

**SAMPLE SURVEY DESIGN AND ANALYSIS:  
A COMPREHENSIVE THREE-DAY COURSE  
WITH APPLICATION TO MONITORING AND EVALUATION**

by

Joseph George Caldwell, PhD  
503 Chastine Drive  
Spartanburg, SC 29301-5977 USA  
(001)(864)439-2772  
[jcaldwell9@yahoo.com](mailto:jcaldwell9@yahoo.com)  
<http://www.foundationwebsite.org>

COURSE NOTES: DAY ONE

BASIC CONCEPTS OF SAMPLE SURVEY

3 April 2007  
(Updated 4 May 2007, 26 March 2009)

***NO RECORDING DEVICES ALLOWED***

© 1980 - 2009 Joseph George Caldwell. All rights reserved.

Posted at Internet website  
<http://www.foundationwebsite.org/SampleSurvey3DayCourseDayOne.pdf> . May be copied or  
reposted for noncommercial use (or for evaluation by those considering attending the course),  
with attribution.

## Introduction

These notes are intended to accompany a lecture, using a board or projector to augment the oral presentation. They have been prepared so that the student may listen to the presentation without having to take notes.

The lecture is accompanied by examples and handouts, which are not included in these notes.

The course also includes in-class student exercises.

The course may be covered in three six-hour days (three hours in morning, three hours in afternoon), or in five half days (three and one-half hours per day). The split-up sessions are intended to accommodate clients whose employees would find it inconvenient or impractical to allocate an entire day, or three days in sequence, to a course.

The course is intended for any class size, but a smaller class size (e.g., 10-30 students) is better for interactive discussion (responses to student questions, clarifications, additional examples).

The topics covered in the three-day course are:

- Day 1: Basic concepts of sample survey
- Day 2: How to design surveys and analyze survey data
- Day 3: Special topics; practical problems in survey design

In Day 1, basic principles of statistics and sampling theory are presented and the major types of sample design are described, and the rationales for selecting each type of design are discussed. Day 2 is concerned with the problem of constructing a design of each major type (i.e., determining sample sizes and sample selection methods). Day 3 is concerned with special topics, such as the use of sample survey design in program monitoring and evaluation.

### The level and scope of the course: managing expectations

This course is an introductory course on the design and analysis of sample surveys. It assumes that the student has taken a prerequisite course in "college math," but it does not require prior knowledge of calculus. For students having knowledge of calculus (or some background in probability and statistics, say from an elementary course in statistics), some additional information is presented. This additional material is marked with the notation **optional**. These optional sections (few in number) are omitted from the course presentation.

Attendees should be somewhat familiar with basic statistical concepts, such as probability, the mean and variance of a distribution, the normal distribution, the binomial distribution, estimation and hypothesis testing, confidence intervals, and regression and correlation. Needed material from these topics is reviewed, but this review is not sufficiently thorough for a person having no previous knowledge of probability or statistics. Ideally, a person attending this course would have previously taken an elementary course in statistics. A person with no previous training in statistics could follow much of the lecture, but it would be expecting a lot to absorb the basic concepts of statistics "on the fly," in addition to the material specific to sample survey.

This course is intended to cover a broad range of topics in sample survey design. To do so, it does not cover each topic in great detail. The concern is with known results and how to apply them, not in proving them.

The course is introductory and elementary, but relatively comprehensive, and certainly intensive. At the end of the course, a person with some mathematical ability should be able to recognize which basic type of sampling is appropriate in a given situation, be able to estimate the sample size required to produce a specified level of precision, and be able to conduct standard analyses of the collected sample data.

There is no way, however, that a three-day course will make an “instant survey statistician” out of anyone. In a survey design situation that is complex or that will involve large amounts of time, effort, or money, the advice of an expert sample survey statistician should be sought.

The course is basically conceptual, with some time spent on working through detailed examples, including numerical calculation of formulas. Someone wishing to construct an actual survey design and analyze the survey data would likely want to consult a reference text to review detailed examples and gain expertise by working through exercises.

This course is an ideal introduction for a project director or government technical (project) officer who wishes to understand the basic concepts of sample survey in order to effectively manage or monitor a project involving sample survey. With the background of this course, the project manager should be able to sense what type of survey design is appropriate in a given situation, and be able to converse meaningfully with a consulting survey statistician on a project involving a sample survey.

The course has been presented a number of times, both on an “advertised” basis at commercial hotels, and on an “in-house” basis at the US Bureau of Labor Statistics. Overall, the evaluation sheets returned by course attendees have been very favorable, but in a few instances it was attended by persons with limited mathematical background and in those cases the material was considered too complicated. While it is possible to present a course on sample survey with virtually no reference to mathematical symbology, such a course would not be of use to a person who actually wanted to design a survey and analyze survey data. This course is not a “no-math” course. While it is elementary and introductory and does not require knowledge of calculus, it does require some familiarity with mathematics at the “college math” level. Persons with little mathematical background could attend the course and understand much of the lecture material, but they would be unable to follow the mathematical formulas and work out numerical examples.

Nobody likes unpleasant surprises. One of the purposes for publishing these notes on the Internet is so that prospective students may quickly peruse them and assess whether the material is too advanced for them, given their present background in mathematics.

The course covers a lot of material in a short time. These notes will enable the student to pay attention to the lecture without having to take notes. It is not expected that everything will “sink in” in a three-day course, and it is recommended that the student who wishes to apply the techniques in practice acquire a reference text for study, or attend a formal course in which many homework exercises will “fix” the concepts.

Each attendee to the course is asked to complete a course evaluation. One of the questions asked is whether the course should spend less time on many topics (as it does), or concentrate

on a small number of designs. The overwhelming response from attendees is that they liked the course as it is – a broad overview of many topics, with less time spent on any particular design or topic.

The course is *comprehensive*, but it is certainly *not exhaustive*. It provides an *introduction* to the major aspects of sample survey design and analysis. There are many specialized topics that it does not cover, and it does not address every possible combination of survey design elements. For example, it includes stratified sampling and ratio estimation, but does not include stratified sampling with ratio estimation – the student is referred to a reference text for the information on that particular combination. Also, the course includes cluster sampling and stratified sampling, but it skips discussion of stratified cluster sampling.

These notes are available for review by anyone considering enrolling in the course. The notes do not contain all of the exercises, examples, and handouts that are included in the presentation.

It is not expected for the student to memorize the various formulas presented, but it is expected that some of the major ones would be familiar and recognized, by the end of the course (e.g., the formulas for a mean, a variance, a weighted average, and a confidence interval). A certain amount of course material (e.g., examples, supplementary material, details) is included as “background,” to place the essential concepts in context. It is not expected that the student remember all of the material presented, and the really important concepts will be identified and stressed.

In a usual academic course, the material covered here would be written out by the professor, over the course of 16 one-hour class sessions. If all of the material covered here were written out, it would not be possible to cover it in a three-day course. Hence, in addition to obviating the need for taking notes, the course notes enable much more material to be covered than would be possible in a usual course. It is recognized that there is a learning advantage to the student’s writing his own notes, but this benefit has been sacrificed in order to cover much material in a short time. The material presented in the notes is available in a variety of reference texts on sample survey. *The essential feature of the course is the lecture and in-class interaction, not the notes.* The notes are made available simply to enable the student to take full advantage of these aspects.

The course lasts 16-18 hours (Day 3 is usually cut a little short, to accommodate travel arrangements). This is about 1/3 of the class time of a “three-unit” college semester (three hours per week for 16 weeks, or 48 hours). The college course, however, would include substantial amounts of homework, which this course does not include.

Sample survey involves a lot of formulas. There are a number of different designs and estimation techniques, and each of them involves its own formulas (or procedures, such as resampling) for calculating estimates and errors of estimation. These course notes include many formulas, for reference, but not a lot of class time is spent in working with the formulas. They are too many and too complicated to learn well in a three-day course. Most of the class time is spent in discussing concepts, examples, and approaches, not with working through complicated estimation formulas. A few detailed numerical examples will be worked out in the early part of the course, so that the student may become familiar with the computational requirements of the estimation formulas. After that, formulas will be shown in order to illustrate concepts and general forms, but no further calculations will be made using them.

Note on course content. If presented on an advertised basis (individual enrollments), the course follows these notes closely. If presented for a single client, the content may be modified somewhat to suit the client's interests. For example, an overseas client may have no interest in information about the process for obtaining omb approval for a questionnaire to be used in a survey funded by the US government, and may wish for more time to be spent on examples.

The pace of the course, the selection of topics, and the time spent on various topics may be adjusted a little by the instructor, in order to address specific concerns or interests of the students.

While these notes parallel the lecture, not every item included in the notes is necessarily included in the lecture, and not every item included in the lecture is included in the notes. The notes are intended to reduce the requirement for the student to take copious notes during the lecture. They are not intended to be a detailed recording of the lecture. For additional detail and examples, the student should consult a sample survey reference textbook.

This course focuses mainly on *estimation* (point and interval estimation), not on *hypothesis testing*. The reason for this focus is that in sampling from finite populations, subpopulations almost always have different parameters, and so the test of the hypothesis of equality of parameters is irrelevant. We do consider hypothesis testing in applications of sample survey to *evaluation*, where the assumption of a conceptually infinite population (which produced the particular finite population) is reasonable. Application of sample survey to monitoring and evaluation is addressed in Day 3 of the course.

(A similar situation (regarding finite and infinite populations) occurs in the field of *statistical quality control*. On the one hand, we may be interested in estimating the percentage of defectives in a *particular lot* of goods, to decide whether to accept the lot. In this case (acceptance sampling), we are interested in estimating the characteristics of a particular finite population (i.e., the lot). On the other hand, a quality control manager will view this lot as a single sample from the process that generated it and many other lots. In this case, the lot is viewed as a single sample from a *conceptually infinite population* of lots, and we are interested in estimating the characteristics of this conceptually infinite population.)

In the past, the course was presented by Dr. Caldwell and his colleague, Dharendra N. Ghosh.

### Course Pricing

The course is not longer given on an advertised basis, but only "in-house" at a client's facility. The current price for the course, conducted over a three-day period at a client's facility, is USD15,000. The price if conducted over five days (one half-day session each day for five days, with some material dropped from the "day 3" syllabus) is USD19,000.

This price is an all-inclusive price, including, subject to the following limitations. Half payment is requested in advance, and half payment upon completion. We will absorb travel and accommodation expense for the presenter(s) (one or two persons) up to USD5,000, but if these limits are exceeded, the client will be billed for travel and per diem (meals, lodging and incidental) expense for the presenters (one or two persons) in accordance with US Government maximum travel per diem allowances (or international-organization allowances) for the travel (from presenter's home base to client's location, time spent at the client's location, and return to the presenter's home base), and requested to pay the amount in excess of USD5,000.

