香港統計學會
Hong Kong Statistical Society
c/o Department of Statistics & Actuarial Science,
The University of Hong Kong, Pokfulam Road, Hong Kong
http://www.hkss.org.hk

Bulletin
Volume 28 No. 2
December 2005 –
March 2006

Editor's Foreword

This is the second Bulletin of this year and is a fruitful one. Last December, we have the Hong Kong Statistical Conference and we have some articles included in this Bulletin. Nevertheless, I would urge you members to contribute some articles for this Bulletin, or, you may inform us some interesting news in statistics.

In this issue, we have our President's Forum and the dinner speech of Professor Man-keung Siu, for the celebration of the 60th birthdays of three outstanding members of our Society, Mr. Frederick W.H. Ho, Professors Tze-leung Lai and Kin Lam. Recently, as the advances of computer and internet, we may deal with massive electronic data effectively and efficiently. We have two articles discussing the impact of the new technology to statistical data collection, for example, the coming by-census 2006. First, we have Annie Y. W. CHAN to show us how to apply the GIS techniques in the areas of sampling frame, fieldwork operation and data dissemination. Also, Parmod K. Sharma and Alan T. L. Cheung discuss with us how to handle such collection of statistical data with e-reporting, particularly for the 2006 population bi-census in Hong Kong. The third article is by Wei Guo and our president, W.K. Fung that adjusting the chi-square test in case-control study for the admixed population using the neutral markers with known allele frequency differences between two ancestral populations. We thank all these authors for allowing us to reprint their articles which were originally published in the Proceedings of the 5th IASC Asian Conference.

Lastly, we would like to remind our members that we are going to have the AGM on March 30, 2006. Please participate this important function!

L.K. Li

		Phone	Fax	E-mail
Editor	: LI, Leong-kwan, PolyU	2766 6927	2362 9045	malblkli@polyu.edu.hk
Associate Editors	: Ms Chan, Jennifer So-kuen, HKU	2857 8316	2858 9041	jskchan@hku.hk
	Ip, Wai-cheung, PolyU	2766 6953	2362 9045	mathipwc@polyu.edu.hk
	Lau, Tai-shing	2609 7927	2609 7927	tslau@hp735a.sta.cuhk.edu.hk
Secretary	: Lam, John Hon-kwan, C&SD	2802 1267	2121 8296	jhklam@censtatd.gov.hk

CONTENTS

(Vol. 28/No.2, December 2005-March 2006)

President's Forum

Professor Tony W.K. FUNG

Time flies. I have been the President of the Society for three terms and it is coming to the end.

This year the Society has organized two major events. The first event is the 2005 Hong Kong Statistical Conference which was held on 17 December 2005. The Society co-organized the 5th IASC Asian Conference on Statistical Computing, which was held in parallel with the HKSS Conference. I have written a report of the two Conferences for your information. A specially invited session in celebrating the 60th birthdays of three outstanding members of our Society, Mr. Frederick W.H. Ho, Professors Tze-Leung Lai and Kin Lam, has been organized in the HKSS Conference. I am most grateful to Professor Man Keung Siu, Department of Mathematics, The University of Hong Kong to give a speech in honour of the 3 outstanding members during the Conference Dinner. The speech is published in this volume of the Bulletin and I am sure you'll find it very interesting. I would also like to thank Annie Chan, P.K. Sharma, Alan Cheung and Wei Guo for allowing us to reprint their articles which were originally published in the Proceedings of the 5th IASC Asian Conference.

The second major event is the 20th Anniversary of the Statistical Project Competition (SPC) for Secondary School Students. We are grateful to Wing Lung Bank Limited in sponsoring the event. The event is still on-going and I am sure that the Organizing Committee will report to members later on details of this special 20th Anniversial event. The Prize Presentation Ceremony will be held at the end of April.

I would like to take this opportunity to express my gratitude to the Council Members, Karen Chan (Vice President, C&SD), Howard Wong (General Secretary, C&SD), Raymond Tam (Treasurer, IVE), Man-Lai Tang (Programme Secretary, Baptist U), Leong-Kwan Li (Publications Secretary, Poly U), Agnes Law (Membership Secretary, City U) and Teresa Ng (Consultation Services Secretary, City U) and many other members of the Society for their strongest assistance and support during my term as the Society's President.

Applications of GIS Techniques in the Hong Kong 2006 Population By-census

Annie Y W CHAN
Census and Statistics Department

Abstract - **Geographic Information System (GIS) technology was applied in the 2001 Population Census through a computer system named "Digital Mapping System (DMS)". This is the first time that GIS techniques were used to support a population census. On field operation, the tailor-made and up-to-date maps of good quality produced by the DMS allowed field operation to be conducted more efficiently. On data dissemination, data analysis on small geographic areas could be performed more conveniently through using thematic maps.**

In the coming By-census scheduled to be conducted in July/August 2006, it is estimated that around 300 000 quarters will be enumerated. GIS techniques will continue to be applied in various aspects of the 2006 Population By-census. This paper discusses the applications of GIS techniques in the By-census in three broad areas: sampling frame, fieldwork operation and data dissemination.

1. Background

It is established practice from 1961 for Hong Kong to conduct a population census once every ten years and a by-census in the middle of the intercensal period. Following this practice, a population by-census will be conducted in Hong Kong in 2006.

Hong Kong covers an area of around 1 100 km^2 including over 200 outlying islands, many large and small towns, villages and rural communities and extensive areas of natural countryside. This diversity poses particular challenges for the organization and collection of census information. In the 2001 Population Census (01C), every single one of over 2 million quarters (accommodating 6.7 million people) was visited by a team of about 22 000 temporary field workers. To facilitate the fieldwork operation, Geographic Information System (GIS)[i] technology was applied in the 01C through the development of a computer system named "Digital Mapping System (DMS)".

In the coming By-census scheduled to be conducted in July/August 2006, it is estimated that around 300 000 quarters will be enumerated. GIS techniques will continue

to be applied in various aspects of the 2006 Population By-census (06BC).

2. The Digital Mapping System

Prior to the development of the DMS, a set of hardcopy base maps was maintained for supporting the population censuses/by-censuses and other household surveys conducted by the Census and Statistics Department (C&SD). Much manual effort was required in the maintenance of the map data and production of maps, involving such tasks as photocopying, cutting-and-pasting, and pencil marking on papers. Such effort, which was cumbersome and time-consuming, had to be repeated each time the base maps from the Survey and Mapping Office of the Lands Department[ii] were updated.

The DMS was developed in 2000 using a GIS software, viz. ArcGIS to support maintenance of spatial data and GIS applications. While making use of the digital maps developed by Lands Department as the base maps, C&SD adds on data layers to supplement information to meet its specific needs. Major functions of the DMS include:

- linkage of spatial data and textual records of buildings/structures to facilitate updating;
- production of tailor-made and up-to-date maps of good quality to support fieldwork operation; and
- dissemination of census/by-census results through thematic maps.

