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 26 No. 1-2
June-September 2003

Editor's Foreword

First of all, I would like to apologize for the belated issue of this bulletin. As mentioned in the President's Forum, articles are lacking for the issue. I repeated requests for your help. Please submit interesting articles to the bulletin and save it. Secondly I would like to thank all the associate editors and particularly, the secretary to serve again.

In this issue, our President reported in his forum that the Noble Prize in Economics goes to two Econometricians this year. More than that, President also contributes an interesting article on the DNA profile. You will find how statistics can be used in forensic science.

The Examination Board provides us up-to-date developments of the HKSS Professional examination since the launching of it in May 2002. It is a very successful project. Many candidates have passed the examinations and their names are listed in the Annex.

We hope that you find this issue enjoyable and informative to read the Bulletin. Once again, the Bulletin won't survive without your support.

P.S. Chan

Editor	:	Chan, Ping-shing, CUHK	Tel.	2609 7920	Fax.	2603 5188
Associate Editors	:	Chan, Wai, CUHK		2609 6241		2603 5019
		Chan, Wai-sum, HKU		2857 8318		2858 9041
		Ms Ng, Teresa, CityU		2788 8565		27888560
Secretary	:	Lam, John Hon-kwan, C&SD		2802 1267		2121 8296

CONTENTS

(Vol. 26/No.1, June-September 2003)

President's Forum

Professor Tony W.K. FUNG

I apologize for the belated issue of the Bulletin.

You know why the issue is late: We just do NOT have ENOUGH ARTICLES!! Since March/April this year, the new Council has been worrying whether we have enough articles for the HKSS Bulletin. Council members have been continuously asking for articles, especially from colleagues of tertiary institutes. The Editorial Board has received (for some time already) only one good article from C&SD, until early/mid October. I wrote something on DNA that may not be of much interest to some of you.

Now I want to call for

Article, Article, Article

(3 will be good enough, not 23! Don't worry.) for the Bulletin, please.

This year the Nobel Prize in economics goes to R.F. Engle and C.W.J. Granger (statisticians!?). I have asked a HKSS member to contribute an article, and I expect it will appear in the coming issue. At least we'll have ≥ 1 article in that issue. Or perhaps we can have a special volume, so friends, students, admirers, of Engle and Granger please think about it (or even better act accordingly).

My forecast (≥ 99% confidence) is that the Nobel Prize winners in economics next year will not be statisticians, and no one will write on Nobel Prize winners next year again. So please submit your article to the future issues of the Bulletin.

I have also urged the new Council members to think about the article matter in the July 3rd (not 1st) Meeting. I would like to take this opportunity to introduce them to you:

Vice-President:
Mr. LEUNG Kwan-chi (C&SD)
General Secretary:
Mr. LI Wing-kin, Ken (IVE)
Treasurer:
Mr. TAM Chi-ho, Raymond (IVE)
Membership Secretary:
Dr. YANG Hailiang (HKU)
Publications Secretary:
Dr. CHAN Ping-shing, Ben (CUHK)
Consultation Services Secretary:
Ms. LAW Ka-yee, Agnes (City U)
Programme Secretary:
Dr. LI Leong-kwan (Poly U)

and I am FUNG Wing-kam, Tony from HKU. The Council members for the 2002-03 term were

Professor LI-Wai-keung (HKU)
Mr. LEUNG Kwan-chi (C&SD)
Miss Clora CHAN Pui-shan (C&SD)
Dr. Boris CHOY Sai-tsang (HKU)
Dr. LI Leong-kwan (Poly U)
Dr. CHAN Ping-shing (CUHK)
Dr. CHAN Wai (CUHK)
Mr. LO Chi-ning (City U)

I want to express my sincere thanks to the in-coming and out-going members for their very hard work.