It is agreed that the client will download the course notes from the Internet website <http://www.foundationwebsite.org> , and print sufficient copies for all attendees. *Note: The Internet version of the course notes does not include all handouts. These supplementary items (as computer files) will be e-mailed to the client prior to the course. If the client does not print the course notes or the supplementary items, the course will be presented without course notes.* This is not the intended format, or the format that has been used successfully in the past. As discussed, much material is presented, and it is not possible to write out this material during a three-day course. At the same time, restricting the course to a lecture, without benefit of the notes, would lose much. *The course is intended to be a lecture supplemented with the Course Notes.*

The client is expected to provide a comfortable environment conducive to learning. If the client does not have suitable accommodations at its own facility, it is recommended that facilities be procured at a local commercial hotel, many of which have excellent facilities for seminars. It is requested that the client provide a computer (with a Microsoft XP or Vista operating system), computer-driven projector and projection screen, for displaying the Course Notes. It is also requested that a medium be provided for ad-hoc classroom presentation by the lecturer. For small groups, this may be a wall board (with chalk or markers) or “flip-chart-and-easel” (with marking pen). For larger groups it is recommended that a “view-graph” projector be available (for displaying writing using markers on clear acetate sheets).

It is requested that the client provide snacks and drinks for the breaks. The client is encouraged to provide lunch to presenters and attendees for full-day sessions, but this is at the client’s discretion. (This was the practice when the course was presented on an advertised basis at a commercial hotel, and it works well (it keeps the class together, and avoids late returns to class after lunch).)

## COURSE SCHEDULE

Sample Survey Design and Analysis:  
A Comprehensive Three-Day Course  
with Application to Monitoring and Evaluation

by

Joseph George Caldwell, PhD

Sample Survey Design and Analysis:  
A Comprehensive Three-Day Course  
with Application to Monitoring and Evaluation

by Joseph George Caldwell, PhD

Course Schedule

Day 1: Basic Concepts of Sample Survey

9:00 - 9:20	Introduction; Course Objectives and Outline; Overview of First Day's Course Content
9:20 -10:00	Review of Basic Statistical Concepts
10:00 -10:30	Simple Random Sampling
10:30 -10:40	Break
10:40 -11:00	Concept of Sample Design
11:00 -11:30	Stratified Sampling
11:30 -12:00	Stratified Sampling
12:00 - 1:00	Lunch
1:00 - 1:30	Cluster Sampling
1:30 - 2:00	Systematic Sampling
2:00 - 2:30	Multistage Sampling
2:30 - 2:40	Break
2:40 - 3:10	Multistage Sampling
3:10 - 3:40	Double Sampling
3:40 - 4:00	Survey of References; Outline of Topics for Second and Third Days; Questions and Answers

Day 2: How to Design Surveys and Analyze Survey Data Part One: How to Design Descriptive Surveys

9:00 - 9:15	Overview of Second Day's Course Content; The Elements of Survey Design; Distinctions between Descriptive and Analytical Surveys
9:15 - 9:30	General Procedure for Designing a Descriptive Sample Survey
9:30 - 9:40	When and How to Use Simple Random Sampling
9:40 - 9:50	When and How to Use Systematic Sampling
9:50 -10:30	When and How to Use Stratification
10:30 -10:40	Break
10:40 -10:50	When and how to Use a Clustered Design
10:50 -11:30	When and How to Use a Multistage Design
11:30 -11:40	When and flow to Use Double Sampling
11:40 -12:00	How to Resolve Conflicting/Multiple Survey Design Objectives
12:00 - 1:00	Lunch

Part Two: How to Design Analytical Surveys

1:00 - 1:30	Review of Regression Analysis
1:30 - 1:45	General Procedure for Designing an Analytical Survey
1:45 - 2:00	How to Use Multiple Stratification for an Analytical Design
2:00 - 2:30	How to Use Controlled Selection for an Analytical Design

2:30 - 2:40 Break

Part Three: How to Analyze Survey Data.

2:40 - 3:20 Standard Estimation Procedures for Descriptive Surveys

3:20 - 3:40 Standard Estimation Procedures for Analytical Surveys

3:40 - 4:00 Computer Programs for Analysis of Survey Data; Outline of Topics for Third Day

Day 3: Special Topics/Practical Problems in Survey Design

9:00 - 10:00 Survey Design for Monitoring and Evaluation

10:00 - 10:30 Instrumentation, Data Collection, and Survey Field Procedures

10:30 - 10:40 Break

10:40 - 11:00 Preparation of OMB Clearance Forms

11:00 - 11:15 Longitudinal Surveys

11:15 - 12:00 Sample Frame Problems

12:00 - 1:00 Lunch

1:00 - 1:15 Sampling for Rare Elements

1:15 - 2:00 Treatment of Nonresponse

2:00 - 2:30 Nonsampling Errors

2:30 - 2:40 Break

2:40 - 3:00 Randomized Responses

3:00 - 3:15 Random Digit Dialing

3:15 - 3:45 Major National Surveys

3:45 - 4:00 Questions and Answers

COURSE SYLLABUS

Sample Survey Design and Analysis:  
A Comprehensive Three-Day Course  
with Application to Monitoring and Evaluation

by

Joseph George Caldwell, PhD

Sample Survey Design and Analysis:  
A Comprehensive Three-Day Course  
with Application to Monitoring and Evaluation

by Joseph George Caldwell, PhD

Course Syllabus

Day 1: Basic Concepts of Sample Survey

1. Introduction
  - Course Objectives and Outline
  - Overview of First Day's Course Content
2. Concepts of a statistical distribution (mean, variance, percentiles; examples: normal, binomial)
3. Types of sampling
  - Purposive (judgment)
  - Haphazard
  - Quota
  - Probability Sampling
4. Concepts of statistical inference from samples
  - Sample
  - Estimators of population parameters (measures of central tendency; other parameters (e.g.,  $p$ ))
  - Properties of estimators: variance, bias; precision vs. trueness; accuracy (mse)
  - Central limit theorem
  - Sample moments vs. population moments
  - Distribution of sample statistic vs. population distribution
5. Simple random sampling
  - When to use
  - How to select a sample
    - Target population, sampling population, sampling frame
    - Random numbers -- how to use, generated vs. tabled
    - Systematic Sampling (from randomly ordered files)
    - Sampling with and without replacement
  - Types of Estimators
    - Simple
    - Ratio
    - Regression
    - Bayes (mention)
    - Resampling (Jackknife, Bootstrap) (mention)
  - Variance formulas
  - Variance estimates
    - Formulas
    - Resampling (mention)
  - Sampling for means vs. sampling for proportions

- Confidence intervals
  - Determining sample sizes
6. The concept of sample design
- Precision/cost ratio; design effect
  - Ways of departing from simple random sampling
    - Variations in the probability of selection
    - Dropping the independence assumption (systematic, cluster, replacement, controlled selection, matching)
  - Optimal design
  - Auxiliary variables
    - Correlated with variables of interest
    - Cost information
7. Stratified sampling
- Description
  - When to use
  - How to select sample
  - Estimation formulas
  - Self-weighting case
  - Variance formulas
  - Variance estimates
  - Construction of strata
  - Multiple stratification
  - Stratification to the limit
  - Cross-stratification
  - Certainty stratum
  - Optimal allocation
  - Determination of sample size
  - Stratification when the variable of stratification is inaccurate
  - Post-stratification
8. Cluster sampling
- Description
  - When to use
  - Intracluster correlation coefficient
9. Systematic random sampling
- Description
  - When to use
  - How to select sample (integer sampling interval, noninteger sampling interval; random start; random starts)
  - Estimation formulas
  - Variance formulas
  - Variance estimation (paired selections, successive differences)
  - Replicated subsamples
10. Multistage sampling
- Description

- When to use
- Intracluster correlation coefficient
- Estimation formulas
- Self-weighting sample
- Methods of sample selection
  - 1st stage: PPS, PPMS, equal probs., w/rep, wo/rep
  - 2nd stage: fixed sample size, variable sample size
  - self-weighting
    - 1st -- PPS, 2nd -- equal probs. (advantages/disad.)
    - 1st -- equal, 2nd -- proportional (adv./disadv.)
- Impact of ICC on selection method
  - If rho fixed (e.g., equal-sized units)
  - If rho variable (e. g., variable-sized units)
- PPS selection
- Certainty stratum
- Variance formulas
- Variance estimation
- Systematic selection ok for 2nd stage units under certain circumstances
- RHC method for sampling wo replacement
- Determination of sample size (design)
  - First stage
  - Second stage
- Need frame only for 1st stage units and selected 2nd stage units
- Generalized variances (mention)
- PPMS

11. Two-phase (double) sampling

- Description
- When to use
- Estimation formulas
- How to select sample
- Variance formulas
- Estimation of variance
- Determination of sample size (1st and 2nd phases)

12. Survey of References; Outline of Topics for 2<sup>nd</sup> and 3<sup>rd</sup> Days; Questions and Answers

Day 2: How to Design Surveys and Analyze Survey Data

Part One: How to Design Descriptive Surveys

1. Introduction
  - Overview of Second Day's Course Content
  - The Elements of Survey Design
  - Distinctions between Descriptive and Analytical Surveys
2. General Procedures for Designing a Descriptive Survey
  - Specify population of interest

- Define estimates of interest
  - Specify precision objectives of survey; resource constraints
  - Specify other variables of interest
  - Develop instrumentation
  - Develop sample design
  - Determine sample size and allocation
  - Specify sample selection procedures
  - Specify field procedures
  - Specify data processing procedures
  - Develop data analysis plan
  - Outline report
3. When and How to Use Simple Random Sampling
    - Nature of situation which warrants use of simple random sample
    - How to select a simple random sample
    - Sampling without replacement
    - How to select a simple random sample without replacement
  4. When and How to Use Systematic. Sampling
    - Reasons for using systematic sampling
    - Nature of situation which warrants use of systematic sampling
    - How to select a systematic sample
  5. When and How to Use Stratification
    - Nature of situation which warrants use of stratified sampling
    - The use of a certainty stratum
    - How to determine the number of strata, and the stratum boundaries
    - Stratification to the limit
    - Collapsed strata
    - Post-stratification
    - Errors in classification
    - Multiple stratification: cross stratification
    - Multiple stratification: nested stratification
    - How to allocate sample sizes to strata, when costs and variances are known
    - How to allocate sample sizes to strata, when costs and variances are unknown
    - Self-weighting design
    - General recommendations regarding stratification
  6. When and How to Use Cluster Sampling
    - Nature of situations which warrants use of cluster sampling
    - The "cluster" effect
    - Determining sample size in cluster sampling (equal-size clusters)
    - Variable-size clusters: sampling with probabilities proportional to size (PPS)
    - Variable-size clusters: sampling with probabilities proportional to a measure of size (PPMS)
    - Stratification of clusters; the use of a certainty stratum of clusters
    - Construction of clusters
    - Variable-size clusters; determination of sample size

- Replacement vs. non-replacement sampling of clusters
  - Situations in which clustering improves precision
  - Self-weighting design
  - Sample frame considerations
  - General recommendations regarding cluster sampling
7. When and how to Use Multistage Sampling (Two-Stage)
    - Nature of situation which warrants use of a multistage design
    - Determining sample sizes in two-stage sample (equal-sized primary units)
    - The use of nonreplacement sampling (equal-size primary units)
    - The use of systematic sampling for selection of second-stage units
    - Determining sample sizes in two-stage sampling (unequal size primary units, selection with equal probabilities)
    - PPS sampling of primary units (unequal-size primary units)
    - Determining sample size in PPS sampling
    - The use of nonreplacement sampling (unequal-size primary units)
    - Stratification of primary units; the use of a certainty stratum
    - Self-weighting design
    - Sample frame considerations
    - General recommendations regarding two-stage designs
  8. When and How to Use Double Sampling
    - Nature of situation which warrants the use of double sampling
    - Determination of sample size in double sampling
  9. How to Resolve Conflicting / Multiple Survey Design Objectives