3. Application of GIS Techniques in the 06BC

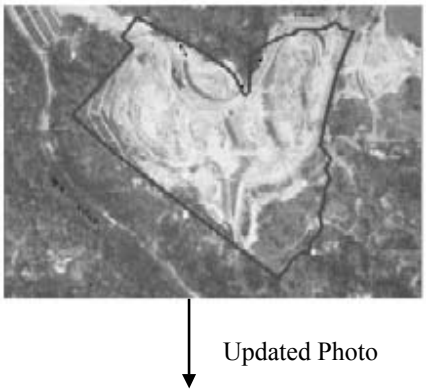
In the 06BC, the GIS techniques will be used more extensively in the following three areas: updating the sampling frame, supporting fieldwork operation and disseminating by-census results.

3.1 Updating the sampling frame

A complete sampling frame is a pre-requisite for obtaining reliable information from population censuses/by-censuses and other household surveys. C&SD maintains a "Frame of Quarters (FOQ)" which contains details of buildings/structures including their address information and units of quarters in Hong Kong. In the 06BC, a one-tenth sample of the quarters in Hong Kong will be selected from the FOQ for enumeration. To eliminate coverage errors like omission of quarters, it is absolutely important to ensure that the FOQ is complete and up-to-date.

With the implementation of the DMS, GIS techniques have been employed in updating the FOQ. Through overlaying the updated base maps or Digital Orthophotos[iii] on a regular basis, information on new buildings, demolished buildings, and other changes can be revealed from the maps and photos, and fed into the FOQ for updating the quarters/buildings records. **Figure 1** illustrates how the building records can be

updated making use of the Digital Orthophotos.



Updated Photo

Delineated building polygon by using Orthophoto

Figure 1: Update Building Records using Digital Orthophotos

3.2 Supporting fieldwork operation

GIS technology will be applied to support a wide range of fieldwork operation activities in the 06BC. Three new applications in the 06BC are highlighted below.

(i) *Delineation of the fieldwork management boundaries* - The 06BC fieldwork will be under a 3-tier management, from the lowest level of "division", to the middle level of "district", and to the highest level of "region". The 300 000 sampled quarters will first be grouped into 290 divisions, then into 19 working districts, and further into 5 regions. Based on the geometry and spatial relationships of buildings/structures and other relevant features (e.g. roads, footpaths, lanes, sea borders and mountains) as shown on the digital maps, the fieldwork management boundaries (i.e. divisions, working districts and regions) can be delineated using the drawing tools in the DMS.

(ii) *Identification of suitable location of field centres* - As mentioned in (i), there will be 19 working districts in the 06BC fieldwork operation. One field centre will be set up in each district to facilitate data collection work during the operation period. Schools will be deployed to serve as the field centres. Apart from meeting some basic criteria, the choice of the field centres should be determined with due reference to the location of the schools. For example, the field centres should preferably be at the centre of the working district; and the convenience and the availability of transportation facilities will also be the prime consideration.

GIS techniques can help in

identifying the suitable field centres based on a number of criteria related to the location of the schools: e.g. (a) High priority will be given to the short-listed schools with short average network distances from the sampled quarters; and (b) Schools with direct and fast means of transport in the proximity are considered more accessible and preferable. The availability of transportation facilities in the proximity of the short-listed schools can be observed from the digital map layers.

(iii) *Allocation of assignments and itinerary planning* – It is estimated that on average, around 60 sampled quarters will be visited by each enumerator in the 06BC. A number of parameters will have to be considered in determining and allocating the optimum number of sampled quarters to each enumerator such as the average traveling distances from the field centre to buildings with sampled quarters and that between buildings with sampled quarters. A new GIS module "Network Analyst"[iv] was plugged in the DMS in the 06BC to meet this purpose.

After allocating the sampled quarters to the enumerators, an itinerary for making household visits will be proposed for each of them. A route will be worked out for the enumerator on how to visit the sampled quarters based on the shortest walking distance calculated from the digital maps using GIS techniques. Such information will help the enumerators to better plan their work, and hence, lead to higher productivity and efficiency.

3.3 Disseminating by-census results

A population census/by-census, with its large scale, provides the most detailed statistical database of the population with detailed spatial reference to the locations where people reside. The extent of detail of both the data and the spatial reference renders statistics from population census/by-census most conducive to the utilization of GIS applications. Statistics from population census/by-census have all along been an important fundamental data layer in many GIS applications, both within the private and public sectors, supporting various modeling functions and better informed decision making with a spatial dimension.

Hong Kong has been using GIS technology to support the data dissemination work of population census/by-census since the 1991 Population Census. Thematic and area maps (see example in **Figure 2**) are produced using GIS technology to disseminate census/by-census results in publications and presentations. CD-ROM products containing detailed census/by-census data and digital maps, both produced by C&SD or in collaboration with private companies, allow users to perform spatial analysis using with GIS techniques.

It is a common practice to apply GIS techniques to disseminate and analyse census/by-census data at different geographical levels. C&SD will explore the wider use of GIS technology in disseminating the 06BC results, making reference to the latest development and past experience.

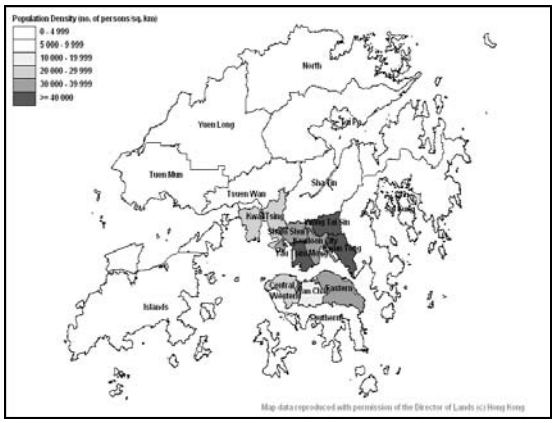


Figure 2: Choropleth Map Showing the Population density by District Council district based on 01C results

4. Conclusions

GIS is a useful tool for helping researchers and managers better understand problems, interpret data and make decisions involving a spatial dimension. For the general public, it enables them to better understand the community that they live in. Applying GIS techniques to perform spatial analysis on statistical data is certainly a potential era for further development in statistical computing.

---

[i] A Geographic Information System can be seen as a system of hardware, software and procedures designed to support the capture, management, manipulation, analysis, modeling and display of spatially-referenced data for solving complex, planning and management problems. (Goodchild & Kemp, 1990)

[ii] The Survey and Mapping Office of the Lands Department is the central authority for land surveys and all types of mapping in Hong Kong. It maintains a comprehensive set of maps in different scales in hard copy and digital form for land administration, town planning, engineering development, education, transportation, election and use by the community.

