STAT 6108 :
Official Statistics and
Structural Equation Modelling
Second Term, 2022-2023

PART 1- OFFICIAL STATISTICS
PowerPoint - 1

Prof Frederick W H HO

January/February 2023

Purpose of this part [Part 1] of the course

- An introduction to the major series of Hong Kong's official statistical data
- Basic concepts and methods of compilation of the data series
- How the statistical data can be obtained by users
- How the statistical data may be analyzed and applied in business and administration.

A BRIEF INTRODUCTION OF THE LECTURER (Professor HO Wing Huen, Frederick 何永煊)

Currently:

>> Adjunct Professor,
Department of Statistics and Actuarial Science,
The University of Hong Kong

>> Adjunct Professor,
Statistics Department,
The Chinese University of Hong Kong;

>> International Statistical Consultant – occasional consulting and lecturing for international statistical organizations;

>> Holding governance/advisory positions (mainly on a volunteering basis) in public/professional/charitable organizations in education, social welfare and health care (hospitals)

A BRIEF INTRODUCTION OF THE LECTURER (Cont'd)

Formerly:

> Commissioner for Census and Statistics (Hong Kong: 1992-2005) :
Head of the Census and Statistics Department of the Government (and concurrently, Head of the Government Statistical Service)
[Was a professional statistician at different ranks from 1972]

> President, Hong Kong Statistical Society (1986-88)

> Chairman, United Nations Committee on Statistics (Asia and the Pacific) (1994-96 and 1996-98)

> Vice President, International Association for Official Statistics (1997-99)

> Council Member, International Statistical Institute (2001-2005)

> Member, Hang Seng Index Advisory Committee (1994-2006)

> Statistical Advisor (2006-2014) and Member of Statistics Advisory Council, National Statistical Bureau (2009-2014), PRC

- During the period April 2012 to October 2012, Prof. HO acted as Programme Host for the Radio Programme “統計人生”（ “Statistics - Life”), which was live-broadcast every Sunday afternoon from 4 to 5 p.m. on Radio One (Chinese channel) of Radio Television Hong Kong
- There are **26 episodes** in the series. They can be downloaded (or listened on-line) from “Podcasts” on the Website of Radio One of Radio Television Hong Kong (http://www.rthk.hk)

  →**See a note** on how to access and listen to the Programme; also, a list of titles and contents of individual episodes

- The purpose of the Radio Programme is to popularize the understanding of the subject of statistics in the community thereby improving "statistical literacy" of Hong Kong citizens.

- One objective is to get people to understand that "Statistics" are not just numbers, but also methodology (and, in fact, one form of logical thinking).
- Listeners are introduced to some statistical concepts and methods. Given the variety of listeners, **the presentation has to be relatively simple and interesting; yet rigour is not to be compromised.** *(This is also an important point to note when statisticians speak with non-statisticians)*

- Given the background of the Programme Host, considerable emphasis is put on official statistics.
- In fact, it is one main objective of the Programme that through it, listeners may get to understand better many aspects of the socio-economic situation of Hong Kong and the application of official statistics in public polices and public administration; in social and economic research; and in business.
- *Although prepared and broadcasted in 2012, the contents of the radio progrmamme remain basically applicable, whether as regards concepts, methods or data. This is because these change very slowly, if at all. The data may change a bit more but the trends also do not change drastically…and listeners/readers are of course encouraged to update them by making reference to current publications*

Acknowledgements

- In this series of PowerPoints there are a considerable volume of statistical data and quotations of methodological notes extracted from the publications of the Hong Kong Census and Statistics Departments. Individual staff members of the Departments have also rendered advice and assistance in updating relevant information. These are gratefully acknowledged.

