. The selected districts are all from eastern region of the state so that reasonable neighbourhood can be constructed. This sample of 307 households was bootstrapped to create a realistic population of $N = 76{,}062$

households by sampling with replacement with probability proportional to a household's sample weight. However, in doing so we divided the survey weights in original sample data by 10 to reduce the overall population size, keeping in mind the computation intensity. Note that this does not change the original layout of the survey data except the population sizes used in SAE. A total of $K = 1000$ independent stratified random samples were then drawn from this bootstrap population, with total sample size equal to that of the original sample and with districts defining the strata. Sample sizes within districts were the same as in the original sample (varies from 16 to 45). The Y variable of interest takes value 1 if the Household's Monthly Per Capita Expenditure (MPCE) is less than median MPCE of these 10 districts and 0 otherwise. We used the household holding (hectares) of the household as the auxiliary variable. The aim is here to estimate the proportion of households below median MPCE class in each district. The results from this simulation are presented in Table 1.

ii) *The Environmental Monitoring and Assessment Program (EMAP) Survey*: The data consist of a sample of 349 plots in the lakes from the North-eastern states of the U.S. The survey is based on a population of 21,028 lakes from which 334 lakes were surveyed, some of which were visited, in different plots, several times during the study period (1991-1995). The total number of measurements is 551. The 349 plots are the result of their grouping by lake and by 6-digit Hydrologic Unit Codes (HUC). Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University supplied this data set, developed by EMAP. The HUCs are considered as regions of interest. These areas were having sample sizes as 1 only. Therefore we decided to combine these regions with their similar regions. Consequently, we left with 23 small areas. Sample sizes in these 23 areas vary from 2 to 45. We generated a population of size $N = 21,028$ by sampling N times with replacement from the above sample of 349 plots (units) and with probability proportional to a unit's sample weight; and then $K = 1000$ independently stratified random samples of the same size as the original sample were selected from this (*fixed*) simulated population. HUC sample sizes were also fixed to be the same as in the original sample. The variable of interest y

Table 1. District-wise performance measures for the NSSO data. Districts are arranged in order of increasing population size.

Districts	Relative Bias, %		Relative RMSE, %		Coverage rates		Mean squared error			
							EBP		SEBP	
	EBP	SEBP	EBP	SEBP	EBP	SEBP	True	Estimated	True	Estimated
1	90.24	60.68	107.95	78.59	0.67	0.90	0.020	0.008	0.015	0.008
2	−4.42	1.94	21.18	19.59	0.97	0.98	0.009	0.011	0.008	0.011
3	−6.06	−3.07	16.63	14.65	0.96	0.97	0.009	0.011	0.007	0.009
4	1.28	0.04	15.37	15.59	0.98	0.97	0.005	0.008	0.005	0.007
5	−0.35	0.91	18.18	18.07	0.97	0.96	0.007	0.010	0.007	0.008
6	4.92	4.80	18.85	19.30	0.98	0.98	0.005	0.009	0.006	0.007
7	–6.70	−6.48	11.25	10.99	0.93	0.91	0.007	0.005	0.007	0.005
8	−1.26	−0.91	12.41	13.68	0.98	0.96	0.004	0.005	0.005	0.005
9	13.10	16.11	29.52	31.24	0.96	0.92	0.005	0.004	0.004	0.004
10	−2.94	−4.57	9.03	9.63	0.96	0.94	0.004	0.004	0.004	0.004
Average	8.78	6.94	26.04	23.13	0.94	0.95	0.0075	0.0075	0.0068	0.0068

takes value 1 if Acid Neutralizing Capacity (ANC) - an indicator of the acidification risk of water bodies- in water resource surveys is less than 500 and 0 otherwise. The elevation of the lake is the auxiliary variable. We are interested in estimation of small area proportion of plots for which ANC less than 500. Results from this simulation experiment are set out in Table 2.

The performance of different small area estimators were evaluated with respect to three basic criteria –the relative bias and the relative root mean squared error both expressed as percentages of estimates of the small area proportions and the coverage rate of nominal 95 per cent confidence intervals for these proportions. In the evaluation of coverage performances intervals are

Table 2. Region-wise performance measures for the EMAP data. Regions are arranged in order of increasing population size.

Regions	Relative Bias, %		Relative RMSE,%		Coverage rates		Mean squared error			
							EBP		SEBP	
	EBP	SEBP	EBP	SEBP	EBP	SEBP	True	Estimated	True	Estimated
1	−8.13	−9.16	8.27	9.47	0.99	0.99	0.0068	0.0199	0.0172	0.0168
2	−1.72	−0.66	1.82	0.79	0.99	0.99	0.0003	0.0016	0.0096	0.0137
3	−14.08	−18.18	14.15	18.65	0.99	0.99	0.0200	0.0290	0.0152	0.0088
4	−4.23	−3.86	4.28	3.95	1.00	0.99	0.0018	0.0026	0.0143	0.0280
5	—	—	—	—	—	—	0.0639	0.0460	0.0201	0.0124
6	−1.06	−2.06	1.10	2.20	1.00	1.00	0.0001	0.0003	0.0087	0.0061
7	2.41	2.25	15.83	15.42	0.87	0.87	0.0143	0.0193	0.0129	0.0119
8	6.43	0.29	75.18	71.60	0.91	0.87	0.0442	0.0430	0.0042	0.0062
9	—	—	—	—	—	—	0.0133	0.0185	0.0083	0.0059
10	0.50	1.10	18.06	17.85	0.94	0.94	0.0131	0.0142	0.0074	0.0033
11	−2.40	−0.81	6.16	5.71	1.00	0.94	0.0033	0.0047	0.0054	0.0057
12	10.84	15.66	28.92	32.03	0.98	0.96	0.0161	0.0236	0.0083	0.0107
13	36.37	28.01	73.68	68.63	0.97	0.97	0.0263	0.0258	0.0093	0.0069
14	−0.35	−0.62	6.45	6.42	0.93	0.94	0.0031	0.0031	0.0040	0.0027
15	4.53	2.96	23.48	22.97	0.95	0.96	0.0071	0.0076	0.0075	0.0099
16	−4.65	−5.03	4.71	5.12	1.00	1.00	0.0022	0.0032	0.0034	0.0026
17	−2.64	−2.60	2.69	2.66	1.00	1.00	0.0007	0.0011	0.0057	0.0058
18	3.48	8.45	24.27	26.52	0.90	0.86	0.0180	0.0134	0.0038	0.0048
19	0.44	0.14	5.91	5.87	0.97	0.97	0.0022	0.0027	0.0052	0.0055
20	2.21	3.69	27.50	27.66	0.87	0.87	0.0163	0.0106	0.0045	0.0051
21	−0.72	−0.55	5.20	5.10	0.96	0.96	0.0021	0.0027	0.0035	0.0045
22	−2.17	−1.39	11.35	11.08	0.93	0.92	0.0087	0.0076	0.0041	0.0048
23	0.52	−0.45	8.43	8.39	0.97	0.97	0.0030	0.0038	0.0040	0.0044
Average	1.22	0.82	17.50	17.53	0.96	0.95	0.0125	0.0132	0.0081	0.0081

defined by the estimate of small area proportion plus or minus twice their standard error. The relative bias was measured by *%AvRB*, where

$$\%AvRB = \underset{i}{mean}\left|\left\{M_i^{-1}\left(K^{-1}\sum_{k=1}^{K}\hat{m}_{ik}\right)-1\right\}\right|\times 100$$

with average over the small areas. The root mean squared error was measured by *%AvRRMSE*, where

$$\%AvRRMSE = \underset{i}{mean}\left[M_i^{-1}\left\{\sqrt{K^{-1}\sum_{k=1}^{K}(\hat{m}_{ik}-m_{ik})^2}\right\}\right]\times 100$$

Coverage performance for prediction intervals was measured by *%AvCR*, where

$$\%AvCR = \underset{i}{mean}\left\{K^{-1}\sum_{k=1}^{K} I\left(\left|\hat{m}_{ik}-m_{ik}\right|\le 2\hat{M}_{ik}^{1/2}\right)\right\}\times 100$$

Note that the subscript k here indexes the K simulations, with m_{ik} denoting the value of the small area i mean in simulation k (this is a fixed population value in the design-based simulations considered here), and $\hat{m}_{ik}$, $\hat{M}_{ik}$ denoting the area i estimated value and corresponding estimated MSE in simulation k. The actual area i mean value (averaged over the simulations) is denoted by $M_i = K^{-1}\sum_{k=1}^{K} m_{ik}$.

In Table 1 we report the relative bias (RB) and relative root mean squared error (RRMSE), coverage rates (CR) for nominal 95% intervals for small area proportions and the mean squared error (both true and estimated) of small area proportion estimates for two methods of small area estimation (i.e., EBP and SEBP) based on repeated sampling from the simulated NSSO population. Analogous results for repeated sampling from the simulated EMAP population are presented in Table 2.

The results in Table 1 show that the average relative bias (*%AvRB*) and average relative root mean squared error (*%AvRRMSE*) of the proposed estimator (i.e., SEBP) is smaller than the EBP. Looking at the region specific results in Table 1 we note that relative biases in 7 out of 10 and relative root mean squared errors in 5 out of 10 regions are smaller for SEBP than the EBP. It seems advantageous to include spatial effects in EBP, with a marginal gain. The average coverage rates (*%AvCR*) are slightly underestimated if spatial effects are ignored in small area models, which again show an advantage of including spatial structure. Table 1 clearly shows a consistently good performance of MSE estimate (5) for both SEBP and EBP estimators. We further note that the average value of true MSE of small area proportions for SEBP is slightly lower than the EBP. In 8 out of 10 districts the values of true MSE of SEBP are either smaller or equal to that of true MSE of EBP. This again indicates the gain in small area estimation by incorporating the spatial dependence between the areas.

In Table 2 we noticed that results for regions 5 and 9 are missing. In these two regions true small area proportions (i.e. population proportions for small areas) are zero. Consequently, we could not calculate the relative performance measures (i.e. relative bias and relative root mean square error) since denominators were zero in these cases. The average results in Table 2 therefore are based on the average of remaining 21 regions. In terms of relative biases and relative RMSEs the conclusions from Table 2 are almost identical to results of NSSO data reported in Table 1. In contrast, ignoring the spatial structure in EMAP data leads to overestimation of coverage rates. From the results in Table 2 too we observed only marginal gain in SAE by incorporating spatial effects in estimation. Overall gain by incorporating spatial effects (when neighbourhood structure is described by a contiguity matrix) in small models for binary variable is marginal. The results in Table 2 show that the MSE estimator (5) performs very well for the EMAP data too. The comparative performance of this estimator for the EBP and SEBP is identical to that of NSSO data.

Overall empirical results reveal that MSE estimator performs well. Only a marginal gain can be achieved by including spatial structure in small area estimation of proportions. It is noteworthy that relatively the gains in SEBP are better for NSSO data than the EMAP data. A critical examination of original sample data reflects that the NSSO data has marginally higher degree of spatial dependence between areas than

the EMAP data. The relative gains in NSSO data are therefore more evident than in the EMAP data.

4. CONCLUDING REMARKS

This paper describes SAE of proportions under the GLMM with spatially correlated random area effects where the neighbourhood structure is defined by a contiguity matrix. The empirical results, based on two real data indicate that the gains from inclusion of spatial structure in SAE do not appear to be large. Note that the spatial models considered in this paper are based on neighbourhoods defined by contiguous areas. It is easy to see that this is just one way of introducing spatial dependence between area effects, and several other options remain to be investigated, e.g. geographical weighted regression etc.

There are many issues that still need to be explored in the context of using unit level models with spatially distributed area effects in SAE of discrete data. The most important of these is identification of situations where inclusion of spatial information does have an impact, and the most appropriate way of then including this spatial information in the small area modelling process. An important practical issue in this regard relates to the computational burden in fitting spatial models to survey data. With the large data sets common in survey applications it can be extremely difficult to fit spatial models without access to high-end computational facilities. Although spatial information is becoming increasingly available in environmental, epidemiological and economic applications, there has been comparatively little work carried out on how to efficiently use this information. A further issue relates to the link between the survey data and the spatial information (Chandra *et al.* 2007).

The development in this paper assumes that the sampling method used is uninformative for the population values of Y given the corresponding values of the auxiliary variables and knowledge of the area affiliations of the population units. As a consequence, same model applies at both sample and population level. However, many often survey data comes from complex sampling designs (e.g., NSSO data illustrated in section 3). There are approaches to incorporate the complex sampling designs for SAE of continuous data (e.g., Pseudo EBLUP under a linear mixed model). However, to my knowledge no such parallel work has been reported for estimation with discrete data. This can be a future research work.

ACKNOWLEDGEMENTS

The constructive and insightful comments from referee are gratefully acknowledged. They resulted in the revised version of the article representing a considerable improvement on the original. The author gratefully acknowledges Dr Nicola Salvati for his help in writing R code for empirical evalutions.

REFERENCES

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall, New York.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *J. Amer. Statist. Assoc.*, **88**, 9-25.

Chandra, H., Salvati, N. and Chambers, R. (2007). Small area estimation for spatially correlated populations - A comparison of direct and indirect model-based methods. *Stat. Trans.,* **8**, 887-906.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada,* 93-105.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 720-729.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, **72**, 320-338.

Manteiga, G.W., Lombardìa, M.J., Molina, I., Morales, D. and Santamarìa, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Comput. Statist. Data Anal.*, **51**, 2720-2733.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall, New York.

McGilchrst, C.A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc.*, **B56**, 61-69.

Pratesi, M. and Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. *J. Off. Statist.*, **25 (1)**, 37–53.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator with autoregressive random area effects. *Statist. Methods Appl.*, **17**, 113-141.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *J. Ag. Biol. Environ. Stat.*, **11**, 169-182.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley, New York.

Saei, A. and Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. *Methodology Working Paper-M03/15*, University of Southampton, United Kingdom.

Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**, 183-195.

Sud U.C., Bathla, H.V.L., Chandra, H. and Singh, J. (2008). Small area estimation - Some application to National Sample Survey data. *Proceedings of the National Seminar on NSS 62th Round Survey Results*, New Delhi, September, 2008.

Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 277-282

Further Results on Diagonal Systematic Sampling Scheme for Finite Populations

J. Subramani*
S. Veerasamy Chettiar College of Engineering and Technology, Tamilnadu

(Received: January 2009, Revised: September 2009, Accepted: October 2009)

SUMMARY

A generalization of diagonal systematic sampling scheme for estimation of a finite population mean is introduced. The assumption of $n \le k$ is relaxed here and hence the proposed method is applicable for all values of n provided $N = kn$. The relative performance of proposed diagonal systematic sample mean along with those of the simple random and systematic sample means is assessed for a natural population.

Key words : Diagonal systematic sampling, Systematic sampling, Natural population, Linear trend, Trend free sampling.

1. INTRODUCTION

Consider a finite population $U = \{U_1, U_2, ..., U_N\}$ of N distinct and identifiable units. Let Y be the study variable and take a real value Y_i measured on U_i, $i = 1, 2,3,...,N$ giving a vector $Y = (Y_1, Y_2,...,Y_N)$. The problem is to estimate the population mean $\bar{Y} = \sum_{i=1}^{N} Y_i$ on the basis of a random sample selected from the population U. Any ordered sequence $S = \{u_1, u_2,..., u_n\} = \{U_{i_1}, U_{i_2},..., U_{i_n}\}$, $1 \le i_l \le N$, and $1 \le l \le n$ is called a random sample of size n.

In the past, several sampling schemes were suggested for selecting a random sample of size n from a finite population of size N. If there is a linear trend among the population units then the systematic sampling is recommended for selecting a sample of size n, which gives the best estimator compared to simple random sampling. Recently Subramani (2000) has introduced the diagonal systematic sampling with fixed sample size, which uses the knowledge of the labels of the population units to provide an unbiased estimator of the population mean. The explicit expressions for the diagonal systematic sample mean and its variance are obtained for certain hypothetical populations. Further, the relative performance of diagonal systematic sampling; simple random sampling and systematic sampling schemes are assessed for certain natural populations. As a result, it has been shown that the diagonal systematic sampling performs better than the simple random sampling and the linear systematic sampling for estimating the finite population mean in the presence of linear trend. For more details the readers are referred to Subramani (2000) and the references cited there in.

It is to be noted here that for the case of simple random sampling there is no restriction either on the sample size n or on the population size N. Similarly in the case of linear systematic sampling scheme it is required to have $N = kn$ and there is no restriction on the sample size n, whether $n \le k$ or $n \ge k$. However, the drawback in diagonal systematic sampling of Subramani (2000) is that it is applicable only when the sample size is $n \le k$, where k is the number of distinct samples and $N = kn$. It has motivated the present study and consequently the generalization of diagonal

* *Corresponding author* : J. Subramani
E-mail address : drjsubramani@yahoo.co.in

systematic sampling scheme is introduced. The proposed sampling scheme requires only $N = kn$ as in the case of linear systematic sampling, which leads to a real comparison between the diagonal systematic sampling and linear systematic sampling schemes. The explicit expressions for the generalized diagonal systematic sample mean and its variance are presented for the hypothetical population with a perfect linear trend among the population values. The relative performance of the proposed diagonal systematic sample mean along with those of the simple random mean and systematic sample mean is assessed for a natural population.

2. GENERALIZED DIAGONAL SYSTEMATIC SAMPLING SCHEME

For the sake of simplicity and for the benefit of the readers, the steps involved in selecting a diagonal systematic sample of size n from a population of size $N = kn$ are reproduced here. Let $N = kn$ where $n \leq k$, be the population size. The population units $U_1, U_2,..., U_N$ are arranged in a $n \times k$ matrix **M** (say) and the j-th row of **M** is denoted by R_j, $j = 1, 2,...n$. The elements of R_j are $\{U_{(j-1)k+i}, i = 1, 2,... k\}$. The diagonal systematic sampling scheme consists of drawing n units from the matrix **M** systematically such that the selected n units are the diagonal elements or broken diagonal elements of the matrix **M**. Hence the selected units are from different rows and from different columns.

The diagonal systematic sampling scheme discussed above is applicable only when $N = kn$ and $n \leq k$. If the sample size $n > k$ then the diagonal systematic sampling discussed above cannot be used to estimate the finite population mean. It is the drawback compared to linear systematic sampling, which is valid for any value of n provided the population size $N = kn$. Hence, an attempt has been made in this paper to relax the condition $n \leq k$ and the resulting sampling scheme is called generalized diagonal systematic sampling. Further, explicit computable expressions for the sample mean and the variance are provided for the hypothetical population with a linear trend among the population values. Here, the sampled units need not be from different columns but are from different rows of **M**. If $n \leq k$ then the generalized diagonal systematic sampling is reduced to the usual diagonal systematic sampling. Hence, we concentrate only for the case where $n > k$.

The steps involved in the generalized diagonal systematic sampling scheme for selecting a random sample of size n are given below:

Let $N = kn$ where $n > k$, be the population size. Since the sample size $n > k$, one can write the sample size $n = pk + m$, where p is any positive integer greater than one.

Step 1. Arrange the N population units $U_1, U_2,...,U_N$ in a $n \times k$ matrix **M** (say).

Step 2. Select a random number r such that $1 \leq r \leq k$.

Step 3. Select downward the diagonal elements from r, in the direction of left to right until reaching the right most column of **M**.

Step 4. Once the right most column of **M** is reached then select the first element in the next row.

Step 5. Repeat the Steps 2 through 4 until selecting a random sample of n elements.

For example to select a generalized diagonal systematic sample of size 5 from a population of size 15 units, consider the following:

$$\mathbf{M} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{bmatrix}$$

As the result, (1,5,9,10,14), (2,6,7,11,15) and (3,4,8,12,13) are the generalized diagonal systematic samples of size 5 with the random starts 1, 2 and 3 respectively. It is to be noted that the first order and second order inclusion probabilities are obtained as given below:

$$\pi_i = \frac{1}{k}, \text{ for } i = 1, 2, 3, \ldots N$$

and $$\pi_{ij} = \begin{cases} \frac{1}{k} & \text{if } i \text{ and } j \text{ are in the same diagonal} \\ 0 & \text{otherwise} \end{cases}$$

The first order inclusion probabilities are the same for both the systematic sampling and the diagonal systematic sampling schemes but the difference is on the second order inclusion probabilities. The two units in the same column will get the same probability $\frac{1}{k}$ in the case of systematic sampling whereas the two units in the same diagonal or broken diagonal will get the

same probability $\frac{1}{k}$ in the case of diagonal systematic sampling and zero for other pair of units. For more details on diagonal systematic sampling and its performance one may refer to Subramani (2000).

3. RELATIVE PERFORMANCE OF GENERALIZED DIAGONAL SYSTEMATIC SAMPLING IN THE PRESENCE OF PERFECT LINEAR TREND

It is well known that the linear systematic sampling is preferred over the simple random sampling whenever there is a linear trend among the population values. For a detailed discussion on estimation of finite population mean one may refer to Cochran (1977), Fountain and Pathak (1989) and the references cited therein. Further, Subramani (2000) has shown that the diagonal systematic sampling is more efficient than the simple random sampling and systematic sampling provided (i) $n \le k$, (ii) $N = kn$ and (iii) There exists a linear trend among the population values. In this section we have compared the relative efficiency of the generalized diagonal systematic sampling scheme with that of simple random sampling and systematic sampling for estimating the mean of finite populations with linear trend among the population values.

3.1 Population with Linear Trend

In this hypothetical population, the values of N population units are in arithmetic progression. That is,

$$Y_i = a + ib \ , \ i = 1, 2,..., N \qquad (3.1)$$

After a little algebra one may obtain the generalized diagonal systematic sample mean for the above hypothetical population with the random start r as given below:

$$\bar{y}_{gdsy} = \begin{cases} a + \frac{1}{n}\left[\frac{pk(pk^2+1)+m(m-1)(k+1)}{2} + m(pk^2+r)\right]b & \text{if} \quad r \le k-m+1 \\ a + \frac{1}{n}\left[\frac{[pk(pk^2+1)+km(m-1) + k(k+1)+(k-m)(k-m+1)]}{2} + mpk^2-(k-m)r\right]b & \text{if} \quad r > k-m+1 \end{cases}$$

where $n = pk + m$ and $p \ge 0$

For the above population with a linear trend, the variances of the simple random sample mean $V(\bar{y}_r)$ systematic sample mean $V(\bar{y}_{sy})$, diagonal systematic sample mean $V(\bar{y}_{dsy})$ and generalized diagonal systematic sample mean $V(\bar{y}_{gdsy})$ are obtained as given below:

$$V(\bar{y}_r) = \frac{(k-1)(N+1)b^2}{12} \qquad (3.2)$$

$$V(\bar{y}_{sy}) = \frac{(k-1)(k+1)b^2}{12} \qquad (3.3)$$

$$V(\bar{y}_{dsy}) = \frac{(k-n)[n(k-n)+2]b^2}{12n} \qquad (3.4)$$

$$V(\bar{y}_{gdsy}) = \frac{m(k-m)[m(k-m)+2]b^2}{12n^2} \qquad (3.5)$$

where $n = pk + m$ and $p \ge 0$

The derivation of the variance of generalized diagonal systematic sample mean is given in the Appendix. Further, one can easily show that diagonal systematic sampling is a particular case of the generalized diagonal systematic sampling. When $m = n$ in the variance expression given in (3.5), it reduces to the variance expression given in (3.4). Hence, we can hereafter consider only the generalized diagonal systematic sampling for assessing the relative performance with that of simple random sampling and systematic sampling schemes. By comparing the various variance expressions given above, one can easily show that generalized diagonal systematic sampling is more efficient than the other sampling schemes. In fact

$$V(\bar{y}_{gdsy}) \le V(\bar{y}_{sy}) \le V(\bar{y}_r) \qquad (3.6)$$

and the equality sign occurs only when $n = 1$, which is the trivial case where $k = N$.

Remark 3.1. If $n = k$ or multiples of k then $V(\bar{y}_{gdsy}) = 0$. In this case the generalized diagonal systematic sampling becomes a completely trend free sampling (See Mukerjee and Sengupta 1990).

4. RELATIVE PERFORMANCE OF GENERALIZED DIAGONAL SYSTEMATIC SAMPLING FOR A NATURAL POPULATION

It has been shown in Section 3 that diagonal systematic sampling performs well, compared to simple random and systematic sampling schemes whenever there exists a perfect linear trend among the population units. However, this is an unrealistic assumption in real life situations. Consequently an attempt has been made to study the efficiency of generalized diagonal systematic sampling for a natural population considered by Subramani (2004). The data were collected for assessing the process capability of a manufacturing process from an auto ancillary manufacturing unit located in Tamil Nadu. The data were pertaining to the measurements taken continuously during turning operation performed on the component, namely Torsion bar in Frontier CNC Lathe Machine. The data were collected for estimating the mean value of the outer diameter of the Torsion bar, one of the key components in integrated power steering system. The measurements were taken continuously for the first 50 components produced in a shift. The 50 measurements based on the order of the production are given below. However, we have taken only the first 48 measurements for assessing the relative performance of the various sampling schemes, which give 8 combinations for (k, n) whereas if we take 50 measurements which lead to 4 combinations only.