[iii] Digital Orthophotos are digital images of ground surface with uniform scale and positional information. The set of Digital Orthophotos currently used by the C&SD is produced from the aerial photographs taken at a flying height of 8000 feet. Distortions of photograph images caused by tilting of aerial camera and terrain relief are rectified except for those of the building structures. It consists of 189 tiles covering all the land area of Hong Kong. The ground pixel size of this set of Digital Orthophotos is 0.5m x 0.5m.

[iv] It provides tools to solve common network problems, such as finding the best route across a city, finding the closest emergency vehicle or facility, or identifying a service area around a site.

Electronic Data Reporting in Statistical Data Collection, with Particular Reference to the 2006 Population By-census in Hong Kong

Mr. Parmod K SHARMA and Mr. Alan T L CHEUNG
Census and Statistics Department

Abstract - **While the Internet has proven to be an effective and efficient channel for disseminating statistics, official statistical agencies are still engaged in finding the best way of using the Internet for data reporting. Low take-up rate and high implementation cost are major challenges in implementing e-reporting solutions.**

The Census and Statistics Department (C&SD) of the Hong Kong Special Administrative Region Government has been making available electronic questionnaires in spreadsheet format for all surveys amenable to e-reporting, without making a huge investment in technical infrastructure. For the 2006 Population By-census, the C&SD will be using a new e-reporting solution which will allow respondents to use the freely available Adobe Acrobat Reader to complete e-questionnaires. An online interface will be provided to enable respondents to download customized e-questionnaires for completion and to upload completed data securely over the Internet.

This paper discusses the considerations leading to the development of the new e-reporting solution, its benefits, cost-effectiveness and some limitations.

Introduction

Official statistical agencies have been using the Internet in different areas of statistical work, particularly in data dissemination and data reporting. While the Internet has proven to be an effective and efficient channel for disseminating official statistics, many official statistical agencies are still engaged in finding the best way of using the Internet for data reporting. In population census/by-census, many countries/ territories have already provided electronic data reporting (e-reporting) option in their respective previous rounds of operation. Despite having gained some practical experience, there are undoubtedly still many challenges faced when planning and implementing e-reporting solutions.

Major Challenges

One of the most significant challenges is the take-up rate. Compared to other reporting modes, the take-up rate for

e-reporting is low. Various factors contribute to this phenomenon. First, respondents may be highly accustomed to traditional modes of data reporting and do not see the benefit of moving to an e-reporting mode. Second, respondents may have concerns over transmitting data over the Internet; such are often more psychological barriers than real security threats since data can always be encrypted during transmission.

A second major challenge is cost. Protection of data confidentiality is a key concern in developing e-reporting solutions but building a secure online e-reporting system (i.e. allowing respondents to retrieve and report data on a real-time basis) can be a very costly venture, not to mention tackling other technical issues such as authenticating respondents and ensuring system compatibility under different computing environments.

Acknowledging that e-reporting is just one of the data reporting options that respondents expect to be available and given the slow adoption rate, the Census and Statistics Department (C&SD) of the Hong Kong Special Administrative Region Government considers it more prudent to adopt a pragmatic approach in developing e-reporting solutions. Without making a huge investment in technical infrastructure, C&SD makes available electronic questionnaires (e-questionnaires) for all surveys (mainly business surveys) amenable to e-reporting as an additional reporting option. These e-questionnaires are currently in spreadsheet format (mainly Microsoft Excel), which has a number of limitations. In particular, the use of a software package that needs to be purchased is considered not very suitable for a large scale household survey like the 2006 Population By-census (06BC) since many people may not have such software on their personal computers (PC) and there could be technical issues when older versions of such software are used to complete the e-questionnaires which often have built-in macros. This called for a review and search for an optimal e-reporting solution for the 06BC.

Reporting Solution for 06BC

After a thorough study, an e-questionnaire approach coupled with an Internet application for uploading/downloading e-questionnaires was selected as the e-reporting approach for the 06BC. Notable changes compared to the existing approach were (a) enhanced features of the e-questionnaires and (b) the online uploading/ downloading module. Key requirements of the new e-questionnaire solution were formulated as follows:-

(i) user-friendliness – the e-questionnaire should be an electronic representation of the paper questionnaire with built-in

validations/checks and question skipping;

(ii) form design flexibility – designing the e-questionnaire should not require complex programming skills. This would facilitate making changes to the questionnaire during the planning and testing stages;

(iii) encryption and password protection – the e-questionnaire should support data encryption during transmission and whilst held on the web server as well as unique passwords for individual e-questionnaires to safeguard data confidentiality;

(iv) multi-mode data/file delivery – the delivery of the e-questionnaires to the respondents and hence, the return of the completed e-questionnaires should be supported by both on-line and off-line modes.

Among the available e-questionnaire solutions, the e-questionnaire in Portable Document File (PDF) format, with extension features to enable respondents to use the freely available Adobe Acrobat Reader to complete the e-questionnaires was chosen. The solution would be customized to streamline the operational flow and will be supported by an Internet application for uploading/downloading the e-questionnaires. The technical infrastructure is illustrated in Figure 1.

Regarding the operational flow, households will first be requested to provide basic information including number of household members, their names and relationship via telephone. Once entered with this basic information, separate e-questionnaires will be generated for each member of the household and each will be protected by a unique password. Using the password that will be delivered to each household member, he/she will be able to login at the designated website to download his/her e-questionnaire and, after completion, submit the data portion via Secure Socket Layer (SSL), by simply clicking a button built-in on the e-questionnaire. As the data portion constitutes just a fraction of the size of the full PDF file, the uploading can be completed more efficiently. The operational flow is summarized in Figure 2.

A trial version of the PDF e-questionnaire and the Internet application had been developed and tested in the 06BC Pilot Survey conducted in July – Aug 2005. A screenshot of the e-questionnaire is at Figure 3. Through this test, C&SD has been able to assess a number of aspects of the e-reporting solution, including public perception and ability to complete the e-questionnaires, the design and operation flow of the e-reporting solution, the quality of data collected and the profile of respondents using the e-questionnaire option.

Findings from the Pilot Survey on the e-reporting solution are quite promising. Respondents using the PDF e-questionnaire were able to complete the questionnaires without much technical difficulties (note: all the technical problems raised by the respondents had since been resolved). The data collected for some items were of high quality, even better than the traditional interviewing method. This is probably because skipping questions and simple data validation features had been built into the e-questionnaires. However, for some items that required complicated and long written answers, the data quality was not that good and follow-up with the respondents was necessary. Based on the valuable experience gained in the Pilot Survey, the e-reporting solution will be further fine-tuned for implementation in the 06BC.