Topics to be covered in the FIVE Lectures

- The uses of official statistics
- The scope of official statistics
- How users can obtain official statistics
- How official statistics are compiled : sources of raw data; methods of compilation of statistics; how quality is assured
- Other features of official statistics
- The analysis and application of official statistics– general introduction
- More detailed study of some major official statistical series

The uses of official statistics

- Official statistics –statistical data produced by Government departments and agencies
- A basis for :- describing , understanding and analyzing the social and economic situations at both the macro-level and the sub-macro-levels
- A basis for :- evidence-based analysis, formulation and decision-making of public policies; and evidence-based formulation and execution of business strategies
- A basis for evaluating the effectiveness and monitoring the progress of action programmes

The scope of official statistics

- SOCIAL STATISTICS
  - > population statistics
  - > household statistics – household composition; household income; accomodation
  - > labour force statistics - labour force participation; employment; unemployment; under-employment
  - > education statistics
  - > health statistics
  - > housing statistics
  - > social welfare statistics
  - > transport statistics
  - > law and order statistics

The scope of official statistics (2)

- ECONOMIC STATISTICS
  - > National Accounts statistics (GDP- Gross Domestic product and GNP- Gross National Product)
  - > Price statistics
  - > Trade statistics (goods and services)
  - > Production statistics (goods and services; industrial structure; operating characteristics of establishments)
  - > Employment statistics and labour cost statistics
  - > Statistics on the property market
  - > Financial statistics
  - > Balance of Payment statistics

The scope of official statistics (3)

- Though grouped into "social statistics" and "economic statistics", many branches of statistics actually **straddle between** the two groups.

- The concept of "***socio-economic***" statistics should be understood and accepted.
  In fact, this use of terminology applies to "social policies", "economic policies" and "socio-economic policies"

How users can obtain official statistics

- Census and Statistics Department (C&SD) is the CENTRAL statistical authority of Hong Kong
- C&SD compiles most of the **'general purpose'** official statistics while various government departments compile most of the **'specific purpose'** official statistics
- Official statistics may be obtained from the C&SD or relevant individual departments
- C&SD publishes **general digests of statistics** covering a selection of all the available statistics
- C&SD also publishes reports on **specific statistical data series**

How users can obtain official statistics (2)

- On those statistics of greatest common interest, C&SD issues **press releases** in which the key data are presented

- More detailed statistical data are contained in the **statistical digests** and the **thematic reports** (hardcopies or softcopies)

- Hardcopy publications can be purchased at C&SD's office. Electronic versions of the publications can be obtained on-line (since June 2006 they can be **read and downloaded free of charge**). In fact, most of the publications are now only published on-line.

How users can obtain official statistics (3)

- The address of the C&SD web-site is www.censtatd.gov.hk/
- If ever in doubt, contact the Department. A list of enquiry points (phone no. and email addresses) is available at:

  https://www.censtatd.gov.hk/en/page_105.html

- Some users may wish to have **statistical data at greater detail than those in the press releases and the reports**. They can contact C&SD (phone or write the relevant section of C&SD – there is information on whom and how to contact in the respective reports).

  The data may be available readily, or have to be compiled from the database specifically for the enquirer. Discussion can be held and some charge may be payable for getting such data. An estimate of the charge is provided to the enquirer before the data are prepared for delivery.

How users can obtain official statistics (4)

- Apart from **statistical data**, C&SD reports normally contain relevant **technical details** about the data as well.

- Whilst not their primary duty to do so, **C&SD staff are happy to discuss with statistics users on the ap*plication* of official statistics in specific studies**.

Learning Official Statistics

- A **career** in official statistics?
  Or, a career in jobs which require **considerable understanding** of the socio-economic environment?

  → Statistical positions in the Civil Service.
  → Some positions (such as planners, marketing managers, general management staff) in both the public and private sectors, wherein possession of knowledge in official statistics is a **definite advantage**.

- **Using statistics** (concepts/methods/data) occasionally or quite frequently in your current/future job? **Or, just wanting to understand** better the developments in society and the economy *in order to perform better in the job*?

Learning Official Statistics (2)

- Wanting to analyze social and economic phenomena in **more depth** in a **scientific manner**, with a view to

  becoming commentators/critics

  of public policies?

  or, just playing a better role

  → as responsible and

  socially conscious citizens ;

  → as teachers;

  → as parents ?