Table 4.1. Data of outer Diameter of Torsion Bar (Spec. 9065 ± 25)

9050	9052	9050	9052	9052	9056	9056	9054	9056	9058
9054	9054	9060	9058	9060	9058	9056	9058	9058	9060
9062	9064	9062	9064	9066	9070	9068	9072	9072	9070
9072	9070	9070	9072	9074	9076	9078	9076	9076	9078
9078	9078	9082	9080	9082	9080	9082	9086	9086	9084

We obtain the variances of simple random sample mean, systematic sample mean and diagonal systematic sample mean for all the possible combinations of (k, n), that is, for (2,24), (3, 16), (4,12), (6,8), (8,6), (12,4), (16,3) and (24,2). Table 4.2 presents these variances for the natural population given above. It is seen now that the generalized diagonal systematic sample mean is the most efficient and also $V(\bar{y}_{gdsy}) \leq V(\bar{y}_{sy}) \leq V(\bar{y}_r)$ in each of the cases. An asterisk indicates the minimum variance in each of the cases in the Table 4.2.

Table 4.2. Comparison of simple random, systematic and generalized diagonal systematic sample means for the population considered by Subramani (2004)

Popln. Number	k	n	$V(\bar{y}_r)$	$V(\bar{y}_{sy})$	$V(\bar{y}_{gdsy})$
1	2	24	2.03531	0.17361	0.00000
2	3	16	4.07063	0.17014	0.00347
3	4	12	6.10594	0.42361	0.20139
4	6	8	10.17657	0.90972	0.61806
5	8	6	14.24719	3.71528	0.43750
6	12	4	22.38844	4.11806	3.07639
7	16	3	30.52970	15.21528	11.10417
8	24	2	46.81221	19.07639	15.82639

ACKNOWLEDGEMENTS

The author wishes to thank the referee and the editor for their useful comments which have helped to improve the presentation of this paper.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques.* 3rd Edition, John Wiley and Sons, New York.

Fountain, R.L. and Pathak, P.L. (1989). Systematic and non-random sampling in the presence of linear trends. *Comm. Statist. – Theory Methods*, **18**, 2511-2526.

Mukerjee, R. and Sengupta, S. (1990). Optimal estimation of a finite population means in the presence of linear trend. *Biometrika*, **77**, 625-630.

Subramani, J. (2000). Diagonal systematic sampling scheme for finite populations. *J. Ind. Soc. Agril. Statist.,* **53(2)**, 187-195.

Subramani, J. (2004). Diagonal systematic sampling for process control in the presence of linear trend. Presented at the International Conference on "Future of Statistical Theory, Practice and Education" held in Indian School of Business, Hyderabad, during 29 December 2004 to 1 January 2005.

APPENDIX

Derivation of Variances of Generalized Diagonal Systematic Sample Means in the Presence of Linear Trend

Let $N = kn$, where $n \geq k$, be the population size. The population units $U_1, U_2,..., U_N$ are arranged in a $n \times k$ matrix **M** (say) and the j-th row of **M** is denoted by R_j, $j = 1, 2,..., n$. The elements of R_j are $\{U_{(j-1)k+i}, i = 1, 2,..., k\}$. The diagonal systematic sampling scheme consists of drawing n units from the matrix **M** systematically such that the selected n units are the diagonal elements or broken diagonal elements of the matrix **M**. Hence the selected units are from different rows and need not be from different columns.

Let Y_{ij} be the observation corresponding to the unit in i-th row and j-th, column, that is corresponding to the unit, $U_{(i-1)k+j}$ then the sample observations are denoted by

$$Sr = \{Y_{1r}, Y_{2(r+1)}, Y_{3(r+2)}, ..., Y_{n(r+n-1)}\}, r = 1, 2,...,k$$

If $r + n - 1 > k$ then $r + n - 1$ has to be reduced to mod k.

A.1 Population with Linear Trend

In this hypothetical population, the values of N population units are in arithmetic progression. That is,

$$Y_i = a + ib,\ i = 1, 2,...,N$$

For the above population with a linear trend, the diagonal systematic sample mean with the random start r and the population mean are obtained as given below:

$$\bar{y}_{gdsy} = \begin{cases} a + \frac{1}{n}\left[\frac{pk(pk^2+1) + m(m-1)(k+1)}{2} + m(pk^2 + r)\right]b & \text{if } r \leq k - m + 1 \\ a + \frac{1}{n}\left[\frac{[pk(pk^2+1) + km(m-1) + k(k+1) + (k-m)(k-m+1)]}{2} + mpk^2 - (k-m)r\right]b & \text{if } r > k - m + 1 \end{cases}$$

$$\bar{Y} = \frac{pk^2 + mk + 1}{2}$$

where $n = pk + m$ and $p \geq 0$.

The variance of generalized diagonal systematic sample mean $\bar{y}_{gdsy}$ is obtained as

$$V(\bar{y}_{gdsy}) = E(\bar{y}_{gdsy} - \bar{Y})^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{gdsy} - \bar{Y})^2$$

$$= \frac{1}{k}\left[\sum_{i=1}^{k-m+1}(\bar{y}_{gdsy} - \bar{Y})^2 + \sum_{i=k-m+2}^{k}(\bar{y}_{gdsy} - \bar{Y})^2\right]$$

$$= \frac{1}{k}\,[SS1 + SS2] \quad \text{(say)}$$

To compute the $V(\bar{y}_{gdsy})$ given above consider the two parts separately. Further the constant b occurs both in sample mean and the population mean as a multiplier and it will be considered at the end of this derivation and is not included in each of the steps involved in the derivation of variance.

When $r \leq k - m + 1$ consider the following:

$$SS1 = \sum_{i=1}^{k-m+1}(\bar{y}_{gdsy} - \bar{Y})^2$$

$$= \sum_{i=1}^{k-m+1}\left(\frac{1}{n}\left[\frac{pk(pk^2+1) + m(m-1)(k+1)}{2} + m(pk^2 + r)\right] - \frac{pk^2 + mk + 1}{2}\right)^2$$

$$= \sum_{i=1}^{k-m+1}\left(\frac{1}{n}\left[\frac{pk(pk^2+1) + m(m-1)(k+1)}{2} + m(pk^2 + r) - \frac{n(pk^2 + mk + 1)}{2}\right]\right)^2$$

After a little algebra we have obtained the value of $SS1$ as given below:

$$SS1 = \frac{m^2}{4n^2}\sum_{i=1}^{k-m+1}[2r-(k-m+2)]^2$$

$$\Rightarrow \quad SS1 = \frac{m^2}{12n^2}(k-m+1)(k-m+2)(k-m)$$

When $r \geq k - m + 2$ consider the following:

$$SS2 = \sum_{i=k-m+2}^{k}(\bar{y}_{gdsy}-\bar{Y})^2$$

$$= \sum_{i=k-m+2}^{k}\left(\frac{1}{n}\left[\frac{[pk(pk^2+1)+km(m-1)+k(k+1)+(k-m)(k-m+1)]}{2} + mpk^2-(k-m)r\right]-\frac{pk^2+mk+1}{2}\right)^2$$

$$= \sum_{i=k-m+2}^{k}\left(\frac{1}{n}\left[\frac{[pk(pk^2+1)+km(m-1)+k(k+1)+(k-m)(k-m+1)]}{2} + mpk^2-(k-m)r-\frac{n(pk^2+mk+1)}{2}\right]\right)^2$$

After a little algebra, we have obtained the value of $SS2$ as given below:

$$SS2 = \frac{(k-m)^2}{4n^2}\sum_{i=k-m+2}^{k}(2k-m+2-2r)^2$$

$$\Rightarrow \quad SS2 = \frac{(k-m)^2}{4n^2}[m(m-1)(m-2)]$$

Now by combining the expressions obtained for $SS1$ and $SS2$, we can obtain the variance of the generalized diagonal systematic sample mean as follows:

$$V(\bar{y}_{gdsy}) = E(\bar{y}_{gdsy}-\bar{Y})^2 = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{gdsy}-\bar{Y})^2$$

$$= \frac{1}{k}\left[\sum_{i=1}^{k-m+1}(\bar{y}_{gdsy}-\bar{Y})^2 + \sum_{i=k-m+2}^{k}(\bar{y}_{gdsy}-\bar{Y})^2\right]$$

$$= \frac{1}{k}[SS1 + SS2]$$

$$= \frac{1}{k}\left[\frac{m^2}{12n^2}(k-m+1)(k-m+2)(k-m) + \frac{(k-m)^2}{12n^2}m(m-1)(m-2)\right]$$

$$= \frac{m(k-m)}{12kn^2}[m(k-m+1)(k-m+2) + (k-m)(m-1)(m-2)]$$

After a little algebra we have obtained the value of $V(\bar{y}_{gdsy})$ as given below:

$$\Rightarrow \quad V(\bar{y}_{gdsy}) = \frac{m(k-m)}{12n^2}[m(k-m)+2]$$

That is, the variance of the generalized diagonal systematic sample mean is obtained as given below:

$$V(\bar{y}_{gdsy}) = \frac{m(k-m)[m(k-m)+2]b^2}{12n^2}$$

where $n = pk + m$ and $p \geq 0$.

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 283-296

On shrinkage Estimation Procedure Combining Direct and Randomized Responses in Unrelated Question Model

Kajal Dihidar*

Indian Statistical Institute, 203, B.T. Road, Kolkata

(Received: January 2009, Accepted: November 2009)

SUMMARY

In this paper, we consider the estimation problem of population proportion, say, θ_A bearing a stigmatizing characteristic in a community. We take into account the unrelated question model with the assumption that the population proportion of unrelated innocuous attribute is not known. We modify that model on combining the direct and randomized responses and an unbiased estimator of θ_A based on randomized responses obtained from persons chosen by simple random sampling with replacement is obtained for this modified model. In this work, a new attempt has been made to construct shrinkage estimator on this modified model based on an adequate prior value θ_{A0} of θ_A. The efficiency properties of the new shrinkage estimator over the usual modified estimator are discussed here theoretically along with some numerical illustrations. In addition the unbiased estimator of the mean squared error is also derived.

Key words : Direct-cum-Randomized responses, Bias, Mean squared error, Improved efficiency, Prior value.

1. INTRODUCTION

Warner (1965) introduced a device of eliciting randomized responses (RR) instead of direct responses (DR) from persons chosen by simple random sampling with replacement (SRSWR) for estimating unbiasedly the proportion θ_A of people bearing a stigmatizing characteristic, say A like habitual alcoholism, durnken driving, drug addiction, induced abortions etc. in a community. In his technique, each respondent is provided with a randomization device by which he/she chooses one of two questions 'Do you belong to A?' or 'Do you belong to A^c?' with respective probabilities p and $1-p$ and then is asked to give a truthful 'yes' or 'no' answer to the question chosen unnoticed by the interviewer. As the interviewer does not see the question chosen, the randomization device with p close to 0.5, protects the privacy of respondent and so he/she may be willing to cooperate by responding truthfully.

Subsequently, the unrelated question model was suggested in Horvitz *et al*. (1967), Greenberg *et al*. (1969) where instead of the question related to A or A^c, the respondent is asked to respond 'yes' or 'no' to whether he/she belongs to A or to another group, say, B which is unrelated in the sense of statistical association to A. The group B and its complement B^c should both be innocuous, e.g. the question could be 'Does your birthday fall in the month of January?' Since this is unrelated to the sensitive attribute A, one can expect that the respondent will be more confident about privacy protection. However, in this approach there is an added difficulty as the true proportion θ_B in group B is also unknown. To remove this difficulty, the respondent is asked to give two randomized resposes from two randomization devices (one from each device) made with questions relating to A and B with probabilities p_1 and $1-p_1$ for 1st device and p_2 and

* *Corresponding author* : Kajal Dihidar
E-mail address: dkajal@isical.ac.in

$1 - p_2$ for 2nd device, ($p_1 \neq p_2$), respectively. Many researchers like Moor (1971), Folsom *et al*. (1973), Lanke (1975), Abul-Ela *et al*. (1967), Mangat *et al*. (1992), Singh *et al*. (1994) and Singh (1994) studied this model in detail. Dowling and Shatchman (1975) compared the Warner's (1965) model with the unrelated question model. Extensive reviews in this regard are available in Chaudhuri and Mukerjee (1985).

Mangat and Singh (1990) gave a method of combining the direct and randomized responses to estimate the proportion of people bearing a stigmatizing attribute in a community and compared the efficiency of their estimator with the estimator given by Warner (1965). In their method a sampled respondent chosen by SRSWR scheme is asked to respond directly his/her true value ('yes' or 'no') regarding the possessing of stigmatizing characteristic with a fixed probability, say, T and to answer ('yes' or 'no') by Warner's (1965) randomized response device with probability $(1 - T)$, unnoticed by the interviewer. We modify the unrelated question model along the line of Mangat and Singh (1990) on assuming that the population proportion of innocuous attribute is not known and an unbiased estimator of population proportion of bearing stigmatizing character is found out. We call this unbiased estimator as the usual estimator in the present context. The variance and an unbiased estimator of that variance of the estimator are also found out.

In usual practice, the experimenter often possesses some prior knowledge about an adequate guess, say, θ_{A0} of the value of θ_A based on some past experience or from some authentical source. Thompson (1968) suggested to construct a shrinkage estimator on using such prior guess. Singh *et al*. (2007) constructed a shrinkage estimator based on such prior guess in unrelated question model developed by Horvitz *et al*. (1967) on assuming that the population proportion of innocuous attribute is known. They also compared the efficiencies of their suggested shrinkage estimators with the estimator given by Horvitz *et al*. (1967).

In our present work, motivated by Thompson (1968) and Singh *et al*. (2007), we make an attempt to construct the new shrinkage estimator in unrelated question model on assuming that the population proportion of innocuous attribute is not known. We show here how the shrinkage estimator performs better in comparison to the usual estimator after combining the direct and randomized response technique in view of the mean squared error. An unbiased estimator of the mean squared error of the shrinkage estimator is also suggested. Finally some numerical computations are done to display the ranges of the required parameters to ensure the gain in effeciencies by our suggested estimator.

2. DIRECT-CUM-RANDOMIZED RESPONSE TECHNIQUE IN UNRELATED QUESTION MODEL ASSUMING θ_B IS UNKNOWN

2.1 Basic Notations and Definition of Unrelated Question Model

Suppose in a finite survey population $U = (1, \ldots, i, \ldots, N)$ a person labelled i has the value y_i on a sensitive variable y. Let

$y_i = 1$ if i bears the stigmatizing feature A.
$= 0$ if he/she bears the complementary feature A^c.

Our problem is to estimate $\theta_A = \dfrac{\sum_{i=1}^{N} y_i}{N}$, the proportion in U bearing A.

An unrelated question RR model studied by Horvitz *et al*. (1967), Greenberg *et al*. (1969) needs another variable x defined on U related to an innocuous human character, say, B not correlated with y in the following way.

$x_i = 1$ if i bears the unrelated innocuous feature B.
$= 0$ if he/she bears the complementary feature B^c.

To apply this RR technique, one box is to be prepared with a number of cards marked as 'Do you belong to A?' with probability $p_1 (0 < p_1 < 1)$ and the rest are marked as 'Do you belong to B?' And a second box is also to be prepared with a number of cards marked as 'Do you belong to A?' with probability $p_2 (0 < p_2 < 1, p_2 \neq p_1)$ and the rest are marked as 'Do you belong to B? These two boxes are offered to a sampled person i. Following Chaudhuri (2001), we note as follows. Using the above two boxes independently, two responses, realizing I_i and J_i are generated leading to the RR's.

$I_i = 1$ if card type matches his/her true feature from 1st box.
$= 0$, else.

J_i = 1 if card type matches his/her true feature from the second box.
= 0, else.

So, generically writing E_R and V_R to denote expectation and variance operator with respect to the personal 'response-making', we have

$$E_R(I_i) = p_1 y_i + (1 - p_1)x_i$$
$$E_R(J_i) = p_2 y_i + (1 - p_2)x_i$$

$$r_i = \frac{(1-p_2)\, I_i - (1-p_1)\, j_i}{p_1 - p_2}, \text{ taking } p_1 \neq p_2,$$

with $E_R(r_i) = y_i$ and on simplification

$$V_R(r_i) = \frac{[(1-p_1)(1-p_2)\{p_1(1-p_2) + p_2(1-p_1)\}]}{(p_1-p_2)^2}\ (y_i - x_i)^2 \tag{2.1}$$

2.2 Modified Definition of Unrelated Question Model

To modify this procedure along the line of Mangat and Singh (1990), our suggestion is that a sampled person i be requested first to draw one card from a box, say, BOXT, containing the cards marked 'True' with probability T and the rest marked as 'RR'. Then if 'RR'-marked card appears, he is requested to produce two independent responses I_i and J_i as above or else give out the true response y_i if 'True'-marked card appears. We write generically

$z_i = y_i$ with probability T using BOXT
= I_i with probability $(1 - T)$ using the 1st box as in Subsection 2.1

$z_i' = y_i$ with probability T using BOXT
= J_i with probability $(1 - T)$ using the 2nd box as in Subsection 2.1

Then

$E_R(z_i) = Ty_i + (1 - T)\,[p_1 y_i + (1 - p_1)x_i]$ and
$E_R(z_i') = Ty_i + (1 - T)\,[p_2 y_i + (1 - p_2)x_i]$

yielding

$$r_i' = \frac{(1-p_2)\, z_i - (1-p_1)\, z_i'}{(p_1 - p_2)} \text{ with } p_1 \neq p_2$$

$E_R(r_i') = y_i$

and on suppressing detailed algebra,

$$V_R(r_i') = \phi(y_i - x_i)^2, \text{ say} \tag{2.2}$$

where

$$\phi = \frac{[(1-T)(1-p_1)(1-p_2)[(1-p_2)(T + p_1 - p_1 T) + (1-p_1)(T + p_2 - p_2 T)]]}{(p_1 - p_2)^2}$$

Based on this modification, an unbiased estimator of θ_A is

$$\hat{\theta}_A = e = \frac{1}{n}\sum_{k=1}^{n} r_k' \tag{2.3}$$

since on writing E_P and V_P to denote the expectation and variance operator with respect to the sampling of respondents

$$E_P E_R(e) = E_P\left(\frac{1}{n}\sum_{k=1}^{n} y_k\right) = \bar{Y} = \theta_A$$

and

$$V(e) = V_P E_R(e) + E_P V_R(e)$$

$$= V_P\left(\frac{1}{n}\sum_{k=1}^{n} y_k\right) + E_P\left[\frac{\phi}{n^2}\sum_{k=1}^{n}(y_k + x_k - 2y_k x_k)\right]$$

$$= \frac{\theta_A(1-\theta_A)}{n} + \frac{\phi}{n}(\theta_A + \theta_B) - \frac{2\phi}{n} E_P\left(\frac{1}{n}\sum_{k=1}^{n} y_k x_k\right)$$

$$= \frac{\theta_A(1-\theta_A)}{n} + \frac{\phi}{n}(\theta_A + \theta_B) - \frac{2\phi}{n}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i X_i\right)$$

$$= \frac{\theta_A(1-\theta_A)}{n} + \frac{\phi}{n}(\theta_A + \theta_B) - \frac{2\phi}{n}\theta_A\theta_B$$

$$= \frac{\theta_A(1-\theta_A)}{n} + \frac{\phi}{n}(\theta_A + \theta_B - 2\theta_A\theta_B) \tag{2.4}$$

It is easy to prove that irrespective of how a sample of respondents is drawn, the estimator based on unrelated question model by above modification perfoms better than the usual estimator based on usual unrelated question model without above modification if one chooses T satisfying

$$T \geq 1 - \frac{p_1(1-p_2)+p_2(1-p_1)}{2(1-p_1)(1-p_2)} \qquad (2.5)$$

2.3 An Unbiased Variance Estimation for Modified Definition

An unbiased variance estimator of $\hat{\theta}_A = e$ may be taken as

$$v(e) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (r_k' - \bar{r}_n')^2$$

where $$\bar{r}_n' = \frac{1}{n} \sum_{k=1}^{n} r_k' \qquad (2.6)$$

because on writing

$$s_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y}_n)^2$$

with $$\bar{y}_n = \frac{1}{n} \sum_{k=1}^{n} y_k$$

$$E_R \left[\frac{1}{n(n-1)} \sum_{k=1}^{n} (r_k' - \bar{r}_n')^2 \right]$$

$$= \frac{1}{n(n-1)} \left[\sum_{k=1}^{n} \{V_R(r_k') + E_R^2(r_k')\} - n\{V_R(\bar{r}_n') + E_R^2(\bar{r}_n')\} \right]$$

$$= \frac{1}{n(n-1)} [\phi \sum_{k=1}^{n} (y_k - x_k)^2 + \sum_{k=1}^{n} y_k^2 - n\frac{\phi}{n^2} \sum_{k=1}^{n} (y_k - x_k)^2 - n\bar{y}_n^2]$$

$$= \frac{\phi}{n^2} \sum_{k=1}^{n} (y_k - x_k)^2 + \frac{s_n^2}{n}$$

Then $E_p \left[\frac{\phi}{n^2} \sum_{k=1}^{n} (y_k - x_k)^2 + \frac{s_n^2}{n} \right]$

$$= \frac{\phi}{nN} \sum_{i=1}^{N} (Y_i - X_i)^2 + \frac{1}{nN} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$= \frac{\phi}{n}(\theta_A + \theta_B - 2\theta_A\theta_B) + \frac{\theta_A(1-\theta_A)}{n} = V(e)$$

3. SHRINKAGE ESTIMATOR ON MODIFIED UNRELATED QUESTION MODEL ASSUMING θ_B IS UNKNOWN

3.1 Estimator (i) and its Mean Squared Error

In usual practice, we may often have some adequate guess, say, θ_{A0} of the value of θ_A based on past knowledge. Let θ_{A0} denotes that prior estimate of θ_A. Motivated by Thompson (1968), we may define a belief parameter, say, δ with $(0 < \delta < 1)$ defined as a value of δ close to 0 implying a strong belief on θ_{A0} and a value of δ close to 1 implying a strong belief on sample estimate $\hat{\theta}_A = e$. Then our suggested shrinkage estimator on using such prior guess and belief parameter is

$$e_{S1} = \delta e + (1 - \delta)\theta_{A0} \qquad (3.1)$$

According to this set-up, bias of e_{S1} is given by

$$\begin{aligned} B(e_{S1}) &= E(e_{S1}) - \theta_A \\ &= E[\delta e + (1-\delta)\theta_{A0}] - \theta_A \\ &= \delta\theta_A + (1-\delta)\theta_{A0} - \theta_A \\ &= (1-\delta)(\theta_{A0} - \theta_A) \end{aligned} \qquad (3.2)$$

The mean squared error of e_{S1} is given by

$$\begin{aligned} MSE(e_{S1}) &= V(e_{S1}) + B^2(e_{S1}) \\ &= V[\delta e + (1-\delta)\theta_{A0}] + (1-\delta)^2(\theta_{A0} - \theta_A)^2 \\ &= V[\delta e] + (1-\delta)^2(\theta_{A0} - \theta_A)^2 \\ &= \delta^2 V[e] + (1-\delta)^2(\theta_{A0} - \theta_A)^2 \\ &= \delta^2 \left[\frac{\phi}{n}(\theta_A + \theta_B - 2\theta_A\theta_B) + \frac{\theta_A(1-\theta_A)}{n} \right] \\ &\quad + (1-\delta)^2(\theta_{A0} - \theta_A)^2 \\ &= \delta^2 \left[\frac{\phi}{n}(\theta_A + \theta_B - 2\theta_A\theta_B) + \frac{\theta_A(1-\theta_A)}{n} \right] \\ &\quad + (1-\delta)^2(\theta_{A0}^2 + \theta_A^2 - 2\theta_{A0}\theta_A) \end{aligned} \qquad (3.3)$$

3.2 An Unbiased Estimator for Mean Squared Error of Estimator (i)

An unbiased estimator for mean squared error (3.3) of above shrinkage estimator (3.1) may be taken as

$$mse(e_{S1}) = \delta^2 v(e) + (1-\theta)^2 (\theta_{A0}^2 + e^2 - v(e) - 2\theta_{A0}e)$$

because of the following reasons.

(*i*) e and $v(e)$ are respectively the unbiased estimators of θ_A and $V(e)$.