On the statistical difference between convenience and “random” samples of DNA profiles

Wing-kam Fung
The University of Hong Kong

Abstract

In this paper, the DNA-STR profiles obtained from ‘random’ and ‘convenience’ sampling are compared. Statistical analyses are conducted on the DNA reference databases collected by the Hong Kong Government Laboratory using these two very different sampling procedures. We are interested in testing the null hypothesis that the population probability or frequency distributions of the STR alleles under different sampling methods are the same. The likelihood ratio, Pearson’s chi-squared, Fisher's exact and simulation tests of significance are employed. None of the p-values are found to be significant at the 5% level. Thus, we are confident to have the conclusion that the reference database collected from ‘convenience’ sampling was not statistically different from that obtained from ‘random’ sampling.

Key Words: DNA profiling; STR loci; Allele frequencies; Convenience sampling.

Background and Introduction

Deoxyribonucleic acid (DNA) profiling or DNA fingerprinting has become a very powerful method for forensic human identification since its inception [1]. It is regarded as one of the most important discoveries in forensic science since the introduction of dermal fingerprinting. In forensic DNA analysis, one issue of interest is whether or not the DNA databases (samples) collected are representative [2, 3]. The DNA database is constructed for the purpose of getting the (relative) frequencies of DNA alleles (distinct types or lengths of DNA). In a local court case, the Hong Kong Government Laboratory has been suggested to collect a representative random sample of DNA profiles/alleles. Thus, the Laboratory collected a ‘random’ sample for that purpose by randomly selecting persons participating in a particular activity: blood donations to the Red Cross. However, such a sample is not truly random and may at most be treated as ‘quasi-random’.

Few, if any, of the existing forensic DNA databases in the world are purely random samples collected from the ethnic groups of interest. The DNA databases were often collected form convenience sampling; for example from blood banks, hospitals and clinics. Thus, the issue of real randomness is difficult to evaluate. Instead, these databases are commonly believed to be representative. This common belief is largely due to population genetic theories and findings [2, 4]. The truthfulness of the belief is, however, hard to be tested statistically. In fact, there often exist a few DNA samples of the same ethnic group collected from different sampling methods. If the samples are significantly different statistically, the belief is bound to be incorrect on the basis of falsification argument. If otherwise, one may have more confidence to use the convenience and/or quasi-random samples.

A very non-random sample of DNA profiles was obtained from the Laboratory Staff of the Hong Kong Government. In this paper, this very convenient sample is compared with the 'quasi-random' sample collected from the Red Cross. Various statistical tests are conducted.

Materials and Methods

The Hong Kong Government Laboratory employed the STR Triplex (CTT) loci (i.e. specific sites in the DNA where each person inherits two alleles from his or her parents) CSF1P0, TPOX and THO1 for human identification. The Laboratory randomly selected Chinese blood donors to the Red Cross and asked them to provide blood specimens for constructing the DNA database. A total of 126 specimens were obtained for this quasi-random (QR) sample. The Laboratory was interested in investigating how different the DNA profiles are under a convenience (CV) sampling and a quasi-random (QR) sampling procedure. The Laboratory designed to collect a non-random sample for comparison. The blood specimens of its Chinese staff were selected for the experiment. The size of this CV sample was 225.

The allelic frequencies of the two databases are compared. Statistical significance tests are employed for testing the null hypothesis that the population frequency distributions of DNA-STR profiles under the two very different sampling procedures, CV and QR, are the same. The bootstrap simulation method [5], the Pearson's chi-squared, likelihood ratio and Fisher's exact tests [6, 7] are used in the study. These significance tests are popular and generally have high asymptotic efficiencies and power [7, 8].

Results and Discussions

Allelic frequencies

Table 1 gives the allelic (relative) frequencies at each locus, and Figures 1-3 plot the histograms of allelic frequencies. It can be observed that the frequencies corresponding to the two databases are pretty close to each other. The two sampling procedures seem to give consistent results, especially after taking into account the random variations of frequency estimates.

The frequency distributions may also be compared by looking at the ratios of allelic frequencies obtained by dividing the more common frequency found in one database by the less common frequency of the other database. For alleles with frequencies less than 1%, the minimum 1% frequency is taken. Alleles with zero frequencies in both databases are not included in the comparison. Table 2 gives the percentages of such ratios of allelic frequencies. 15.9% of the ratios are exactly equal to 1, which correspond to the alleles with minimum 1% frequencies in both databases. The remaining 84.1% of the ratios are between 1 and 2.