### Part Two: How to Design Analytical Surveys

1. Review of Regression Analysis
2. General Procedures for Designing an Analytical Survey
  - Sample survey design for analysis
  - Essential problems in design of an analytical survey
  - Two conceptual approaches to design of analytical surveys
  - Methods for the design of analytical surveys
3. Illustration of Methods for the Design of Analytical Surveys

### Part Three: How to Analyze Survey Data

1. Standard Estimation Procedures for Descriptive Surveys
  - Preliminary analysis
  - Planned analysis
  - Special analysis
2. Standard Estimation Procedures for Analytical Surveys
  - Preliminary analysis
  - Planned analysis

- Tests of model adequacy/model revision
3. Computer Programs for Analysis of Survey Data; Outline of Topics for Third Day

Day 3: Special Topics/Practical Problems in Survey Design

1. Survey Design for Monitoring and Evaluation
2. Instrumentation, Data Collection, and Survey Field Procedures
  - Selection of Data Collection Procedures
  - Questionnaire Development
  - Development of Field Procedures (Treatment of Nonresponse, Inplace Interviews vs. Travelling Team, Incentive Payments)
  - Pretesting and Pilot Testing
  - Editing, Coding, Data Base Design and Development
3. Preparation of OMB Clearance Forms
4. Longitudinal Surveys
5. Sample Frame Problems
6. Sampling for Rare Elements
7. Treatment of Nonresponse
8. Nonsampling Errors
9. Randomized Responses
10. Random Digit Dialing
11. Major National Surveys
12. Questions and Answers

Sample Survey Design and Analysis:  
A Comprehensive Three-Day Course  
with Application to Monitoring and Evaluation

by Joseph George Caldwell, PhD

Course Critique Form

Dear Participant:

We appreciate your attendance and are interested in your comments in order to improve our course. Please answer the following questions, adding additional comments as necessary, and send the form back in the attached envelope. Thank you.

Date of course \_\_\_\_\_ Location of course \_\_\_\_\_

Course Content

1. How useful do you consider the information? \_\_\_\_\_
2. Was the material presented in sufficient detail? \_\_\_\_\_
3. Were there some topics you would have preferred more discussion on? Yes\_\_ No\_\_  
If so, which ones? \_\_\_\_\_

Course Delivery

1. Were the presentations effective? \_\_\_\_\_
2. Were the visual aids helpful? \_\_\_\_\_
3. Were the course notes sufficiently detailed? \_\_\_\_\_

Facilities

1. Was the seating arrangement satisfactory? \_\_\_\_\_
2. Were the meals satisfactory? \_\_\_\_\_
3. Was parking adequate? \_\_\_\_\_
4. Is the location convenient? \_\_\_\_\_

General

1. How did you find out about this course? \_\_\_\_\_  
Brochure in mail \_\_\_\_\_  
Organizational channels \_\_\_\_\_  
Associate \_\_\_\_\_  
Internet \_\_\_\_\_  
Other (specify) \_\_\_\_\_
2. Did you have sufficient registration time? \_\_\_\_\_
3. Did you feel the course was as you expected it to be, from the flyer?  
\_\_\_\_\_

4. Did you feel the course was as you expected it to be, from the Course Notes (if examined on the Internet)? \_\_\_\_\_
5. If from out of town: Did you stay at the hotel where the course was presented? \_\_\_\_\_
6. This course was presented to provide a *broad overview* of Sample Survey Design Techniques. Would you have preferred to concentrate on a *few specific designs*? \_\_\_\_\_
7. Have you ever attended a course on sampling before?  
Yes \_\_\_\_\_ No \_\_\_\_\_
8. Would you prefer a more detailed course of 5 days, \_\_\_\_\_  
or a less detailed course of 2 days? \_\_\_\_\_
9. Would you prefer a more advanced course, \_\_\_\_\_  
or a less advanced course? \_\_\_\_\_
10. Compared to other short courses of which you are familiar, was the cost of this course:  
About right \_\_\_\_\_  
Rather high \_\_\_\_\_  
Lower than expected \_\_\_\_\_
11. What additional seminars might you be interested in?  
Time Series Analysis, Forecasting and Control \_\_\_\_\_  
Biostatistics \_\_\_\_\_  
Experimental Design \_\_\_\_\_  
Quality Control \_\_\_\_\_  
Evaluation Research \_\_\_\_\_  
Introduction to Statistics and Data Analysis \_\_\_\_\_  
Simulation and Modeling \_\_\_\_\_  
Optimization \_\_\_\_\_  
Other (specify) \_\_\_\_\_

Additional Comments: \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_

Name (optional) \_\_\_\_\_

Organization (optional) \_\_\_\_\_

## DAY 1: BASIC CONCEPTS IN SAMPLE SURVEY

INTRODUCTION; COURSE OBJECTIVES AND OUTLINE; OVERVIEW OF FIRST DAY'S COURSE CONTENT

- BASIC STATISTICAL CONCEPTS
- SIMPLE RANDOM SAMPLING
- CONCEPT OF SAMPLE DESIGN
- STRATIFIED SAMPLING
- CLUSTER SAMPLING
- SYSTEMATIC SAMPLING
- MULTISTAGE SAMPLING
- DOUBLE SAMPLING
- SURVEY OF REFERENCES; OUTLINE OF TOPICS FOR SECOND AND THIRD DAYS; QUESTIONS AND ANSWERS

## BASIC CONCEPTS IN SAMPLE SURVEY

POPULATION: THE POPULATION IS A WELL-DEFINED COLLECTION OF ELEMENTS (MEMBERS, ITEMS, OBJECTS),  $a_1, a_2, \dots, a_N$  (POPULATION SIZE =  $N$ ).

IN MOST OF THIS COURSE, THE POPULATIONS WILL BE FINITE. AT ONE POINT (DEALING WITH EVALUATION RESEARCH) WE WILL CONSIDER CONCEPTUALLY INFINITE POPULATIONS.

EXAMPLES:

1. ALL RESIDENTS OF A SPECIFIED COUNTRY (THE USUAL "POPULATION")
2. ALL SCHOOLS IN THE COUNTRY
3. ALL TEACHERS IN THE COUNTRY
4. ALL PUPILS IN THE COUNTRY
5. ALL HOSPITALS IN A REGION OF THE COUNTRY
6. ALL PERSONS INFECTED WITH HIV IN THE COUNTRY
7. ALL EXPORTERS OF NONTRADITIONAL COMMODITIES
8. ALL ELEPHANTS

THE POPULATION IS DEFINED BY FOUR QUANTITIES: CONTENT, UNITS, EXTENT AND TIME (E.G., THE INCOME, OF US CITIZENS, RESIDING OVERSEAS, IN THE PAST YEAR).

WE ARE INTERESTED IN DESCRIBING CERTAIN CHARACTERISTICS (ATTRIBUTES, FEATURES, PROPERTIES) OF THE POPULATION. LET  $X_i$  DENOTE AN ARBITRARY NUMERICAL ATTRIBUTE THAT CAN BE DETERMINED FOR A POPULATION ELEMENT (SUCH AS GENDER, AGE, INCOME, HIV STATUS, SCHOOL SIZE, HOSPITAL OWNERSHIP).

FOR EXAMPLE, IN EXAMPLE (1), WE MAY WISH TO DESCRIBE THE PREVIOUS YEAR'S EARNINGS AND CURRENT EMPLOYMENT STATUS OF ALL RESIDENTS (ON JULY 1), BY AGE CATEGORY, GENDER, AND MARITAL STATUS.

THE PROBLEM OF SAMPLE SURVEY ("SAMPLING") IS TO ESTIMATE THE VALUE OF POPULATION CHARACTERISTICS (E.G., A MEAN, PROPORTION OR TOTAL) FROM A SUBSET (PART, PORTION, "SAMPLE") OF THE POPULATION.

WHY A SUBSET?

- PRACTICAL ADVANTAGES
- IMPOSSIBILITY OF A CENSUS IN SOME CASES
- THE EXACT VALUE IS SELDOM NECESSARY
- CERTAIN AMOUNT OF ERROR IS TOLERATED
- EVEN A COMPLETE ENUMERATION (CENSUS) WILL NOT PRODUCE THE EXACT VALUE
- CAN EXAMINE A SUBSET MORE CAREFULLY THAN THE ENTIRE POPULATION

THE POPULATION TO BE SAMPLED (THE SAMPLED POPULATION) MAY DIFFER FROM THE POPULATION OF INTEREST (THE TARGET POPULATION), FOR PRACTICAL REASONS.

BEFORE SELECTING A SUBSET, THE SAMPLED POPULATION IS DIVIDED INTO SAMPLING UNITS (NONOVERLAPPING, EXHAUSTIVE). A LIST OF ALL OF THE SAMPLING UNITS IS CALLED A FRAME (OR SAMPLE FRAME OR SAMPLING FRAME). A SAMPLE (TECHNICAL DEFINITION) IS A COLLECTION OF SAMPLING UNITS DRAWN FROM A FRAME.

EXAMPLE: WANT A SAMPLE OF PUBLIC-SCHOOL STUDENTS. ALL STUDENTS ARE IN SCHOOLS, SO WE MAY DEFINE THE SAMPLING UNIT AS A SCHOOL, AND SELECT A SAMPLE OF SCHOOLS TO OBTAIN A SAMPLE OF STUDENTS. WE ARE MUCH MORE LIKELY TO BE ABLE TO OBTAIN A LIST OF SCHOOLS (SCHOOL FRAME) THAN A LIST OF STUDENTS (STUDENT FRAME).

AFTER SELECTION OF THE SAMPLE, MEASUREMENTS ARE MADE ON THE SAMPLE ELEMENTS (AND ALSO PERHAPS ON THE SAMPLING UNITS) (E.G., A STUDENT'S AGE; A TEACHER'S LEVEL OF EDUCATION; A SCHOOL'S TYPE OF OWNERSHIP; A HOSPITAL'S ANNUAL INCOME).

TWO MAJOR TYPES OF MEASUREMENT SCALES ("VARIABLES"): DISCRETE AND CONTINUOUS.

DISCRETE (NOMINAL/CATEGORICAL, ORDINAL/RANKING): CAN BE COUNTED (E.G., INTEGERS). EXAMPLES: GENDER (M OR F); EMPLOYMENT STATUS (EMPLOYED OR UNEMPLOYED); FAMILY SIZE; EDUCATIONAL LEVEL.

SPECIAL CASE: FOR A BINARY VARIABLE THE  $X_i$ 's ARE 0 OR 1 (E.G., MALE=0, FEMALE=1; ABSENCE OF SOME CONDITION = 0, PRESENCE OF THE CONDITION = 1).

CONTINUOUS (INTERVAL, RATIO): DISTANCES / DIFFERENCES CAN BE MEASURED ON AN INTERVAL SCALE (REAL NUMBERS); EXAMPLES: AGE, HEIGHT, TEMPERATURE, BLOOD COUNT, INCOME

STATISTICAL THEORY GUIDES US IN SUMMARIZING AND ANALYZING THE SAMPLE, TO MAKE INFERENCES ABOUT THE POPULATION. IT ALSO GUIDES US IN THE DESIGN OF THE SURVEY, THE SAMPLE SELECTION PROCEDURES, AND THE SURVEY INSTRUMENTS (QUESTIONNAIRES, DATA COLLECTION FORMS).