(*ii*) $E_P E_R[e^2 - v(e)] = E_P[V_R(e) + E_R^2(e)] - V(e)$

$$= E_p\left[V_e(e) + \bar{y}_n^2\right] - V(e)$$

$$= E_P(V_R(e)] - V_P(\bar{y}_n) + E_P^2(\bar{y}_n) - V(e)$$

$$= E_P[V_R(e)\} + V_P E_R(e) + E_P^2(\bar{y}_n) - V(e)$$

$$= V(e) + \bar{Y}^2 - V(e) = \bar{Y}^2 = \theta_A^2$$

Hence after simplification the above expression reduces to

$$mse\ (e_{S1}) = (1 - \delta^2)\ (\theta_{A0} - e)^2 - (1 - 2\delta)\ v(e) \tag{3.4}$$

for which $v(e)$ is to be substituted from (2.6).

3.3 Efficiency Comparison between e_{s1} and e

The new estimator e_{S1} in (3.1) will be more efficient than the usual estimator e in (2.3) if

$$MSE\ (e_{S1}) \le V\ (e)$$

On suppressing considerable algebra this condition reduces to

$e_{S1} \succ e$ if

$$\delta \ge \frac{n(\theta_{A0} - \theta_A)^2 - [\theta_A(1-\theta_A) + \phi(\theta_A + \theta_B - 2\theta_A\theta_B)]}{n(\theta_{A0} - \theta_A)^2 + [\theta_A(1-\theta_A) + \phi(\theta_A + \theta_B - 2\theta_A\theta_B)]} \tag{3.5}$$

This lower bound of δ for different values of ϕ (i.e. for different values of p_1, p_2, T) and θ_A, θ_B and θ_{A0} are presenented in Tables 1a and 1b. The bounds yielding negative values are presented as zero.

3.4 Optimum Value of δ to Minimize MSE(e_{s1})

The optimum value of δ which minimizes $MSE(e_{S1})$ can be obtained by solving

$$\frac{\partial}{\partial \delta}\ [MSE(es_1)] = 0$$

And this gives rise to the following solution which also after substitution to the $MSE(e_{S1})$ yields a positive quantity.

$$\delta_{\text{optimum}} = \frac{n(\theta_{A0} - \theta_A)^2}{\left[\theta_A(1-\theta_A) + \phi(\theta_A + \theta_B - 2\theta_A\theta_B) + n(\theta_{A0} - \theta_A)^2\right]} \tag{3.6}$$

But, we note that this optimum value of δ depends on two unknown quantities, θ_A and θ_B. So, to use this optimum value in practical survey situation, we may take into consideration the guessed value θ_{A0} of θ_A and also one more guessed value of θ_B. Suppose, θ_{B0} denotes an initial estimate or prior guessed value of θ_B. In order to obtain a usable value of δ_{optimum} utilizing the prior guesses of θ_A and θ_B, we may replace θ_A by $\alpha_A\theta_{A0}$ and θ_B by $\alpha_B\theta_{B0}$ in (3.6), where α_A and α_B are positive constants ($0 < \alpha_A\ ,\ \alpha_B < 1$). We shall also examine the efficacy of this new estimator on changes of these two constants. The new value of δ thus obtained as $\hat{\delta}_{opt}$, say, is

$$\hat{\delta}_{opt} = \frac{n(\theta_{A0} - \alpha_A\theta_{A0})^2}{[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} + \alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0}) + n(\theta_{A0} - \alpha_A\theta_{A0})^2]} \tag{3.7}$$

3.5 Estimator (ii) and its Mean Squared Error

Substitution of above $\hat{\delta}_{opt}$ from (3.7) in place of δ in (3.1) yields a second shrinkage estimator as

$$e_{S2} = \theta_{A0} + \frac{(e - \theta_{A0})n\theta_{A0}^2(1-\alpha_A)^2}{[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} + \alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0}) + n(\theta_{A0} - \alpha_A\theta_{A0})^2]}. \tag{3.8}$$

The bias of e_{S2} is given by

$$B(e_{S2}) = (\theta_{A0} + \theta_A)\left[\frac{[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} + \alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0})]}{[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} + \alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0}) + n(\theta_{A0} - \alpha_A\theta_{A0})^2]}\right] \tag{3.9}$$

The mean squared error of this estimator (ii) is given by

$$MSE(e_{S2}) = \frac{\begin{array}{l}[n^2\theta_{A0}^4(1-\alpha_A)^4\,V(e) + (\theta_{A0}-\theta_A)^2 \\ [\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} \\ +\alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0})]^2]\end{array}}{\begin{array}{l}[[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} \\ +\alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0}) \\ +n(\theta_{A0}-\alpha_A\theta_{A0})^2]^2]\end{array}}$$

This after substitution of the expression for $V(e)$ from (2.4) reduces to

$$MSE(e_{S2}) = \frac{\begin{array}{l}[n\theta_{A0}^4(1-\alpha_A)^4\{\theta_A(1-\theta_A) \\ +\phi(\theta_A+\theta_B-2\theta_A\theta_B)\}]\end{array}}{\begin{array}{l}[[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} \\ +\alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0}) \\ +n(\theta_{A0}-\alpha_A\theta_{A0})^2]^2]\end{array}}$$

$$+ \frac{\begin{array}{l}[(\theta_{A0}-\theta_A)^2[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) \\ +\phi(\alpha_A\theta_{A0}+\alpha_B\theta_{B0}-2\alpha_A\theta_{A0}\alpha_B\theta_{B0})]^2]\end{array}}{\begin{array}{l}[[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} \\ +\alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\,\alpha_B\theta_{B0}) \\ +n(\theta_{A0}-\alpha_A\theta_{A0})^2]^2]\end{array}} \qquad (3.10)$$

3.6 An Unbiased Estimator for Mean Squared Error of Estimator (ii)

An unbiased estimator for mean squared error (3.10) of above shrinkage estimator (3.8) may be taken as

$$mse\ (e_{S2}) = \frac{\begin{array}{l}[n^2\theta_{A0}^4(1-\alpha_A)^4\,v(e) + (\theta_{A0}^2 - 2\theta_{A0}e \\ +e^2 - v(e))\,[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) \\ +\phi(\alpha_A\theta_{A0}+\alpha_B\theta_{B0}-2\alpha_A\theta_{A0}\alpha_B\theta_{B0})]^2]\end{array}}{\begin{array}{l}[[\alpha_A\theta_{A0}(1-\alpha_A\theta_{A0}) + \phi(\alpha_A\theta_{A0} \\ +\alpha_B\theta_{B0} - 2\alpha_A\theta_{A0}\alpha_B\theta_{B0}) \\ +n(\theta_{A0}-\alpha_A\theta_{A0})^2]^2]\end{array}}$$

This after simplification reduces to

$$mse\ (e_{S2}) = (1-\hat{\delta}_{opt}^2)\,(\theta_{A0}-e)^2 - (1-2\hat{\delta}_{opt})\,v(e) \qquad (3.11)$$

for which $v(e)$ is to be substituted from (2.6).

3.7 Efficiency Comparison between e_{S2} and e

The new estimator e_{S2} in (3.8) will be more efficient than the usual estimator e in (2.3) if

$$MSE\ (e_{S2}) \le V(e)$$

Let $k = \theta_A - \theta_A^2 + \phi\theta_A + \phi\theta_B - 2\phi\theta_A\theta_B$. Then on suppressing considerable algebra the required condition for e_{S2} to be more efficient than e is that

$$\alpha_A^2\ P_A - \alpha_A Q_A + R_A \ge 0 \qquad (3.12)$$

where

$$P_A = \theta_{A0}^2\,[n(\theta_{A0}-\theta_A)^2 + (2n-1)k]$$

$$Q_A = \theta_{A0}[(1+\phi-2\phi\alpha_B\theta_{B0})\ \{n(\theta_{A0}-\theta_A)^2 - k\} + 4n\theta_{A0}k]$$

and

$$R_A = 2nk\theta_{A0}^2 - \alpha_B\theta_{B0}\phi\{n(\theta_{A0}-\theta_A)^2 - k\}$$

To get a real solution of this inequality the necessary condition $Q_A^2 - 4P_AR_A \ge 0$ gives us the admissible ranges of α_B. On suppressing considerable algebra, this condition reduces to

$$\alpha_B^2\,P_B + \alpha_B Q_B + R_B \ge 0 \qquad (3.13)$$

where

$$P_B = 4\phi^2\theta_{B0}^2\ [n(\theta_{A0}-\theta_A)^2 - k]$$

$$Q_B = 4\theta_{B0}\phi\,[2nk - \phi\ \{n(\theta_{A0}-\theta_A)^2 - k\} - 4n\theta_{A0}k]$$

and

$$R_B = (1+\phi)^2\{n(\theta_{A0}-\theta_A)^2 - k\} + 8nk\theta_{A0}\ \{1+\phi-\theta_{A0}\}$$

For different values of ϕ (i.e. for different values of p_1, p_2, T), θ_A, θ_B, θ_{A0}, θ_{B0} and n, the ranges of α_B are displayed in Tables 2*a* and 2*b*. Then corresponding to a particular value of $\alpha_B = 0.40$ satisfying the most of the solution ranges of (3.13), the required solution in terms of upper bounds of α_A satisfying (3.12) for $e_{S2} \succ e$ are presented in Tables 3*a* and 3*b*.

4. NUMERICAL ILLUSTRATIONS

We compute the lower bound of δ as proved in (3.5) for efficiency comparison between e_{S1} and e and our illustrative findings are presented in Tables 1*a* and 1*b*. We also compute the bounds of α_A as proved in (3.12) for efficiency comparison between e_{S2} and e and our illustrative findings are presented in Tables 3*a* and 3*b*. But prior to that we first compute the ranges of α_B

Table 1a. Illustrative lower bound of δ for assured gain in efficiency of e_{S1} over e

	p_1	p_2	Range of T		T	p_1	p_2	Range of T		T	p_1	p_2	Range of T		T
	0.3	0.45	≥ 0.377		0.40	0.5	0.3	≥ 0.286		0.30	0.5	0.6	Any		0.2
	n					n					n				
θ_{A0}	10	25	50	75	100	10	25	50	75	100	10	25	50	75	100
$\theta_A = 0.20$ and $\theta_B = 0.15$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.086	0.226	0.0	0.0	0.0	0.0	0.0
$\theta_A/8$	0.0	0.0	0.0	0.0	0.100	0.0	0.0	0.037	0.236	0.366	0.0	0.0	0.0	0.0	0.036
$\theta_A/10$	0.0	0.0	0.0	0.0	0.128	0.0	0.0	0.065	0.262	0.390	0.0	0.0	0.0	0.0	0.064
$\theta_A = 0.30$ and $\theta_B = 0.15$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.118	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.100	0.237	0.0	0.0	0.175	0.363	0.481	0.0	0.0	0.0	0.033	0.176
$\theta_A/8$	0.0	0.0	0.0	0.247	0.376	0.0	0.0	0.320	0.489	0.590	0.0	0.0	0.0	0.185	0.320
$\theta_A/10$	0.0	0.0	0.077	0.273	0.400	0.0	0.013	0.345	0.510	0.608	0.0	0.0	0.013	0.213	0.345
$\theta_A = 0.40$ and $\theta_B = 0.15$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.036	0.0	0.0	0.0	0.174	0.310	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.095	0.290	0.415	0.0	0.032	0.362	0.524	0.620	0.0	0.0	0.031	0.229	0.361
$\theta_A/8$	0.0	0.0	0.244	0.424	0.534	0.0	0.184	0.488	0.626	0.706	0.0	0.0	0.183	0.369	0.487
$\theta_A/10$	0.0	0.0	0.271	0.447	0.554	0.0	0.211	0.509	0.643	0.720	0.0	0.0	0.210	0.394	0.508
$\theta_A = 0.20$ and $\theta_B = 0.25$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.002	0.144	0.0	0.0	0.0	0.0	0.0
$\theta_A/8$	0.0	0.0	0.0	0.0	0.012	0.0	0.0	0.0	0.155	0.291	0.0	0.0	0.0	0.0	0.0
$\theta_A/10$	0.0	0.0	0.0	0.0	0.040	0.0	0.0	0.0	0.182	0.317	0.0	0.0	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.25$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.072	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.048	0.190	0.0	0.0	0.130	0.321	0.444	0.0	0.0	0.0	0.0	0.127
$\theta_A/8$	0.0	0.0	0.0	0.200	0.333	0.0	0.0	0.277	0.452	0.559	0.0	0.0	0.0	0.137	0.275
$\theta_A/10$	0.0	0.0	0.028	0.227	0.358	0.0	0.0	0.303	0.474	0.578	0.0	0.0	0.0	0.165	0.300
$\theta_A = 0.40$ and $\theta_B = 0.25$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.015	0.0	0.0	0.0	0.155	0.291	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.074	0.270	0.397	0.0	0.012	0.344	0.509	0.608	0.0	0.0	0.009	0.209	0.342
$\theta_A/8$	0.0	0.0	0.224	0.406	0.519	0.0	0.165	0.472	0.614	0.696	0.0	0.0	0.162	0.351	0.470
$\theta_A/10$	0.0	0.0	0.251	0.429	0.539	0.0	0.192	0.494	0.631	0.710	0.0	0.0	0.189	0.375	0.492

Table 1b. Illustrative lower bound of δ for assured gain in efficiency of e_{S1} over e

	p_1	p_2	Range of T		T	p_1	p_2	Range of T		T	p_1	p_2	Range of T		T
	0.3	0.45	≥ 0.377		0.40	0.5	0.3	≥ 0.286		0.30	0.5	0.6	Any		0.2
	n					n					n				
θ_{A0}	10	25	50	75	100	10	25	50	75	100	10	25	50	75	100
$\theta_A = 0.20$ and $\theta_B = 0.55$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/8$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.108	0.0	0.0	0.0	0.0	0.0
$\theta_A/10$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.136	0.0	0.0	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.55$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.0	0.068	0.0	0.0	0.012	0.211	0.344	0.0	0.0	0.0	0.0	0.003
$\theta_A/8$	0.0	0.0	0.0	0.078	0.219	0.0	0.0	0.165	0.353	0.472	0.0	0.0	0.0	0.013	0.156
$\theta_A/10$	0.0	0.0	0.0	0.106	0.245	0.0	0.0	0.192	0.377	0.494	0.0	0.0	0.0	0.041	0.183
$\theta_A = 0.40$ and $\theta_B = 0.55$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.099	0.239	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.015	0.214	0.347	0.0	0.0	0.294	0.466	0.571	0.0	0.0	0.0	0.152	0.288
$\theta_A/8$	0.0	0.0	0.168	0.356	0.474	0.0	0.110	0.427	0.578	0.666	0.0	0.0	0.104	0.298	0.423
$\theta_A/10$	0.0	0.0	0.195	0.380	0.496	0.0	0.137	0.450	0.596	0.681	0.0	0.0	0.132	0.323	0.445
$\theta_A = 0.20$ and $\theta_B = 0.70$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/8$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.035	0.0	0.0	0.0	0.0	0.0
$\theta_A/10$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.063	0.0	0.0	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.70$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.0	0.016	0.0	0.0	0.0	0.163	0.299	0.0	0.0	0.0	0.0	0.0
$\theta_A/8$	0.0	0.0	0.0	0.026	0.169	0.0	0.0	0.115	0.308	0.432	0.0	0.0	0.0	0.0	0.104
$\theta_A/10$	0.0	0.0	0.0	0.055	0.196	0.0	0.0	0.143	0.334	0.455	0.0	0.0	0.0	0.0	0.132
$\theta_A = 0.40$ and $\theta_B = 0.70$															
θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.074	0.214	0.0	0.0	0.0	0.0	0.0
$\theta_A/4$	0.0	0.0	0.0	0.189	0.323	0.0	0.0	0.270	0.446	0.553	0.0	0.0	0.0	0.125	0.263
$\theta_A/8$	0.0	0.0	0.141	0.332	0.453	0.0	0.084	0.406	0.560	0.651	0.0	0.0	0.077	0.273	0.400
$\theta_A/10$	0.0	0.0	0.169	0.357	0.475	0.0	0.112	0.429	0.579	0.667	0.0	0.0	0.105	0.298	0.423

Table 2a. Illustrative lower bound of α_B satisfying (3.13) for assured gain in efficiency of e_{S2} over e

		p_1	p_2	Range of T	T	p_1	p_2	Range of T	T	p_1	p_2	Range of T	T
		0.3	0.45	≥ 0.377	0.40	0.5	0.3	≥ 0.286	0.30	0.5	0.6	Any	0.2
		n				n				n			
θ_{B0}	θ_{A0}	10	25	50	75	100	10	25	50	75	100	10	25	50	75	100
$\theta_A = 0.20$ and $\theta_B = 0.15$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.130	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.300	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	—	0.0	0.0	0.0	0.0	0.205	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	—	0.375	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0	—	0.690	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.760	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0	—	—	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.15$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.530	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.855	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	—	0.0	0.0	0.0	0.0	0.230	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.0	0.0	0.0	.740	0.0	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_A = 0.40$ and $\theta_B = 0.15$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.055	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	—	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_A = 0.20$ and $\theta_B = 0.25$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.085	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.185	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.795	0.0	0.0	0.0	0.0	0.155	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	—	0.300	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0	—	0.485	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.555	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0	—	0.780	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.25$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.345	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.540	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	—	0.0	0.0	0.0	0.0	0.225	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.0	0.0	0.0	0.0	0.555	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_A = 0.40$ and $\theta_B = 0.25$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	0.865	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	—	0.0	0.0	0.0	0.0

Table 2b. Illustrative lower bound of α_B satisfying (3.13) for assured gain in efficiency of e_{S2} over e

		p_1	p_2	Range of T		T	p_1	p_2	Range of T		T	p_1	p_2	Range of T		T
		0.3	0.45	≥ 0.377		0.40	0.5	0.3	≥ 0.286		0.30	0.5	0.6	Any		0.2
		n					n					n				
θ_{B0}	θ_{A0}	10	25	50	75	100	10	25	50	75	100	10	25	50	75	100
$\theta_A = 0.20$ and $\theta_B = 0.55$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.090	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.380	0.0	0.0	0.0	0.0	0.095	0.0	0.0	0.0	0.0	0.450	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	—	0.195	0.0	0.0	0.0	0.535	0.0	0.0	0.0	0.0	—	0.275	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.325	0.0	0.0	0.0	0.750	0.0	0.0	0.0	0.0	—	0.425	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.55$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.180	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.265	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	0.770	0.0	0.0	0.0	0.0	0.185	0.0	0.0	0.0	0.0	0.915	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.0	0.0	0.0	0.0	0.355	0.0	0.0	0.0	0.0	—	0.055	0.0	0.0	0.0
$\theta_A = 0.40$ and $\theta_B = 0.55$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/4$	0.445	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.615	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	0.690	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.895	0.0	0.0	0.0	0.0
$\theta_A = 0.20$ and $\theta_B = 0.70$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.070	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.300	0.0	0.0	0.0	0.0	0.080	0.0	0.0	0.0	0.0	0.355	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	0.810	0.165	0.0	0.0	0.0	0.435	0.0	0.0	0.0	0.0	0.910	0.230	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	—	0.270	0.0	0.0	0.0	0.605	0.0	0.0	0.0	0.0	—	0.350	0.0	0.0	0.0
$\theta_A = 0.30$ and $\theta_B = 0.70$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.150	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.215	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	0.620	0.0	0.0	0.0	0.0	0.170	0.0	0.0	0.0	0.0	0.735	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	0.850	0.0	0.0	0.0	0.0	0.310	0.0	0.0	0.0	0.0	—	0.075	0.0	0.0	0.0
$\theta_A = 0.40$ and $\theta_B = 0.70$																
θ_B	θ_A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$3\theta_B/4$	$3\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/2$	$\theta_A/2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/4$	$\theta_A/4$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\theta_B/8$	$\theta_A/8$	0.365	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.500	0.0	0.0	0.0	0.0
$\theta_B/10$	$\theta_A/10$	0.565	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.720	0.0	0.0	0.0	0.0

Table 3a. Illustrative upper bound of α_A for assured gain in efficiency of e_{S2} over e for a particular value $\alpha_B = 0.4$

		p_1	p_2	Range of T	T	p_1	p_2	Range of T	T	p_1	p_2	Range of T	T
		0.3	0.45	≥ 0.377	0.40	0.5	0.3	≥ 0.286	0.30	0.5	0.6	Any	0.2
		n				n				n			
θ_{B0}	θ_{A0}	10	25	50	75	100	10	25	50	75	100	10	25	50	75	100
$\theta_A = 0.20$ and $\theta_B = 0.15$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	—	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.655	0.515	—	1.0	1.0	1.0	1.0
$\theta_B/8$	$\theta_A/8$	—	1.0	1.0	1.0	0.460	—	1.0	0.625	0.305	0.235	—	—	1.0	1.0	0.630
$\theta_B/10$	$\theta_A/10$	—	—	1.0	1.0	0.355	—	1.0	0.485	0.230	0.165	—	—	1.0	1.0	0.485
$\theta_A = 0.30$ and $\theta_B = 0.15$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.790	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	—	1.0	1.0	0.625	0.500	1.0	1.0	0.540	0.435	0.395	—	1.0	1.0	0.755	0.545
$\theta_B/8$	$\theta_A/8$	—	1.0	0.570	0.295	0.230	1.0	1.0	0.255	0.195	0.170	—	1.0	1.0	0.340	0.250
$\theta_B/10$	$\theta_A/10$	—	1.0	0.440	0.220	0.165	—	0.720	0.190	0.135	0.115	—	.1.0	0.715	0.260	0.185
$\theta_A = 0.40$ and $\theta_B = 0.15$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.750	0.695	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	0.620	0.460	0.410	1.0	0.755	0.425	0.380	0.355	1.0	1.0	0.760	0.495	0.430
$\theta_B/8$	$\theta_A/4$	—	1.0	0.620	0.460	0.185	1.0	0.335	0.195	0.165	0.155	—	1.0	0.340	0.230	0.195
$\theta_B/10$	$\theta_A/10$	—	1.0	0.220	0.155	0.130	1.0	0.260	0.140	0.120	0.105	—	1.0	0.260	0.170	0.140
$\theta_A = 0.20$ and $\theta_B = 0.25$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	—	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.580	—	1.0	1.0	1.0	1.0
$\theta_B/8$	$\theta_A/8$	—	1.0	1.0	1.0	0.775	—	1.0	1.0	0.365	0.245	—	—	1.0	1.0	1.0
$\theta_B/10$	$\theta_A/10$	—	—	1.0	1.0	0.565	—	1.0	1.0	0.270	0.160	—	—	1.0	1.0	1.0
$\theta_A = 0.30$ and $\theta_B = 0.25$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.830	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	1.0	0.715	0.520	1.0	1.0	0.580	0.440	0.390	—	1.0	1.0	1.0	0.585
$\theta_B/8$	$\theta_A/8$	—	1.0	1.0	0.305	0.215	1.0	1.0	0.250	0.170	0.140	—	1.0	1.0	0.375	0.245
$\theta_B/10$	$\theta_A/10$	—	1.0	0.605	0.220	0.140	—	1.0	0.170	0.105	0.075	—	1.0	1.0	0.280	0.170
$\theta_A = 0.40$ and $\theta_B = 0.25$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	0.910	1.0	1.0	1.0	0.760	0.695	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	0.650	0.455	0.395	1.0	0.840	0.415	0.360	0.335	1.0	1.0	0.855	0.495	0.420
$\theta_B/8$	$\theta_A/8$	—	1.0	0.275	0.180	0.145	1.0	0.330	0.165	0.130	0.115	—	.1.0	0.335	0.205	0.160
$\theta_B/10$	$\theta_A/10$	—	1.0	0.200	0.115	0.090	1.0	0.245	0.105	0.075	0.060	.—	.1.0	0.245	0.135	0.100

Table 3b. Illustrative upper bound of α_A for assured gain in efficiency of e_{S2} over e for a particular value $\alpha_B = 0.4$

		p_1	p_2	Range of T		T	p_1	p_2	Range of T		T	p_1	p_2	Range of T		T
		0.3	0.45	≥ 0.377		0.40	0.5	0.3	≥ 0.286		0.30	0.5	0.6	Any		0.2
		n					n					n				
θ_{B0}	θ_{A0}	10	25	50	75	100	10	25	50	75	100	10	25	50	75	100
$\theta_A = 0.20$ and $\theta_B = 0.15$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	—	1.0	1.0	1.0	1.0
$\theta_B/8$	$\theta_A/8$	—	1.0	1.0	1.0	1.0	—	1.0	1.0	—	0.420	—	1.0	1.0	1.0	1.0
$\theta_B/10$	$\theta_A/10$	—	1.0	1.0	1.0	1.0	—	1.0	1.0	—	0.295	—	—	1.0	1.0	1.0
$\theta_A = 0.30$ and $\theta_B = 0.15$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	1.0	1.0	0.665	1.0	1.0	0.845	0.485	0.400	1.0	1.0	1.0	1.0	0.925
$\theta_B/8$	$\theta_A/8$	—	1.0	1.0	0.450	0.230	.1.0	1.0	0.295	0.130	0.075	—	1.0	1.0	0.740	0.305
$\theta_B/10$	$\theta_A/10$	—	1.0	1.0	0.320	0.125	1.0	1.0	0.185	0.0	0.0	—	.1.0	.1.0	0.520	0.190
$\theta_A = 0.40$ and $\theta_B = 0.15$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.790	0.705	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	0.815	0.455	0.370	1.0	1.0	0.400	0.320	0.285	1.0	1.0	1.0	0.515	0.400
$\theta_B/8$	$\theta_A/4$	—	1.0	0.260	0.105	0.055	1.0	0.355	0.080	0.0	0.0	—	1.0	0.365	0.140	0.075
$\theta_B/10$	$\theta_A/10$	—	1.0	0.155	0.0	0.0	1.0	0.240	0.0	0.0	0.0	—	1.0	0.245	0.050	0.0
$\theta_A = 0.20$ and $\theta_B = 0.25$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/8$	$\theta_A/8$	—	.1.0	1.0	1.0	1.0	—	1.0	1.0	1.0	0.635	—	1.0	1.0	1.0	1.0
$\theta_B/10$	$\theta_A/10$	—	1.0	1.0	1.0	1.0	—	1.0	1.0	1.0	0.465	—	1.0	1.0	1.0	1.0
$\theta_A = 0.30$ and $\theta_B = 0.25$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	1.0	1.0	0.825	1.0	1.0	1.0	0.520	0.415	1.0	1.0	1.0	1.0	1.0
$\theta_B/8$	$\theta_A/8$	—	1.0	1.0	0.645	0.270	1.0	1.0	0.360	0.130	0.055	—	1.0	1.0	1.0	0.380
$\theta_B/10$	$\theta_A/10$	—	1.0	1.0	0.455	0.150	1.0	1.0	0.235	0.0	0.0	—	1.0	1.0	1.0	0.250
$\theta_A = 0.40$ and $\theta_B = 0.25$																
θ_B	θ_A	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$3\theta_B/4$	$3\theta_A/4$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\theta_B/2$	$\theta_A/2$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.815	0.715	1.0	1.0	1.0	1.0	1.0
$\theta_B/4$	$\theta_A/4$	1.0	1.0	1.0	0.460	0.360	1.0	1.0	0.395	0.305	0.265	1.0	1.0	1.0	0.535	0.400
$\theta_B/8$	$\theta_A/8$	1.0	1.0	0.270	0.080	0.0	1.0	0.390	0.050	0.0	0.0	—	.1.0	0.410	0.120	0.0
$\theta_B/10$	$\theta_A/10$	—	1.0	0.155	0.0	0.0	1.0	0.260	0.0	0.0	0.0	.—	.1.0	0.265	0.0	0.0

admitting the real solution of the equation (3.12) and our solution ranges of α_B are presented in Tables 2*a* and 2*b*. Then corresponding to a particular illustrative value of $\alpha_B = 0.4$ falling in most of the solution ranges, we obtain the ranges of α_A for assured gain in efficiency of e_{S2} over e.