We also examine via significance tests to determine whether the alleles have different population frequencies. The bootstrap significance test is employed here since the underlying distributions of alleles are unknown. The bootstrap frequencies for each locus are obtained from a sample of size 126 individuals (the database size) randomly selected with replacement from the QR database [5]. The process is then repeated independently 1000 times to generate the bootstrap confidence intervals of the estimated allelic frequencies. A similar procedure can be carried out for the 225 individuals of the CV database.

We can then test, based on the constructed bootstrap confidence intervals, whether the population allelic frequencies corresponding to the two sampling procedures are the same. The null hypothesis of equal population allelic frequencies is rejected at the 5% level of significance if the two corresponding 83% confidence intervals do not overlap. The number of alleles with significantly different frequencies for each locus is reported in Table 3. None of the alleles (out of a total of 21 different alleles) have different frequencies at the 5% level. In other words, the apparent difference in sample frequencies of the two sampling methods as shown in Figures 1-3 may be due to sampling variation.

Chi-squared, likelihood ratio and Fisher's exact tests

The above analyses consider the frequencies at individual alleles (i.e. at allelic level). Next we are going to investigate all alleles at a locus and see whether the two population allelic frequency distributions at that locus (i.e. at locus level) are the same. At each locus, a two-way contingency table of counts can be constructed with the allele being one factor and the sampling method (i.e. QR or CV) being another. These counts can be obtained easily by multiplying the frequencies in Table 1 to the corresponding sizes of the databases. The null hypothesis of interest is that the population allelic frequencies of the locus corresponding to the two different sampling methods are the same. The common Pearson's chi-squared test for independence [6] is well suited for our purpose. The *p*-values of the test can be determined based on the asymptotic χ^2 distribution, and they are reported in the second column of Table 4. It can be observed that these values are all larger than the nominal 5% significance level.

There is a rule-of-thumb for using the

chi-squared test: the expected counts should be larger than five [6]. However, this rule-of-five is not satisfied in our case. The complete enumeration can be taken to compute the exact p-value, but it is intractable here. Instead, we use the simulation method to approximate the exact p-value. A Monte Carlo simulation of 4,000 repetitions is randomly taken for each locus [9], and the approximate p-values (with standard errors of about 0.01) are reported in the third column of Table 4. These values are also much higher than the 5% significance level. Moreover, they are fairly close to those obtained from the asymptotic χ^2 distribution. This is in line with the findings of that the rule-of-five is often too conservative [10].

The null hypothesis of equal population allelic frequencies is also examined using two other popular statistical tests: the likelihood ratio and Fisher's exact tests [6]. The p-values are obtained from the asymptotic distribution method and Monte Carlo simulation with 4,000 repetitions. These values are also reported in Table 4. In general, they are rather close to those obtained from the chi-squared tests. Thus, we may conclude that the population allelic frequency distributions based on the two sampling methods are not different statistically.

Conclusion

DNA-STR profiles obtained from the convenience (CV) and the quasi-random (QR) sampling methods are compared. Analyses are conducted to investigate whether the databases collected by different sampling methods are statistically equivalent. It is found that the databases give very similar profile frequencies. We have also constructed a number of significance tests and the results are promising that none of the p-values are smaller than 5%. This gives us more confidence in using STR reference databases collected by 'random' and/or convenience sampling.

References

Jeffreys, A.J., Wilson, V. and Thein, S.L. Individual-specific 'fingerprints' of human DNA. *Nature*, 1985; **316**:76-79.

Roeder K. DNA fingerprinting: A review of the controversy (with discussion). *Statist. Sci.*, 1994; **9**:222-278.

Geisser S. Some statistical issues in forensic DNA profiling. In *Modelling and Prediction Honoring Seymour Geisser.* Ed. Lee J.C., Johnson W.O. and Zellner A. 1996; pp3-18. New York: Springer.

National Research Council (NRC). *The Evaluation of Forensic DNA Evidence*. Washington D.C: National Academy Press, 1996.