## THE ELEMENTS OF SURVEY DESIGN

1. SPECIFY POPULATION OF INTEREST
2. SPECIFY UNITS OF ANALYSIS AND ESTIMATES OF INTEREST
3. SPECIFY PRECISION OBJECTIVES OF THE SURVEY; RESOURCE CONSTRAINTS; POLITICAL CONSTRAINTS
4. SPECIFY OTHER VARIABLES OF INTEREST (EXPLANATORY VARIABLES, STRATIFICATION VARIABLES)
5. REVIEW POPULATION CHARACTERISTICS (DISTRIBUTIONAL, COST)
6. DEVELOP INSTRUMENTATION (DEVELOPMENT, PRETEST, PILOT TEST, RELIABILITY AND VALIDITY ANALYSIS)
7. DEVELOP SAMPLE DESIGN
8. DETERMINE SAMPLE SIZE AND ALLOCATION
9. SPECIFY SAMPLE SELECTION PROCEDURE
10. SPECIFY FIELD PROCEDURES
11. DETERMINE DATA PROCESSING PROCEDURES
12. DEVELOP DATA ANALYSIS PLAN
13. OUTLINE FINAL REPORT

(FROM "VISTA'S APPROACH TO SAMPLE SURVEY DESIGN," AT <http://www.foundationwebiste.org/ApproachToSampleSurveyDesign.htm> .)

## DESCRIPTION (CHARACTERISTICS) OF A FINITE POPULATION OF SIZE N

LET  $X$  DENOTE A (NUMERICAL-VALUED) CHARACTERISTIC, SUCH AS AGE OR INCOME ( $X$  IS A “CONCEPT”). LET  $x$  DENOTE A PARTICULAR VALUE OF  $X$  (SUCH AS AN AGE OF 43).

1. MEAN (ARITHMETIC AVERAGE) =  $\mu_x = \bar{X} = \frac{x_1 + x_2 + \dots + x_N}{N} = \sum_{i=1}^N x_i$

(FOR BINARY DATA,  $\mu_x = P_x$ , WHERE  $P_x$  DENOTES THE PROPORTION OF 1's)

2. MEDIAN: THE MIDDLE VALUE WHEN THE  $x_i$ 's ARE ARRANGED IN ORDER.

3. PERCENTILES, E.G., THE 95-TH PERCENTILE,  $p_{95}$  FOR INCOME IS THE INCOME SUCH THAT 95 PERCENT OF THE POPULATION HAS INCOME LESS THAN OR EQUAL TO  $p_{95}$

4. VARIANCE =  $\sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu_x^2$  (THE “COMPUTATIONAL” FORM)

ALSO  $S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)^2$

5. STANDARD DEVIATION: SQUARE ROOT OF VARIANCE =  $\sigma_x$

6. COEFFICIENT OF VARIATION:  $\sigma_x/\mu_x$

7. TOTAL =  $\tau_x = X = N \mu_x$

THE MEAN AND MEDIAN ARE MEASURES OF LOCATION, OR CENTRAL TENDENCY; THE VARIANCE AND STANDARD DEVIATION ARE MEASURES OF SPREAD, VARIATION, OR DISPERSION.

THE PRECEDING QUANTITIES ARE *SINGLE-VALUED* ATTRIBUTES (CHARACTERISTICS, “PARAMETERS”) THAT SUMMARIZE THE LOCATION AND SPREAD OF THE ATTRIBUTE. IN ADDITION, WE CAN SUMMARIZE THE POPULATION USING MORE COMPLEX REPRESENTATIONS, SUCH AS *FREQUENCY DISTRIBUTIONS*, *CROSSTABLATIONS*, AND *TABLES OF MEANS*.

POPULATION PARAMETERS ARE USUALLY DENOTED BY LOWER-CASE GREEK LETTERS (E.G.,  $\mu, \sigma, \tau$  OR  $\mu_x, \sigma_x, \tau_x$ ) OR BY UPPER-CASE LATIN LETTERS (E.G.,  $\bar{X}, S, X$  OR  $\bar{X}, S_x, X$ ). USE OF AN UPPER-CASE LATIN LETTER FOR THE POPULATION TOTAL ( $X$ ) MAY BE CONFUSING, HOWEVER, SINCE THAT IS THE SAME SYMBOL USED TO DENOTE THE UNDERLYING RANDOM VARIABLE (ALSO  $X$ ). WE WILL USUALLY USE GREEK LETTERS TO DENOTE PARAMETERS, BUT NOT ALWAYS, IN ORDER TO FAMILIARIZE THE STUDENT WITH ALTERNATIVE NOTATION THAT IS IN COMMON USE.

## NOTE ON FONTS

NOTE ON FONTS: TO ENHANCE READABILITY (ON THE COMPUTER SCREEN AND ON WALL PROJECTIONS), THESE NOTES ARE PRESENTED IN BLOCK LETTERS, USING THE MICROSOFT ARIEL FONT. MATHEMATICAL SYMBOLS ARE *ITALICIZED*, TO MAKE THEM EASIER TO DISTINGUISH FROM NORMAL TEXT.

MATHEMATICAL EXPRESSIONS ARE CONSTRUCTED USING MICROSOFT EQUATION EDITOR 3.0, WHICH USES THE MICROSOFT *TIMES NEW ROMAN* FONT, ITALICIZED. THERE ARE HENCE SOME SLIGHT DIFFERENCES BETWEEN SYMBOL FONTS IN THE TEXT AND IN THE FORMULAS (E.G.,  $E(x)$  IN THE TEXT VS.  $E(X)$  IN A FORMULA;  $f(x)$  AND  $g(x)$  IN TEXT VS.  $f(x)$  AND  $g(x)$  IN A FORMULA).

(THE USE OF TIMES FONT FOR THE TEXT WOULD DECREASE READABILITY, AND THE USE OF THE EQUATION EDITOR TO REPRESENT ALL SYMBOLS IN THE TEXT WOULD INTRODUCE VARIATIONS IN LINE SPACING, GREATLY EXPAND THE COMPUTER FILE SIZE OF THIS DOCUMENT, SIGNIFICANTLY INCREASE THE TIME REQUIRED TO TYPE THESE NOTES, AND SIGNIFICANTLY INCREASE THE FILE SIZE AND INTERNET DOWNLOAD TIME (SINCE FORMULAS ARE STORED AS SEPARATE FILES IN .htm DOCUMENTS).)

DESCRIPTION OF A FINITE POPULATION (CONT.)

FREQUENCY DISTRIBUTION, TABULAR FORM:

<u>INTERVAL</u>	<u>FREQUENCY</u>	<u>RELATIVE FREQUENCY</u>
$a_0 - a_1$	$f_1$	$f_1/N$
$a_1 - a_2$	$f_2$	$f_2/N$
$a_2 - a_3$	$f_3$	$f_3/N$
...		
$a_{k-1} - a_k$	$f_k$	$f_k/N$

$N$  (POPULATION SIZE) =  $f_1 + f_2 + \dots + f_k$

VALUES FALLING ON AN INTERVAL BOUNDARY ARE ASSIGNED TO THE LOWER INTERVAL (I.E., THE VALUE  $a_i$  IS ASSIGNED TO THE CATEGORY  $a_0 - a_1$ , NOT TO  $a_1 - a_2$ ).

EXAMPLE: AGE DISTRIBUTION OF THE POPULATION

<u>INTERVAL</u>	<u>FREQUENCY</u>	<u>RELATIVE FREQUENCY (PROPORTION)</u>
0-18	247	.27
19-64	549	.61
65+	113	.12
TOTAL	909	1.00

SPECIAL CASE: DISCRETE VARIABLE HAVING A SMALL NUMBER OF CATEGORIES (SUCH AS GENDER, EMPLOYMENT STATUS, OR HOUSEHOLD SIZE). IN THIS CASE THE INTERVALS MAY INCLUDE A SINGLE NUMBER:

EXAMPLE: GENDER DISTRIBUTION OF THE POPULATION

<u>GENDER</u>	<u>FREQUENCY</u>	<u>RELATIVE FREQUENCY</u>
MALE	110	.48
FEMALE	117	.52
TOTAL	227	1.00

EXAMPLE: DISTRIBUTION OF HOUSEHOLD SIZE

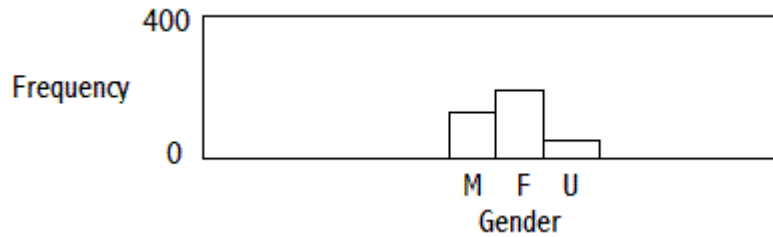
<u>HOUSEHOLD SIZE</u>	<u>FREQUENCY</u>
1	$f_1$
2	$f_2$
3	$f_3$
4	$f_4$
5	$f_5$
6	$f_6$
7	$f_7$
8	$f_8$
9	$f_9$
10	$f_{10}$
11, 12, 13,....	$f_{11}, f_{12}, f_{13}, \dots$

DESCRIPTION OF A FINITE POPULATION (CONT.)

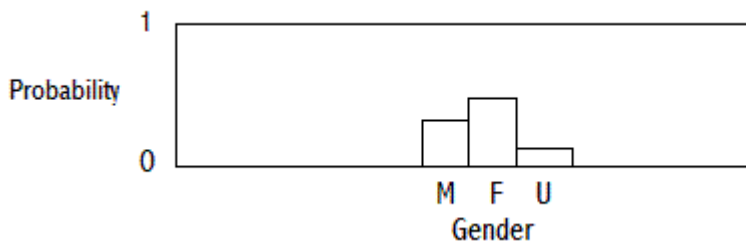
FREQUENCY DISTRIBUTIONS, GRAPHICAL FORM:

DISCRETE VARIABLES

FREQUENCY DISTRIBUTION OF GENDER (THE SUM OF THE FREQUENCIES IS N)

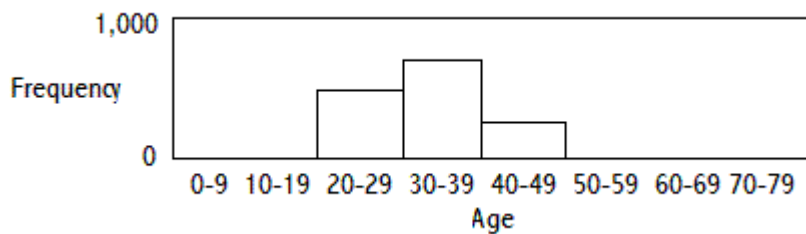


PROBABILITY DENSITY FUNCTION OF GENDER (THE SUM OF THE PROBABILITIES IS 1)

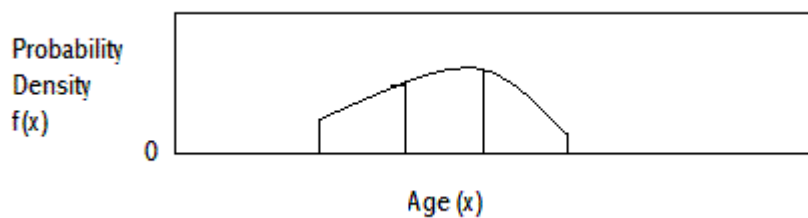


CONTINUOUS VARIABLES (OR ORDERED DISCRETE VARIABLES HAVING MANY VALUES)

FREQUENCY DISTRIBUTION OF AGE (HISTOGRAM)



PROBABILITY DENSITY FUNCTION OF AGE



DESCRIPTION OF A FINITE POPULATION (CONT.)