For numerical illustration, we use three sets of values of p_1 and p_2 and the value of T satisfying (2.5) for assured gain in efficiency in combining direct and randomized responses irrespective of how a sample of respondents is chosen. These three sets of device parameters are

(i) $p_1 = 0.3$, $p_2 = 0.45$, Range of $T \geq 0.377$ and the used value of $T = 0.4$,

(ii) $p_1 = 0.5$, $p_2 = 0.3$, Range of $T \geq 0.286$ and the used value of $T = 0.3$ and

(iii) $p_1 = 0.5$, $p_2 = 0.6$, Range of T = any value in between (0.0, 1.0) and the used value of T = 0.2.

For the sake of simplicity, we use the sample sizes as n = 10, 25, 50, 75, 100; θ_A = 0.2, 0.3, 0.4 and θ_B = 0.15, 0.25, 0.55, 0.70. We also use the prior guess values $\theta_{A0} = \theta_A$, $3\theta_A / 4$, $\theta_A / 2$, $\theta_A / 4$, $\theta_A / 8$, $\theta_A / 10$ and $\theta_{B0} = \theta_B$, $3\theta_B / 4$, $\theta_B / 2$, $\theta_B / 4$, $\theta_B / 8$, $\theta_B / 10$.

It may be observed from Tables 1a and 1b that our proposed estimator e_{S1} in (3.1) is better than the usual estimator e in (2.3) for full range of δ in most situations showing the lower bound as zero specially when the sample sizes are small (≤ 25). The range of δ decreases as the sample size increases and also as the prior guess value θ_{A0} departs much from the true value θ_A.

It is observed from Tables 2*a* and 2*b* that here also in most cases α_B attains the full range to assure the availability of real solution in (3.12) showing the lower bound as zero. The cases when sample sizes are small (≤ 25) and when the prior gueses θ_{A0} and θ_{B0} depart much from their true value, the findings are divided in two types of results: the solution being not available (presenting as — in respective boxes) in some cases and in some cases, the ranges become smaller than the full range.

Based on the results of full range of α_B from the two tables (Tables 2*a* and 2*b*), we decide to present the illustrative solution ranges of α_A in terms of upper bound, corresponding to $\alpha_B = 0.4$ in Tables 3*a* and 3b. From Tables 3*a* and 3*b*, it is clear that here also, in most situations, α_A attains the full range (showing upper bound as 1.0) to ensure the gain in efficiency of e_{S2} in (3.8) over e in (2.3). Whenever the value $\alpha_B = 0.4$ is not consistent with Tables 2*a* and 2*b*, the ouput is presented as — in respective boxes of Tables 3a and 3b. However, the ranges become smaller for higher sample sizes and for prior guess values departing much from true values. Of course, in very few cases, we do not get the admissible solution ranges of α_A and in those cases the upper bounds are presented as zero in respective boxes.

5. CONCLUDING REMARKS

We observe from the results of above numerical illustration that after profitable application of Mangat and Singh (1990)'s technique in unrelated question model, there is enough scope to choose δ to generate more efficient estimators. In practice, since the survey on sensitive questions are very expensive and time consuming, in many frequent cases, the sample sizes are kept small. In those cases and also even when the prior guess values depart much from their true values, our proposed estimator performs better than the usual estimator.

REFERENCES

Abul-Ela, Abdel-Latif, A., Greenberg, B.G. and Horvitz, D.G. (1967). A multiproportions randomized response models. *J. Amer. Statist. Assoc.,* **62**, 990-1008.

Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *J. Statist. Plann. Inf.,* **94**, 37-42.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques.* Marcel Dekker, New York.

Dowling, T.A. and Shachtman, R. (1975). On the relative efficiency of randomized response models. *J. Amer. Statist. Assoc*., **70**, 84-87.

Folsom, R.E., Greenberg, B.G., Horvitz, D.G. and Abernathy, J.R. (1973). The two alternate questions randomized response model for human surveys. *J. Amer. Statist. Assoc*., **68**, 525-530.

Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question RR model: Theoretical framework. *J. Amer. Statist. Assoc*., **64**, 520-539.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question RR model. *Proceedings of the Social Statistics Section, Amer. Statist. Assoc*., 65-72.

Lanke, J. (1975). Some contributions to the theory of survey sampling. Ph.D. thesis, University of Lund.

Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, **77(2)**, 439-442.

Mangat, N.S., Singh, R., Singh, S. (1992). An improved unrelated question randomized response strategy. *Cal. Statist. Assoc. Bull.,* **42**, 277-281.

Moor, J.J.A. (1971). Optimization of the unrelated question randomized response model. *J. Amer. Statist. Assoc*., **66**, 627-629.

Singh, S. (1994). Unrelated question randomized response sampling using continuous distribution. *J. Ind. Soc. Agril. Statist.,* **46**, 349-361.

Singh, S., Mangat, N.S. and Singh, R. (1994). On estimation of mean/total of stigmatizing quantitative variable. *Statistica*, **54**, 383-386.

Singh, H.P, Shukla, S. and Mathur, N. (2007). Shrinkage estimation of proportion of population possessing stigmatizing character in unrelated question randomized response technique. *J. Amer. Statist. Assoc*., **61(1)**, 1-13.

Thompson, J.R. (1968). Some shrinkage techniques for estimating the mean. *J. Amer. Statist. Assoc*., **63**, 113-122.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc*., **60**, 63-69.

Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 297-304

Estimation of Population and Domain Totals under Two-phase Sampling in the Presence of Non-response

Raj S. Chhikara[1*] and U.C. Sud[2]

[1]*University of Houston-Clear Lake, Houston, Texas, USA*

[2]*Indian Agricultural Statistics Research Institute, New Delhi*

(Received: July 2009, Revised: November 2009, Accepted: November 2009)

SUMMARY

This paper considers estimation of both domain and population totals for an item of interest using a two-phase sampling where the domain identity is realized, but the item response is not necessarily available from a phase I sampled unit. The optimality of sample design is studied considering the probability of item response, the cost of phase I vs. phase II sampling, and the item variability in the domains. Numerical evaluations made using a simulation study show that the proposed sampling and estimation method is more efficient than an alternative method given by Agrawal and Midha (2007).

Key words : Cost function, Domain estimation, Optimal sample design, Probability of item response.

1. INTRODUCTION

A common practice in survey sampling is not only to estimate a population mean or total for an item of interest, but also its estimate is desired at the domain level. The domain size is invariably unknown and varies across domains. When the response for a sample unit is subject to a chance mechanism, a standard sampling approach is to employ a two phase sample design as discussed in Cochran (1977, p. 370) and Sarndal, *et al*. (1992, p. 566). It consists of a single or multistage phase I sample of a fixed number of units drawn from the population. For phase II sampling, the construction of an efficient sample design would depend upon whether or not the response for a sample unit in phase I lacks completely or partially. If the sampled unit has no response at all as in mail surveys, it is a complete non-response, termed as unit non-response, whereas if the unit identity for its domain is realized, but not its value for an item, it is a partial response, termed as item non-response. In this paper we assume the latter so that in phase I the sample unit response for its domain identity is realized, but not necessarily its value for the item of interest. Next, it is assumed that the item response would be ascertained for the phase II sample units. A subset of the item non-responding phase I sample units is considered for phase II sub-sampling.

It is worth while to mention here that the problem of estimation of domain means using two-phase sampling was considered as early as 1973 by Degraft-Johnson and Sedransk. However, the necessary theory was developed under the assumption of no non-response. Kun He (1995) obtained a minimax estimator under a squared error loss function for the domain totals when the number of sample units falling into the different domains is random.

Recently, Agrawal and Midha (2007) discussed estimation of domain total using a two-phase sample design that consisted using the phase I samples to estimate the domain size and the phase II sub-sample to estimate the domain total for the item observed. Since the item values may be available for some phase I

[*]*Corresponding author* : Raj S. Chhikara
E-mail address : chhikara@uhcl.edu

sample units, as in telephone surveys, their domain estimator can be improved upon by incorporating the item responses obtained during phase I sampling. Sarndal and Swensson (1987) employed basically the same two phase sample design as presently considered. Although they considered the item responses only from the phase II sub-sample drawn from strata (fixed or not) using stratified sampling, yet they assumed to have available the unit values for an auxiliary variable at the first phase sampling. As such they proposed and discussed regression estimation for both the population and strata/domains. The response homogeneity groups were assumed for domains so as to achieve higher precision for the estimator and reduce its bias due to item non-response. However, they did not discuss the optimality of their sample design as they did not raise the issue of cost of sampling in their study.

The estimator and its variance for the domain and population totals can be stated in terms of inclusion probabilities for the two phases of sampling under a general framework as given in Sarndal *et al*. (1992); but we skip it. Instead, we focus on the estimators and their variances for the case of simple random sampling without replacement, which is the frequently employed sampling method in sample surveys.

The two phase sample design and the estimation of domain totals are discussed in the next section. The optimality of phase II sample design is also developed. The estimator for population total as a sum of the domain total estimators is utilized to optimize phase I sample size and is described in Section 3. This estimator of the population total is contrasted with the standard estimator obtained directly without considering the domains as given in Sarndal *et al*. (1992) in the context of non-response. A simulation study is made to evaluate the properties of the proposed estimators of domain and population totals, and then these are compared with those given in Agrawal and Midha (2007).

2. DOMAIN ESTIMATION

2.1 Two-Phase Sampling Scheme

Consider a partitioning of the population U into sub-sets, $U_1,\ldots,U_d,\ldots,U_D$, called domains. Let N_d be the size of U_d, where N_d is assumed unknown, and the domain identity of units is also not known a priori. We have the partitioning equations

$$U = \bigcup_{d=1}^{D} U_d \ , \ N = \sum_{d=1}^{D} N_d$$

Let y denote the variate for the item of interest. The objective is to estimate the domain totals, $t_d = \sum_{i}^{N_d} y_{id}$, $d = 1,\ldots., D$, or the domain means $\bar{Y}_d = t_d / N_d$, $d = 1,\ldots., D$, as well as the population total $t = \sum_{1}^{D} t_d$ or mean $\bar{Y} = t/N$.

Assume that a survey is carried out for the population U where a sample s of size n is drawn from U according to SRSWOR sampling design. We assume that the selected unit is subject to item non-response; however the unit response in terms of its domain identity is obtained. The sample of size n is post-stratified into D domains on the basis of the unit domain identity as observed. Let s_d denote the part of s that happens to fall in U_d, that is, $s_d = s \cap U_d$. Denote by n_d the size of s_d. Hence

$$s = \bigcup_{d=1}^{D} s_d \ ; \ n = \sum_{d=1}^{D} n_d$$

Here n_d is random. We need to avoid for n_d to be quite small. As such a large sample of size n may have to be drawn if D is not small. Let n_{1d} of n_d units falling in the d-th domain respond for the item of interests, while the remaining n_{2d} units do not respond, $n_d = n_{1d} + n_{2d}$. Further, we assume that the item response set s_{d1} is generated as a result of n_d independent Bernoulli trials, one for each element k in s_d, with constant probability θ_d of 'success', i.e., item response. Thus for any $k \in s_d$, and every k and $l \in s_d$, $Pr(k \in s_{d1}/s_d) = \theta_d$; $Pr(k \ \& \ l \in s_{d1}/s_d) = \theta_d^2$.

For phase II sampling, a stratified sample is considered, where from the d-th domain ($d = 1, 2, \ldots, D$) a random sub-sample of size m_{2d} out of n_{2d} non-responding units is randomly drawn, and the responses are obtained for all m_{2d} sampled units through specialized efforts. Thus, the cost of obtaining response

in the second phase is expected to be much more than that in the first phase. With this limitation for the phase II sample of observations, the statistician's task is to minimize the phase II sampling, and to make the best possible use of the sample observations to estimate the domain totals or means.

2.2 Estimator of t_d

Define

$$y_i = \begin{cases} y_{id} & \text{if } i \in U_d \\ 0 & \text{otherwise} \end{cases}$$

Let $\bar{y}_{n_{1d}} = \dfrac{\sum^{n_{1d}} y_{id}}{n_{1d}}$ and $\bar{y}_{m_{2d}} = \dfrac{\sum^{m_{2d}} y_{id}}{m_{2d}}$

Then we propose

$$\hat{t}_d = \frac{N}{n}\left[n_{1d}\bar{y}_{n_{1d}} + n_{2d}\bar{y}_{m_{2d}} \right] \qquad (2.1)$$

for an estimator of the item total for the d-th domain, d=1, 2,…,D.

Remark: The response mechanism is sometimes considered deterministic in the following way. The population consists of two groups U_1 and U_2. All elements in U_1 respond with probability 1 if selected, whereas for all elements in U_2 the response probability is 0. Thus the composition of the groups is fixed, once for all. This case has been dealt with by Sud *et al*. (2009).

The estimator $\hat{t}_d$ is subjected to variability due to (1) the first phase sampling of the population and the number of sample units falling in domain d, (2) the sample units drawn from the domain (3) the chance mechanism for item response for a unit as modeled by Bernoulli sampling, and (4) SRSWOR for the second phase sampling in each domain. Let E_1, E_2, E_3 and E_4 denote expectations with conditionals in the order of variability stated above. Then it easily follows that $\hat{t}_d$ is an unbiased estimator of t_d since

$$E(\hat{t}_d) = \frac{N}{n}E_1E_2E_3\left(n_{1d}\bar{y}_{n_{1d}} + n_{2d}\bar{y}_{n_{2d}}\right)$$

$$= \frac{N}{n}E_1E_2\left(n_d\bar{y}_{n_d}\right) = N_d\bar{Y}_d$$

To compute the variance of $\hat{t}_d$, denote here V_1, V_2, V_3 and V_4 to be the variances with proper prior conditionals in the variability due to (2)-(4) as stated above. Then the variance of $\hat{t}_d$ can be obtained by considering

$$V(\hat{t}_d) = V_1E_2E_3E_4(\hat{t}_d) + E_1V_2E_3E_4(\hat{t}_d) + E_1E_2V_3E_4(\hat{t}_d) + E_1E_2E_3V_4(\hat{t}_d)$$

Let

$$P_d = \frac{N_d}{N};\ Q_d = 1 - P_d;\ f_{2d} = \frac{m_{2d}}{n_{2d}}$$

$$S_d^2 = \frac{1}{(N_d - 1)}\sum_i^{N_d}\left(y_i - \bar{Y}_d\right)^2$$

$$S_{n_{2d}}^2 = \frac{1}{(n_{2d} - 1)}\sum_i^{n_{2d}}\left(y_i - \bar{y}_{n_{2d}}\right)^2$$

$$S_{n_d}^2 = \frac{1}{(n_d - 1)}\sum_i^{n_d}\left(y_i - \bar{y}_{n_d}\right)^2$$

It follows that

$$E_1E_2E_3V_4(\hat{t}_d) = \frac{N^2}{n^2}E_1E_2E_3\left(n_{2d}\left(\frac{1}{f_{2d}} - 1\right)S_{n_{2d}}^2\right)$$

$$= \frac{N^2}{n^2}(1-\theta_d)E_1E_2\left(n_d\left(\frac{1}{f_{2d}} - 1\right)S_{n_d}^2\right)$$

$$= \frac{N^2}{n}P_d(1-\theta_d)\left(n_d\left(\frac{1}{f_{2d}} - 1\right)S_d^2\right)$$

$$E_1E_2V_3E_4(\hat{t}_d) = \frac{N^2}{n^2}E_1E_2V_3\left(n_d\bar{y}_d\right) = 0$$

$$E_1V_2E_3E_4(\hat{t}_d) = \frac{N^2}{n^2}E_1V_2\left(n_d\bar{y}_d\right)$$

$$= \frac{N^2}{n^2}E_1\left[n_d\left(\frac{N_d - n_d}{N_d - 1}\right)P_dS_d^2\right]$$

$$= \frac{N^2}{n}\left[1 - \frac{(n-1)}{(N_d - 1)}P_d\right]P_dS_d^2$$

$$V_1E_2E_3E_4(\hat{t}_d) = \frac{N^2}{n}V_1E_2\left(n_d\bar{y}_d\right)$$

$$= \frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_dQ_d\bar{Y}_d^2$$

Adding the above four terms and simplifying we get

$$V(\hat{t}_d) = \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+Q_d\left(\frac{N-n}{N-1}\right)\left(\frac{1}{\left(CV_d\right)^2}-\frac{1}{N_d}\right)\right.$$

$$\left.+\left(1-\theta_d\right)\left(\frac{1}{f_{2d}}-1\right)\right]P_dS_d^2 \qquad (2.2)$$

where, $CV_d = S_d/\bar{Y}_d$ is the coefficient variation for the d-domain.

A variance estimator of $V(\hat{t}_d)$ is obtained by

$$\hat{V}(\hat{t}_d) = \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+q_d\left(\frac{N-n}{N-1}\right)\left[\frac{1}{\left(cv_d\right)^2}-\frac{n}{Nn_d}\right]\right.$$

$$\left.+(1-\hat{\theta}_d)\left(\frac{1}{f_{2d}}-1\right)\right]\left(\frac{n_d}{n}\right)s_d^2 \qquad (2.3)$$

Where $q_d = 1 - \frac{n_d}{n}$, $cv_d = \frac{s_d}{\bar{y}_d}$

$$s_d^2 = \frac{1}{(n_{1d}+m_{2d})-1}\sum_{1}^{n_{1d}+m_{2d}}(y_i-\bar{y}_d)^2$$

$$\bar{y}_d = \frac{1}{(n_{1d}+m_{2d})}\sum_{1}^{n_{1d}+m_{2d}}y_i \text{ and } \hat{\theta}_d = \frac{n_{1d}}{n_d}$$

If N and N_d are large, ignoring terms of order $1/N$ and $1/N_d$, $V(\hat{t}_d)$ is approximately equal to

$$V(\hat{t}_d) \sim \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+Q_d\left(\frac{N-n}{N}\right)\left(\frac{1}{\left(CV_d\right)^2}\right)\right.$$

$$\left.+(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)\right]P_dS_d^2 \qquad (2.4)$$

A variance estimate can then be obtained by replacing the population quantities by their sample estimate as done in (2.3).

2.3 Optimization under a Cost Function

For domain d, consider the sample unit cost c_{1d} for the response at phase I and c_{2d} at phase II, where $c_{2d} >> c_{1d}$, possibly dominated by a large multiple. Then the cost function for domain d, except for a fixed overall cost, is

$$C_d = n_d c_{1d} + m_{2d}c_{2d}$$

Then its expected cost is given by

$$E(C_d) = nP_d[c_{1d} + f_{2d}(1-\theta_d)c_{2d}] \qquad (2.5)$$

The optimum value of f_{2d} is obtained by fixing the variance of $\hat{t}_d$ say equal to $N_d^2V_{0d}$ and minimizing the expected cost. This can be determined by minimizing the function

$$\phi = nP_d[c_{1d} + f_{2d}(1-\theta_d)c_{2d}]$$

$$+\lambda\left[\frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+Q_d\left(\frac{N-n}{N}\right)\left(\frac{1}{(CV_d)^2}\right)\right.\right.$$

$$\left.\left.+(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)\right]P_dS_d^2 - N_d^2V_{0d}\right]$$

where λ is the Lagrangian multiplier.

Differentiating with respect to n, λ, f_{2d}, equating the resultant equations to 0 and solving for f_{2d} gives the optimum value of f_{2d}, after ignoring $(1/N_d)$ term, as

$$f_{2d(\text{opt})} = \sqrt{\frac{c_{1d}/c_{2d}}{\theta_d + \frac{Q_d}{(CV_d)^2}}} \qquad (2.6)$$

To determine an optimum value of n one has to look at the objective of the sample survey. If the objective is simply to achieve a desired level of precision for the estimator of domain total or mean, then we need to determine n so that $V(\hat{t}_d) \le N_d^2V_{0d}$ as considered in the determination of $f_{2d(\text{opt})}$. The optimum value of n, say $n^{(d)}$, is given by

$$n^{(d)} = \frac{\left[N^2\left[(1-\theta_d)\left(\frac{1}{f_{2d(opt)}}-1\right)\right.\right.}{\left.\left.+1-N\frac{P_d}{Q_d}+Q_d(CV_d)^{-2}\right]\right]}$$

$$\frac{}{\left[N^2\frac{V_{0d}}{P_dS_d^2}+N-\frac{Q_d}{P_d}+NQ_d(CV_d)^{-2}\right]} \qquad (2.7)$$

Eq.(2.7) leads to an optimal value for n corresponding to each domain. One might choose $n = \max(n^{(d)})$ for the optimum phase I sample size. However, it is more appropriate and hence preferable to optimize the phase I sample size so that the estimator of the population total has a desired precision. The variance of the estimator of population total is discussed in the next section and so is determination of n.