Benefits

Being in a PDF format which is very popular nowadays, it is expected that the e-questionnaires will be well accepted by respondents. More importantly, since PDF e-questionnaires can retain the look and feel of the original questionnaire, this will increase their intuitive appeal to the respondents. Respondents will also find the PDF e-questionnaires user-friendly and easy to complete since they are only required to use the freely available Adobe Acrobat Reader which they probably already have on their PCs. In addition, secure transmission of data can be ensured with the use of the PDF's password protection and encryption functions. Last but not the least, the implementation cost will be comparatively low. For the case of 06BC, it is estimated that the total cost of the proposed solution (including development of the PDF e-questionnaire and the Internet application) will be around one-third of that required for developing a full fledged on-line e-reporting system.

Limitations and Shortcomings

The PDF e-questionnaire approach is not without limitations and implementation difficulties. First, not all potential respondents have access to the Internet and those without such access (though expected to be small in number) cannot use the PDF e-questionnaire in the present context. Second, there are some operational and technical issues still to be resolved. For example, there may be a chance that one member of a household can mistakenly get hold of the letter (containing the username and password) of the other members of the household and get access to the other members' pre-filled e-questionnaires. Thirdly, being a rather novel solution, "teething problems" can be expected in the implementation process. Potential users are forewarned to make provision for sufficient time and manpower with the necessary skill

set to develop and implement the solution.

Potential Deployment to Other Surveys

The PDF e-questionnaire is considered to be a versatile solution that can be applied in other surveys (both household and business).

For complex surveys, the PDF e-questionnaires can be offered in the extension mode as in the case of the 06BC where respondents can download the e-questionnaires, complete them offline at their leisure and submit the data portion securely online. For simpler surveys, a more cost effective approach is to offer PDF e-questionnaires without the extension mode in which case respondents will access, complete and submit the PDF e-questionnaire, all online through a dedicated, secure website.

Conclusion

While promoting user friendliness and encouraging respondents to use e-reporting, it is very important to manage the implementation cost at a reasonable level so as to achieve good "value-for-money".

The PDF e-questionnaire solution looks quite promising in respect of its ease of use, user-friendliness and cost effectiveness and is therefore worth to be pursued.

The C&SD will adopt the PDF e-reporting solution in the 06BC. Depending on the results of the implementation, consideration would be given to adopting the PDF e-reporting solution as a standard e-reporting option for other business and household surveys.