CROSSTABULATIONS (TABLES OF COUNTS AND MEANS)

(JOINT) FREQUENCY DISTRIBUTION OF POPULATION BY GENDER AND AGE

<u>AGE</u>	<u>MALE</u>	<u>FEMALE</u>	<u>BOTH SEXES</u>
<u>0-18</u>	20	150	170
<u>19-64</u>	150	550	700
<u>65+</u>	30	100	130
<u>ALL AGES</u>	200	800	1,000

TABLE OF MEAN ANNUAL INCOME BY GENDER AND AGE

<u>AGE</u>	<u>GENDER</u>		<u>TOTAL</u>
	<u>MALE</u>	<u>FEMALE</u>	
<u>0-18</u>	1,000	800	900
<u>19-64</u>	30,000	35,000	34,000
<u>65+</u>	10,000	10,000	10,000
<u>TOTAL</u>	20,000	22,000	30,000

STATISTICAL MODELS: REGRESSION EQUATIONS:

INCOME AS A FUNCTION OF EDUCATION: FORMULA OR TABLE

$$y = \bar{y} + b_1x_1 + b_2x_2 + \dots + b_7x_7 + \dots + b_{11}x_{11} + e$$

WHERE

- y = AGE
- x<sub>1</sub> = HAS HIGH SCHOOL DIPLOMA (0 OR 1)
- x<sub>2</sub> = HAS COLLEGE DEGREE (0 OR 1)
- x<sub>7</sub> = PARENTS HAVE COLLEGE DEGREE (0 OR 1)
- x<sub>11</sub> = NUMBER OF YEARS OF WORK EXPERIENCE
- e = ERROR TERM

		Education								
		<12 years	HSD	BA/BS	MS	PhD	MD	Other Prof Degree	Other Degree	Other
Income	<50K									
	50K-100K									
	100K-200K									
	>200K									

SOCIAL AND ECONOMIC IMPACT OF AN ECONOMIC DEVELOPMENT PROGRAM:  
FORMULA OR TABLE

		Program Participation	
		Non-Participant	Participant
Gender	Female		
	Male		

## FACTORS AFFECTING SAMPLE SURVEY DESIGN

THE DESIGN OF THE SAMPLE SURVEY (E.G., CHOICE OF SAMPLING UNITS, SAMPLE SIZES) WILL DEPEND ON WHAT THE ESTIMATION OBJECTIVES ARE, AND THE COSTS INVOLVED (E.G., INSTRUMENT PREPARATION, PRETESTING, SAMPLE DESIGN COSTS, SAMPLING COSTS, ANALYSIS COSTS).

THE OBJECTIVE OF SAMPLE SURVEY DESIGN IS TO ENABLE THE PRODUCTION OF ESTIMATES, OF DESIRED QUANTITIES, THAT ARE OF ADEQUATE ACCURACY (HIGH PRECISION, LOW BIAS) AND ACCEPTABLE COST, TO SUPPORT DECISIONS / ACTIONS.

TWO MAIN CLASSES OF SAMPLE SURVEYS: DESCRIPTIVE SURVEYS (ENUMERATIVE SURVEYS) AND ANALYTICAL SURVEYS (TO SUPPORT MODEL DEVELOPMENT – SIMILAR TO DESIGN OF EXPERIMENTS). THIS COURSE WILL ADDRESS BOTH TYPES OF SURVEYS.

DESCRIPTIVE SURVEYS FOCUS ON ESTIMATION OF OVERALL POPULATION (OR SUBPOPULATION) CHARACTERISTICS (SUCH AS MEANS OR TOTALS). ANALYTICAL SURVEYS FOCUS ON ESTIMATION OF RELATIONSHIPS AMONG VARIABLES AND ON TESTS OF HYPOTHESIS (E.G., IS IT REASONABLE TO CONCLUDE THAT TWO POPULATIONS COULD HAVE BEEN GENERATED BY THE SAME PROBABILITY DISTRIBUTION; OR, DOES AN ECONOMIC DEVELOPMENT PROGRAM HAVE A POSITIVE ECONOMIC IMPACT).

## TYPES OF SAMPLING

1. HAPHAZARD: WITHOUT ANY SCHEME
2. PURPOSIVE, OR JUDGMENT: REPRESENTATIVE
3. RANDOM SAMPLING, OR PROBABILITY SAMPLING: STATISTICAL THEORY CAN BE USED TO MAKE USEFUL STATEMENTS (INFERENCES). STATISTICAL THEORY IS APPLICABLE ONLY IN THIS CASE.

SIMPLEST FORM OF RANDOM SAMPLING: SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

## SOME BASIC CONCEPTS OF PROBABILITY AND STATISTICS

### PROBABILITY THEORY

CONSIDER AN EXPERIMENT, WHICH, WHEN PERFORMED, HAS AN OUTCOME (THE RESULT OF THE EXPERIMENT)

SAMPLE SPACE: THE SET OF ALL POSSIBLE OUTCOMES OF AN EXPERIMENT

EXAMPLES:

COIN-TOSSING EXPERIMENT: HEAD, TAIL

A PERSON SELECTED IN A SURVEY: JOHN SMITH, MARY JONES,...

THE GENDER OF A PERSON SELECTED IN A SURVEY: M, F

A HOUSEHOLD SELECTED IN A SURVEY: THE SMITH FAMILY, THE JONES FAMILY

THE SIZE OF A HOUSEHOLD SELECTED IN A SURVEY: 0, 1, 2, 3, 4,...

AN OPINION: DISAGREE STRONGLY, DISAGREE MILDLY, NEITHER AGREE NOR DISAGREE, AGREE MILDLY, AGREE STRONGLY

IN SAMPLE SURVEY, THE EXPERIMENT IS THE SELECTION ("DRAWING") OF A SAMPLE UNIT. THE SAMPLE SPACE IS THE SET (COLLECTION) OF ALL SAMPLING UNITS.

THE PROBABILITY ASSOCIATED WITH A SAMPLE UNIT IS THE RELATIVE FREQUENCY WITH WHICH THAT UNIT WOULD BE SELECTED, IN REPEATED DRAWINGS.

IN THE SIMPLEST CASE, THE PROBABILITY OF SELECTION OF EACH SAMPLE UNIT IS THE SAME (I.E.,  $1/N$  IN THE CASE OF A SINGLE DRAW). THIS IS REFERRED TO AS SAMPLING WITH EQUAL PROBABILITIES.

IN SAMPLE SURVEY, IT IS FREQUENTLY THE CASE THAT THE SAMPLE UNITS ARE SELECTED WITH UNEQUAL PROBABILITIES. HOW TO SPECIFY THOSE PROBABILITIES, AND HOW TO SELECT THE SAMPLE ACCORDINGLY, IS THE CENTRAL PROBLEM OF SAMPLE SURVEY DESIGN.

**OPTIONAL (SOME ADDITIONAL INFORMATION ABOUT PROBABILITIES, INCLUDED FOR STUDENTS HAVING MATHEMATICAL BACKGROUND):**

EVENT: A SUBSET (PART) OF THE SAMPLE SPACE (A COLLECTION OF OUTCOMES).  
EXAMPLES: HEAD (IN A COIN-TOSSING EXPERIMENT). AN INCOME OF \$50,000 (OF A RESPONDENT TO A SURVEY). USUALLY DENOTED BY  $A, B, C, \dots$

EVENT SPACE: THE COLLECTION OF ALL EVENTS.

AN OUTCOME IS REFERRED TO AS A "SIMPLE EVENT"

THE PROBABILITY OF AN EVENT: THE RELATIVE FREQUENCY WITH WHICH A PARTICULAR OUTCOME OCCURS IN REPETITIONS OF AN EXPERIMENT.

NOTATION:

OUTCOMES ("SIMPLE EVENTS")  $a, b, c, \dots$  or  $A, B, C, \dots$

PROBABILITY OF AN EVENT,  $A = \text{Prob}(A) = \text{Pr}(A) = P(A)$

COMPOUND EVENTS: UNION OF  $A$  AND  $B$  ("A OCCURS OR B OCCURS"),  
INTERSECTION OF  $A$  AND  $B$  ("A OCCURS AND B OCCURS") COMPLEMENT OF  $A$  ("A DOES NOT OCCUR")

PROBABILITY OF  $A$  OR  $B = P(A \text{ UNION } B) = P(A + B)$

PROBABILITY OF  $A$  AND  $B = P(A \text{ INTERSECT } B) = P(AB)$

THE PROBABILITY FUNCTION,  $P(\cdot)$ , SPECIFIES THE PROBABILITY OF EACH EVENT. ITS VALUES RANGE FROM 0 TO 1, AND IF EVENTS  $A$  AND  $B$  ARE MUTUALLY EXCLUSIVE, THEN  $P(A \text{ OR } B) = P(A) + P(B)$ .

DEFINITION OF CONDITIONAL PROBABILITY:

PROBABILITY OF  $A$  GIVEN (CONDITIONAL ON)  $B = P(A|B) = P(AB)/P(B)$  IF  $P(B) > 0$

DEFINITION OF INDEPENDENT EVENTS:

EVENTS  $A$  AND  $B$  ARE INDEPENDENT IF ANY ONE OF THE FOLLOWING THREE CONDITIONS HOLDS:

$$P(AB) = P(A)P(B)$$

$$P(A|B) = P(A) \text{ IF } P(B) > 0$$

$$P(B|A) = P(B) \text{ IF } P(A) > 0$$

RULES FOR WORKING WITH PROBABILITIES:

$$P(A + B) = P(A) + P(B) - P(AB)$$

$$P(AB) = P(A|B)P(B) = P(A)P(B|A)$$

## RANDOM VARIABLES AND DISTRIBUTION FUNCTIONS

RANDOM VARIABLE: A NUMERICAL-VALUED FUNCTION WHOSE VALUE DEPENDS ON THE OUTCOME OF AN EXPERIMENT. (IN MATHEMATICS: “A REAL-VALUED FUNCTION DEFINED ON A SAMPLE SPACE”; IT IS NOT A VARIABLE, BUT A FUNCTION.) DENOTED BY  $X(\cdot)$  OR  $X$ .

A PARTICULAR VALUE OF THE RANDOM VARIABLE WILL BE DENOTED IN LOWER CASE (E.G.,  $x$  IS A PARTICULAR VALUE (RESULT OF A PARTICULAR EXPERIMENT) OF THE RANDOM VARIABLE  $X$ ).