3. ESTIMATOR OF POPULATION TOTAL

Theorem. The estimator

$$\hat{t} = \sum_1^D \hat{t}_d = \frac{N}{n}\sum_1^D \left[n_{1d}\bar{y}_{n_{1d}} + n_{2d}\bar{y}_{m_{2d}} \right]$$

is an unbiased estimator of population total t with variance

$$V(\hat{t}_d) = N^2 \frac{S^2}{n}\left(1-\frac{n}{N}\right) + \frac{N^2}{n}\sum_1^D P_d\left(1-\theta_d\right)\left(\frac{1}{f_{2d}}-1\right)S_d^2 \quad (3.1)$$

where

$$S^2 = \frac{1}{(N-1)}\left[\sum_1^D N_d(\bar{Y}_d-\bar{Y})^2 + \sum_1^D (N_d-1)S_d^2\right]$$

$$\bar{Y} = \sum_1^D P_d\bar{Y}_d$$

Proof. From Section 2, $E_1E_2E_3E_4(\hat{t}_d) = t_d$. Therefore,

$$E_1E_2E_3E_4(\hat{t}) = \sum_1^D t_d = t$$

Next,

$$V(\hat{t}_d) = V_1E_2E_3E_4(\hat{t}) + E_1V_2E_3E_4(\hat{t}) + E_1E_2V_3E_4(\hat{t}) + E_1E_2E_3V_4(\hat{t})$$

where

$$E_1E_2E_3V_4(\hat{t}) = E_1E_2E_3\frac{N^2}{n^2}\sum_1^D n_{2d}\left(\frac{1}{f_{2d}}-1\right)S^2_{n_{2d}}$$

$$= \frac{N^2}{n^2}E_1E_2\sum_1^D(1-\theta_d)(n_d)\left(\frac{1}{f_{2d}}-1\right)S^2_{n_d}$$

$$= \frac{N^2}{n}\sum_1^D P_d(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)S_d^2$$

From Section 2, both $E_1E_2V_3E_4(\hat{t})$, $E_1V_2E_3E_4(\hat{t})$, can be shown to be equal to 0.

$$V_1E_2E_3E_4(\hat{t}) = V_1\frac{N}{n}\sum_1^D\left(n_d\bar{y}_{n_d}\right)$$

$$= V_1(N\bar{y})$$

$$= \frac{N}{n}\left(1-\frac{N}{n}\right)S^2$$

Here $\bar{y} = \dfrac{\sum_1^D n_d\bar{y}_{n_d}}{n}$

Adding the terms we obtain

$$V(\hat{t}) = N^2\frac{S^2}{n}\left(1-\frac{n}{N}\right) + \frac{N^2}{n}\sum_1^D P_d(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)S_d^2 \quad (3.1)$$

The optimum value of n is obtained by fixing the variance of $\hat{t}$, say, equal to N^2V_0, where V_0 is pre-specified. Thus, the optimum value of n, ignoring $1/N$ terms, is given by

$$n_{opt} = \frac{\left[S^2 + \sum_1^D P_d(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)S_d^2\right]}{V_0} \quad (3.2)$$

Next, a variance estimator is obtained by replacing the population quantities in (3.1) by the corresponding sample statistics

$$\hat{V}(\hat{t}) = N^2\frac{s^2}{n}\left(1-\frac{n}{N}\right) + \frac{N^2}{n}\sum_1^D p_d\left(1-\hat{\theta}_d\right)\left(\frac{1}{f_{2d}}-1\right)s_d^2$$

where

$$s^2 = \frac{1}{(n-1)}\left[\sum_1^D(n_{1d}+m_{2d})(\bar{y}_d-\bar{y})^2 + \sum_1^D(n_{1d}+m_{2d})s_d^2\right]$$

$$\bar{y} = \frac{1}{\sum_1^D(n_{1d}+m_{2d})}\sum_1^{n_{1d}}\sum_1^{m_{2d}} y_i$$

Alternatively, one may utilize the following unbiased estimator of t as given in Sarndal *et al*. (1992) obtained without consideration of domains

$$\hat{t}_0 = \frac{N}{n}\left[n_1\bar{y}_1 + n_2\bar{y}_{m_2}\right] \quad (3.4)$$

where

$$n_1 = \sum_1^D n_{1d}\ ,\ n_2 = n - n_1$$

$$m_2 = \sum_1^D m_{2d} \text{ and } \bar{y}_1 = \frac{1}{n_1}\sum_i^{n_1} y_{1i}$$

and $$\bar{y}_{m_2} = \frac{1}{m_2}\sum_i^{m_2} y_{2i}$$

Here y_{1i} (i = 1, 2, ..., n_1) and y_{2i} (i = 1, 2, ..., m_2) represent the observed values of y from the responses of units obtained using SRSWOR in phase I and II, respectively. Note that it ignores the partition of the population and hence the sample distribution by domain. Next, the variance of $\hat{t}_0$ is given by

$$V(\hat{t}_0) = \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+(1-\theta)\left(\frac{1}{f_2}-1\right)\right]S^2 \quad (3.5)$$

where, $\theta = \sum_d^D P_d\theta_d$ and S^2 is the population variance. Comparing it with the variance given in Eq (3.1) and ignoring the first term which is common in both, we have $V(\hat{t}_0) > V(\hat{t})$ unless the domain partition of the population is completely random, in which case the two variances are equal.

Another alternative estimator is to utilize the one proposed by Agrawal and Midha (2007) as given by

$$\hat{t}^* = \sum_1^D \hat{t}_d^*$$

$$\hat{t}_d^* = \frac{N}{n} n_d \bar{y}_d \quad (3.6)$$

where

$$\bar{y}_d = \frac{1}{m_d}\sum_i^{m_d} y_{id} \quad (3.7)$$

Here in Eq (3.7), m_d denotes the number of sample units falling in domain d from the phase II sampling of $m' = n(m_2/n_2)$ out of n phase I sample units, drawn using SRSWOR. Note that $m' \geq m_2$, and so a larger sample may be required in phase II under this approach, resulting an increase in the cost of sampling. It assumes that n_d will be known from phase I sampling; however it does not exploit the known identity of phase I samples in selecting phase II samples by domain as well as the use of item responses that might be made available for some of the phase I sampled units, as for example in telephone surveys. Clearly, the estimator in (3.6) is unbiased; however, we demonstrate later using simulations that for the population total estimators, $V(\hat{t}^*) > V(\hat{t})$.

4. OPTIMAL SAMPLE DESIGN

Optimality for phase I and phase II sampling was investigated considering several populations with different domain configurations as listed in Table 1. The sample size n and the sampling fraction f_{2d} (d = 1, 2, 3) were found mainly influenced by the item response rate, the cost of phase I sampling relative to phase II sampling and the variability for the domains. Considering the cases of low, medium and high response probability θ_d, and a low value for the cost ratio, c_{1d}/c_{2d}, and a low to high coefficient of variation, CV_d, d = 1, 2, 3, the following conclusions were drawn: n_{opt} and $f_{2d,opt}$ (d =1, 2, 3) increase as θ_d (d = 1, 2, 3) decreases; $f_{2d,opt}$ increases as c_{1d}/c_{2d}

Table 1. Parametric values and optimal phase II sampling fractions for domains

Case 1						
d	N_d	$\bar{Y}_d$	S_d	θ_d	c_{1d}/c_{2d}	Sampling fraction, f_{2d}
1	2500	10	5	0.2	0.1	0.125
2	5000	20	15	0.5	0.1	0.268
3	2500	30	10	0.8	0.1	0.163
Case 2						
1	1500	10	5	0.1	0.1	0.169
2	3000	20	15	0.3	0.2	0.360
3	5500	30	10	0.6	0.3	0.254
Case 3						
1	4000	10	5	0.4	0.1	0.189
2	3000	20	15	0.6	0.2	0.329
3	3000	30	10	0.8	0.3	0.206

(d = 1, 2, 3) increases and as CV_d (d = 1, 2, 3) increases; the increase in $f_{2d,opt}$ is approximately proportional to CV_d (d = 1, 2, 3) and thus CV_d could substantially influence $f_{2d,opt}$ than perhaps the cost ratio c_{1d}/c_{2d} (d = 1, 2, 3).

The optimal values of the sampling fraction f_{2d} computed for the domains using Eq. (2.5) are listed in the last column of Table 1 that describes the three population cases considered. For the population, we have N = 10,000, $\bar{Y}$ = 20 and S = 13.32. The overall probability of response θ= 0.5, 0.435, 0.58 for the three cases considered. Letting the desired precision with CV = 0.05 in estimation of population total, the corresponding optimal phase I sample sizes computed from Eq (3.2) are given as n_{opt} = 409, 240, 311.

5. NUMERICAL EVALUATION USING SIMULATIONS

A simulation study was made to evaluate $\hat{t}_d$ (d = 1, 2, 3) and $\hat{t}$ for the three population cases considered above in Section 4. The simulations were performed using R package. For the three domains, item values for the units were simulated with $\bar{Y}_d$ and S_d (d = 1, 2, 3) as in Table 1. The normal distribution was used in generating the values for each domain. The domain means and standard deviations were chosen to have a substantial overlap between the middle domain and each of the remaining two. This is to reflect the most likely situation when the domains are considered in an ascending order of their means.

In each case of phase I sample size n = 409, 240, 311, sample units were drawn using SRSWOR. It was then followed by stratified sampling at phase II as described using the sample sizes, m_{2d}, d = 1, 2, 3, resulting from the sampling fractions given in Table 1 and the phase I non-response samples that were in domain d (d = 1, 2, 3). The estimates, $\hat{t}_d$ (d = 1, 2, 3) and $\hat{t}$ were computed for each simulation. The sampling and estimation process was repeated 2000 times. The empirical average, variance and root mean square error (RMSE) were determined for $\hat{t}_d$ and $\hat{t}$ from the 2000 estimates.

To compare the proposed estimator $\hat{t}$ to $\hat{t}_0$, the total phase II sample of size $m_2 = \sum_d m_{2d}$, was drawn using SRSWOR for computation of $\hat{t}_0$. Similarly, $\hat{t}^*$ was computed using the comparable phase II sample size of $m' = m_2\left(\frac{n}{n_2}\right)$ needed for the implementation of the Agrawal and Midha's (2007) SRSWOR sampling scheme at phase II. Agrawal and Midha (2007) provide estimates primarily at the domain level denoted by $\hat{t}_d^*$. The coefficient of variation was computed for each estimator from dividing the observed RMSE by the actual value of the corresponding total. The numerical results obtained are presented in Table 2 for the three cases discussed earlier.

These results show that the proposed estimator has much lower CV than the other two alternative estimators of the population total. Note that the CV for the proposed estimator is in close agreement with the specified CV= 0.05. For the domains, it consistently has smaller CV and hence performs better than the domain estimator of Agrawal and Midha (2007). One exception that occurs is for Domain 1 in Case 2. The reason for it is that this domain is relatively small and also has a much lower variability for the item value than the other two domains and hence, it gets allocated much fewer samples in phase II under the optimal subsampling than the sampling considered by Agrawal and Midha. Moreover, the overall sampling cost is expected to be higher under their approach, particularly if the item response in phase I sampling is moderate to high or the cost of sampling in phase II relative to that in phase I is high.

Table 2. CVs for Domain and Population Total Estimators

Case 1					
Domain	$CV(\hat{t}_d)$	$CV(\hat{t}_d^*)$	$CV(\hat{t})$	$CV(\hat{t}^*)$	$CV(\hat{t}_0)$
1	0.1366	0.1637	0.0489	0.1150	0.3427
2	0.0940	0.1147			
3	0.1001	0.1787			
Case 2					
1	0.2773	0.1726	0.0545	0.1119	0.2083
2	0.0877	0.1239			
3	0.1229	0.1864			
Case 3					
1	0.1565	0.1810	0.0532	0.1238	0.3673
2	0.1045	0.1211			
3	0.1097	0.1985			

ACKNOWLEDGEMENTS

The research work of Dr. Raj S. Chhikara was conducted during his visit to IASRI, New Delhi, while he was on Faculty Development Leave from the University of Houston-Clear Lake and was partially supported by research funding from the National Agricultural Statistics Service of the U.S. Department of Agriculture, Washington, D.C.

The authors are grateful to the referee for constructive suggestions which led to improvement in the paper.

REFERENCES

Agrawal, M.C. and Midha, C.K. (2007). Some efficient estimators of the domain parameters. *Statist. Probab. Lett*., **77**, 704-709.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd ed. Wiley, New York.

Degraft-Johnson, K.T. (1973). Estimation of domain means using two-phase sampling. *Biometrika*, **60(2)**, 387-393.

Kun He. (1995). On estimating domain totals over a subpopulation. *Ann. Inst. Statist. Math.*, **47(4)**, 637-643.

Rao, P.S.R.S. (2000). *Sampling Methodologies with Applications*. Chapman & Hall/CRC, New York.

Sarndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *Internat. Statist. Rev*., **55**, 279-294.

Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York.

Sud, U.C., Chandra, Hukum and Chhikara, Raj S. (2009). Domain estimation in the presence of non-response. Paper submitted to *J. Ind. Soc. Agril. Statist*. for publication.

Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 305-310

Optimum Designs for Stress Strength Reliability

Manisha Pal* **and N.K. Mandal**
Calcutta University, Kolkata

(Received: August 2009, Revised: November 2009, Accepted: December 2009)

SUMMARY

In a stress-strength model, the reliability of a system is measured by the probability of the strength of the system exceeding the environmental stress. As the strength of the system depends on a number of controllable factors, the accuracy in estimating the reliability can be enhanced by a proper choice of these factors. In this paper, we have assumed stress and strength to be independent exponential variables, with mean strength being a function of the controllable factors. The optimum design for the estimation of the system reliability has been proposed using suitable optimality criterion.

Key words : Stress-strength reliability, Exponential distribution, Optimum design, Central composite design.

1. INTRODUCTION

According to the simple stress-strength model of failure, a system fails when and only when the applied stress exceeds its strength. Since the stress is a function of the environment to which the system is subjected, it can be regarded as a random variable. Similarly, the strength of a system that is mass-produced depends on a number of factors, some of which are controllable, like material properties, manufacturing procedures etc., while the others are uncontrollable, and hence can also be treated as random. In such a stochastic environment the designer is interested in the reliability of the system subject to a stress. However, since the distributions of stress and strength are generally unknown, either fully or partially, estimation of reliability of the system becomes important. In estimating reliability, one may take into account the factors influencing stress and strength. To date, considerably little work has been done along this line, see Guttman *et al*. (1988), Weerahandi and Johnson (1990), Guttman and Papandonatos (1997). However, the existing literature does not include any study on choice of the controllable factors affecting the strength so as to improve the accuracy in estimation.

In this paper, we attempt to find an optimum continuous design for the controllable factors $\mathbf{z} = (z_1, z_2, \ldots, z_p)'$ that should be used in the experiment to achieve maximum accuracy in estimation of system reliability. We assume stress and strength to have independent exponential distributions, with the expected strength being a function of $\mathbf{z}$. In Section 2, we derive the asymptotic variance of the maximum likelihood estimate (MLE) of the system reliability as a function of the design; in Section 3 the optimum design is derived using suitable criteria, and a discussion, including applications, is given in Section 4.

2. MLE OF THE SYSTEM RELIABILITY AND ITS ASYMPTOTIC VARIANCE

Let X and Y denote the stress and strength, respectively, which are assumed to be independently distributed with density functions $f(x)$ and $g(y)$, and means μ_X and μ_Y. We assume $\mu_Y = \eta(\mathbf{z}; \beta)$, a function of the controllable factors $\mathbf{z} = (z_1, z_2, \ldots, z_p)'$, where β denotes the vector of regression coefficients associated with $\mathbf{z}$.

* *Corresponding author* : Manisha Pal
E-mail address : manishapal2@gmail.com

Then, the reliability of the system is given by

$$R(\mathbf{z};\, \mu_X,\, \beta) = P\,(X < Y \mid z) \tag{2.1}$$

For μ_X unknown, we may find its MLE $\hat{\mu}_X$, based on a random sample of stresses ($X_1, X_2, \ldots, X_m$). Similarly, to obtain an estimate of μ_Y, we consider a random sample of strengths ($Y_1, Y_2, \ldots, Y_n$), associated with a n-point continuous design

$$\xi = \begin{Bmatrix} z_1,\ z_2,\ \ldots,\ z_n \\ w_1, w_2, \ldots, w_n \end{Bmatrix}$$

in the domain Ξ of $\mathbf{z}$, where $z_1, z_2, \ldots, z_n$ are the design points with weights $w_1, w_2, \ldots, w_n$, respectively, $w_i \geq 0$, $i = 1(1)n$, $\sum_{i=1}^{n} w_i = 1$.

Based on this, we obtain the MLE $\hat{\beta}$ of β. The MLE of reliability is, therefore, $\hat{R}(\mathbf{z}) = R(\mathbf{z}; \hat{\mu}_X, \hat{\beta})$.

The asymptotic variance of $\hat{R}(\mathbf{z})$ is given by

$$\tilde{V}\,(\hat{R}(\mathbf{z})) = A'(\mu_X, \beta, \mathbf{z})\mathbf{I}^{-1}(\mu_X, \beta; \xi)\, A(\mu_X, \beta, \mathbf{z})$$

where

$$A'(\mu_X, \beta, \mathbf{z}) = \left(\frac{\partial}{\partial \mu_X} R(\mathbf{z}; \mu_X, \beta), \frac{\partial}{\partial \beta'} R(\mathbf{z}; \mu_X, \beta) \right)$$

and

$$\mathbf{I}(\mu_X,\, \beta,\, \xi) = \begin{pmatrix} \dfrac{1}{\mathrm{var}(\hat{\mu}_X)} & 0 \\ 0 & M(\xi) \end{pmatrix}$$

is the information matrix of (μ_X, β), with $\mathbf{M}(\xi)$ being the information matrix of β, given ξ, where information matrix is the variance of the score function, which is the gradient of the log-likelihood function.

We shall assume X and Y to follow independent exponential distributions having densities

$$f(x) = \frac{1}{\mu_X} e^{-\frac{x}{\mu_X}},\ x \geq 0,\ \mu_X > 0$$

$$g(y) = \frac{1}{\mu_Y} e^{-\frac{y}{\mu_Y}},\quad y \geq 0,\ \mu_Y > 0,$$

$$\mu_Y = \eta\,(\mathbf{z},\, \beta)$$

Then,

$$R(\mathbf{z};\, \mu_X,\, \beta) = \frac{1}{1 + \dfrac{\mu_X}{\eta(\mathbf{z};\beta)}} \tag{2.2}$$

and

$$\hat{R}(z) = \frac{1}{1 + \dfrac{\hat{\mu}_X}{\eta(\mathbf{z};\hat{\beta})}} \tag{2.3}$$

Let us assume $\eta\,(\mathbf{z};\, \beta)$ to be a polynomial of certain degree in the elements of $\mathbf{z}$ and write $\eta\,(\mathbf{z};\, \beta) = h'(\mathbf{z})\beta$, where $h(\mathbf{z})$ is a column vector involving powers of z_i's and their products.

Then,

$$\tilde{V}\,(\hat{R}(\mathbf{z})) = \frac{\mu_X^2}{[\ \eta(\mathbf{z};\beta)]^4 \left(1 + \dfrac{\mu_X}{\eta(\mathbf{z};\beta)}\right)^4} \left[\frac{\{\eta(\mathbf{z};\, \beta)\}^2}{m} + h'(\mathbf{z})\mathbf{M}^{-1}(\xi)h(\mathbf{z}) \right]$$

$$= a(\mathbf{z}) + c(\mathbf{z})\ b(\mathbf{z},\, \mathbf{M}) \tag{2.4}$$

where

$$\mathbf{M}(\xi) = \sum_{i=1}^{n} w_i h(z_i) h'(z_i) \tag{2.5}$$

$$a(\mathbf{z}) = \frac{\mu_X^2}{m\{\eta(\mathbf{z};\, \beta)\}^2 \left(1 + \dfrac{\mu_X}{\eta(\mathbf{z};\, \beta)}\right)^4}$$

$$b(z,\, I) = h'(z)\mathbf{M}^{-1}\,(\xi)\ h(\mathbf{z})$$

$$c(\mathbf{z}) = \frac{\mu_X^2}{\{\eta(\mathbf{z};\, \beta)\}^4 \left(1 + \dfrac{\mu_X}{\eta(\mathbf{z};\, \beta)}\right)^4} \tag{2.6}$$

As expected, the asymptotic variance of $\hat{R}(\mathbf{z})$ is a function of the unknown parameters and $\mathbf{z}$. However, $\tilde{V}(\hat{R}(\mathbf{z}))$ depends on the design ξ only through $b(\mathbf{z}, \mathbf{M})$. To tackle this problem, we have used a simplified criterion of minimizing $b(\mathbf{z}, \mathbf{M})$ by a proper choice of design.

3. OPTIMUM DESIGNS

In this section, we attempt to find the optimum design for the estimation of $R(\mathbf{z}; \mu_X, \beta)$ by minimizing $b(\mathbf{z}, \mathbf{M})$, for the cases where $\eta(\mathbf{z}; \beta)$ is linear and quadratic in $\mathbf{z}$, respectively.

We shall assume the domain of $\mathbf{z}$ to be $\mathbf{Z} = \{\mathbf{z}: \mathbf{z}'\mathbf{z} \le p\}$. With this assumption, the vertices of the symmetrized unit cube, $[-1; 1]^p$ come to lie on the boundary of the sphere $\mathbf{z}'\mathbf{z} \le p$. This is the appropriate generalization of the experimental domain [–1; 1] for the case of a single factor (cf. Pukelsheim 1993).

Case 1: Linear Regression

Let us assume

$$\eta(\mathbf{z}; \beta) = \beta_0 + \sum_{i=1}^{p} \beta_i z_i \qquad (3.1)$$

Here, $h'(\mathbf{z}) = (1, \mathbf{z}')$
so that $h'(\mathbf{z})\, h(\mathbf{z}) = 1 + \mathbf{z}'\mathbf{z}$

Now, under linear regression, for a design x the moment matrix is given by

$$\mathbf{M}(\xi) = \begin{pmatrix} 1 & [1] & [2] & \ldots & [p] \\ & [11] & [12] & \ldots & [1p] \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & & [pp] \end{pmatrix}$$

with

$$[i] = \sum_{k=1}^{n} w_k z_{ik}\,, \; [i\,j] = \sum_{k=1}^{n} w_k z_{ik} z_{jk}$$

where z_{ik} is the i-th element of the k-th design point z_k.

Since $b(\mathbf{z}, \mathbf{M})$ depends on $\mathbf{z}$, to find the optimum design, one approach would be to minimize $\max_{\mathbf{z} \in Z} b(\mathbf{z}, \mathbf{M})$ with respect to the design. But, by the Equivalence Theorem (cf. Kiefer and Wolfowitz 1960), this amounts to maximizing the determinant of the information matrix for the linear model (3.1) above. Now, the problem of maximizing the determinant of $\mathbf{M}(\xi)$ is invariant with respect to permutations and sign changes of the factors. Hence, a D-optimum design will be invariant, and we may therefore confine our attention to the class of invariant designs (cf. Pal and Mandal 2008). Further, we may make use of the fact that for the factor space $Z = \{\mathbf{z}: \mathbf{z}'\mathbf{z} \le p\}$, a design ξ whose moment matrix has the form

$$\mathbf{M}(\xi) = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{I}_p \end{bmatrix}$$

is Kiefer optimal (cf. Pukelsheim 1993) and hence is necessarily D-optimal. Such a design can always be constructed with the help of a Hadamard matrix. Similar dominance results are also available in Liski *et al*. (2002).

Example. Let $p = 3$. Consider a Hadamard matrix of order 4 in the standard form:

$$\mathbf{H}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} = (1\ D)$$

Since, $\mathbf{H}_4'\, \mathbf{H}_4 = 4\, \mathbf{I}_4$, the moment matrix of the optimum design, which puts equal weights ¼ at the four design points $(1, 1, 1)'$, $(-1, 1, -1)'$, $(1, -1, -1)'$ and $(-1, -1, 1)'$, is given by $\mathbf{M} = \mathbf{I}_4$ and hence is optimum.

Case II: Quadratic Regression

Let

$$\eta(\mathbf{z}; \beta) = \beta_0 + \sum_{i=1}^{p} \beta_i z_i + \sum_{i=1}^{p} \sum_{j=1}^{p} \beta_{ij} z_i z_j \qquad (3.2)$$
$$\beta_{ij} = \beta_{ji}, \; \forall i \ne j$$

This is a non-singular model.

Here, let us write

$$\beta = (\beta_0, \beta_{11}, \beta_{22}, \ldots, \beta_{12}, \beta_{13}, \ldots, \beta_{p-1,\,p}, \beta_1, \beta_2, \ldots, \beta_p)'$$

$$h(z) = (1,\ z_1^2, z_2^2, \ldots, z_p^2, z_1 z_2,\ z_1 z_3,\ \ldots,\ z_{p-1} z_p,\ z_1, z_2, \ldots, z_p)'$$

To find the optimum design using the minimax criterion considered in Case I, we again take help of the celebrated *Equivalence Theorem* (cf. Kiefer and Wolfowitz 1960) which states that the problem of minimizing $\max_{\mathbf{z}\in Z} h'(\mathbf{z})\mathbf{M}(\xi)^{-1} h(\mathbf{z})$ w.r.t. the design ξ is equivalent to maximizing $|\mathbf{M}(\xi)|$ w.r.t. ξ, where $|\mathbf{M}(\xi)|$ is the determinant of the moment matrix $\mathbf{M}(\xi)$.

It may be noted that the problem of maximizing $|\mathbf{M}(\xi)|$ remains invariant under

(a) permutation of the co-ordinates of $\mathbf{z}$;

(b) sign change of co-ordinates of $\mathbf{z}$.

Let $\mathfrak{I}$ denote the class of all permutation matrices and sign change matrices of order $p \times p$.

Let ξ^T be the design obtained from ξ by virtue of the transformation $\mathbf{z} \to T\mathbf{z}$, $T \varepsilon \mathfrak{I}$, and let,

$$\bar{\mathbf{M}} = \frac{1}{t}\sum_{T\in\mathfrak{I}} \mathbf{M}(\xi^T)$$

where t denotes the cardinality of $\mathfrak{I}$.