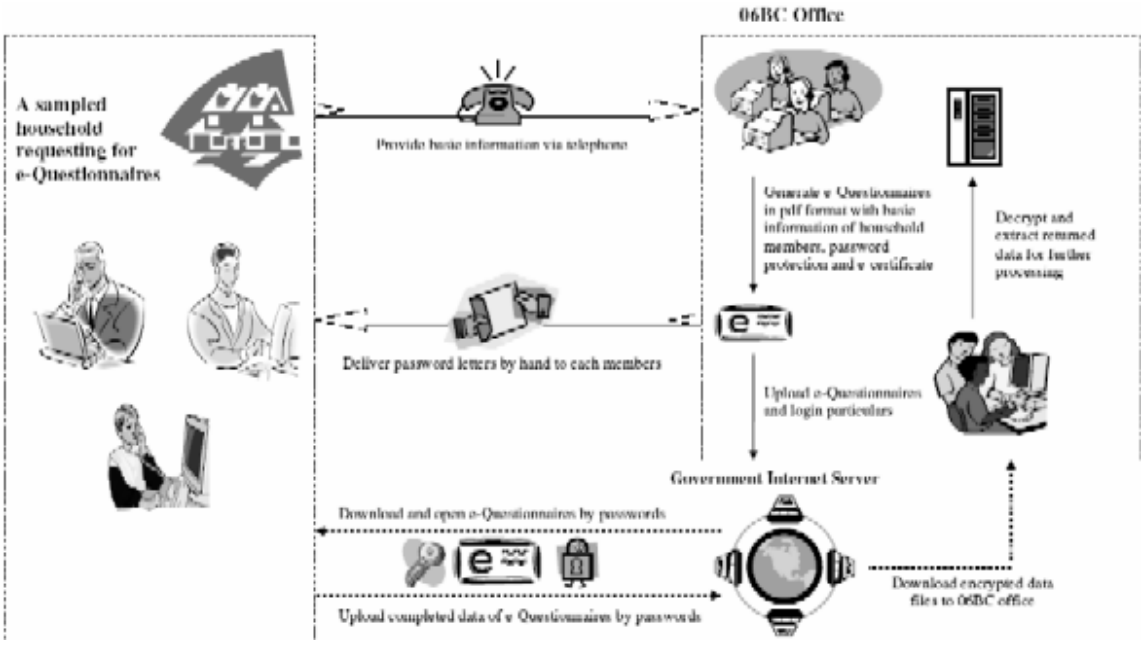


Figure 1: Technical infrastructure for the e-questionnaire in the 06BC

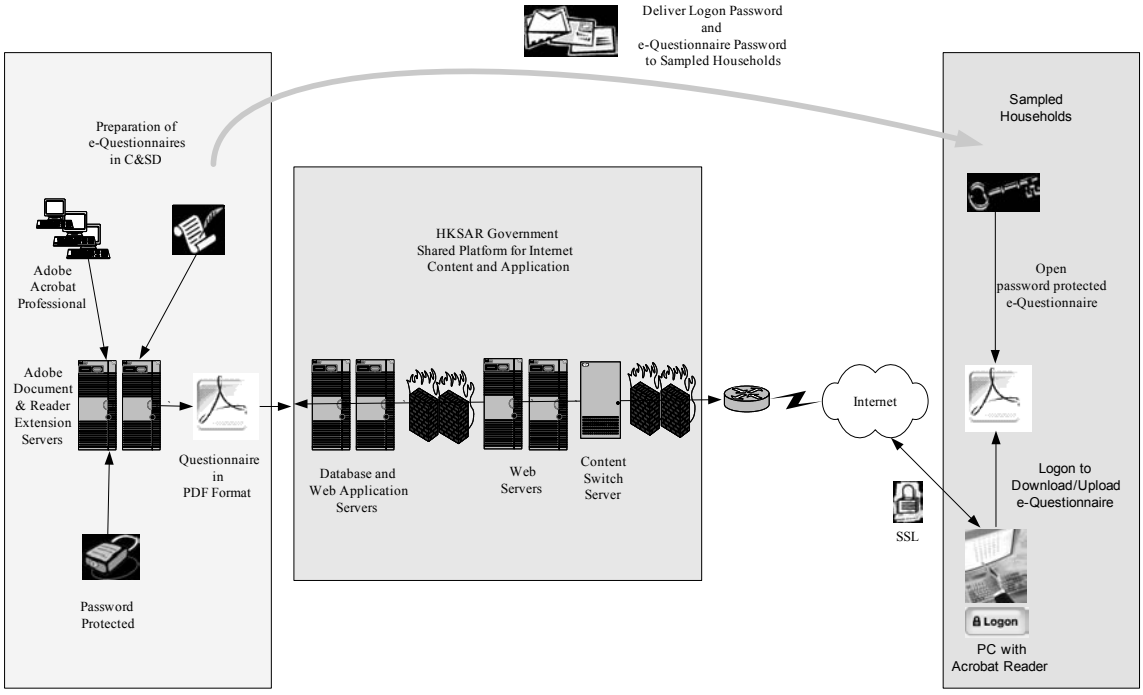


Figure 2: Operational flow for the e-questionnaire in the 06BC

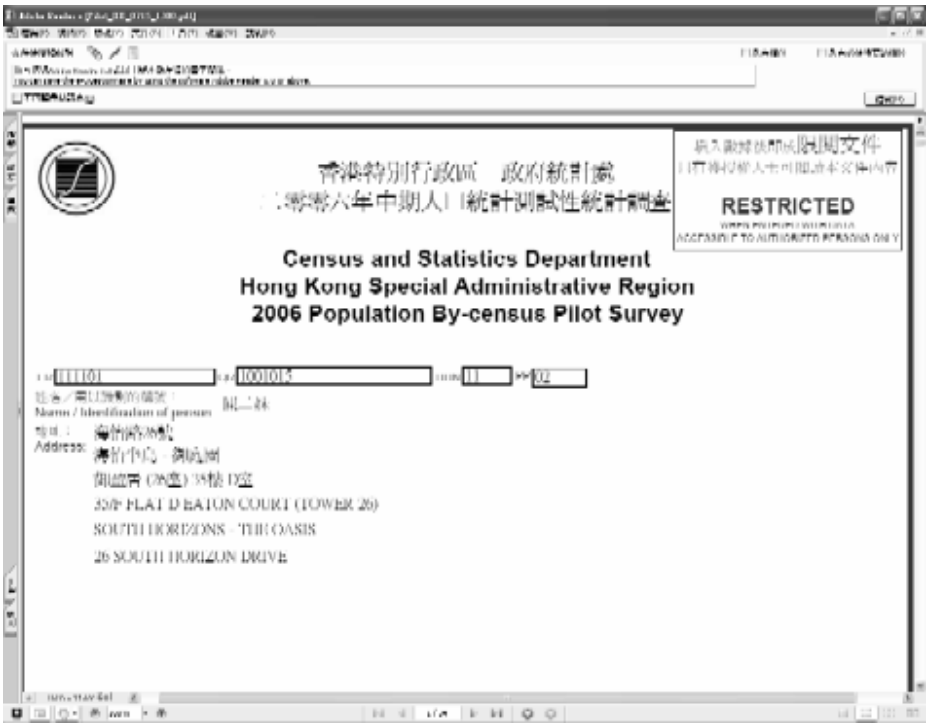


Figure 3: A screenshot of the e-questionnaire in the 06BC Pilot Survey

An Adjusted Chi-Square Test for Case-Control Study in an Admixed Population

Wei GUO and, Wing K. FUNG
The University of Hong Kong

Abstract - **Case-control study has been advocated as a powerful method for mapping complex-trait loci. However, when a case-control study is used to identify the disease-susceptible loci, spurious disease-marker association may be caused by population stratification or population admixture. Genomic control (GC) using neutral loci has been suggested to correct for spurious association, but it does not control the size of the test well in some situations because of the unknown allele frequencies in subpopulations. Here, we propose a way to adjust the chi-square test in case-control study for the admixed population using the neutral markers with known allele frequency differences between two ancestral populations.**

Introduction

Case-control association study is a powerful tool for inferring non-random association between the disease-susceptible loci and some complex-trait loci [1]. However, in case-control study, population stratification or population admixture can lead to the false positive problem, or called 'spurious associations', even at marker loci completely unlinked to the disease locus [2,3]. The basic approach of a case-control study is to compare the allele frequencies between the affected cases and the unaffected controls. This procedure rests on a key assumption that the case and control groups are sampled from an identical population without substructure.

In order to control the false positive problem in case-control study, the family data can be used to match the ethnic backgrounds of patients to controls carefully [4-6]. Another strategy is by GC, which is to measure and correct for population substructure by a moderate number of unlinked markers in the same set of cases and controls [7-9]. However, Shmulewitz [10] commented that GC can lead to a notable loss of power to detect a true association (conservative) in many circumstances or may fail to eliminate the spurious associations (anticonservative). Moreover, Pfaff [11] suggested that, in the case of population admixture, we would better do some adjustments for the case-control tests based on the allele frequency differences between two parental populations, which could be available in some

admixed populations. In this article, we propose a way to adjust the chi-square test in case-control study for the admixed population using the neutral markers with known allele frequency differences between two ancestral populations.

Methods

Consider a case-control experiment with n_{1*} affected cases and n_{2*} unaffected controls sampled independently from an admixed population. We assume a disease locus with alleles D and d, and a number of unlinked marker loci, designated Mi and mi as two alleles on the ith marker locus, $i = 0, L$, where '0' indicates the candidate marker and others indicate the neutral loci. Suppose the penetrances of the disease given genotypes DD, Dd and dd are f_2, f_1 and f_0, respectively, and $A = f_2q^2 + 2f_1q(1 - q) + f_0(1 - q)^2$ is the prevalence of the disease in the population. Let θ be the recombination fraction between the candidate marker and disease locus, and the recombination fractions between any neutral marker locus and disease locus are 0.5.

	Allele M_i	Allele m_i	+
Cases	$n_{11}^{(i)}$	$n_{12}^{(i)}$	n_{1*}
Controls	$n_{21}^{(i)}$	$n_{22}^{(i)}$	n_{2*}
+	$n_{*1}^{(i)}$	$n_{*2}^{(i)}$	n_{**}

Table 1: Allele distribution on the ith marker in the case-control experiments.

For each marker locus, the data can be summarized via a 2 × 2 allelic table (Table 1). Based on the allelic data, the classical chi-square statistic with 1 degree of freedom can then be calculated:

$$x_{(i)}^2 = n_{**}\frac{n_{1*}n_{2*}}{n_{*1}^{(i)}n_{*2}^{(i)}}\left(\frac{n_{11}^{(i)}}{n_{1*}} - \frac{n_{21}^{(i)}}{n_{2*}}\right)^2, \qquad (1)$$

$$i = 0, 1, ..., L.$$

A high $x_{(0)}^2$ value may suggest that the candidate marker is linked to, or itself is a disease susceptibility locus. But all other chi-square statistics, $x_{(1)}^2$, ..., $x_{(L)}^2$, are used only as genomic control in the following tests when admixture exists.

Consider an admixed population (e.g. Africa American) which is descended from two ancestral populations (e.g. Africans and Europeans). Let p_i and q be the allele frequencies of M_i and D in the current admixed population, respectively. We assume that the penetrances are the same in both parental populations.

To control the false positive problem in an admixed population, one of the GC methods is to use a modified chi-square statistic

$$x_{(0)}^2 / \lambda, \qquad (2)$$

where $\lambda = \frac{1}{L}\sum_{i=1}^{L} x_{(i)}^2$ is the mean chi-square value across L unlinked markers [7-9].