### EXAMPLES OF RANDOM VARIABLES:

#### DISCRETE RANDOM VARIABLES (NOMINAL, ORDINAL; SMALL INTEGERS):

HIV STATUS OF A PERSON IN A SURVEY: NOT INFECTED: 0; INFECTED: 1  
GENDER OF INDIVIDUALS IN A SURVEY: FEMALE: 0; MALE: 1  
SIZE OF A FAMILY IN A SURVEY: 0, 1, 2, 3, 4, 5, ...  
AGE CATEGORY: 0-17: 0; 18-64: 1; 65+: 2  
INCOME CATEGORY: 0-\$50,000 / YR: 1; 50,001 – 100,000 /YR: 2; 100,001 + /YR: 3  
OPINION RESPONSE (“LIKERT SCALE”): DISAGREE STRONGLY: 1;  
DISAGREE MILDLY: 2; NEITHER AGREE NOR DISAGREE: 3; AGREE MILDLY: 4; AGREE STRONGLY: 5

#### CONTINUOUS RANDOM VARIABLES (INTERVAL OR RATIO SCALES OF MEASUREMENT):

THE AGE OF SOMEONE SELECTED IN A SURVEY (IN YEARS)  
THE ANNUAL INCOME OF A FAMILY SELECTED IN A SURVEY (IN DOLLARS; NOT REALLY CONTINUOUS (DOLLARS OR THOUSANDS OF DOLLARS), BUT CLOSE ENOUGH – IT IS THE CONCEPTUAL MEASUREMENT SCALE THAT COUNTS)

A RANDOM VARIABLE IS A FUNCTION (OF THE OUTCOME) THAT HAS A NUMERICAL VALUE, WHEREAS THE OUTCOME OF AN EXPERIMENT MAY SIMPLY BE A NON-NUMERICAL ABSTRACT CONCEPT, SUCH AS A “HEAD” OR “TAIL” IN A COIN-TOSSING EXPERIMENT, A SAMPLE UNIT (SCHOOL, HOSPITAL, PERSON) SELECTED IN A SURVEY, OR GENDER (MALE, FEMALE) OF A PERSON IN A SURVEY.

RANDOM VARIABLES ARE USUALLY DENOTED BY UPPER-CASE LETTERS NEAR THE END OF THE ALPHABET, SUCH AS  $X$ ,  $Y$ ,  $Z$ ,....

NEXT: PROPERTIES OF RANDOM VARIABLES:

EXPECTATION

VARIANCE

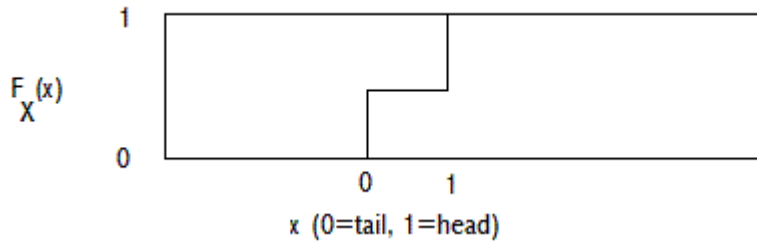
PROBABILITY DISTRIBUTION

**OPTIONAL: CUMULATIVE DISTRIBUTION FUNCTION,  $F_X(\cdot)$ , OF A RANDOM VARIABLE,  $X$ :**  
 $F_X(x) = P(X \leq x) = \text{Prob}(\text{the set of all outcomes, } \omega, \text{ such that } X(\omega) \leq x)$  FOR EVERY REAL NUMBER  $x$ .

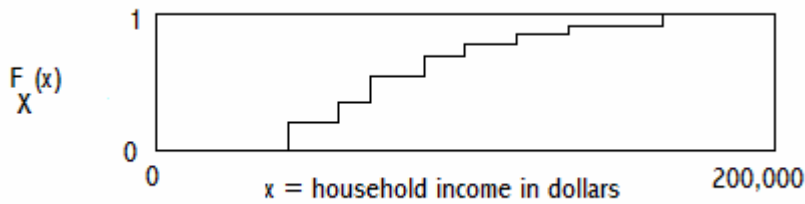
$$F_X(x) = P(X \leq x) = P(\omega : X(\omega) \leq x) \text{ for every real number } x$$

EXAMPLES OF CUMULATIVE DISTRIBUTION FUNCTIONS:

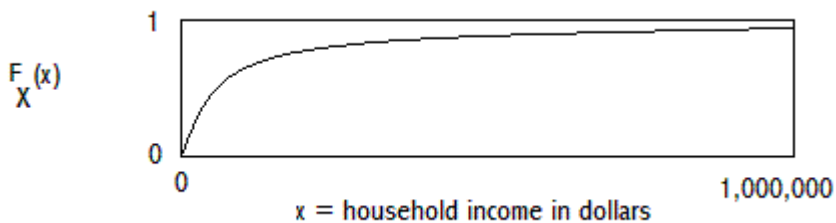
EXAMPLE 1, DISCRETE DISTRIBUTION: COIN TOSSING (TAIL=0, HEAD=1):



EXAMPLE 2, DISCRETE DISTRIBUTION: HOUSEHOLD INCOME IN A SURVEY OF HOUSEHOLDS



EXAMPLE 3, CONTINUOUS DISTRIBUTION: HOUSEHOLD INCOME IN A SURVEY OF HOUSEHOLDS



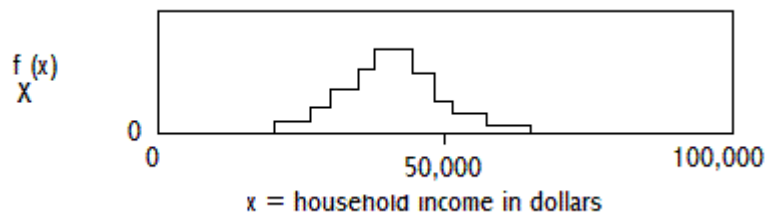
## PROBABILITY DENSITY FUNCTIONS

### DISCRETE RANDOM VARIABLES

THE PROBABILITY FUNCTION, OR DISCRETE DENSITY FUNCTION, OF A RANDOM VARIABLE,  $X$ , HAVING VALUES  $x_1, x_2, x_3, \dots$  IS DEFINED AS:

$$f_x(x) = \begin{cases} P(X = x_i) & \text{if } x = x_i, i = 1, 2, \dots \\ 0 & \text{if } x \neq x_i \end{cases}$$

EXAMPLE: PROBABILITY FUNCTION OF HOUSEHOLD SIZE IN SURVEY OF HOUSEHOLDS



THE SUM OF THE PROBABILITIES IS EQUAL TO ONE.

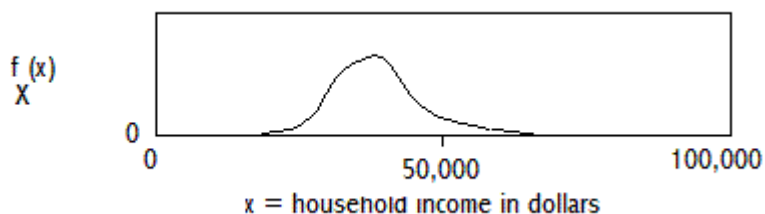
**OPTIONAL: CONTINUOUS RANDOM VARIABLES**

LET  $F_X(\cdot)$  BE THE CUMULATIVE DISTRIBUTION FUNCTION OF THE RANDOM VARIABLE  $X$ . THE RANDOM VARIABLE  $X$  IS CONTINUOUS IF THERE EXISTS A FUNCTION  $f_X(\cdot)$  SUCH THAT

$$F_X(x) = \int_{-\infty}^x f_X(u) du$$

FOR EVERY REAL NUMBER  $x$ . THE FUNCTION  $f_X(\cdot)$  IS CALLED THE PROBABILITY DENSITY FUNCTION OF  $X$ .

EXAMPLE: PROBABILITY DENSITY FUNCTION OF HOUSEHOLD INCOME IN SURVEY OF HOUSEHOLDS



THE AREA UNDER THE CURVE IS EQUAL TO ONE.

## EXPECTATION AND VARIANCE OF A RANDOM VARIABLE

EXPECTATION, OR MEAN, OR EXPECTED VALUE:

DISCRETE CASE:  $E(X) = \mu_x = \sum_i x_i f_X(x_i)$

**OPTIONAL:** CONTINUOUS CASE:  $E(X) = \mu_x = \int_{-\infty}^{\infty} x f_X(x) dx$

THE MEAN IS THE CENTER OF GRAVITY (CENTROID) OF THE UNIT MASS DETERMINED BY THE DENSITY FUNCTION.

VARIANCE: (EXPECTATION OF SQUARED DEVIATIONS FROM THE MEAN):

DISCRETE CASE:

$\text{var}(X) = V(X) = \sigma_x^2 = E(x - \mu)^2 = \sum_i (x_i - \mu_x)^2 f_X(x_i) = \sum_i x_i^2 f_X(x_i) - \mu_x^2$

**OPTIONAL:** CONTINUOUS CASE:  $\text{var}(X) = V(X) = \sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_X(x) dx$

STANDARD DEVIATION = SQUARE ROOT OF VARIANCE:  $\sigma_x = \sqrt{\sigma_x^2}$

NOTE: THE ABOVE FORMULAS FOR THE MEAN AND VARIANCE PRODUCE THE SAME RESULTS (IN THE DISCRETE CASE) AS THE FORMULAS GIVEN EARLIER FOR THE MEAN AND VARIANCE OF THE FINITE POPULATION (WHERE THE PROBABILITY ASSIGNED TO EACH MEMBER OF THE POPULATION IS  $1/N$ ). ALL THAT IS DIFFERENT IS THAT THE ABOVE FORMULAS ARE BASED ON THE PROBABILITY DENSITY FUNCTION OF A RANDOM VARIABLE, WHEREAS THE ORIGINAL FORMULAS WERE INTRODUCED BEFORE THE CONCEPTS OF PROBABILITY, RANDOM VARIABLE, AND THE PROBABILITY DENSITY FUNCTION OF A RANDOM VARIABLE WERE INTRODUCED. THE FORMULAS ARE DIFFERENT, BUT THE RESULTS ARE EXACTLY THE SAME.

THE REASON FOR INTRODUCING THE STATISTICAL THEORY IS NOT TO COMPLICATE THINGS UNNECESSARILY, BUT TO LEAD TO A BETTER UNDERSTANDING OF THE CONCEPTS TO BE INTRODUCED NEXT: STATISTICS, ESTIMATORS, AND SAMPLING DISTRIBUTIONS.

THE PRIMARY GOAL OF SAMPLE SURVEY IS TO OBTAIN ESTIMATES OF THE MEAN AND VARIANCE (AND OTHER QUANTITIES) OF POPULATIONS OF INTEREST, BASED ON SAMPLES FROM THOSE POPULATIONS, AND TO MAKE STATEMENTS ABOUT THE ACCURACY OF THOSE ESTIMATES. THE THEORY OF STATISTICS ENABLES US TO DO THIS.

**OPTIONAL: SOME RULES FOR WORKING WITH RANDOM VARIABLES:**

IF X AND Y ARE TWO RANDOM VARIABLES, THEN

$$\begin{aligned} E(cX) &= c E(X) \\ \text{var}(cX) &= c^2 \text{var}(X) \\ E(X + Y) &= E(X) + E(Y). \end{aligned}$$

IF X AND Y ARE INDEPENDENT, THEN  $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ .

IF  $g(\cdot)$  IS A FUNCTION, THEN THE EXPECTATION OF  $g(X)$  IS DEFINED AS

$$E(g(X)) = \sum_i g(x_i) f_X(x_i).$$

IF X IS DISCRETE, AND

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

IF X IS CONTINUOUS.

CHEBYCHEV (TCHEBYCHEFF) INEQUALITY:

$$P(g(X) \geq k) \leq \frac{E(g(X))}{k} \text{ for every } k > 0.$$

IF WE SET  $g(x) = (x - \mu_x)^2$  and  $k = r^2 \sigma_x^2$ , WE OBTAIN:

$$P(|\bar{X} - \mu_{\bar{X}}| \geq r \sigma_{\bar{X}}) = P((\bar{X} - \mu_{\bar{X}})^2 \geq r^2 \sigma_{\bar{X}}^2) \leq \frac{1}{r^2} \text{ FOR EVERY } r > 0$$

OR

$$P(|\bar{X} - \mu_{\bar{X}}| < r \sigma_{\bar{X}}) \geq 1 - \frac{1}{r^2}$$

FOR  $r = 2$ , WE HAVE  $P(\mu_x - 2\sigma_x < X < \mu_x + 2\sigma_x) \geq 3/4$ , OR FOR ANY RANDOM VARIABLE X HAVING FINITE VARIANCE AT LEAST THREE-FOURTHS OF THE PROBABILITY FALLS WITHIN TWO STANDARD DEVIATIONS OF THE MEAN. (THIS IS NOT A VERY USEFUL RESULT, SINCE IT HOLDS FOR ALL FINITE-VARIANCE RANDOM VARIABLES.)

JENSEN'S INEQUALITY. IF  $g(\cdot)$  IS A CONVEX FUNCTION, THEN  $E(g(X)) \geq g(E(X))$ .

TAYLOR'S APPROXIMATION FOR THE VARIANCE:  $\text{var}(g(X)) \approx (g'(\mu))^2 \text{var}(X)$ , WHERE  $\mu$  DENOTES THE EXPECTATION OF X,  $E(X)$ .

IF  $a_i$  DENOTES THE  $i$ -TH ELEMENT OF A FINITE POPULATION (SAMPLE SPACE), AND  $\text{Prob}(a_i) = p_i$ , LET US DEFINE THE RANDOM VARIABLE  $X(\cdot)$  AS  $X(a_i) = x_i/p_i$ , WHERE  $x_i$  DENOTES SOME ATTRIBUTE (SUCH AS INCOME). THEN  $E(X) =$

$$\sum_i p_i (x_i / p_i) = \sum_i x_i = \tau_X, \text{ THE POPULATION TOTAL.}$$

## (DESCRIPTIVE) SAMPLING THEORY: SAMPLE; PROBABILITY SAMPLE

WE WILL DRAW CONCLUSIONS (MAKE INFERENCES) ABOUT THE POPULATION BASED ON A SAMPLE SELECTED FROM THE POPULATION. THIS IS INDUCTIVE INFERENCE, NOT DEDUCTIVE INFERENCE, SINCE OUR CONCLUSIONS ARE NOT MADE WITH CERTAINTY.

A SAMPLE IS A COLLECTION OF SAMPLING UNITS,  $X_1, X_2, \dots, X_n$  DRAWN FROM A FRAME. THE SIZE (NUMBER OF UNITS) OF THE SAMPLE IS DENOTED BY  $n$ . THE SAMPLE WILL BE DRAWN IN A SPECIAL WAY, DEPENDING ON THE OBJECTIVES OF THE SURVEY.

MOST AREAS OF STATISTICS (EXPERIMENTAL DESIGN, QUALITY CONTROL, RELIABILITY) APPLICATIONS DEAL WITH RANDOM SAMPLES. A RANDOM SAMPLE IS USUALLY DEFINED AS A SET OF INDEPENDENT AND IDENTICALLY DISTRIBUTED RANDOM VARIABLES (ALL DEFINED ON THE SAME SAMPLE SPACE).

THIS IS NOT THE TYPE OF SAMPLE THAT IS TYPICALLY USED IN (DESCRIPTIVE) SAMPLE SURVEY. IN SAMPLE SURVEY, SOME SAMPLE ELEMENTS MAY BE SELECTED FROM SUBSETS OF THE SAMPLE SPACE (POPULATION), AND THEY ARE NOT NECESSARILY INDEPENDENT. FURTHERMORE, THE PROBABILITY DENSITY FUNCTION USUALLY HAS NO SPECIFIC FORM (E.G., NORMAL, LOGNORMAL, EXPONENTIAL, BINOMIAL).

THE SAMPLES IN SAMPLE SURVEY ARE PROBABILITY SAMPLES (EACH UNIT OF THE SAMPLE IS SELECTED WITH A KNOWN, NONZERO PROBABILITY), BUT NOT THE USUAL "RANDOM SAMPLES" OF STATISTICS (INDEPENDENT AND IDENTICALLY DISTRIBUTED RANDOM VARIABLES).

THE INFERENCES ARE BEING MADE ABOUT THE PARTICULAR FINITE POPULATION AT HAND, NOT ABOUT A HYPOTHETICAL PROCESS THAT MAY HAVE GENERATED IT (CREATED IT AS A SAMPLE, OR "REALIZATION," FROM A "SUPERPOPULATION" OF POPULATIONS). (IN DAY 2 WE WILL CONSIDER ANALYTICAL SURVEY DESIGNS, WHICH ARE BASED ON A MODEL OF A HYPOTHETICAL PROCESS THAT MAY BE CONSIDERED TO HAVE GENERATED THE PARTICULAR POPULATION AT HAND.)

(FOR DISCUSSION OF THIS CONCEPT, SEE "HISTORY AND DEVELOPMENT OF THE THEORETICAL FOUNDATIONS OF SURVEY BASED ESTIMATION AND ANALYSIS," BY J. N. K. RAO AND D. R. BELLHOUSE, SURVEY METHODOLOGY, VOL. 16, NO. 1, PP. 3-29 (JUNE 1990) STATISTICS CANADA. RAO AND BELLHOUSE DISCUSS THREE APPROACHES TO SAMPLE SURVEY: DESIGN-BASED (CORRESPONDING TO DESCRIPTIVE SURVEYS), MODEL-DEPENDENT AND MODEL-BASED (OR MODEL-ASSISTED) (THE LATTER TWO CORRESPONDING TO ANALYTICAL SURVEYS).)

## SAMPLING THEORY (CONT.): STATISTIC; ESTIMATOR

FOR EACH POPULATION QUANTITY OF INTEREST (E.G., POPULATION MEAN, SUBPOPULATION MEANS), WE WISH TO ESTIMATE THE QUANTITY FROM THE SAMPLE.

A STATISTIC IS ANY QUANTITY THAT CAN BE CALCULATED FROM THE SAMPLE (I.E., A FUNCTION OF THE SAMPLE). (A STATISTIC IS A RANDOM VARIABLE; IT DOES NOT DEPEND ON ANY UNKNOWN PARAMETERS, SUCH AS  $\mu$  OR  $\sigma_2$ .)

AN ESTIMATOR IS A STATISTIC USED TO ESTIMATE A POPULATION CHARACTERISTIC (PARAMETER, SUCH AS A MEAN OR TOTAL). (IF A STATISTIC IS USED TO ESTIMATE A POPULATION PARAMETER, THEN IT IS AN ESTIMATOR.)

EXAMPLE OF AN ESTIMATOR: USE THE SAMPLE MEAN TO ESTIMATE THE POPULATION MEAN. (THE VALUE OF THE ESTIMATOR, CALCULATED FROM A PARTICULAR SAMPLE, IS CALLED THE ESTIMATE. THE ESTIMATOR IS A FUNCTION (FORMULA); THE ESTIMATE IS A NUMBER.)

TWO TYPES OF ESTIMATION (ESTIMATORS): POINT ESTIMATION AND INTERVAL ESTIMATION. WE CONSIDER POINT ESTIMATION FIRST.

THE ACADEMIC DISCIPLINE OF SAMPLE SURVEY IS CONCERNED WITH IDENTIFYING "GOOD" ESTIMATORS (ESTIMATORS THAT ARE IN SOME SENSE "CLOSE" TO THE POPULATION VALUES THEY ARE INTENDED TO ESTIMATE), AND IN DETERMINING SAMPLE DESIGNS THAT ASSURE THAT THE ESTIMATES ARE AS "CLOSE" AS DESIRED.

THE APPLIED FIELD OF SAMPLE SURVEY DESIGN AND ANALYSIS IS CONCERNED WITH KNOWING THOSE ESTIMATORS AND SAMPLE DESIGN PROCEDURES.

SAMPLING THEORY (CONT.): PRECISION, TRUENESS/BIAS, ACCURACY

PROPERTIES OF ESTIMATORS:

PRECISION: HOW MUCH VARIATION IS THERE IN THE ESTIMATE, FROM SAMPLE TO SAMPLE

TRUENESS: ON AVERAGE, HOW “CLOSE” IS THE ESTIMATE TO THE POPULATION CHARACTERISTIC BEING ESTIMATED

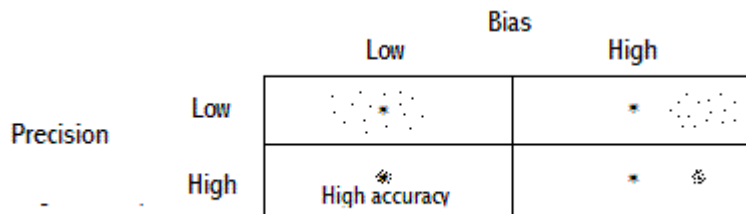
ACCURACY: A COMBINATION OF PRECISION AND TRUENESS

ISO-5725 (“ACCURACY (TRUENESS AND PRECISION) OF MEASUREMENT METHODS AND RESULTS”) USES THE TERM “TRUENESS”; STATISTICIANS USUALLY USE THE TERM “BIAS” (WHICH HAS THE REVERSE HIGH/LOW SENSE OF TRUENESS).

OTHER TERMS FOR THE CONCEPT OF PRECISION ARE REPEATABILITY AND RELIABILITY (ALL IN THE SAME HIGH/LOW SENSE); AND VARIABILITY, SPREAD AND DISPERSION (REVERSE SENSE).

OTHER TERMS FOR THE CONCEPT OF TRUENESS ARE VALIDITY AND UNBIASEDNESS (SAME HIGH/LOW SENSE); AND BIAS (REVERSE SENSE).

GRAPHIC ILLUSTRATION OF RELATIONSHIP BETWEEN PRECISION, BIAS AND ACCURACY.



THE PRECISION (OF AN ESTIMATOR,  $X$ ) WILL BE MEASURED BY THE VARIANCE:

$$\text{var}(X) = V(X) = \sigma_x^2 = E((X - \mu_x)^2);$$

WHERE  $\mu_x = E(X)$ , OR THE STANDARD DEVIATION:

$$SD(X) = \sigma_x = \sqrt{\text{var}(X)}.$$

THE TRUENESS OF AN ESTIMATOR WILL BE MEASURED BY THE BIAS, WHICH IS DEFINED AS THE DIFFERENCE BETWEEN THE EXPECTATION OF THE ESTIMATOR AND THE POPULATION PARAMETER OF WHICH IT IS AN ESTIMATE:

$$B_x = E(X) - \mu_x.$$

AN ESTIMATOR IS UNBIASED IF ITS EXPECTATION IS EQUAL TO THE POPULATION PARAMETER OF WHICH IT IS AN ESTIMATE.