Compilation of official statistics: **Sources** of raw data

→ "Individual Data" vs

"Statistical data" (aggregated data)

→Where are the "**raw data**" ? How are they obtained ?

They are :

>> Administrative Records in Administrative Systems

>> Information in Special Returns

>> Information collected from Censuses or Sample Surveys

(I) Administrative records

(1) Records kept in an administrative system which relate to the ***"status"*** in time [the "stock" concept], ***or***

(2) Records arising from ***activities*** within an administrative system [the "flow" concept]

>> **Example of (1)**: Register of persons

One record is kept for one person on his details; and the record is updated for changes from time to time

>> **Example of (2)** : Applications for certain government services (public housing; social welfare benefits)

>>>**Another example of (2)** :

Arrival cards filled in by visitors to the territory

(II) "Special returns"

- Information in "special returns" submitted by persons or organizations to government ministries and agencies in accordance with certain legislations **just for the purpose of compiling statistics**

- ***Example:*** the Trade Statistics System in Hong Kong.

  1. Hong Kong is a free port. Traders **have to submit** "trade declarations" only AFTER the importation or exportation of commodities

  2. The "trade declarations" contain details on the goods: description of commodities imported/exported; source or destination; quantities; prices; and so on. **They are used for statistical purposes only.**

  3. Trade statistics are compiled by the Census and Statistics Department based on these. [[There exist safeguards **to ensure complete submission**]]

(II) "Special returns"

Another example :

- Private medical practitioners **have to fill in and submit to the Health Ministry/Department** a BRIEF form in respect of each of his patients (**or** a summary form for all patients each week, say) diagnosed for a certain **communicable disease**.

- The purpose is ***not*** to TRACK each and every patient;

- they are only for the compilation of statistics to show some general trend.

- (**If the disease is a serious one such that *each and every* patient has to be tracked**, then the completed from is an *administrative record* and ***NOT*** a "special return").

(III) Surveys (Censuses and sample surveys)

- Persons or organizations are contacted and asked to fill in ***questionnaires*** to supply required information (the “raw data”).

- Based on the “raw data”,

  “**statistical data** “ (or just, “ **statistics** “)

  are produced.

(III) Surveys (Censuses and sample surveys)

- "Census" **---** a "full coverage" survey.

  For example –

  (1) Census of the Population (**all** persons in the Population are covered) ;

  (2) Census of Manufacturing Establishments

  (**all** factories are included) [*note*-HK does not conduct CME]

- "Sample survey" **---** only a sample is selected for surveying. The sample has to be selected in a scientific manner to enable the **production of reliable estimates** on the characteristics of the ENTIRETY of the Population (Example (1) above), or, the ENTIRETY of the Factories (Example(2) above).

(III) Surveys (Censuses and sample surveys)

Sample surveys are commonly used because they require much less time, effort and other resources **than** censuses. But censuses allow

(1) finer details to be studied

(2) data on more detailed geographical units to be presented.

Some sample surveys are conducted for a specific purpose (e.g. to find out about the situation on a specific topic or theme--- a "**special purpose survey**") ; **and** some are conducted for general purposes (e.g. to find out about the general situation --- a "**general purpose survey**").

(III) Surveys (Censuses and sample surveys)

- Designed for a special purpose, a "special purpose survey" may sometimes still provide statistics useful for purposes which are ***not originally intended***.

- For a "general purpose survey", it is intended for use for many purposes. Thus, during the design stage, **wide consultation of potential users** is desirable.

(Thus, a Population Census is actually a (full coverage) " general purpose survey", and the data serve studies on the population, labour matters, household characteristics, education attainment of persons, income distribution,….)