Efron B. and Tibshirani R. *An Introduction to the Bootstrap Methods*. New York:John Wiley, 1993.

Conover, W.J. *Practical Nonparametric Statistics*, *2nd ed.*, New York:John Wiley, 1980.

Gibbons J.D. and Chakraborti S. *Nonparametric Statistical Inference*, *3rd ed.* New York:Marcel Dekker, 1992.

Hajek J. and Sidak Z. *Theory of Rank Tests*. New York:Academic Press, 1967.

StatXact, *Statistical Software for Exact Nonparametric Inference*. CYTEL Software Corporation, Cambridge, Massachusetts.

Fienberg S.E. The use of chi-squared statistics for categorical data problems. *J.R. Stat. Soc. B*. 1991; **41**:54-64.

Table 1. Allelic frequency distributions at loci CSF1P0, TPOX and THO1 for databases collected from quasi-random (QR) and convenience (CV) sampling procedures

	CSF1P0		TPOX		THO1	
Allele	QR	CV	QR	CV	QR	CV
6	0	0	0	0	0.119	0.091
7	0.008	0.007	0	0	0.294	0.329
8	0	0.002	0.560	0.564	0.056	0.053
9	0.032	0.038	0.087	0.118	0.448	0.442
9.3	0	0	0	0	0.024	0.033
10	0.266	0.227	0.028	0.020	0.056	0.051
11	0.234	0.244	0.306	0.278	0.004	0
12	0.365	0.391	0.020	0.020	0	0
13	0.087	0.080	0	0	0	0
14	0.004	0.011	0	0	0	0
15	0.004	0	0	0	0	0
Total size	252	450	252	450	252	450

Table 2. Percentage of ratios of allelic frequencies between databases collected by the two sampling methods

Ratio	CSF1P0	TPOX	THO1	Average
1	33.3%	0%	14.3%	15.9%
> 1-2	66.7%	100%	85.7%	84.1%

Table 3. Number of alleles with statistically different population frequencies, using the 5% level of significance

CSF1P0	TPOX	THO1
0/9	0/5	0/7

Table 4. *P*-values for three significance tests for equal population frequency distributions obtained from the asymptotic method (P_1) and simulation method (P_2)

Locus	Chi-squared		Likelihood ratio		Fisher's exact	
	P_1	P_2	P_1	P_2	P_1	P_2
CSF1P0	0757	0.814	0.679	0.793	0.779	0.817
TPOX	0.686	0.694	0.680	0.691	0.668	0.691
THO1	0.644	0.663	0.609	0.681	0.657	0.678

Fig.1. Histogram of STR profiles at CSF1P0 for samples collected from quasi-random (QR) and convenience (CV) sampling procedures