Then, if $\phi(\mathbf{M}) = \log|\mathbf{M}|$, which is convex in $\mathbf{M}$, we have

$$\begin{aligned}\phi(\bar{\mathbf{M}}) &= \phi(\mathbf{M}(\bar{\xi})) \\ &\le \frac{1}{t}\sum_{T\in\mathfrak{I}} \phi(\mathbf{M}(\xi^T)) \\ &= \phi(\mathbf{M}(\xi))\end{aligned}$$

where $\bar{\xi}$ denotes a design with information matrix $\bar{\mathbf{M}}$ (cf. Pal and Mandal 2008).

Thus, we may restrict our search for optimum design within the subclass of symmetric designs. Consider the following designs:

$$\xi_0 = \{\mathbf{z} \mid \mathbf{z}'\mathbf{z} = 0\} \tag{3.3}$$

$$\tilde{\xi}_{\sqrt{p}} = \frac{n_c\xi_c + n_s\xi_s}{n} \tag{3.4}$$

$$n_c = p^2,\ n_s = 2^{p-k},\ n = n_c + n_s$$

$\xi_c = \dfrac{1}{2^k}$ fraction of a 2^p factorial experiment with levels ± 1

ξ_s = set of star points of the form $(\pm\sqrt{p},0,0,...,0)$, $(0,\pm\sqrt{p},0,...,0)$,..., $(0, 0,..., \pm\sqrt{p})$

Let, $\mathbf{M}_c$ and $\mathbf{M}_s$ denote the moment matrices corresponding to the designs ξ_c and ξ_s, respectively. Then, it can be easily seen that

$$\mathbf{M}_c = \begin{pmatrix} \mathbf{1} & \mathbf{1}_p' & 0 & 0 \\ & \mathbf{J}_p & 0 & 0 \\ & & \mathbf{I}_{\binom{p}{2}} & 0 \\ & & & \mathbf{I}_p \end{pmatrix}$$

$$\mathbf{M}_s = \begin{pmatrix} \mathbf{1} & \mathbf{1}_p' & 0 & 0 \\ & p\mathbf{I}_p & 0 & 0 \\ & & 0 & 0 \\ & & & \mathbf{I}_p \end{pmatrix} \tag{3.5}$$

where

$$\mathbf{J}_p = \mathbf{1}_p\mathbf{1}_p'$$

Hence, the moment matrix of the design $\tilde{\xi}_{\sqrt{p}}$, given by (3.4), is

$$\mathbf{M}_P = \frac{n_c\mathbf{M}_c + n_s\mathbf{M}_s}{n} \tag{3.6}$$

Now, we may use the following result for the full model

$$\eta(\mathbf{z};\beta) = \beta_0 + \sum_{i=1}^{p}\beta_i z_i + \sum_{i=1}^{p}\sum_{j=1}^{p}\beta_{ij} z_i z_j \tag{3.7}$$

which is singular (see Pukelsheim 1993):

Kiefer Optimality : Given a symmetric design ξ, there exists a design $\xi^* = (1-\alpha)\xi_0 + \alpha\tilde{\xi}_{\sqrt{p}}$, concentrated at $\mathbf{z}'\mathbf{z} = 0$ and $\mathbf{z}'\mathbf{z} = p$, such that $\xi^* \succ \xi$ in the sense of Loewner Order Dominance.

Such a design ξ^* is called a central composite design (CCD). Note that the design ξ^* is completely characterized by α.

Since the Loewner Order Dominance of ξ^* for the full model (3.7) implies the same for ξ^* in the non

singular set-up (3.2), we may utilize the above result to reduce the class of symmetric designs substantially.

The moment matrix of the CCD $\xi^* = (1 - \alpha)\xi_0 + \alpha\tilde{\xi}_{\sqrt{p}}$, where ξ_0 and $\tilde{\xi}_{\sqrt{p}}$ are given by (3.3) and (3.4) respectively, comes out to be

$$\mathbf{M}(\xi^*) = (1-\alpha)\mathbf{M}(\xi_0) + \alpha\mathbf{M}(\tilde{\xi}_{\sqrt{p}})$$

$$= \begin{bmatrix} \mathbf{1} & \alpha\mathbf{1}'_p & 0 & 0 \\ & \alpha\{b\mathbf{J}_p + (1\text{-}b)p\mathbf{I}_p\} & 0 & 0 \\ & & \alpha b\mathbf{I}_{\binom{p}{2}} & 0 \\ & & & \alpha\mathbf{I}_p \end{bmatrix} \qquad (3.8)$$

where

$$b = \frac{n_c}{n}$$

By virtue of Kiefer Optimality, our problem thus reduces to finding α, $0 \le \alpha \le 1$, such that $|\mathbf{M}(\xi^*)|$ is maximized.

Clearly,

$$|\mathbf{M}(\xi^*)| = \alpha^{2p+\binom{p}{2}} b^{\binom{p}{2}} |p(1-b)\mathbf{I}_p + (b-\alpha)\mathbf{J}_p|$$

$$= \alpha^{2p+\binom{p}{2}} b^{\binom{p}{2}} p[p(1-b)]^{p-1}(1-\alpha)$$

which is maximum at

$$\alpha = \alpha_0 = \frac{2p + \binom{p}{2}}{2p + \binom{p}{2} + 1} \qquad (3.9)$$

Thus, we have the following theorem:

Theorem 3.1. The optimum design for estimating the stress-strength reliability in a model with exponentially distributed stress and strength, and expected strength defined by the quadratic regression model (3.2) in the domain $Z = \{\mathbf{z}: \mathbf{z}'\mathbf{z} \le .p\}$, is given by the central composite design $\xi^* = (1 - \alpha_0)\xi_0 + \alpha_0\tilde{\xi}_{\sqrt{p}}$, where ξ_0 and $\tilde{\xi}_{\sqrt{p}}$ are as stated in (3.3) and (3.4), respectively, and $\alpha_0 = \dfrac{2p + \binom{p}{2}}{2p + \binom{p}{2} + 1}$

Example. Let $p = 4$. Then $\tilde{\xi}_{\sqrt{p}}$ is given by

$$\tilde{\xi}_{\sqrt{p}} = \frac{n_c\xi_c + n_s\xi_s}{n}$$

where

$$n_c = p^2,\ n_s = 2^{p-k},\ n = n_c + n_s$$

For $k = 1$, ξ_c = ½ fraction of 2^4 factorial experiment with levels ±1, corresponding to the identifying equation I = ABCD, where A, B, C and D represent the four factors:

A	B	C	D
–1	–1	–1	–1
1	1	–1	–1
1	–1	1	–1
1	–1	–1	1
–1	1	1	–1
–1	1	–1	1
–1	–1	1	1
1	1	1	1

ξ_s corresponds to the eight star points given by (± 2,0,0,0) (0, ± 2,0,0,0), (0,0, ± 2,0) and (0, 0, 0, ± 2)

Since n =16 and n = 8,

$\tilde{\xi}_{\sqrt{4}} = \dfrac{16\xi_c + 8\xi_s}{24}$, and the optimum design is given by $\xi^* = (1 - \alpha_0)\xi_0 + \alpha_0\tilde{\xi}_{\sqrt{4}}$, with $\alpha_0 = 14/15$.

4. DISCUSSION

In this investigation, we have considered the problem of determining the optimum design for estimating the system reliability in a stress-strength model, when the strength of the system is known to depend on a number of controllable factors, which are the covariates. A suitable criterion has been used to derive the optimum design when the mean strength of the system is taken to be a function of the covariates.

Two functional forms have been considered, viz. linear and quadratic.

Applications of the models can be found in agriculture, as well as in industry. In agriculture, for example, the yield of a crop depends, among others, on the soil quality. The soil needs to have the elastic behaviour for good yield. The strength of the soil is its capacity to retain the property of elasticity, and this depends on a number of factors like soil composition, (grain size, shape of particles, mineralogy etc.), state (loose, dense, over-consolidated, stiff, soft etc.), structure (arrangement of particles within soil mass, the manner in which the particles are packed or distributed) and loading conditions. Some of these factors are controllable. The stress on the soil is the prevailing shear stress. In industry, the strength of a manufactured product depends on its design, which in turn is dependent on a number of controllable parameters. These parameters are, therefore, the covariates affecting the strength of the product. For example, the strength of alkali activated slag concrete depends on the type of alkaline activator used, type and fineness of the slag, etc. The product breaks down when the environmental stress actng on it exceeds its strength.

ACKNOWLEDGEMENTS

The authors thank the anonymous referee for his/her comments, which immensely improved the presentation of the paper. They also acknowledge with thanks the fruitful suggestions received from Professor S.P. Mukherjee.

REFERENCES

Guttman, I., Johnson, R.A., Bhattacharya, G.K. and Reiser, B. (1988). Confidence limits for stress-strength models with explanatory variables. *Technometrics*, **30(2)**, 161-168.

Guttman, I. and Papandonatos, G.D. (1997). A Bayesian approach to a reliability problem: Theory, analysis and interesting numerics. *Canad. J. Statist.*, **25(2)**, 143-158.

Keifer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.*, **12**, 363-366.

Liski, E.P., Mandal, N.K., Shah, K.R. and Sinha, B.K. (2002). *Topics in Optimal Design*. Springer Verlag.

Pal, M. and Mandal, N.K. (2008). Minimax designs for optimum mixtures. *Statist. Probab. Lett.*, **78(6)**, 608-615.

Pukelsheim, F. (1993). *Optimal Design of Experiments.* Wiley, New York.

Silvey, S.D. (1980). *Optimal Design.* Chapman & Hall, London.

Weerahandi, S. and Johnson, Richard A. (1990). Testing reliability in a stress-strength model when X and Y are normally distributed. *Technometrics*, **34(1)**, 83-91.

Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 311-319

Construction of Optimal Mixed-Level Supersaturated Designs

V.K. Gupta[1], Poonam Singh[2], Basudev Kole[1*] and Rajender Parsad[1]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Department of Statistics, University of Delhi, Delhi*

(Received: November 2009, Revised: December 2009, Accepted: December 2009)

SUMMARY

This article describes some methods of construction of mixed level f_{NOD}-optimal supersaturated designs. The methods of construction exploit the layout and the property of Uniform designs and Hadamard matrices. Mathematical expression for $E(f_{NOD})$ and $E(\chi^2)$ criteria have been obtained for many designs constructed in this paper. Some examples are given to illustrate the methods of construction. A catalogue of 67 optimal mixed level supersaturated designs with at most 60 runs and 60 factors is prepared. Some other important features of the designs are also given in the catalogue. All designs are f_{NOD}- optimal while some designs are χ^2- optimal too.

Key words : Uniform design, Hadamard matrices, Efficiency criteria.

1. INTRODUCTION

Supersaturated Design (SSD) is essentially a fractional factorial design in which the number of runs is not enough to estimate the main effects of all the factors in the experiment. SSDs are mainly used in experimental situations where a large number of factors are to be tested but only few of the factors are active and the experimentation is expensive and also time consuming.

A common application of SSDs is the screening experiment. In screening experiment there are usually a large number of factors to be investigated, but it is believed that only a few of them will be active, and the few active factors have significant influence on the response. Identifying these few active factors correctly and economically is the main purpose of screening experiments. This phenomenon is commonly recognized as effect sparsity.

The construction of SSDs started with the use of random balance experiments by Satterthwaite (1959). Booth and Cox (1962) proposed an algorithm to construct systematic SSDs. A two-level SSD is balanced when the number of times each level appears in a column is same. Many methods have been proposed for constructing SSDs [see e.g., Nguyen (1996), Gupta and Chatterjee (1998), Butler *et al*. (2001), Xu and Wu (2003), Bulutoglu and Cheng (2004), Bulutoglu (2007), Liu *et al*. (2007a), Ryan and Bulutoglu (2007), Das *et al*. (2008), Gupta *et al*. (2008b) and Nguyen and Cheng (2008)].

Two-level SSDs have been studied extensively because of their application and ease of generation. But multi-level SSDs are often required in industrial and scientific experimentation for exploring nonlinear effects of the factors. It is never advised to reduce the factor levels to two if it would result in severe loss in information. Some references on multi-level factor designs include Yamada and Lin (1999), Fang *et al.* (2000), Lu and Sun (2001) and Liu *et al*. (2007b).

Mixed-level SSDs are also requested frequently in experimentation. In an experimental situation when several factors have same number of levels and one or two factors have different number of levels than the rest

[*] *Corresponding author* : Basudev Kole
E-mail address : basudevkole@gmail.com

of the factors, mixed level SSDs become useful. The few factors with different levels may be called as factors of asymmetry and are important factors. Some recent references on mixed level SSDs are Fang *et al*. (2003), Li *et al*. (2004), Koukouvinos and Mantas (2005), Ai *et al*. (2007), Tang *et al*. (2007), Gupta *et al*. (2008a, 2009) and Chen and Liu (2008).

Most of the research on SSDs has been restricted to balanced designs. A mixed level SSD is said to be balanced if all the levels of all the factors (in every column of the design) appear a constant number of times in the design runs. In this paper also, we restrict ourselves to balanced SSDs only. This paper introduces some methods of constructing balanced, mixed level SSDs, derived essentially from the juxtaposition of Uniform designs and Hadamard matrices. The methods of construction are illustrated with the help of examples. $E(f_{NOD})$ and $E(\chi^2)$ criteria have been used to investigate the efficiency of the designs constructed. Mathematical expressions of $E(f_{NOD})$ and $E(\chi^2)$ have been obtained for the constructed SSDs. All designs obtained are f_{NOD}-optimal while some designs are also χ^2- optimal. A catalogue of 67 $E(f_{NOD})$-optimal designs is prepared and the $E(\chi^2)$-efficiency of the designs is also given.

We begin with some preliminaries in Section 2. Different non-orthogonality criteria for evaluation of SSDs are also defined in Section 2 for the sake of completeness. Proposed construction methods, mathematical expression of $E(f_{NOD})$ and $E(\chi^2)$ and some examples are given in Section 3. Some concluding remarks and a catalogue of optimal designs are given in Section 4.

Hadamard matrices used for generation of the designs have been taken from "http://www.iasri.res.in/WebHadamard/WebHadamard.htm" available at Design Resources Server, IASRI, New Delhi and Uniform designs used have been taken from http://www.math.hkbu.edu.hk/UniformDesign/ maintained by Chang-Xing Ma. The Uniform designs used are those with centered L_2-discrepancy. The 67 SSDs obtained are available at http://iasri.res.in/design Supersaturated_Design/Supersaturated.html.

2. PRELIMINARIES

This section is devoted towards giving some useful definitions for the sake of completeness.

2.1 Uniform Design

Uniform design is an efficient fractional factorial design originally proposed by Fang and Yuan (1980). A Uniform design, denoted as $\mathbf{U}_n(n^s)$, is an $n \times s$ array in n symbols with each column having n symbols appearing once, where n is the number of runs and also the number of levels of each factor and s is the number of factors in the design. The uniform designs used in this paper for construction of SSDs are those with centered L_2-discrepancy.

2.2 Hadamard Matrix

A square matrix $\mathbf{H}_n$ of order n and with entries +1 and –1 is said to be a Hadamard matrix of order n if and only if $\mathbf{H}_n\mathbf{H}_n' = \mathbf{H}_n'\mathbf{H}_n = n\mathbf{I}_n$. A necessary condition for the existence of a Hadamard matrix is that n must be an integer and n, n/12 or n/20 must be a power of 2. A Hadamard matrix is invariant with respect to permutation of rows and / or columns and also with respect to any scalar multiplication with –1 of any row or column. We shall write a Hadamard matrix of order n in semi normalized form as $\mathbf{H}_n = [\mathbf{1}_n\ \mathbf{L}_n]$, where $\mathbf{1}$ is a vector of all ones and $\mathbf{L}_n$ is an $n \times (n-1)$ matrix of the remaining columns of $\mathbf{H}_n$. For maintaining uniformity of notations, we shall recode the symbols –1 and +1 in $\mathbf{L}_n$ as 1 and 2, respectively, wherever required.

2.3 Evaluation Criteria of Mixed-level SSDs

Consider a mixed-level SSD represented as SSD-$(n; q_1 \times q_2 \times ... \times q_m)$ with m factors, the j^{th} factor being experimented with q_j levels, $j = 1, 2, ..., m$ and number of design points or runs as n. Define $v = \dfrac{\sum_{j=1}^{m}(q_j - 1)}{n-1}$.

For an SSD, $v > 1$. We denote an SSD as an $n \times m$ matrix $\mathbf{X}$ with $\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^j, ..., \mathbf{x}^m$ as its m columns, the j^{th} column $\mathbf{x}^j$ containing q_j symbols, $j = 1, 2, ..., m$. Fang *et al*. (2003) defined the following $E(f_{NOD})$ criterion for measuring non-orthogonality of the mixed-level SSDs:

$$f_{NOD}^{ij} = \sum_{u=1}^{q_i}\sum_{v=1}^{q_j}\left[n_{uv}^{(ij)} - \frac{n}{q_i q_j}\right]^2 \qquad (2.1)$$

Here the subscript *NOD* stands for non-orthogonality of the design; $n_{uv}^{(ij)}$ is the number of (u, v)-pairs in $(\mathbf{x}^i, \mathbf{x}^j)$ and n/q_iq_j stands for the average frequency of level-combination in each pair of columns $\mathbf{x}^i$ and $\mathbf{x}^j$, $i \neq j = 1, 2, ..., m$.

A criterion $E(f_{NOD})$ is defined as minimizing

$$E(f_{NOD}) = \sum_{1\le i< j\le m} f_{NOD}^{ij} \Big/ \binom{m}{2} \qquad (2.2)$$

Fang *et al*. (2004) obtained a lower bound to $E(f_{NOD})$ which is sharper than the one obtained earlier by Fang *et al*. (2003) and is given in Theorem 2.1.

Theorem 2.1 (Fang *et al*. 2004)**.** For any balanced SSD-$(n; q_1 \times q_2 \times ... \times q_m)$

$$E(f_{NOD}) \ge \frac{n(n-1)}{m(m-1)}[(\gamma + 1 - \psi)(\psi - \gamma) + \psi^2] + C(n, q_1, ..., q_m) = L[E(f_{NOD})] \qquad (2.3)$$

where

$$C(n; q_1, q_2, ..., q_m) = \frac{nm}{m-1} - \frac{1}{m(m-1)}\left(\sum_{i=1}^{m}\frac{n^2}{q_i} + \sum_{i,j=1, j\neq i}^{m}\frac{n^2}{q_iq_j}\right)$$

depends on $\mathbf{X}$ only through $n, q_1, q_2, ...,q_m$.

Here $\psi = \dfrac{\sum_{i=1}^{m} n/q_i - m}{(n-1)}$, $\gamma = [\psi]$ and

$[x]$ denotes the integer part of x.

For a balanced mixed level SSD-$(n; q_1, q_2, ..., q_m)$-SSD, Yamada and Lin (1999) defined the following $E(\chi^2)$ criterion. For every pair of columns $(\mathbf{x}^i, \mathbf{x}^j)$, $\mathbf{i \neq j} = 1, 2, ..., m$

$$\chi^2(\mathbf{x}^i, \mathbf{x}^j) = \frac{q_iq_j}{n}\sum_{u=1}^{q_i}\sum_{v=1}^{q_j}\left(n_{uv}^{(ij)} - \frac{n}{q_iq_j}\right)^2 \qquad (2.4)$$

Obviously, the value of $\chi^2(\mathbf{x}^i, \mathbf{x}^j)$ measures the non-orthogonality between two columns $\mathbf{x}^i$ and $\mathbf{x}^j$. Then the $E(\chi^2)$ value can be used to evaluate the overall non-orthogonality between the columns of $\mathbf{X}$, where $E(\chi^2)$ is defined as

$$E(\chi^2) = \frac{2}{m(m-1)}\sum_{1\le i< j\le m}\chi^2(\mathbf{x}^i, \mathbf{x}^j) \qquad (2.5)$$

For a balanced design, Ai *et al*. (2007) obtained lower bound to the value of $E(\chi^2)$ as given in Theorem 2.2.

Theorem 2.2 (Ai, Fang and He 2007). For any balanced SSD-$(n; q_1 \times q_2 \times ... \times q_m)$,

$$E(\chi^2) \ge \frac{1}{m(m-1)(n-1)}\left(nm - \sum_{k=1}^{m} q_k\right)^2 + C_1(n; q_1, q_2, ..., q_m) = L[E(\chi^2)] \qquad (2.6)$$

where $C_1(n; q_1, q_2, ..., q_m)$

$$= \frac{1}{m(m-1)}\left[\left(\sum_{k=1}^{m} q_k\right)^2 - n\sum_{k=1}^{m} q_k\right] - n$$

In this paper we have used lower bounds $L[E(f_{NOD})]$, and $L[E(\chi^2)]$, given in Theorem 2.1 and Theorem 2.2, respectively.

For any design $d \in$ SSD-$(n; q_1 \times q_2 \times ... \times q_m)$, we define f_{NOD}-efficiency and χ^2-efficiency as

$$f_{NOD} - \text{efficiency} = \frac{L\left[E(f_{NOD})\right]}{E_d(f_{NOD})}$$

$$\chi^2 - \text{efficiency} = \frac{L\left[E(\chi^2)\right]}{E_d(\chi^2)} \qquad (2.7)$$

where $E_d(f_{NOD})$ and $E_d(\chi^2)$ are the values of $E(f_{NOD})$ and $E(\chi^2)$, respectively, for the design d. A design with high efficiency is acceptable and a design with f_{NOD}–efficiency [χ^2 – efficiency] one is f_{NOD} [χ^2]-optimal.

3. CONSTRUCTION OF F_{NOD}-OPTIMAL MIXED-LEVEL SUPERSATURATED DESIGNS

This Section gives a method of constructing f_{NOD}-optimal mixed level SSDs by juxtaposing Uniform designs with Hadamard matrices.

Theorem 3.1. Consider a Hadamard matrix $\mathbf{H}_n$ of order $n \ge 4$, written in semi-normal form as $\mathbf{H}_n = [\mathbf{1}_n \mid \mathbf{L}_n]$,

and a Uniform design with n runs and s factors each at n levels as $\mathbf{U}_n(n^s) = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_s]$. Select any $p \le s$ columns of $\mathbf{U}_n(\mathbf{n}^s)$ and write them as an $n \times p$ array $\mathbf{M}_{n\times p} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_p]$ The mixed level SSD- $(n; n^p \times 2^{n-1})$ given by $\mathbf{A} = [\mathbf{M}_{n\times p} \vdots\ \mathbf{L}_n]$, has $E(f_{NOD})$ and $E(\chi^2)$ as follows:

$$\mathrm{E}(f_{NOD}) = \frac{p[nm-n-p+1]}{m(m-1)}$$

$$\mathrm{E}(\chi^2) = \frac{np\left[(p-1)(m-1)+2(n-1)\right]}{m(m-1)}$$

where $m = p + n - 1$.

Proof. The SSD in **A** can be written as

$$\mathbf{A} = \left[\underbrace{\mathbf{u}_1\ \ \mathbf{u}_2\ \ \cdots\ \ \mathbf{u}_p}_{\text{Group 1}} \vdots\ \underbrace{\mathbf{L}_n}_{\text{Group 2}}\right]$$

Case I. $E(f_{NOD})$ for **A**

For the design **A**, we can rewrite (2.2) as

$$E(f_{NOD}) = \frac{1}{\binom{m}{2}}\left[\sum_{1\le i<j\le p} f_{NOD}^{ij} + \sum_{p+1\le i<j\le m} f_{NOD}^{ij} + \sum_{i=1}^{p}\sum_{j=1}^{n-1} f_{NOD}^{ij}\right] \tag{3.1}$$

Since Group 1 consists of p columns of $\mathbf{U}_n(n^s)$,

$$\sum_{1\le i<j\le p} f_{NOD}^{ij} = \frac{p(p-1)(n-1)}{2} \tag{3.2}$$

Similarly, for Group 2 consisting of n–1columns of $\mathbf{H}_n$

$$\sum_{p+1\le i<j\le m} f_{NOD}^{ij} = 0 \tag{3.3}$$

The third component of (3.1) is $\sum_{i=1}^{p}\sum_{j=1}^{n-1} f_{NOD}^{ij}$, which has contribution from p columns of $\mathbf{U}_n(n^s)$ and n–1columns of $\mathbf{H}_n$. There will be $2n$ pairs of symbols but only n pairs appear together once. It is not required to know which n pairs appear together once. The remaining n pairs do not appear together.