Compilation of official statistics: Matters to address in the statistical plan

- Definitions, classifications, counting rules
- Instruments of data collection – the design of the data record *or* the questionnaire
- Sampling to be applied?
- Confidentiality of data pertaining to individual firms, households or persons
- Respondent burden in the case of censuses/surveys
- Actual data collection work – the work plan and its implementation (in particular field workers in surveys)
- Precision of data (where sampling is involved)
- Data processing (editing, tabulation, computation)

Survey sampling-
Probability Sampling

- Commonly used Sampling Methods
  - >Simple Random Sampling
  - >Systematic Sampling
  - >Stratified Sampling
  - >Cluster Sampling
  - >Multistage Sampling

→ Note: though we are talking about "survey sampling", the methods may also be applicable to sampling from "populations" of administrative records, [whereby the work on processing can be reduced]

Probability Sampling *vs* Non-probability Sampling

- Probability Sampling

  > **Elements** are selected into the sample through a certain **Random Process** whereby every **element** in the population has a **known**, **non-zero** probability of being selected

  > **Precision of the estimates** can be assessed

  > **Equal probability of selection method (EPSEM) [whereby each and every element in the population is selected with equal probability] is often desirable, but not a necessity**

- Non-probability Sampling

  > Haphazard Sampling (Convenience Sampling)

  > Judgemental Sampling

  > Quota Sampling

  The probability of an element being selected is unknown; and actually some elements have zero probability of being selected. Bias often result. Also, precision of the estimate cannot be assessed.

Probability sampling (cont'd)

> Simple Random Sampling

1. A list of members of the population.

   [[the list is the "**sampling frame**"]]

2. Assign no. 1, 2, 3,…to the members
3. Lot drawing. (There are different methods of doing so – computer lot drawing; using the table of random numbers;)

> Systematic Sampling

1. A list of members of the population.
2. Assign no. 1, 2, 3,…to the members
3. Select a random start (e.g. lot drawing)
4. Take (say) the 3rd, the 13th, the 23rd ….elements

   [[In this case, the **sampling interval** is 10]]

Probability sampling (cont'd)

> Stratified Sampling

1. Divide the list into a number strata using available information (e.g. male stratum, female stratum)
2. Select a probability sample from **each and every** stratum(normally the same sampling method is used in each of the strata)
3. The "sample" is the totality of all the samples thus selected

Example: Stratifying a population of students by sex (separate the list of all students into two lists) and choose one sub-sample from ***each*** list.

>Cluster Sampling

1. Divide the list of members into clusters.
2. Number the clusters 1, 2, 3…..
3. Select a number of clusters
4. Take all members in **each** of the selected clusters to form the "sample"

Example: Each class in the school is taken as a cluster; and ***some*** classes are selected, to form the sample.

……..

Important to distinguish between "clusters" and "strata"

– actually the distinction is quite obvious, if one reads the above carefully

Probability sampling (cont'd)

> Multi-stage Sampling

TWO-STAGE SAMPLING

1. Apply cluster sampling to select a number of clusters.
2. Within **each** selected cluster, instead of taking all members therein, take a sample among the members in the cluster

Example:

1. Population: All students in THE SCHOOL.
2. Each class is taken as a cluster. Within **each** of the selected classes, select a sample of students.

Probability sampling (cont'd)

THREE-STAGE SAMPLING:

1. Apply cluster sampling once (the clusters are "Primary sampling units")
2. Within **each** of the **selected** PSUs, form clusters ("Secondary sampling units"); select a sample of clusters.
3. Within **each** of the **selected** SSU's, take a sample among the "Tertiary sampling units" therein

Example:

Population: All students in HK,

PSU's- Schools ; SSU's - classes ; TSU's – students

We can combine the use of stratification and cluster sampling (or even multi-stage sampling)

For example, we stratify schools by DISTRICTS,

and then apply 3-stage sampling as described above.

Getting data on the units selected

- Conduct a survey to collect data from the elements which have been selected into the sample
- From the data collected in respect of each of the selected members. A suitable formula can be applied to get an estimate of, *for example*, the mean of the variable in question.
- (For example, the variable of interest is the

  " weight of a student". We can then estimate the "mean weight of students".