Before considering our adjusted chi-square test in an admixed population, it needs to figure out the admixture linkage disequilibrium pattern first. Let $\Delta^{(i)}$ be the coefficient of linkage disequilibrium (LD) between the marker allele Mi and the disease allele D, defined as $\Delta^{(i)}$= *freq*(M_iD) - *freq*(M_i)*freq*(D). Irrespective of the hybrid isolation model or the continuous gene flow model, one of the important properties is that the coefficient of LD ($\Delta^{(i)}$) is proportional to the allele frequency difference at Mi between two parental populations, denoted by δ_i, and furthermore, it is a strict monotonic function of the recombination fraction between the *i*th marker locus and the disease locus [12-14]. So $\Delta^{(i)}/\delta_i$ are identical for all unlinked markers due to the same recombination fractions (θ = 0.5).

The conditional probabilities of the marker allele Mi in the cases and controls can be calculated as follows,

$$\begin{aligned} P(M_i \mid \text{case}) &= p_i + \Delta^{(i)}[q(f_2 - f_1) \\ &\quad + (1-q)(f_1 - f_0)]/A, \\ P(M_i \mid \text{control}) &= pi - \Delta^{(i)}[q(f_2 - f_1) \\ &\quad + (1-q)(f_1 - f_0)]/(1-A). \end{aligned} \tag{3}$$

Then the expectation of the allele frequency difference at Mi between cases and controls is

$$E\left(\frac{n_{11}^{(i)}}{n_{1*}} - \frac{n_{21}^{(i)}}{n_{2*}}\right) = \Delta^{(i)}[q(f_2 - f_1) + (1-q)(f_1 - f_0)/[A(1-A)], \tag{4}$$

When there is no association, i.e. linkage equilibrium, it is expected that the allele frequencies do not vary across cases and controls. However, in the case of an admixed population, the association may arise due to admixture even for unlinked marker loci. Nevertheless, we notice that

$$E\left[\frac{1}{\delta_0}\left(\frac{n_{11}^{(i)}}{n_{1*}} - \frac{n_{21}^{(i)}}{n_{2*}}\right)\right] = \frac{\Delta^{(i)}}{\delta_i}[q(f_2 - f_1) + (1-q)(f_1 - f_0)/[A(1-A)], \tag{5}$$

are identical for any $i \geq 1$ since $\Delta^{(i)}/\delta_i$ are equal for all unlinked markers. Thus, an adjusted chi-square test can be constructed as follows,

$$T = \frac{\left[\frac{1}{\delta_0}\left(\frac{n_{11}^{(0)}}{n_{1*}} - \frac{n_{21}^{(0)}}{n_{2*}}\right) - Mean\right]^2}{Variance}, \tag{6}$$

where the mean and variance can be approximated using the unlinked markers and they are given as,

$$\begin{aligned} \text{Mean} &= \frac{1}{L}\sum_{i=1}^{L}\frac{1}{\delta_i}\left(\frac{n_{11}^{(i)}}{n_{1*}} - \frac{n_{21}^{(i)}}{n_{2*}}\right), \\ \text{Variance} &= \frac{1}{\delta_0^2}\left(\frac{n_{11}^{(0)}n_{12}^{(0)}}{n_{1*}^3} + \frac{n_{21}^{(0)}n_{22}^{(0)}}{n_{2*}^3}\right) \\ &\quad + \frac{1}{L^2}\sum_{i=1}^{L}\frac{1}{\delta_i^2}\left(\frac{n_{11}^{(i)}n_{12}^{(i)}}{n_{1*}^3} + \frac{n_{21}^{(i)}n_{22}^{(i)}}{n_{2*}^3}\right). \end{aligned} \tag{7}$$

Simulation

We investigate the performance of the proposed method in a recent admixed population. In generation 1, the offspring of

two linkage equilibrium subpopulations combine into the first generation according to the contribution proportions α and 1 - α, and then random mating is maintained in the next two generations. 200 cases and 200 controls are sampled form this admixed population. Consider L unlinked loci that the allele frequencies in subpopulations 1 and 2 are independently generated from a uniform distribution between 0 and 1, but the absolute values of the allele frequency differences are taken to be larger than a fixed value β. Three disease models of inheritance are considered: (1) recessive model $f_2 = 1$, $f_1 = f_0 = 0$; (2) additive model $f_2 = 1$, $f_1 = 0.5$, $f_0 = 0$; (3) dominant model $f_2 = f_1 = 1$, $f_0 = 0$. The parameter values are taken as $\beta = 0$, 0.1 and $\alpha = 0.3$, 0.5, 0.7, with $L = 20$, 50, 100, 500. The frequency of the disease allele D is taken as 0.1 in subpopulation 2, and the frequency in subpopulation 1 is determined by the relative risk, which is the ratio of the disease prevalence of subpopulation 1 to that of subpopulation 2 and this ratio is taken as 3 under all three disease models.

When the allele frequency difference at the candidate marker ($\delta_{(0)}$) is large, the size of the existing $x^2_{(0)}/\lambda$ test is around 10% - 25% (Table 2), which is much larger than the nominal significance level of 5%. However, when $\delta_{(0)}$ is moderate, the $x^2_{(0)}/\lambda$ test is conservative, of which the size is mainly around 0.3% - 4%. In fact, because of the large magnitude of the noncentral parameter caused by a strong LD, the $x^2_{(0)}/\lambda$ test, having the adjustment using a constant, cannot work well whatever the constant is [10]. On the contrary, the size of the chi-square test T is close to the nominal significance level of 0.05 irrespective of the parameter values such as the admixture proportion, number of unlinked markers, β value and so on. When the LD decays rapidly for a longer history of admixture, though it is still not as satisfactory as the adjusted chi-square test T, the size of the $x^2_{(0)}/\lambda$ test is closer to the nominal 5% level. These results and the simulated results on the power of the tests are not shown here for brevity.

Discussion

The spurious disease-marker association may be caused by population stratification or population admixture. The admixture linkage disequilibrium can result in excessive false positives though the admixed population is a random mating population. Moreover, irrespective of the hybrid isolation model or the continuous gene flow model, one important property is that the LD is proportional to the difference of marker allele frequencies between two parental populations. So we propose a way to adjust the chi-square test for case-control study in an admixed population using the neutral markers with known allele frequency differences between two ancestral populations. From the

simulation results, the method we proposed is found to be valid as it controls the size well.

Acknowledgements

This project was partly supported by the Natural Science Foundation of China for Distinguished Young Scholars (Grant Number 10329102).

References

[1] Risch N. and Merikangas K. (1996). The future of genetic studies of complex human diseases. Science, 273, 1516-1517.

[2] Chakraborty R. and Weiss K.M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA, 85, 9119-9123.