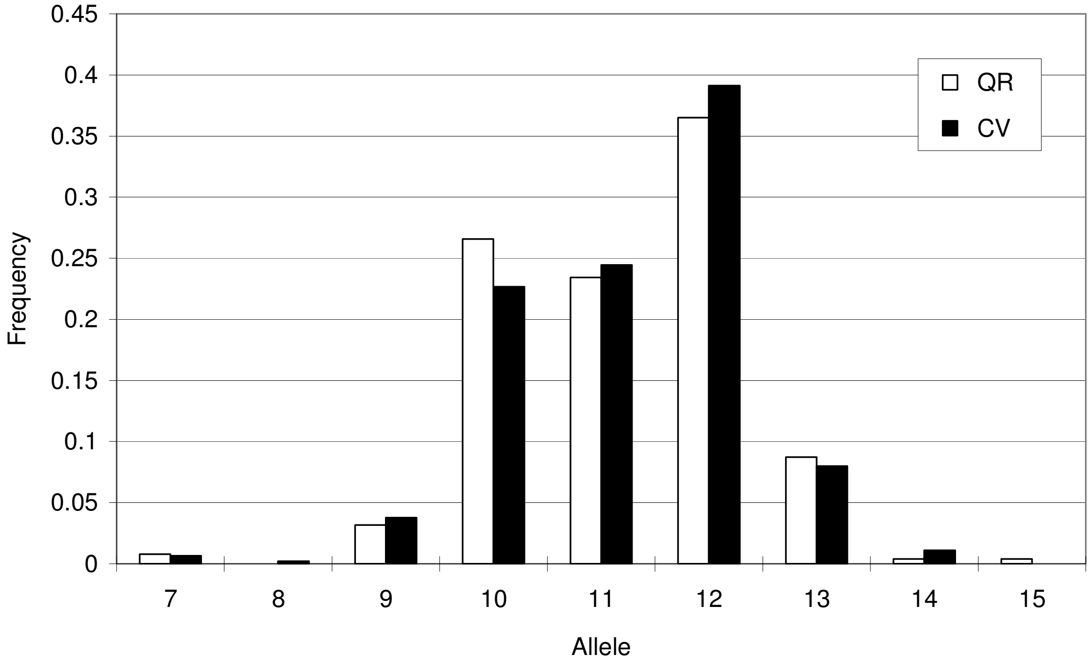


Fig.2. Histogram of STR profiles at TPOX for samples collected from quasi-random (QR) and convenience (CV) sampling procedures

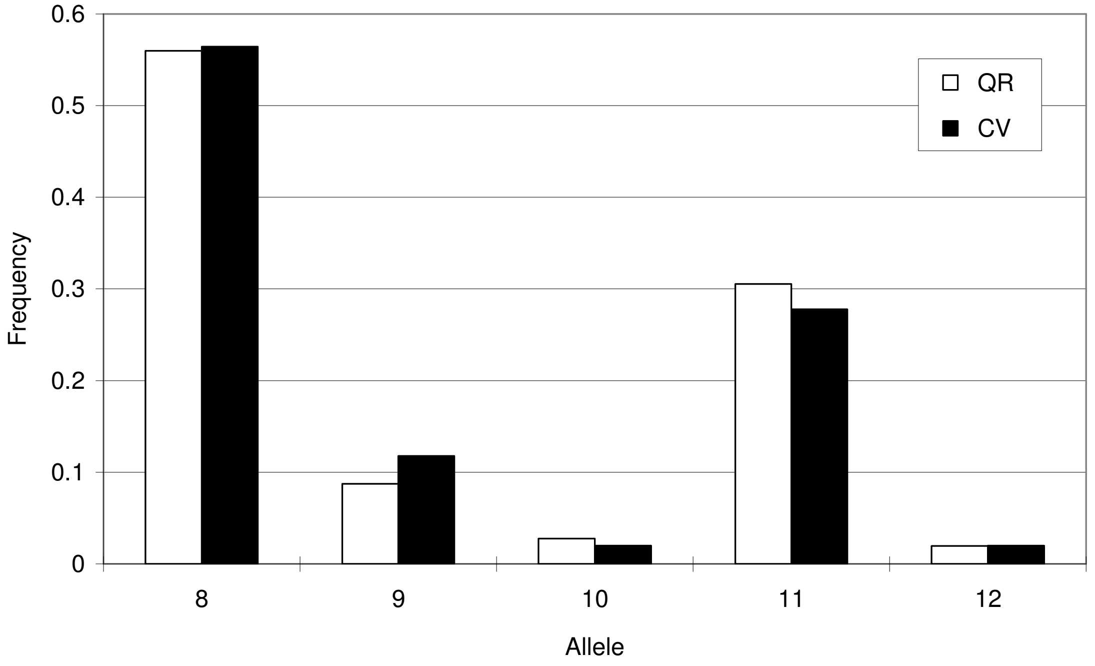
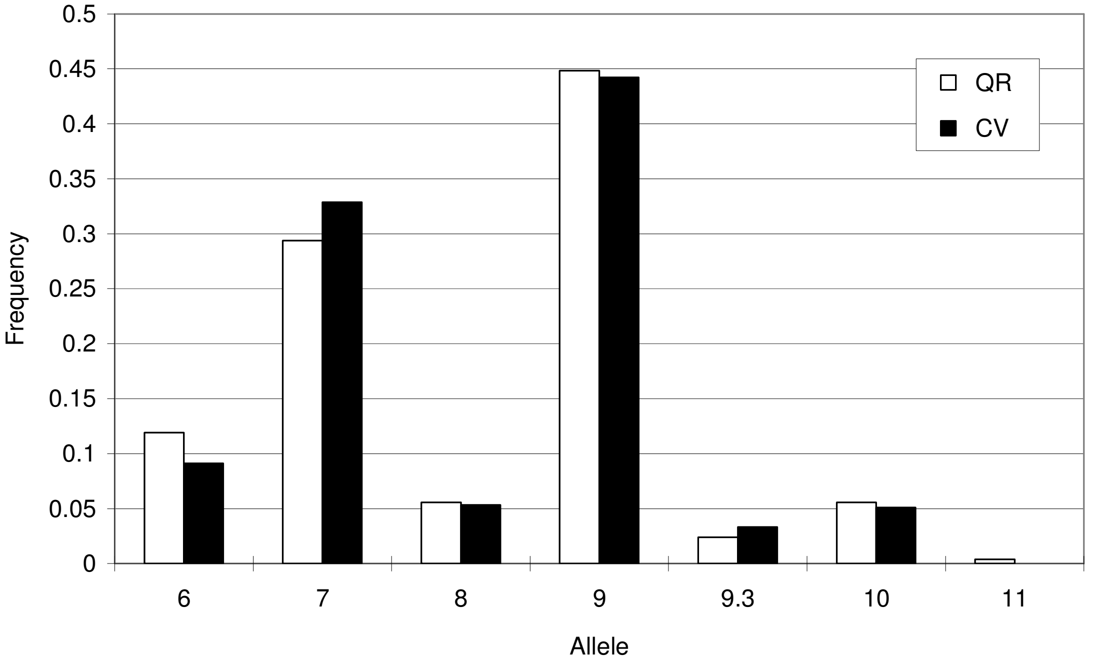