  The study can cover more than one variable at the same time—e.g. " height" may be another variable being studied)

Estimating the population mean from the sample mean

- Let us say we have a population under study, and we are interested in a certain variable " the parameter")
- Population mean= μ *(example: the variable* is *"weight of a student". μ is the "mean weight" of the population of students)*

- A sample (with size n) is selected according to a predetermined scheme
- **Theoretically, many possible samples can be selected**
- **In practice,** we only select one sample and conduct the survey, collecting data on the sampling units having been selected
- Let the "sample mean" (i.e. the mean of the units having been selected) be m

Estimating the population mean from the sample mean (Cont'd)

- Theoretically, there is a probability distribution of the "means" of all the possible samples – " the **sampling distribution** of the MEAN"

- There is P% (say 95%) probability that

$$\mu - \varepsilon < m < \mu + \varepsilon \quad (\text{the theory will give us } \varepsilon)$$

Look at : $\mu - \varepsilon < m$ and we have

$$\mu < m + \varepsilon$$

Look at : $m < \mu + \varepsilon$ and we have

$$m - \varepsilon < \mu$$

>>Wtaking them together,

we can infer, with P% confidence, that μ is "close" to m,

but subject to a margin of error ε ,

i.e. $m - \varepsilon < \mu < m + \varepsilon$

Estimating the population parameters

*** *"ε" is the "* ***margin of error*** *" in our statistical process. It is related to*

(1) the "confidence level" that we have chosen ***and***

*(2) the "****sampling error****"* ***

- The smaller is the "margin of error", the higher is the **"precision"** of the estimate

- The factors affecting the value of **sampling error** are:

(1) **the design of the sampling scheme** (– a **better** "sample design" is one which, with the same n, yields a smaller sampling error ; but, note, the design is not necessarily *easy* to implement, and hence the "best" design may often not be used)

(2) **size of the sample** – under a given design, a larger sample size (n) normally gives a smaller sampling error

(Note: it is **the *size* of the sample (n) that matters,**

and **not** the *sampling fraction* (n/N) -- N being the population size.

[It may be noted, though, in some designs, the factor (n/N) **does** come in; but this is usually of minor effect only])

(3) **variability of the population** in regard to the parameter under study : under a given sample design and a given sample size, smaller variability of the population would give a smaller sampling error ['Variability' is represented by the 'standard deviation' of the population]

Sample size determination

- Under a chosen sample design

  ε = F (P, Variability(of the population), n)
- Make **an assumption** about Variability (pop.) — based on a past survey, a pilot survey, or experience from other people's surveys of a similar nature…
- Set ε , the *tolerable* margin of error
- Set P , the desired confidence level (usually 95%)
- The **function** F is related to the sample design; its form is available from statistical theory
- Solve the equation to get n.

……

- →→So, we use the sample size n. (Often we need to adjust ε *and P to get another n because the originally obtained n is rather large and we nay not afford the COST.)* After conducting the survey, we will calculate ("estimate") the sample error based on the survey findings. The value of "ε" that comes out may not be the same as **that originally planned** i.e. it can be bigger or smaller).

More sophisticated ESTIMATORS

- The **sample mean** is the natural, and usually the simplest, ***estimator*** for the **population mean.** ("estimator" refers to the formula for working out the "estimate")

- But it is ***not necessarily*** the best one (in terms of precision of the estimate to be obtained)

- In ***somewhat more advanced*** work, some other formulae (other than the simple arithmetic mean) can be applied --- i.e. some other **estimator** can be used (e.g. with the use of available ***auxiliary information***). Correspondingly, the formulae for calculating the "sampling error" and the "margin of error" will have to be separately worked out, according to relevant theory.

Questionnaire Design

- Clear definitions

- Clear specification of requirements

  (for quantitative questions – even for age, income, there can be ambiguity. For example, for "income", the coverage may not be taken the same by all people so this should be clearly spelt out)

  (for qualitative questions (e.g. on attitude, opinion). They are necessarily more subjective : "like it *very much* " :"agree *strongly"*. **We can only ask respondents to be serious in responding and to try to be *consistent* among his own responses**.)