[3] Lander E.S. and Schork N.J. (1994). Genetic dissection of complex traits. Science, 265, 2037-2048.

[4] Horvath S. and Laird N.M. (1998). A discordant-sibship test for disequilibrium and linkage, no need for parental data. Am J Hum Genet, 63,1886-1897.

[5] Spielman R.S. and Ewens W.J. (1998). A sibship test for linkage in the presence of association, the sib transmission / disequilibrium test. Am J Hum Genet, 62, 450-458.

[6] Ewens W.J. and Spielman R.S. (1995). The transmission/disequilibrium test, history, subdivision, and admixture. Am J Hum Genet, 57, 455-464.

[7] Devlin B. and Roeder K. (1999). Genomic control for association studies. Biometrics, 55, 997-1004.

[8] Pritchard J.K. and Rosenberg N.A. (1999). Use of unlinked genetic markers to detect population stratification in

association studies. Am J Hum Genet, 65, 220-228.

[9] Reich D.E. and Goldstein D.B. (2001). Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol, 20, 4-16.

[10] Shmulewitz D., Zhang J.Y. and Greenberg D.A. (2004). Case-control association studies in mixed populations, correctiong using GC. Hum Hered, 58, 145-153.

[11] Pfaff C.L. (2002). Adjusting for population structure in Admixed populations. Genet Epidemiol, 22, 196-201.

[12] Pfaff C.L., Parra E.J., Bonilla C., Hiester K., McKeigue P.M., Kamboh M.I., Hutchinson R.G., Ferrell R.E., Boerwinkle E. and Shriver M.D. (2001). Population structure in admixture populations, effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet, 68, 198-207.

[13] Guo W., Fung W.K., Shi N.Z. and Guo J.H. (2005). On the formula for admixture linkage disequilibrium. Hum Hered 60, 177-180.

[14] Guo W. and Fung W.K. (2006). The admixture linkage disequilibrium and genetic linkage inference on the gradual admixture population. Acta Genetica Sinica, 33, 12-18.

Table 2: The empirical sizes of the T test and $x^2_{(0)}/\lambda$ test under the recessive, additive and dominant models.

			p_1=0.9, p_2=0.1						p_1=0.4, p_2=0.1					
			Recessive		Additive		Dominant		Recessive		Additive		Dominant	
β	α	L	T	$x^2_{(0)}/\lambda$	T	$x^2_{(0)}/\lambda$	T	$x^2_{(0)}/\lambda$	T	$x^2_{(0)}/\lambda$	T	$x^2_{(0)}/\lambda$	T	$x^2_{(0)}/\lambda$
	0.3	20	0.042	0.205	0.026	0.191	0.039	0.246	0.046	0.05	0.037	0.037	0.053	0.034
		50	0.052	0.168	0.045	0.191	0.047	0.199	0.044	0.033	0.054	0.035	0.055	0.024
		100	0.048	0.18	0.04	0.168	0.054	0.216	0.043	0.039	0.057	0.026	0.051	0.022
		500	0.042	0.152	0.051	0.212	0.046	0.224	0.053	0.038	0.045	0.036	0.053	0.018
0	0.5	20	0.041	0.202	0.038	0.218	0.044	0.25	0.038	0.047	0.039	0.042	0.044	0.033
		50	0.038	0.146	0.037	0.196	0.059	0.202	0.037	0.031	0.042	0.032	0.049	0.026
		100	0.05	0.158	0.045	0.161	0.04	0.211	0.044	0.029	0.052	0.027	0.061	0.024
		500	0.045	0.156	0.035	0.184	0.055	0.199	0.046	0.036	0.055	0.03	0.044	0.018
	0.7	20	0.037	0.158	0.037	0.168	0.049	0.224	0.044	0.051	0.047	0.045	0.043	0.052
		50	0.053	0.134	0.046	0.145	0.048	0.171	0.047	0.029	0.05	0.041	0.048	0.026
		100	0.063	0.137	0.042	0.163	0.044	0.176	0.052	0.035	0.056	0.028	0.054	0.019
		500	0.04	0.153	0.044	0.146	0.044	0.189	0.052	0.032	0.048	0.031	0.044	0.021
	0.3	20	0.045	0.143	0.049	0.128	0.04	0.126	0.041	0.037	0.045	0.032	0.034	0.009
		50	0.039	0.116	0.041	0.136	0.044	0.145	0.042	0.026	0.058	0.034	0.048	0.007
		100	0.057	0.132	0.048	0.135	0.039	0.124	0.038	0.019	0.046	0.018	0.039	0.003
		500	0.044	0.131	0.046	0.147	0.059	0.158	0.048	0.028	0.045	0.016	0.07	0.013
0.1	0.5	20	0.037	0.119	0.03	0.151	0.045	0.177	0.048	0.029	0.041	0.022	0.032	0.014
		50	0.045	0.124	0.045	0.134	0.046	0.157	0.05	0.028	0.044	0.021	0.049	0.009
		100	0.049	0.113	0.055	0.128	0.05	0.126	0.053	0.023	0.042	0.014	0.055	0.005
		500	0.05	0.135	0.055	0.163	0.05	0.139	0.044	0.02	0.046	0.013	0.051	0.011
	0.7	20	0.049	0.134	0.044	0.134	0.045	0.154	0.043	0.037	0.044	0.039	0.045	0.017
		50	0.047	0.117	0.048	0.144	0.041	0.127	0.052	0.029	0.035	0.021	0.045	0.013
		100	0.054	0.116	0.033	0.108	0.049	0.11	0.041	0.02	0.042	0.021	0.044	0.012
		500	0.049	0.106	0.053	0.119	0.047	0.147	0.053	0.021	0.051	0.023	0.043	0.013

Speech during the HKSS Conference Dinner

Professor Man-keung SIU

It is my great pleasure and honour to speak on this happy occasion. The three distinguished members the Hong Kong Statistical Society honours tonight are my old friends for over forty years. I hope they regard me the same reciprocally!

The trio --- Kin, Tze-Leung and Wing-Huen --- form a triangle. I am like the Fermat point of this triangle, that is, the point with minimum total distance from the three vertices. Unlike the Fermat point, my proximity to each of the three differs with time and space.

I got to know Wing-Huen the earliest. We were classmates in school since the late 1950s. (PHOTO 1 – a class picnic in the Spring of 1960)

(PHOTO 2 – in a less playful mood after the graduation ceremony in December of 1961.)

Wing-Huen is an articulate fellow who is good at both Chinese and English. This did not escape the notice of our Chinese Master, who always praised him for this strength of his --- perhaps he was sometimes a bit too articulate in class! We entered Hong Kong University as undergraduates in the Faculty of Science together, and both of us belonged to St. John's College. Wing-Huen is definitely a much more lively and able person than I. He became the Chairman of the College Association, demonstrating the quality of leadership one would witness later

in his long career in the civil service. Since 1989 he took the helm of the Census and Statistics Department of the Hong Kong Government. In recent years I continue to benefit from his scholarship, as I did in our school days, by reading the interesting educational package on *Living With Statistics* produced by the Census and Statistics Department under his supervision.