Fig.3. Histogram of STR profiles at THO1 for samples collected from quasi-random (QR) and convenience (CV) sampling procedures



Recent Developments of the HKSS Professional Statistical Examination

HKSS Examination Board

A bilingual press release announcing the new developments of the HKSS examination was issued to all major newspapers on 25th September 2003. This article highlights the major developments and reports on the briefing seminar held on 29th September 2003.

Two rounds of the examination successfully conducted

Since the taking over of the professional statistical examination by the HKSS from the Royal Statistical Society (RSS), two rounds of examinations have been successfully conducted in May 2002 and 2003. For the 2003 round of examination, altogether 64 examination associates/ members of the Society have registered for it, with 13 for the Ordinary Certificate (OC) level, 37 for the Higher Certificate (HC) level and 14 for the Graduate Diploma (GD) level. Examination results for the OC and HC levels of the examination were especially encouraging, with nearly half of the candidates getting through and around one-third scoring credit or distinctions.

Options paper on "Social, Economic and Financial Statistics" continuously be offered

As from the May 2003 examination, taking into account the latest trend of development in international standard for statistics, a new subject titled "Social, Economic and Financial Statistics" has been added as another Options Paper to the syllabus of the GD level of the examination. The Society will continue to offer this Options Paper in the 2004 round of the examinations.

The syllabus for this new subject was specifically designed to address the growing demand for professional knowledge in the analysis of social, economic and financial issues. The RSS has endorsed the proposed syllabus for the new subject and agreed that it is of an intellectual standard equivalent to those of the syllabuses of the other papers offered for the GD by the two societies. This illustrates the equivalence of professional standard advocated by both the RSS and HKSS.

Chinese papers newly introduced in 2004

Another significant development of the examination is that Chinese papers will be offered for the OC and HC levels of examination from the May 2004 round of examination.

Candidates taking the HKSS examination may opt to sit for OC/HC papers in either English or Chinese. For the GD examination, candidates must answer the questions in English. Candidates who plan to take the English papers should use the English application/registration forms, whilst candidates who plan to take the Chinese papers, the corresponding Chinese forms. The language used in the examination will be shown in both the result notices and examination certificates.

The HKSS website has accordingly been enhanced to allow for viewing and downloading the Chinese version of all relevant examination materials.