Therefore,

$$\sum_{i=1}^{p}\sum_{j=1}^{n-1} f_{NOD}^{ij} = \frac{pn(n-1)}{2} \tag{3.4}$$

Using (3.2), (3.3) and (3.4) in (3.1) gives

$$E(f_{NOD}) = \frac{p\left[nm-n-p+1\right]}{m(m-1)} \tag{3.5}$$

Case II. $E(\chi^2)$ for **A**

For the design **A**, we can write (2.5) as

$$E(\chi)^2 = \frac{1}{\binom{m}{2}}\left[\sum_{1\le i<j\le p} \chi^2(\mathbf{x}^i,\mathbf{x}^j) + \sum_{p+1\le i<j\le m} \chi^2(\mathbf{x}^i,\mathbf{x}^j) + \sum_{i=1}^{p}\sum_{j=1}^{n-1} \chi^2(\mathbf{x}^i,\mathbf{x}^j)\right] \tag{3.6}$$

Following on the lines of Case I above, we have

$$\sum_{1\le i<j\le p} \chi^2(\mathbf{x}^i,\mathbf{x}^j) = \frac{p(p-1)(n-1)n}{2} \tag{3.7}$$

$$\sum_{p+1\le i<j\le m} \chi^2(\mathbf{x}^i,\mathbf{x}^j) = 0 \tag{3.8}$$

and

$$\sum_{i=1}^{p}\sum_{j=1}^{n-1} \chi^2(\mathbf{x}^i,\mathbf{x}^j) = pqn \tag{3.9}$$

Using (3.7), (3.8) and (3.9) in (3.6) gives

$$E(\chi^2) = \frac{np\left[(p-1)(n-1)+2(n-1)\right]}{m(m-1)} \tag{3.10}$$

The proof is thus complete.

Theorem 3.2. For any balanced mixed level SSD- $(n; n^p \times 2^{n-1})$, denoted as an $n \times m$ array, $m = p + n - 1$ the lower bounds to $E(f_{NOD})$ and $\mathrm{E}(\chi^2)$ are the following:

$$L[E(f_{NOD})] = \frac{p\left[nm-n-p+1\right]}{m(m-1)}$$

$$L[E(\chi^2)] = \frac{np\left[(p-1)(m-1)+2(n-1)\right]}{m(m-1)}$$

Proof. We first obtain the lower bound to $E(f_{NOD})$.

Case I. $L[E(f_{NOD})]$

$L[E(f_{NOD})]$ for the designs constructed by Method 3.1, can be obtained by using Theorem 2.1 and equation (2.3).

For the designs under consideration, $q_1 = q_2 = ... = q_p = n$ and $q_{p+1} = q_{p+2} = ... = q_m = 2$ and number of runs is also n. Now we calculate the following:

$$\psi = \frac{\sum_{i=1}^{m} n/q_i - m}{(n-1)} = \frac{n}{2} - 1 \qquad (3.11)$$

$\gamma = [\psi] = \frac{n}{2} - 1$, because n is a Hadamard number and, therefore, $\frac{n}{2} - 1$ is an integer. (3.12)

Now

$$\sum_{i=1}^{m} \frac{n^2}{q_i} + \sum_{i,j=1, j\neq i}^{m} \frac{n^2}{q_i q_j}$$

$$= \sum_{i=1}^{p} \frac{n^2}{q_i} + \sum_{i=p+1}^{m} \frac{n^2}{q_i} + \sum_{i=1}^{p} \sum_{j=1,\, i\neq j}^{p} \frac{n^2}{q_i q_j}$$

$$+ \sum_{i=p+1}^{m} \sum_{j=p+1,\, i\neq j}^{m} \frac{n^2}{q_i q_j} + 2\sum_{i=1}^{p} \sum_{j=p+1}^{m} \frac{n^2}{q_i q_j}$$

$$= n^2 p + (n-1)\frac{n^3}{4} + p(p-1)$$

Therefore

$$C(n; n^p, 2^{n-1}) = \frac{nm}{(m-1)} - \frac{1}{m(m-1)}\left[n^2 p + (n-1)\frac{n^3}{4} + p(p-1)\right] \qquad (3.13)$$

Now using (3.11) and (3.12) in $[(\gamma + 1 - \psi)(\psi - \gamma^2) + \psi^2]$ we get

$$[(\gamma + 1 - \psi)(\psi - \gamma) + \psi^2] = \left(\frac{n}{2} - 1\right)^2 \qquad (3.14)$$

Now using (3.13) and (3.14) in (2.3) we get the following:

$$L[E(f_{NOD})] = \frac{n(n-1)}{m(m-1)}\left(\frac{n}{2} - 1\right)^2 + \frac{nm}{(m-1)} - \frac{1}{m(m-1)}\left[n^2 p + (n-1)\frac{n^3}{4} + p(p-1)\right]$$

$$= \frac{p}{m(m-1)}\,[nm - n - p + 1] \qquad (3.15)$$

Case II. $L[E(\chi^2)]$

$L[E(\chi^2)]$ for the designs constructed by Method 3.1, can be obtained by using Theorem 2.2 and equation (2.6).

For the designs under consideration, $q_1 = q_2 = ... = q_p = n$ and $q_{p+1} = q_{p+2} = ... = q_m = 2$ and number of runs is also n. Now we calculate the following:

$$C_1(n; n^p \times 2^{n-1}) = \frac{1}{m(m-1)}\left[\left(\sum_{k=1}^{m} q_k\right)^2 - n\sum_{k=1}^{m} q_k\right] - n$$

$$= \frac{[(pn + 2n - 2)(pn + n - 2)]}{m(m-1)} - n \qquad (3.16)$$

Using (3.16) in (2.6) gives

$$L[E(\chi^2)] = \frac{1}{m(m-1)(n-1)}\left(nm - \sum_{k=1}^{m} q_k\right)^2 + C_1(n; np \times 2^{n-1})$$

$$= \frac{1}{m(m-1)(n-1)}\,[nm - np - (n-1)\,2]^2 + \frac{[(pn + 2n - 2)(pn + n - 2)]}{m(m-1)} - n$$

$$= \frac{np[(p-1)(n-1) + 2(n-1)]}{m(m-1)} \qquad (3.17)$$

The proof is thus complete.

Corollary 3.1. Let $\mathbf{H}_n = [\mathbf{1}_n \mid \mathbf{L}_n]$ and $\mathbf{U}_n(n^s) = [\mathbf{u}_1\ \mathbf{u}_2\ ...\mathbf{u}_s]$ be as in Theorem 3.1. The mixed level SSD-$(n; n \times 2^{n-1})$ given by $\mathbf{A} = [\mathbf{u}_1 \vdots \mathbf{L}_n]$, has $E(f_{NOD}) = 1$ and $E(\chi^2) = 2$.

Proof. Follows by taking $p = 1$ in Theorem 3.1.

Result 3.1. Applying Theorem 3.1, Corollary 3.1 and Theorem 3.2 in (2.7) gives that the mixed level SSD-$(n; n^p \times 2^{n-1})$ and SSD-$(n; n \times 2^{n-1})$, are both f_{NOD}-optimal and γ^2-optimal.

Example 3.1. An f_{NOD}- optimal mixed level SSD $(8; 8^3 \times 2^7)$ obtained from $\mathbf{U}_8(8^7)$ and $\mathbf{H}_8$. Consider the following Hadamard matrix $\mathbf{H}_8$ in semi-normalized form, obtained from http://www.iasri.res.in/WebHadamard/WebHadamard.htm:

2	2	2	2	2	2	2	2
2	1	2	1	2	1	2	1
2	2	1	1	2	2	1	1
2	1	1	2	2	1	1	2
2	2	2	2	1	1	1	1
2	1	2	1	1	2	1	2
2	2	1	1	1	1	2	2
2	1	1	2	1	2	2	1

Further, obtain the following Uniform design $\mathbf{U}_8(8^7)$ from http://www.math.hkbu.edu.hk/UniformDesign/:

7	7	1	4	2	5	2
8	5	6	5	5	8	8
3	4	3	8	8	7	3
5	2	8	3	6	4	1
4	6	2	2	7	2	7
2	3	5	1	1	6	5
1	8	7	6	4	3	4
6	1	4	7	3	1	6

Juxtaposing the first three columns of $\mathbf{U}_8(8^7)$ and the last seven columns of $\mathbf{H}_8$ above gives the following desired mixed level SSD $(8; 8^3 \times 2^7)$:

1	7	1	2	2	2	2	2	2	2
8	5	6	1	2	1	2	1	2	1
3	4	3	2	1	1	2	2	1	1
5	2	8	1	1	2	2	1	1	2
4	6	2	2	2	2	1	1	1	1
2	3	5	1	2	1	1	2	1	2
1	8	7	2	1	1	1	1	2	2
6	1	4	1	1	2	1	2	2	1

Here we can take any of the three columns of $\boldsymbol{U}_8(8^7)$ instead of taking the first three columns. For this design $E(f_{NOD}) = 2.33$ and $E(\chi^2) = 7.47$ and the design is both f_{NOD}-optimal and χ^2-optimal.

Example 3.2. An f_{NOD}-optimal mixed level SSD $(8; 8 \times 2^7)$ obtained from $\mathbf{U}_8(8^7)$ and $\mathbf{H}_8$. Consider the Hadamard matrix $\mathbf{H}_8$ in semi-normalized form and the Uniform design $\mathbf{U}_8(8^7)$ as in Example 3.1.

Juxtaposing the first column of $\mathbf{U}_8(8^7)$ with the last seven columns of $\mathbf{H}_8$ gives the desired mixed level SSD $(8; 8 \times 2^7)$ as follows:

7	2	2	2	2	2	2	2
8	1	2	1	2	1	2	1
3	2	1	1	2	2	1	1
5	1	1	2	2	1	1	2
4	2	2	2	1	1	1	1
2	1	2	1	1	2	1	2
1	2	1	1	1	1	2	2
6	1	1	2	1	2	2	1

Here we can take any of the seven columns of $\mathbf{U}_8(8^7)$ instead of taking the first column. For this design, $E(f_{NOD}) = 1.00$ and $E(\chi^2) = 2.00$, and the design is both f_{NOD}-optimal and χ^2-optimal.

Method 3.1. Consider again a Hadamard matrix $\mathbf{H}_n$ of order $n \geq 4$, written in semi-normal form as $\mathbf{H}_n = [\mathbf{1}_n \mid \mathbf{L}_n]$. Consider a Uniform Design $\mathbf{U}_t(t^s)$, with s factors each at t levels in t runs, where $t = \frac{n}{2}$. The $t \times s$ array, $\mathbf{U}_t(t^s)$, may be written as $\mathbf{U}_t(t^s) = [\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_s]$, where $\mathbf{u}_1, \mathbf{u}_2 \ldots \mathbf{u}_s$ are the columns of $\mathbf{U}_t(t^s)$. Select first two columns of $\mathbf{U}_t(t^s)$ and write them as an $n \times 1$ array $\mathbf{D}_{n\times 1} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}$. Then, $\mathbf{A} = [\mathbf{D}_{n\times 1} \vdots \mathbf{L}_n]$ is an f_{NOD}- optimal mixed level SSD-$(n; t \times 2^{n-1})$ with one factor having t $(=n/2)$ levels and $(n-1)$ factors having two levels and n runs.

Remark 3.1. For the mixed level SSD-$(n; t \times 2^{n-1})$ given by $\mathbf{A} = [\mathbf{D}_{n\times 1} \vdots \mathbf{L}_n]$, obtained by Method 3.1, it is not possible to obtain mathematical expressions for $E(f_{NOD})$ and $E(\chi)^2$, similar to the ones obtained for the designs generated in Theorem 3.1 and Corollary 3.1. The reason is that in both the equations (3.1) and (3.6), it is not possible to evaluate the third term. The values of $n_{uv}^{(ij)}$ will be 0, 1 or 2, but it is not possible to know their frequencies, although the total of their frequencies is n. But we have used this method of construction using $\mathbf{H}_n$, $n \leq 60$ and $\mathbf{U}_t(t^s)$, $t \leq 30$ and obtained f_{NOD}-optimal mixed level SSDs-$(n; t \times 2^{n-1})$. These designs are given in Table 4.2. The χ^2–efficiency of these designs is also given in the Table.

Example 3.3. An f_{NOD}-optimal Mixed level SSD $(8; 4 \times 2^7)$ obtained from $\mathbf{U}_4(4^3)$ and $\mathbf{H}_8$. A uniform design $\mathbf{U}_4(4^3)$ obtained from http://www.math.hkbu.edu.hk/UniformDesign/ is the following:

3	4	3
2	1	4
1	3	1
4	2	2

Using the first two columns of $\mathbf{U}_4(4^3)$ and juxtaposing the new column with the last seven columns of $\mathbf{H}_8$ from Example 3.1 gives the desired mixed level SSD $(8; 4 \times 2^7)$ as follows.

3	2	2	2	2	2	2	2
2	1	2	1	2	1	2	1
1	2	1	1	2	2	1	1
4	1	1	2	2	1	1	2
4	2	2	2	1	1	1	1
1	1	2	1	1	2	1	2
3	2	1	1	1	1	2	2
2	1	1	2	1	2	2	1

Instead of using the first two columns of $\mathbf{U}_4(4^3)$, one could use any two columns of $\mathbf{U}_4(4^3)$. For this design, $E(f_{NOD}) = 0.86$ and the design is f_{NOD}-optimal. Further, the design has $E(\chi^2) = 0.86$ with χ^2 – efficiency = 0.71.

4. DISCUSSION

This article introduces construction of mixed level SSDs by juxtaposition of Uniform designs and the Hadamard matrices. All the designs generated through this method are f_{NOD}-optimal. Some series of designs are both f_{NOD} and χ^2-optimal. Such designs are catalogued in Table 4.1. It is not possible to establish algebraically the optimality of designs obtained in Method 3.1. However, using this method f_{NOD}-optimal SSDs have been obtained and are given in Table 4.2. The χ^2–efficiency of these designs is also given. All the designs given in the Tables 4.1 and 4.2 are available at

Table 4.1. f_{NOD} and χ^2-optimal mixed level SSDs obtained using Theorem 3.1 and Corollary 3.1

Sl. No.	Design	$E(f_{NOD})$	$E(\chi^2)$
1	$(4;4.2^3)$	1.00	2.00
2	$(4;4^2.2^3)$	1.50	3.60
3	$(8;8.2^7)$	1.00	2.00
4	$(8;8^2.2^7)$	1.75	4.67
5	$(8;8^3.2^7)$	2.33	7.47
6	$(8;8^4.2^7)$	2.80	10.18
7	$(8;8^5.2^7)$	3.18	12.73
8	$(8;8^6.2^7)$	3.50	15.08
9	$(8;8^7.2^7)$	3.77	17.23
10	$(12;12.2^{11})$	1.00	2.00
11	$(12;12^2.2^{11})$	1.83	5.08
12	$(12;12^3.2^{11})$	2.54	8.70
13	$(12;12^4.2^{11})$	3.14	12.57
14	$(12;12^5.2^{11})$	3.67	16.50
15	$(12;12^6.2^{11})$	4.13	20.38
16	$(12;12^7.2^{11})$	4.53	24.16
17	$(12;12^8.2^{11})$	4.89	27.79
18	$(12;12^9.2^{11})$	5.21	31.26
19	$(12;12^{10}.2^{11})$	5.50	34.57
20	$(12;12^{11}.2^{11})$	5.76	37.72
21	$(16;16.2^{15})$	1.00	2.00
22	$(16;16^2.2^{15})$	1.88	5.29
23	$(16;16^3.2^{15})$	2.65	9.41
24	$(16;16^4.2^{15})$	3.33	14.04
25	$(16;16^5.2^{15})$	3.95	18.95
26	$(16;16^6.2^{15})$	4.50	24.00
27	$(16;16^7.2^{15})$	5.00	29.09
28	$(16;16^8.2^{15})$	5.45	34.15
29	$(16;16^9.2^{15})$	5.87	39.13
30	$(16;16^{10}.2^{15})$	6.25	44.00
31	$(16;16^{11}.2^{15})$	6.60	48.74
32	$(16;16^{12}.2^{15})$	6.92	53.33
33	$(16;16^{13}.2^{15})$	7.22	57.78
34	$(16;16^{14}.2^{15})$	7.50	62.07
35	$(16;16^{15}.2^{15})$	7.76	66.21
36	$(20;20.2^{19})$	1.00	2.00
37	$(20;20^2.2^{19})$	1.90	5.43
38	$(20;20^3.2^{19})$	2.71	9.87
39	$(20;20^4.2^{19})$	3.45	15.02
40	$(20;20^5.2^{19})$	4.13	20.65
41	$(20;20^6.2^{19})$	4.75	26.60
42	$(20;20^7.2^{19})$	5.32	32.74
43	$(20;20^8.2^{19})$	5.85	38.97
44	$(20;20^9.2^{19})$	6.33	45.23
45	$(20;20^{10}.2^{19})$	6.79	51.47
46	$(20;20^{11}.2^{19})$	7.21	57.66
47	$(20;20^{12}.2^{19})$	7.60	63.75
48	$(20;20^{13}.2^{19})$	7.97	69.73
49	$(20;20^{14}.2^{19})$	8.31	75.59
50	$(20;20^{15}.2^{19})$	8.64	81.31
51	$(20;20^{16}.2^{19})$	8.94	86.89
52	$(20;20^{17}.2^{19})$	9.23	92.33
53	$(20;20^{18}.2^{19})$	9.50	97.62
54	$(20;20^{19}.2^{19})$	9.76	102.76

Table 4.2. f_{NOD}-optimal mixed level SSDs obtained using Method 3.1

Sl. No.	Design	$E(f_{NOD})$	$E(\chi^2)$	$L[E(\chi^2)]$	χ^2–efficiency
1	$(8;4.2^7)$	0.86	0.86	0.61	0.71
2	$(12;6.2^{11})$	0.91	0.91	0.66	0.73
3	$(16;8.2^{15})$	0.93	0.93	0.68	0.73
4	$(20;10.2^{19})$	0.95	0.95	0.70	0.74
5	$(24;12.2^{23})$	0.96	0.96	0.71	0.74
6	$(28;14.2^{27})$	0.96	0.96	0.71	0.74
7	$(32;16.2^{31})$	0.97	0.97	0.72	0.74
8	$(36;18.2^{35})$	0.97	0.97	0.72	0.74
9	$(40;20.2^{39})$	0.97	0.97	0.72	0.74
10	$(44;22.2^{43})$	0.98	0.98	0.73	0.74
11	$(52;26.2^{51})$	0.98	0.98	0.73	0.75
12	$(56;28.2^{55})$	0.98	0.98	0.73	0.75
13	$(60;30.2^{59})$	0.98	0.98	0.73	0.75

http://iasri.res.in/design/Supersaturated_Design/Supersaturated.html. The Hadamard matrix used to construct designs in the catalogue have been taken from the link http://www.iasri.res.in/WebHadamard/WebHadamard.htm maintained at Design Resources Server. The Uniform designs are taken from http://www.math.hkbu.edu.hk/UniformDesign/ maintained by Chang-Xing Ma.

We now describe an interesting problem. Consider again the mixed level SSD-(n; $n \times 2^{n-1}$), obtained as $\mathbf{A} = [\mathbf{u}_1 \vdots \mathbf{L}_n]$ in Corollary 3.1. Suppose that the n levels of $\mathbf{u}_1$ are such that $n = s_1 \times s_2 \times ... \times s_r$. Using the replacement technique one can get from SSD-(n; $n \times 2^{n-1}$) another SSD-(n; $s_1 \times s_2 \times ... \times s_r \times 2^{n-1}$) on replacing the n level factor with r factors at s_1, s_2, ..., s_r levels respectively. The design has n runs. Intuitively, it appears that the new design will also be f_{NOD}-optimal and χ^2-optimal. But this may not be so. Consider the following example:

Example 4.1. Suppose we have an SSD (12; 12×2^{11}) constructed using Theorem 3.1. Therefore by Corollary 3.1, the design has $E(f_{NOD}) = 1$ and $E(\chi^2) = 2$, and is both f_{NOD}-optimal and χ^2-optimal. The design is

5	1	1	1	1	1	1	1	1	1	1	1
4	2	1	2	1	1	1	2	2	2	1	2
3	2	2	1	2	1	1	1	2	2	2	1
7	1	2	2	1	2	1	1	1	2	2	2
1	2	1	2	2	1	2	1	1	1	2	2
6	2	2	1	2	2	1	2	1	1	1	2
11	2	2	2	1	2	2	1	2	1	1	1
10	1	2	2	2	1	2	2	1	2	1	1
9	1	1	2	2	2	1	2	2	1	2	1
8	1	1	1	2	2	2	1	2	2	1	2
2	2	1	1	1	2	2	2	1	2	2	1
12	1	2	1	1	1	2	2	2	1	2	2

Now using replacement technique we can construct an SSD (12; $4 \times 3 \times 2^{11}$). The constructed design is as follows:

2	2	1	1	1	1	1	1	1	1	1	1
2	1	2	1	2	1	1	1	2	2	2	1
1	3	2	2	1	2	1	1	1	2	2	2
3	1	1	2	2	1	2	1	1	1	2	2
1	1	2	1	2	2	1	2	1	1	1	2
2	3	2	2	1	2	2	1	2	1	1	1
4	2	2	2	2	1	2	2	1	2	1	1
4	1	1	2	2	2	1	2	2	1	2	1
3	3	1	1	2	2	2	1	2	2	1	2
3	2	1	1	1	2	2	2	1	2	2	1
1	2	2	1	1	1	2	2	2	1	2	2
4	3	1	2	1	1	1	2	2	2	1	2

This design has $E(f_{NOD}) = 1.31$ and $E(\chi^2) = 0.77$. The design is f_{NOD}-optimal but not χ^2-optimal. In fact, the χ^2–efficeincy of the design is 0.73. This problem needs to be investigated further and work in this direction is in progress.

ACKNOWLEDGEMENTS

The authors are grateful to the Associate Editor and the referees for making very good suggestions, which has led to considerable improvements in the presentation of the results.

REFERENCES

Ai, Mingyao, Fang, Kai, Tai and He, Shuyuan (2007). $E(\chi^2)$-optimal mixed-level supersaturated designs. *J. Statist. Plann. Inf.*, **137**, 306–316**.**

Booth, K.H.V. and Cox, D.R. (1962). Some systematic supersaturated designs. *Technometrics*, **4**, 489–495.

Butler, N., Mead, R., Eskridge, K.M. and Gilmour, S.G. (2001). A general method of constructing $E(s^2)$-optimal supersaturated designs. *J. Roy. Statist. Soc.*, **B63**, 621–632.

Bulutoglu, D.A. (2007). Cyclicly Constructed $E(s^2)$-optimal supersaturated designs. *J. Statist. Plann. Inf.*, **137**, 2413–2428.

Bulutoglu, D.A. and Cheng, C.S. (2004). Construction of $E(s^2)$-optimal supersaturated designs. *Ann. Statist.*, **32**, 1662–1678.

Chen, J. and Liu, M. Q. (2008). Optimal mixed-level supersaturated design with general number of runs. *Statist. Probab. Lett.*, **78**, 2496–2502.

Das, A., Dey, A., Chan, L.Y. and Chatterjee, K. (2008). $E(s^2)$ - optimal supersaturated designs. *J. Statist. Plann. Inf.*, **138**, 3749–3757.

Fang, K.T. (1980). Experimental design by uniform distribution, *Acta Mathematicae Applicatae Sinica*, **3**, 363-372.

Fang, K.T., Lin, D.K.J., and Liu, M.Q. (2003). Optimal mixed-level supersaturated design. *Metrika*, **58**, 279–291.

Fang, K.T., Lin, D.K.J. and Ma, C.X. (2000). On the construction of multi–level supersaturated designs. *J. Statist. Plann. Inf.*, **86**, 239–252**.**

Gupta, S. and Chatterjee, K. (1998). Supersaturated designs: A review. *J. Comb. Inf. Sys. Sci.*, **23(1–4)**, 475–488.

Gupta, V.K., Parsad, R., Bhar, L.M. and Kole, B. (2008a). Supersaturated designs for asymmetrical factorial experiments. *J. Statist. Theo. Prac.*, **2**, 95–108.

Gupta, V.K., Parsad, R., Kole, B. and Bhar, L.M. (2008b). Computer-aided construction of efficient two level supersaturated designs. *J. Ind. Soc. Agril. Statist.*, **62(2)**, 183–194.

Gupta, V.K., Singh P., Kole, B. and Parsad, R. (2009). Construction of efficient unbalanced mixed-level supersaturated designs. *Statist. Probab. Lett.*, **79(22)**, 2359–2366.

Koukouvinos, C. and Mantas, P. (2005). Construction of some $E(f_{NOD})$ optimal mixed–level supersaturated designs. *Statist. Probab. Lett.*, **74(4)**, 312–321.

Li, P.F., Liu, M.Q. and Zhang, R.C. (2004). Some theory and the construction of mixed-level supersaturated designs. *Statist. Probab. Lett.*, **69**, 105–116.