I got to know Kin somewhat later, after he joined the same school, one year my junior. (PHOTO 3 – at Columbia University in the summer of 1968.)

However, his reputation in mathematics already spread throughout the school. We became closer friends after he entered Hong Kong University, first as an undergraduate in the Faculty of Science, later switched to the Faculty of Arts as a maths major. As a double major in mathematics and physics I spent one more year to learn more mathematics like the maths majors did, so we became real classmates in the academic year 1966-67. Before that we collaborated closely during a Science Exhibition. (PHOTO 4 - Science Exhibition in December 1965.)

As classmates we collaborated again closely in pairing up to prepare a seminar in an algebra course.

The third time we worked closely was to study together a Chinese textbook on real analysis and functional analysis in the summer just before we went abroad for graduate study. It served me in good stead, for it enabled me to pass my qualifying examination eight months later, thanks to the clear elucidation by Kin. He performed a similar task again in giving me a crash course on probability in the summer of 1975 when I returned to teach at Hong Kong University, only to find that I was assigned to teach a course on queueing theory, even the name of which was then foreign to me. That started many enjoyable years of further collaboration between us, me always receiving more than I could give. However, I can never attain his height in versatility. Kin moves freely from one field to another with high accomplishment.

Once Kin and I spent a summer in Chicago doing joint research with another friend in computer science, Clement Yu, on a topic new to both of us. One evening we whiled away the hours by composing playful 'pseudo-poems' and 'pseudo-couplets' (打油詩, 打油聯). Under a spell of reluctant resignation I composed a couplet :

有志圖當巨擘, 也曾躍躍欲試.
無意竟成雜家, 何需耿耿於懷.

Actually, I initially used the term "無奈" instead of "無意", but Kin offered the right criticism that it would sound too negative and too resigned. He is of course right, because he is an exemplar of a rare combination of "雜家" and "巨擘", while I am merely a "雜家" --- moreover in the Cantonese version ("jar kar") rather than in the Putonghua version ("za(2) jia(1)") !

Now I come to another "巨擘" (master) who needs no introduction --- Tze-Leung. (PHOTO 5 – in the United Nations Garden in New York in August of 1968.)

He entered Hong Kong University as a maths major in the Faculty of Arts in the same year as Kin. We got to know each other again as co-workers in the Science Exhibition. Again we became real classmates in the next year. (PHOTO 6 – class of mathematics major in the academic year 1966-67, with both Tze-Leung and Kin there.)

I still remember vividly how Tze-Leung amazed the class as well as the teacher in the algebra seminar I referred to a minute ago. He gave his own proof on the decomposition of finitely generated modules over a principal ideal domain. At that moment I knew I had the honour of being a classmate of a master, only that I did not know at the time he would become a master in probability and statistics instead of in algebra. But I am sure he would be if he only steps into the field.

Tze-Leung came to Columbia University to study mathematical statistics one year after I was there. We spent three years in the same dormitory and two summers in sharing a sublet apartment (with a third friend Clement Yu, then an undergraduate in computer science at Columbia). Besides his calibre and dedication in mathematics, which I cannot hope to emulate to satisfactory degree, I also witness his concentration, tenacity and diligence, particularly during those months as his roommate. He worked for days and weeks, or I should say for years, on end. We used to ask him everyday when he came back in the early morning to catch a few hours of sleep, "How many theorems have you produced last night?" My one contribution to statistics, and the only one, is to do the cooking in those two summers, to satisfy the nutritious need of Tze-Leung so that he could concentrate on his research! Despite the many hours he spent on his research he was always kind and willing to listen to my outpouring of frustration whenever I made no progress in my own study in algebra. If I had spent lesser hours in his office doing that, he would have produced even more theorems!

Hence, to all three of the trio I owe a debt of gratitude, for their inspiring and supportive companionship in school, in university, in graduate school and in my career. Thank you, Kin, Tze-Leung and Wing-Huen! All the best to the three of you, and Happy 60th Birthday!

Reports of the 2005 Hong Kong Statistical Conference and The 5th IASC Asian Conference on Statistical Computing

Professor FUNG Wing-kam, Tony
Conference Chairman

The Hong Kong Statistical Society held its conference on 17 December 2005. In parallel, the Society co-organized the 5th International Association for Statistical Computing (IASC) Asian Conference on 15-17 December, 2005. Both conferences were held in The University of Hong Kong.

There were about 200 participants for the Conferences, and 70 of them were from local. There were some 20 presentations in the HKSS 2005 Conference and over 100 papers were presented in both conferences. We are honored to have Professor Tze-Leung Lai, Stanford University, to be our Keynote Speaker and its talk was "A new approach to

The 2005 Hong Kong Statistical Conference (HKSS2005)
17 December 2005, HONG KONG

generalized additive and nonlinear mixed effects models: theory, applications and computational issues". A specially invited session has been organized in the HKSS 2005 Conference to celebrate the 60th birthdays of three outstanding members of our Society, Mr. Frederick W.H. Ho, Professors Tze-Leung Lai and Kin Lam. We are grateful to the specially invited speakers, Mr. H.W. Fung and Professors N.H. Chan and W.K. Li, for speaking about the contributions of the three outstanding members.

The 5th IASC Asian Conference on Statistical Computing (IascAsian05)
15-17 December 2005, HONG KONG

The HKSS and IASC Conferences held a joint Conference Dinner at the YWCA, Tsimshatsui on 17 December, 2005. There were about 50 registrants from each conference. Unfortunately the IASC registrants, including myself and some of my colleagues, could not attend the Dinner due to the heavy traffic jam arising from the WTO demonstrations while we returned from the Peak Tour in that evening.

I would like to take this opportunity to thank committee members of the conferences, P.S. Chan (CUHK), Y.K. Chung (HKU), K.T. Fang (Baptist U), H.W. Fung (C&SD), W.C. Ip (Poly U), Anges Law (City U), Stephen Lee (HKU), P.K. Leung (Poly U), Alvin Li

(C&SD), L.K. Li (Poly U), Albert Lo (HKUST), H.P. Lo (City U), Mike So (HKUST), Raymond Tam (IVE), M.L. Tang (Baptist U), Howard Wong (C&SD), K.H. Wu (CUHK), Iris Yeung (City U), Philip Yu (HKU) and K.C. Yuen (HKU), for their kindest support and assistance.

The HKSS 2005 Conference gratefully acknowledges the patronage of the SAS Institute Limited, Wing Lung Bank Limited and the University of Hong Kong.

More details on the HKSS 2005 Conference can be found in the following web-site.

www.hku.hk/statistics/HKSS2005