Accreditation

The exercise on establishing an accreditation system between the HKSS examination and various statistics programmes offered by tertiary and vocational institutes in Hong Kong has commenced. The Examination Board has exchanged essential information with the Department of Applied Mathematics of the HK Polytechnic University and started to examine details of the Department's relevant statistics programmes. Similar arrangement is being made with other universities and relevant vocational institutes.

Briefing seminar

To introduce the above major developments of the examinations to members and to facilitate experience sharing amongst them, a briefing seminar was held on 29 September 2003 at the HKU SPACE, Admiralty. About 70 examination associates/ members of the Society and students from tertiary and vocational institutes attended the seminar.

Prof Tony Fung, the HKSS President and Mr. HW FUNG, Chairman of the HKSS

Examination Board, briefed participants of the recent developments of the examination. Senior lecturers from various tertiary and vocational institutes including Ms May WONG from the HKU School of Professional and Continuing Education, Mr. Patrick KWAN of the HK Polytechnic University, Ms Teresa NG of the City University, Mr. Raymond TAM of the Hong Kong Institute of Vocational Education and Dr. William LEUNG of the Hong Kong Technical College of Technology introduced relevant statistics courses to participants.

The opportunity was also taken to present certificates to candidates who had successfully completed the OC / HC Certificate levels of the HKSS examination.

A name list of candidates passing the OC/ HC examination is given at the Annex. Congratulations to all of them and wish them every success in their further study !

At the end of the seminar, some senior members of the Society and candidates who have obtained very good results in the 2003 round of the examination shared with participants their experience in sitting the examination.

Most of the participants found the seminar particularly the sharing sessions informative. Quite a number of them expressed interest in knowing more about relevant statistics courses. They were advised to obtain such information from the respective institutes.

Registration

The coming round of the examination will take place in Hong Kong on 18 - 20 May 2004. Interested members who wish to sit the examination should note that the deadlines for registration as examination associates and application for academic assessment are 31st January 2004 and 1st February 2004 respectively, whilst that for examination registration, 1 March 2003. Details of the registration procedures, application forms and past papers etc. can be found from the HKSS website at www.hkss.org.hk.

Annex

List of Candidates who have awarded the Ordinary Certificates and Higher Certificates of HKSS Examination in 2003

The following candidates have successfully passed the Ordinary Certificate (OC) and Higher Certificate (HC) levels of the HKSS examination in May 2003. They have been awarded the respective OC / HC certificates, duly signed by both the HKSS and RSS. Congratulations to all of them and wish them every success in their further study !

- *Ordinary Certificates*

CHAN Man-fai (with credit)
LEUNG Miu-ling (with distinction)
NG Suk-yin
SHUM Kwok-shuen
SO Chung-pan (with credit)
SO Kwok-yin (with credit)
TAM Wing-kwan (with distinction)

- *Higher Certificates*

CHAN Tsz-shing (with distinction)
CHEUNG Chun-kit (with distinction)
CHIM Mei-ling
CHOW Wing-cheong (with credit)
FU Wing-suet (with credit)
KWAN Pui-ki (with credit)
KWOK Chun-yu
LAU Cheuk-man (with distinction)
LI Kam-ling
LO Tak-kim
NG Chun-wai (with distinction)
SHUM Po-cheung
WONG Pik-ha
WONG Pui-chi (with credit)
YEUNG Hin

News

Graduate Statistician membership recently endorsed by the Council

The Council has recently endorsed 5 applications for the Graduate Statistician Membership status of the Hong Kong Statistical Society. Congratulation!!! Name of these Graduate Statistician members include:

Mr. KWOK Ping-hung
Mr. HO Wai-ip
Miss HO Po-sze
Mr. WONG King-tai
Mr. WONG Chong-yung

University of Hong Kong

Professor WK LI and Professor Tony WK FUNG of the Department of Statistics and Actuarial Science, University of Hong Kong were elected Fellow of the American Statistical Association

Dr LX ZHU of the Department of Statistics and Actuarial Science, University of Hong Kong was elected Fellow of Institute of Mathematical Statistics