Liu, Y., Ruan S. and Dean A.M. (2007a). Construction and analysis of $E(s^2)$ efficient supersaturated designs. *J. Statist. Plann. Inf.*, **137**, 1516–1529.

Liu, Y., Liu, M. and Zhang, R. (2007b). Construction of multi-level supersaturated design via Kronecker product. *J. Statist. Plann. Inf.*, **137(9)**, 2984–2992.

Lu, X. and Sun, Y. (2001). Supersaturated design with more than two levels. *Chinese Ann. Maths.*, **B22**, 183–194.

Nguyen, N.K. (1996). An algorithmic approach to constructing supersaturated designs. *Technometrics,* **38**, 69-73.

Nguyen, N.K. and Cheng C.S. (2008). New $E(s^2)$-Optimal supersaturated designs constructed from incomplete block designs. *Technometrics*, **50**, 26–31.

Ryan, K.J. and Bulutoglu, D.A. (2007). $E(s^2)$-Optimal supersaturated designs with good minimax properties. *J. Statist. Plann. Inf.*, **137**, 2250–2262.

Satterwaite, F. (1959). Random balance experimentation. *Technometrics*, **1**, 111–137.

Tang, Yu., Ai, M., Ge, G. and Fang, K.T. (2007). Optimal mixed–level supersaturated designs and a new class of combinatorial designs. *J. Statist. Plann. Inf.*, **137**, 2249–2301.

Xu, H. and Wu, C.FJ. (2003). Construction of optimal multilevel supersaturated designs. UCLA, Department of Statistics, Electronic Publication.(http://repositories.cdlib.orgluclastat).

Yamada, S. and Lin, D.K.J. (1999). Three level supersaturated designs. *Statist. Probab. Lett.*, **45**, 31–39.

ISAS

भारतीय कृषि सांख्यिकी संस्था की पत्रिका

अंक 63 खंड 3 दिसम्बर 2009 321-325

हिन्दी परिशिष्ट : इस अंक में प्रकाशित शोधपत्रों के सारांश

सुरेश चन्द्र राय

अंक 63 अगस्त 2009 खंड 3

अनुक्रमणिका

डॉ राजेन्द्र प्रसाद स्मारक भाषण :

भारत में खाद्य एवं पोषण सुरक्षा - कुछ समकालीन मुद्दे

एच.एस. गुप्ता
भारतीय कृषि अनुसंधान संस्थान, नई दिल्ली

इस लेख में भारतीय खाद्य योजनाओं का परीक्षण अभिनव आर्थिक एवं कृषि विकास की दृष्टि से किया गया है। मुख्य रूप से अभिनव खाद्य संकट, जलवायु परिवर्तन तथा कृषि वृद्धि में कमी पर विचार किया गया है। विश्व स्तर पर जलवायु परिवर्तन से उपज दर में कमी के विरुद्ध खाद्य पदार्थों की मांग बढ़ेगी। यह मानते हुए आत्मनिर्भरता तथा प्राकृतिक और आर्थिक वृद्धि के लिए सरकारी हस्तक्षेप को बढ़ावा देना होगा। देश में पोषण मुद्दों पर अधिक ध्यान की आवश्यकता है क्योंकि यहाँ एक तिहाई लोग कुपोषण से पीड़ित हैं। हमें खाद्यानों की उपज दर में वृद्धि के लिए अथक प्रयास करना होगा। यह केवल खाद्य आवश्यकताओं की पूर्ति के लिए ही नहीं बल्कि कुछ क्षेत्रों की अधिक मूल्यवान फसल जैसे फल, सब्जी आदि के लिए भी बचाना होगा। उपज दर में वृद्धि के लिए उपलब्ध ज्ञान तथा तकनीक का प्रयोग विशेष रूप से देश के पूर्वी क्षेत्रों में तत्काल प्रभाव डाल सकता है। फसल के विभिन्न किस्मों का विकास, प्रबंधन योजनाओं का प्रयोग तथा जलवायु परिवर्तन के प्रभाव को कम करने के लिए एक दूरगामी योजना का निर्माण करना चाहिए।

डॉ वी.जी. पान्से स्मारक भाषण :

प्रायोगिक अभिकल्पनाओं के निर्माण में सहविचरित चरों के चयन पर परावर्तन

बिकास के. सिन्हा
भारतीय सांख्यिकीय संस्थान, कोलकाता

प्रायोगिक अभिकल्पनाओं में सहविचरित चरों के उपयोग के सन्दर्भ में विचार किया गया है। इष्टतम सहविचरित अभिकल्पनाओं के सिद्धान्त में आधुनिक प्रोन्नति यह सुझाव देती है कि सहविचरित चरों के मान का इष्टतम चयन इसमें सार्थक सुधार कर सकता है। इसके सिद्धान्त तथा उपयोग पर प्रकाश डाला गया है ।

पूर्वी क्षेत्र के प्रदेशों में सामाजिक-आर्थिक विकास की विभिन्नता का मूल्यांकन

प्रेम नारायण, वी.के. भाटिया तथा एस.सी. राय
भारतीय कृषि सांख्यिकी संस्था, नई दिल्ली

देश के पूर्वी क्षेत्र के विभिन्न प्रदेशों के विकास स्तर का आकलन संयुक्त सूचकांक द्वारा किया गया है। संयुक्त सूचकांक अनेक सामाजिक-आर्थिक संकेतकों के आधार पर निर्मित है। इस क्षेत्र के पाँच बड़े प्रदेश तथा सात छोटे प्रदेश इस अध्ययन में लिए गए हैं। वर्ष 2001-02 के विभिन्न संकेतकों के आंकड़ों को विश्लेषण में प्रयोग किया गया है। विकास स्तर का आकलन कृषि क्षेत्र, अवसंरचना, सुविधाएँ तथा कुल सामाजिक-आर्थिक क्षेत्र के लिए अलग अलग किया गया है। सामाजिक-आर्थिक विकास में बड़े प्रदेशों में पश्चिम बंगाल तथा छोटे प्रदेशों में मिज़ोरम प्रथम स्थान पर पाए गए। विभिन्न प्रदेशों के विकास स्तर में विभिन्नता पाई गई। बड़े तथा छोटे प्रदेशों में अवसंरचना सुविधाएँ सामाजिक-आर्थिक विकास को धनात्मक रूप से प्रभावित करती हैं। एक समान क्षेत्रीय विकास के लिए कम विकसित छोटे प्रदेशों के विभिन्न संकेतकों के विभव लक्ष्य का आकलन किया गया।

बाजार आवक के आधार पर फूलों के उत्पादन के आकलन की पद्धति

ए.के. गुप्ता, एच.वी.एल. बठला, यू.सी. सूद तथा के.के. त्यागी
भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली

राष्ट्रीय सांख्यिकीय आयोग ने फूलों के क्षेत्रफल तथा उत्पादन के आकलन के लिए एक सुयोग्य प्रतिचयन पद्धति का विकास करने का सुझाव दिया। इस संदर्भ में भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली ने एक प्रतिचयन पद्धति द्वारा दिल्ली में महत्वपूर्ण फूलों के क्षेत्रफल तथा उत्पादन के आकलन के लिए एक प्रारम्भिक अध्ययन किया। इस अध्ययन में बाजार सर्वेक्षण तथा ग्राम सर्वेक्षण पद्धतियों को अपनाया गया। बाजार सर्वेक्षण पद्धति दिल्ली में फूलों की तीन मंडियाँ जो खारी-बावली, हनुमान मंदिर तथा महरौली में हैं, वहाँ अपनाई गईं। ग्राम सर्वेक्षण पद्धति दिल्ली

के याद्दच्छिक रूप से चयनित फूलों के उत्पादक ग्रामों में अपनाई गई। यह अध्ययन दर्शाता है कि प्रतिचयन पद्धति द्वारा फूलों के उत्पादन का आकलन अपेक्षित परिशुद्धता के साथ किया जा सकता है।

इकाई-स्तरीय आकाशीय मॉडल के अन्तर्गत लघु क्षेत्रीय अनुपात का आकलन

हुकुम चन्द्र
भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली

व्यापक रैखिक मिश्रित निदर्श (GLMM) जो स्थिर तथा याद्दच्छिक क्षेत्र संबंधी प्रभावों वाला होता है, उसका प्रयोग असतत चरों के लघु क्षेत्रीय आकलन (SAE) के लिए प्रायः किया जाता है, (मैक गिलक्रस्ट 1994 तथा राव 2003) GLMM में याद्दच्छिक क्षेत्रीय प्रभाव में क्षेत्र के मध्य परिवर्तनशीलता जो इस निदर्श में सम्मिलित सहायक चरों द्वारा प्रकट किया जाता है उसके अतिरिक्त स्पष्ट करता है। ये क्षेत्रीय प्रभाव प्रायः SAE से स्वतंत्र माने जाते हैं। व्यावहारिक रूप से क्षेत्रीय प्रभाव निकटवर्ती क्षेत्रों से सहसंबंधित होते हैं और सहसंबंध दूर के क्षेत्रों से शून्य हो जाता है। इस लेख में SAE आधारित GLMM जो आकाशीय सहसंबंधित याद्दच्छिक क्षेत्रीय प्रभाव पर अन्वेषण किया गया है, जहाँ निकटवर्ती ढांचा समीपस्थ आव्यूह द्वारा प्रकट किया गया है। आनुभविक सर्वोत्तम प्राग्वता की तुलना लघु क्षेत्रीय अनुपात से जो आकाशीय सहसंबंधित क्षेत्रीय प्रभाव के साथ हो अथवा उसके न होने पर अनुकार अध्ययनों से की गई है। अनुकार अध्ययन दो आंकड़ों के समूह पर किया गया है। जब आकाशीय सम्बद्धता लघु क्षेत्रीय SAE निदर्श में ली जाती है तो आनुभविक परिणाम केवल सीमान्त लाभ-दर्शाते हैं।

परिमित समष्टियों के लिए विकर्ण क्रमबद्ध प्रतिचयन योजना के कुछ और परिणाम

जे. सुब्रमणि
एस. वीरास्वामी चेट्टियार इन्जिनियरिंग तथा टेक्नालोजी कालेज, पुलियागुडी, तमिलनाडु

परिमित समष्टि के माध्य के आकलन के लिए विकर्ण क्रमबद्ध प्रतिचयन योजना के व्यापीकरण का प्रस्ताव किया गया है। यहाँ पर $n \le k$ की मान्यता में कुछ ढ़ील दी गई है; इसलिए प्रस्तावित पद्धति n के सभी मानों के लिए उपयुक्त है जब $N = Kn$ हो। प्राकृतिक समष्टि के लिए प्रस्तावित विकर्ण क्रमबद्ध प्रतिदर्श माध्य की आपेक्षिक दक्षता का आकलन सरल याद्दच्छिक तथा क्रमबद्ध प्रतिदर्श माध्य के साथ किया गया है।

असंबंधित प्रश्न निदर्श के अन्तर्गत सीधे तथा याद्दच्छीकृत अनुक्रियाओं के संयोग से संकुचन आकलन पर

काजल दिहिदार
भारतीय सांख्यिकीय संस्थान, कोलकाता

इस लेख में बिन्दुकता लक्षण वाली समष्टि अनुपात θ_A पर विचार किया गया है। इस मान्यता के साथ कि असंबंधित अबाधक गुण के समष्टि अनुपात अज्ञात हैं, असंबंधित प्रश्न निदर्श को लिया गया है। निदर्श जिसमें सीधी तथा याद्दच्छीकृत अनुक्रियाओं को परस्पर संबंधित किया गया है, उसमें सुधार करके θ_A का अनभिनत आकलन प्राप्त किया गया है। यहाँ पर इस संशोधित निदर्श के अन्तर्गत θ_A का पर्याप्त प्रारम्भिक मान θ_{Ao} के आधार पर संकुचन आकलक के निर्माण का प्रयास किया गया है। सामान्य संशोधित आकलक की तुलना में इस नवीन संकुचन आकलक की दक्षता का आकलन सैद्धान्तिक रूप से तथा कुछ संख्यात्मक उदाहरणों द्वारा किया गया है। इसके अतिरिक्त त्रुटि वर्ग माध्य का अनभिनत आकलक भी प्राप्त किया गया है।

द्वि-प्रावस्था प्रतिचयन में अननुक्रिया के साथ समष्टि तथा प्रक्षेत्र योग का आकलन

राज एस छिकारा तथा यू.सी. सूद*
हॉस्टन विश्वविद्यालय, हॉस्टन, यू.एस.ए.

इस लेख में द्वि-प्रावस्था प्रतिचयन विधि के अन्तर्गत, दोनों प्रक्षेय तथा समष्टि योग का आकलन, जहाँ प्रावस्था प्रथम के प्रतिदर्श इकाइयों से प्रक्षेत्र असात है, किया गया है। प्रतिचयन अभिकल्पना के इष्टतमत्व का अध्ययन, इकाइयों के अनुक्रिया की प्रायिकता, प्रथम तथा द्वितीय प्रावस्था

प्रतिचयन में व्यय तथा प्रक्षेत्र में इकाइयों की परिवर्तनशीलता को ध्यान में रख कर किया गया है। अनुकार अध्ययन के संख्यात्मक आंकड़ों के आधार पर यह पाया गया कि प्रस्तावित प्रतिचयन तथा आकलन विधि वैकल्पिक अग्रवाल तथा मिश्रा (2007) द्वारा दी गई विधि से अधिक दक्ष है।

* भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012

प्रतिबल सामर्थ्य की विश्वसनीयता के लिए इष्टतम अभिकल्पना

मनिषा पाल तथा एन.के. मन्डल
कोलकाता विश्वविद्यालय, कोलकाता

प्रतिबल सामर्थ्य निदर्श के अन्तर्गत किसी पद्धति की विश्वसनीयता को उस पद्धति के सामर्थ्य से जो वातावरण के प्रतिबल से अधिक हो, मापा जाता है। चूँकि पद्धति का सामर्थ्य अनेक नियंत्रित कारकों पर आधारित होता है, इसलिए विश्सनीयता के आकलन की यथार्थता इन कारकों के उचित चुनाव से की जा सकती है। इस लेख में प्रतिबल तथा सामर्थ्य स्वतंत्र चरघातांकी चर माने गए हैं जिसमें सामर्थ्य का माध्य नियन्त्रित कारकों का एक फलन है। सुयोग्य इष्टतमत्व कसौटी के उपयोग से पद्धति के विश्वसनीयता के आकलन के इष्टतम अभिकल्पना का प्रस्ताव किया गया है।

इष्टतम मिश्रित स्तरीय अतिसंतृप्त अभिकल्पनाओं का निर्माण

वी.के. गुप्ता, पूनम सिंह*, बासुदेव कोले तथा राजेन्द्र प्रसाद
भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली

इस लेख में मिश्रित स्तरीय f_{NOD} इष्टतम अतिसंतृप्त अभिकल्पनाओं के निर्माण की कुछ विधियाँ दी गई हैं। विधियों में एक समान अभिकल्पनाओं के निर्माण तथा उनके गुण एवं हाडामर्ड आव्यूह का उपयोग किया गया है। इस लेख में वर्णित अनेक अभिकल्पनाओं के $E(f_{NOD})$ तथा $E(\chi^2)$ के लिए गणितीय सूत्र प्राप्त किए गए हैं। निर्माण की कुछ पद्धतियों के उदाहरण दिए गए हैं। 67 इष्टतम मिश्रित स्तरीय अतिसंतृप्त अभिकल्पनाओं जिनमें अधिकतम 60 रन तथा 60 कारक हैं, उनकी सूची दी गई है। इस सूची में अभिकल्पना के कुछ अन्य दूसरे गुण भी दिए गए हैं। सभी अभिकल्पनाएँ f_{NOD} इष्टतम हैं जबकि कुछ अभिकल्पनाएँ χ^2 इष्टतम हैं।

* दिल्ली विश्वविद्यालय, दिल्ली

Available online at www.isas.org.in

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 327-328

Acknowledgements to the Reviewers

The Executive Council of the Indian Society of Agricultural Statistics and the Editorial Board of the Journal of the Indian Society of Agricultural Statistics in general and the Chair Editor in particular greatly appreciate and duly acknowledge the help and support received from all the Associate Editors as well as reviewers appointed by the Associate Editors as well as Chair Editor for critically reviewing the papers submitted for possible publication in the Journal of the Indian Society of Agricultural Statistics. By and large, the Associate Editors and reviewers have tried to maintain the time schedule. Their comments have always helped in improving the academic contents of the papers which in turn has helped in improving the quality of research publication in the Journal. This help is gratefully acknowledged.

Ranjana Agrawal, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Tauqueer Ahmad, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Kashinath Chatterjee, Visva Bharati University, Santiniketan, West Bengal

R.S. Chhikara, School of Science & Computer Engineering, University of Houston-Clear Lake, Bay Area Blvd Houston, Texas 77058, USA

Subir Ghosh, Department of Statistics, 2662, Statistics Computer Building, University of California, Riverside, CA 92521, USA

A.P. Gore, Statistical Services, CYTEL Software Services (P) Ltd., 8th Floor, Siddharth Tower, Off Karved, Kothrud, Pune - 411 029

P.C. Gupta, Sector 4 Ja 8, Jawahar Nagar, Jaipur – 302 004

Sat Gupta, Department of Mathematics & Statistics, University of North Carolina at Greensboro, NC 27402, USA

V.K. Gupta, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Seema Jaggi, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Rajni Jain, National Centre for Agricultural Economics & Policy Research, New Delhi – 110 012

D.K. Jain, National Dairy Research Institute, Karnal – 132 001

Sanpei Kageyama, Department of Environmental Design, Faculty of Environmental Studies, Hiroshima Institute of Technology, 2-1-1- Miyake Saeki-ku, Hiroshima 731-5193, Japan

Pranesh Kumar, Department of Mathematics, College of Science & Management, University of Northern British Columbia, 3333 University Way, Prince George, BC V2N 49, Canada

B.S. Kulkarni, Department of Statistics & Mathematics, Acharya NG Ranga Agricultural University, Rajendra Nagar, Hyderabad – 500 030

Dibyen Majumdar, Department of Mathematics, Statistics & Computer Science, University of Illinois, Chicago, IL 60607-7045, USA

P.K. Malhotra, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Ying Miao, Department of Social Systems and Management, Faculty of Systems and Information Engineering University of Tsukuba, Tennodai 1-1-1, Tsukuba 305-8573, Japan

Monica Pratesi, Dipartimento di Statistica e Matematica Applicata all'Economia, University of Pisa, Via Ridolfi, 10, 56124 - Pisa, Italy

Prajneshu, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Rajender Parsad, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Shyamal D. Peddada, Biostatistics Branch, NIEHS, Alexander Dr. RTP, NC 27709, USA

Anil Rai, IASRI, Library Avenue, Pusa, New Delhi – 110 012

B. Singh, Indian Veterinary Research Institute, Izatnagar – 243 122

Randhir Singh, 10, Avtar Enclave, Paschim Vihar, New Delhi – 110 063

M. Srinath, National Research Centre for Women in Agriculture, Opp. Kalinga Studio, Khandagiri, Bhubaneswar – 751 003

U.C. Sud, IASRI, Library Avenue, Pusa, New Delhi – 110 012

Girma Taye, Biometrics & Informatics, Ethiopian Institute of Agricultural Research (EIAR), PO Box 2003, Addis Ababa, Ethiopia

Zhiwu Yan, Abbott Laboratories, Dept R43V, Bldg AP9A-1, 100 Abbott Park Road, Abbott Park, IL 60064-6124, USA

OBITUARY

- The members of the Indian Society of Agricultural Statistics deeply mourn the sad and sudden demise of Prof. Jogabrata Roy, a life member of the Society, who left for heavenly abode in Kolkata. Prof. Roy took keen interest in the activities of the Society.

  Holding a Master's and D. Phil Degree in Statistics of Calcutta University, Prof. Roy spent sometime at University of North Carolina, Chapel Hill, working with Late Professor SN Roy on some aspects of Multivariate Testing. His work on Step-wise Regression is very well known and widely acclaimed. Prof. Roy had served in numerous capacities at Indian Statistical Institute (ISI) and also in the Government (CSO/NSSO/State Bureau) and in the UNDP. He had served as Acting Director of ISI for a short time.

  With his demise, the scientific community in general and the Indian Society of Agricultural Statistics in particular have lost a great statistician.

  The members of the Indian Society of Agricultural Statistics convey their heartfelt sympathies and condolence to the members of the bereaved family and pray to the Almighty to give eternal peace to the departed soul and enough strength to the bereaved family to bear this irreparable loss.

- The members of the Indian Society of Agricultural Statistics deeply mourn the sad and sudden demise of Shri Shanti Sarup, a life member of the Society, who left for heavenly abode on May 01, 2009 in Delhi. He took keen interest in the activities of the Society.

  Shri Shanti Sarup was born on September 01, 1936. He worked as Senior Scientist at Indian Agricultural Statistics Research Institute, New Delhi. He retired from the service on August 31, 1996.

  With his demise, the scientific community in general and the Indian Society of Agricultural Statistics in particular have lost a great statistician.

  The members of the Indian Society of Agricultural Statistics convey their heartfelt sympathies and condolence to the members of the bereaved family and pray to the Almighty to give eternal peace to the departed soul and enough strength to the bereaved family to bear this irreparable loss.

OBITUARY

◆ The members of the Indian Society of Agricultural Statistics deeply mourn the sad and sudden demise of Dr. K.B. Dutta, a life member of the Society, who left for heavenly abode on December 01, 2009 in Sambalpur, Orissa. He took keen interest in the activities of the Society.

Dr. K.B. Dutta was born on March 07, 1949. Starting his career as Lecturer in Statistics in Sambalpur University, Sambalpur, Orissa in 1977, Dr. Dutta rose to the position of Professor of Statistics in 1993. His fields of specialization were Statistical Inference, Discrete Probability Distributions, Statistical Genetics and Stochastic Process. He had published a number of research papers and articles in national and international journals. He was a life member of six Mathematical and Statistical Associations including Indian Society of Agricultural Statistics.

At the time of his demise, he was the Professor and Head, Department of Mathematics and Statistics, Synergy Institute of Engineering and Technology, Dhenkanal, Orissa.

With his demise, the scientific community in general and the Indian Society of Agricultural Statistics in particular have lost a great statistician.

The members of the Indian Society of Agricultural Statistics convey their heartfelt sympathies and condolence to the members of the bereaved family and pray to the Almighty to give eternal peace to the departed soul and enough strength to the bereaved family to bear this irreparable loss.

OTHER PUBLICATIONS OF THE SOCIETY

I. SAMPLING THEORY OF SURVEYS WITH APPLICATIONS

P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme and C. Asok

It is the third Revised Edition containing all the principal developments in the theory of sampling with examples and exercises.

The book contains 11 chapters

I. Introduction and Basic Concepts
II. Simple Random Sampling without Replacement
III. Sampling with Varying Probabilities
IV. Regression Methods of Estimation
V. Ratio Type Methods of Estimation
VI. Stratified Sampling
VII. Choice of Sampling Unit
VIII. Sub-Sampling
IX. Sub-Sampling (continued)
X. Systematic Sampling
XI. Non-Sampling Errors

Price : Rs. 150.00

II. IMPACT OF P.V. SUKHATME ON AGRICULTURAL STATISTICS AND NUTRITION

Edited by Prem Narain

It contains articles by eminent statisticians and other scientists in the country and abroad covering topics on Agricultural Statistics and Nutrition in which Prof. Sukhatme has made significant contributions.

Price : Rs. 35.00 (Paperback) and Rs. 50.00 (Hardbound)

III. SYMPOSIA ON

(i) Measurement of Impact of Green Revolution, and

(ii) Statistical Assessment of Intensive Cattle Development Programme.

Price : Rs. 10.00 (Inland) and $3.00 (Foreign)

IV. STATISTICAL DATA - THEIR CARE AND MAINTENANCE

David J. Finney

This bulletin is extremely useful for students and research workers engaged in data collection and analysis. It describes in a lucid manner how data can be scientifically gathered for drawing sound inference. The various topics dealt with are: acquisition of data, design of data gathering, care for data, types and units of data analysis and databases, copying, statistical ethics, data-entry to the computer, data scrutiny, integrity and some illustrations.

Price : Rs. 10.00 (Inland) and $3.00 (Foreign)

V. DIET, DISEASE AND DEVELOPMENT

Edmundson, Sukhatme and Edmundson

Price : Rs. 145.00

VI. SYMPOSIA ON

(i) Computer Intensive Techniques in Agricultural Research, and

(ii) Economic Reforms in Agriculture Sector— A Statistical Assessment.

Edited by Prem Narain and R.K. Pandey

Price : Rs. 100.00

Please order your copies to:

The Secretary
INDIAN SOCIETY OF AGRICULTURAL STATISTICS
IASRI Campus, Library Avenue, Pusa
New Delhi 110 012

EDITORIAL BOARD: 2009

Continued on next page towards left

CONTENTS

ISSN 0019-4344