- Beware of memory errors

- Use terms understood by potential respondents

Questionnaire Design (Cont'd)

- Ask only questions that respondents are knowledgeable of
- Avoid leading questions
- Avoid loaded questions
- Avoid composite questions
- Screening questions are included, where required
- Smooth flow of questions
- Clear instructions

Questionnaire Design (cont'd)
Examples of problematic questions

>>"It is the government's responsibility to xxx. Do you agree that we should see xxx done ?"
[leading question]

>>"As everybody knows, yyy. Do you agree that we should have yyy ?" [leading question]

>>"Did you watch any TV programmes in the evening last week, such as A and B ?" [loaded question]
(similar to leading question, but "leading in a more subtle manner)

>>"Do you plan to leave your job and look for another one within the next one month?" [composite question]

Other issues regarding sample surveys

- Response rates
  >Foreign elements;
  >Non-contacts;
  >Refusals
- Total Survey Error:
  >Sampling Bias
  >Sampling Error
  >Non-sampling Bias
  >Non-sampling Error
- Mandatory vs voluntary survey
- Legislation

Other features of official statistics

- Timeliness *vs* accuracy of statistical data
  (provisional figure vs revised figures)
- Level of detail of statistical data
- Resources required for the production of statistics
- Behavioural standards and Credibility –

>> C&SD's Statement of Vision, Mission and Values

>> United Nations Fundamental Principles of Official Statistics

>> International Monetary Fund's 'Special data Dissemination Standard"

VMV STATEMENTS of the Census and Statistics Department (C&SD)

- Vision- To provide high-quality statistical services, contributing to the social and economic developments of Hong Kong.

- Mission- 1.To provide adequate, relevant, reliable and timely statistics to facilitate research, discussion, planning and decision making within the government and in the community.

  2.To ensure that the compilation and dissemination of statistics are in accordance with scientific principles, professional ethics and international standards.

  3.To promote a user-based culture, ensuring that users can obtain effective and convenient services.

- Values- Professionalism; Objectivity and neutrality;
  Cost-effectiveness; Respect for privacy;
  Progressing with the times; Commitment to excellence

UN Fundamental Principles of Statistics-
(adopted 1994, re-affirmed 2013)

*-- The phrases given below are to **assist** in the understanding of the Principles. The **full** meaning of the Principles **may not** be covered. Please read a document which contains a full statement of the Principles; the original being at the **UN website**: --*
https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf

- P1-Relevance, impartiality and equal access
- P2-Professional standard and professional ethics
- P3-Accountability; transparency;
  Presenting "meta-data" (e.g. data collection method,…)
  scientifically (i.e. systematically, accurately and clearly)
- P4-Prevention of misuse
- P5-Sources of official statistics
- P6-Confidentiality of individual data
- P7-Promulgation of relevant procedures,
  including Legislation
- P8-National Co-ordination
- P9-Use of international standards
- P10-International co-operation

SDDS (Special Data Dissemination Standard) of the IMF (International Monetary Fund)

--Please read a document on this subject for more details--

- The data dimension: coverage, periodicity, and timeliness;
- Access by the public;
- Integrity of the disseminated data; and
- Quality of the disseminated data.

DQAF---Data Quality Assessment Framework

……

- Hong Kong has been a subscriber of the SDDS since 1997.
- There is now a simplified version of SDDS - the GDDS (**General Data Dissemination System**) for subscription by economies which may not be advanced enough in their statistical infrastructure. They will nevertheless abide by the behavioural standards despite still not able to meet the requirements of the Data Dimension.

The analysis and application of official statistics – general introduction

Social and economic situations

THE ECONOMY -- size, growth, inflation, productivity, employment, income, consumption, industrial structure, investment, external orientation (trade and investment)

SOCIETY– demographic structure, birth, death, fertility, gender, ageing, education, employment and unemployment, housing, social welfare, transport, law and order, civic society

STATISTICS support

>>the analysis of situations ; the formulation and decision of policies

>>the design of measures and action plans (and the monitoring and evaluation of their subsequent implementation) both in the public and private sectors

The analysis and application of official statistics– general introduction (2)

- Some key terms
  - > tabulations
  - > percentages; rates; ratio; proportions/shares
  - > index
  - > indicators

- Care in the use of graphical presentations

(End of PowerPoint-1)