ACCURACY WILL BE MEASURED BY THE MEAN SQUARED ERROR:

$$MSE(X) = E((X - \mu_x)^2) = \sigma_x^2 + B_x^2.$$

WE WOULD LIKE ESTIMATORS THAT HAVE LOW VARIANCE (HIGH PRECISION) AND LOW BIAS (COMPARED TO OTHER ESTIMATORS THAT HAVE THE SAME SAMPLE SIZE OR SAMPLING COST).

THERE ARE OTHER PROPERTIES OF ESTIMATORS, SUCH AS CONSISTENCY (THE TENDENCY FOR AN ESTIMATE OF A PARAMETER TO APPROACH THE PARAMETER VALUE AS THE SAMPLE SIZE INCREASES). THE MOST IMPORTANT PROPERTIES OF ESTIMATORS ARE PRECISION AND BIAS.

PRINCIPAL ITEMS OF INTEREST FOR EACH SAMPLE DESIGN AND ESTIMATION METHOD:

IN WHAT FOLLOWS, WE SHALL PRESENT FORMULAS FOR VARIOUS SAMPLE ESTIMATES OF POPULATION PARAMETERS.

WE SHALL INDICATE WHETHER AN ESTIMATOR IS BIASED OR UNBIASED. WE SHALL ALSO PRESENT FORMULAS FOR THE TRUE VARIANCE OF THE ESTIMATE AND THE SAMPLE ESTIMATE OF THE VARIANCE OF THE ESTIMATE (AND ITS SQUARE ROOT, THE ESTIMATED STANDARD ERROR OF THE ESTIMATE).

THE TRUE VALUE IS OF INTEREST TO HELP US DECIDE ON SAMPLE SIZES, DURING THE COURSE OF DESIGNING A SURVEY.

THE ESTIMATED VALUE IS OF INTEREST TO INDICATE THE PRECISION OF AN ESTIMATE, AFTER THE SURVEY IS COMPLETED AND THE DATA ANALYZED. THE ESTIMATED STANDARD ERROR IS USED TO CONSTRUCT CONFIDENCE INTERVALS.

## SAMPLING THEORY (CONT.): NOTES

THERE ARE VARIOUS METHODS FOR DETERMINING ESTIMATORS (METHOD OF MOMENTS, LEAST-SQUARES, MAXIMUM LIKELIHOOD, BAYESIAN METHODS, RAO-BLACKWELL METHOD, MINIMUM CHI-SQUARE, MINIMUM-DISTANCE) AND VARIOUS CRITERIA FOR COMPARING ESTIMATORS (BIAS, VARIANCE, MEAN SQUARED ERROR, CONSISTENCY, SUFFICIENCY, LOCATION/SCALE INVARIANCE). THESE METHODS WILL NOT BE ADDRESSED IN THIS COURSE.

NOTE ON SCOPE OF COURSE: IN THIS COURSE, WE RESTRICT ATTENTION TO STANDARD ESTIMATORS OF COMMON POPULATION PARAMETERS SUCH AS MEANS AND TOTALS, OR RATIOS OR DIFFERENCES AMONG THEM, AND USE LINEAR ESTIMATION TECHNIQUES (LINEAR COMBINATIONS OF THE SAMPLE VALUES). FOR MORE COMPLEX PARAMETERS, SUCH AS SIMPLE AND PARTIAL CORRELATION COEFFICIENTS, MEDIANS, QUANTILES, REGRESSION COEFFICIENTS, MORE ADVANCED METHODS (NONLINEAR ESTIMATION PROCEDURES) ARE REQUIRED. SINCE THE SAMPLING METHODS USED IN SAMPLE SURVEY ARE COMPLEX, THESE ESTIMATION METHODS ARE COMPLICATED (E.G., TAYLOR SERIES EXPANSION, BALANCED REPEATED REPLICATION, THE "JACKKNIFE" METHOD, RESAMPLING).

### SAMPLING THEORY (CONT.): EXAMPLE

CONSIDER THE ESTIMATOR THAT IS THE SAMPLE MEAN, FROM A SIMPLE RANDOM SAMPLE DRAWN WITH REPLACEMENT ("SRSWR").

POPULATION ELEMENTS:  $x_1, x_2, \dots, x_N$  (LOWER CASE SIGNIFIES ACTUAL NUMERICAL VALUES)

SAMPLE:  $X_1, X_2, \dots, X_n$  (UPPER CASE SIGNIFIES RANDOM VARIABLES; FUNCTIONS; CONCEPTUAL)

SAMPLE:  $x_1, x_2, \dots, x_n$  (LOWER CASE SIGNIFIES NUMBERS, IN A SPECIFIC CASE)

NOTE: THE ITEMS  $x_1, x_2, \dots, x_n$  OF THE SAMPLE ARE NOT (NECESSARILY) THE FIRST  $n$  ITEMS ( $x_1, x_2, \dots, x_n$ ) OF THE POPULATION.

POPULATION MEAN:  $\mu_x = \frac{\sum_{i=1}^N x_i}{N} = \bar{X} = \bar{X}_N$

SAMPLE MEAN:  $\hat{\mu}_x = \frac{\sum_{i=1}^n X_i}{n} = \hat{\bar{X}} = \bar{X}_n$  (conceptual, a random variable)

OR  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$  (a number, calculated for a particular sample)

## SAMPLING THEORY (CONT.): EXAMPLE (CONT.)

NOTE: THERE IS INCONSISTENCY IN THE FIELD OF STATISTICS IN THE USE OF CAPITAL LETTERS AND LOWER-CASE (“SMALL”) LETTERS. SOME AUTHORS USE CAPITAL LETTERS TO DENOTE RANDOM VARIABLES (FUNCTIONS), AND SMALL LETTERS TO DENOTE REAL NUMBERS (ELEMENTS OF A POPULATION OR SAMPLE, OBSERVED VALUES OF AN ESTIMATOR). OTHERS USE CAPITAL LETTERS TO REFER TO POPULATION PARAMETERS (MEAN, TOTAL) AND SMALL LETTERS TO REFER TO POPULATION ELEMENTS, SAMPLE ELEMENTS, AND SAMPLE ESTIMATORS, WITHOUT DISTINGUISHING BETWEEN RANDOM VARIABLES (FUNCTIONS) AND OBSERVED NUMERICAL VALUES.

WHAT IS EVEN MORE CONFUSING IS THAT MANY AUTHORS USE THE SAME SYMBOL INTERCHANGEABLY AS A RANDOM VARIABLE OR A REAL NUMBER, WITHOUT

COMMENT. FOR EXAMPLE, IN THE EXPRESSION  $\bar{y} = \sum_{i=1}^n y_i$ ,  $\bar{y}$  AND  $y_1, y_2, \dots, y_n$  ARE

USED EITHER AS RANDOM VARIABLES OR AS NUMBERS FROM A PARTICULAR SAMPLE. THIS PRACTICE MUST BE VERY CONFUSING TO THE NEW STUDENT, BUT IT IS NOT UNUSUAL. (ANOTHER CONFUSING ITEM IS THE USE OF THE TERM “RANDOM VARIABLE” TO DESCRIBE A FUNCTION.)

IN MATHEMATICAL STATISTICS, THIS DISTINCTION IS VERY IMPORTANT. FOR EXAMPLE, IN THE STATEMENT,  $Prob(X = x)$ ,  $X$  REFERS TO A RANDOM VARIABLE (A FUNCTION), AND  $x$  REFERS TO A REAL NUMBER. FOR EXAMPLE,  $Prob(AGE = 27)$ . IN THIS CASE, THE EXPRESSION  $E(X)$  REFERS TO THE EXPECTATION (EXPECTED VALUE, MEAN VALUE, AVERAGE VALUE) OF THE RANDOM VARIABLE  $X$ , WHICH IS THE MEAN AGE OF THE MEMBERS OF THE POPULATION. THE EXPRESSION  $E(x)$  IS SIMPLY THE EXPECTATION OF THE REAL NUMBER,  $x$ , WHICH IS  $x$ . IN THE EXAMPLE WHERE  $X = AGE$  AND  $x = 27$ ,  $E(AGE) = \mu_x = 40.3$ , SAY, BUT  $E(27) = 27$ .

IN THIS COURSE, CAPITAL LETTERS WILL REFER TO POPULATION PARAMETERS AND TO RANDOM VARIABLES, AND LOWER-CASE LETTERS WILL REFER TO POPULATION ELEMENTS, SAMPLE ELEMENTS, AND THE CALCULATED VALUES OF SAMPLE STATISTICS. LOWER-CASE LETTERS WILL NOT REFER TO RANDOM VARIABLES (UNLESS SPECIFICALLY STATED). ALTERNATIVE NOTATIONS WILL BE PRESENTED, TO FAMILIARIZE THE STUDENT WITH NOTATIONS FOUND IN DIFFERENT REFERENCE TEXTS.

WHILE THIS DISTINCTION IS IMPORTANT CONCEPTUALLY (MATHEMATICALLY), IT OFTEN IS IGNORED IN PRACTICAL APPLICATIONS. FOR EXAMPLE, IN RECALLING THE FORMULA USED TO CALCULATE THE SAMPLE MEAN, IT DOES NOT MATTER WHETHER

ONE RECALLS  $\bar{X} = \sum_{i=1}^n X_i / n$  (A FORMULA INVOLVING RANDOM VARIABLES) OR

$\bar{x} = \sum_{i=1}^n x_i / n$  (A FORMULA INVOLVING REAL NUMBERS FROM A PARTICULAR SAMPLE).

IN THE INTEREST OF SIMPLICITY, WE SHALL OFTEN USE THE LATTER TYPE OF EXPRESSION (LOWER-CASE LETTERS, SAMPLE VALUES) FOR FORMULAS.

SAMPLING THEORY (CONT.): EXAMPLE (CONT.)

IT CAN BE SHOWN THAT:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu_x$$

AND

$$\sigma_{\bar{X}}^2 = \text{var}(\bar{X}) = V(\bar{X}) = \frac{\sigma_x^2}{n}$$

WHERE  $\mu_x$  IS THE POPULATION MEAN AND  $\sigma_x^2$  IS THE POPULATION VARIANCE.

$\mu_x$  IS ESTIMATED (IN SRSWR) BY  $\hat{\mu}_x = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ .

$\sigma_x^2$  IS ESTIMATED (IN SRSWR) BY:

$$\hat{\sigma}_x^2 = \hat{V}(\bar{X}) = v(\bar{X}) = v(\hat{\mu}_x) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n (X_i - \hat{\mu}_x)^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 - n\hat{\mu}_x^2}{n-1}.$$

(THE DIVISOR  $n-1$  IS USED INSTEAD OF  $n$  SO THAT  $\hat{\sigma}_x^2$  IS UNBIASED. THE BEST FORMULAS FOR SAMPLE ESTIMATES OF POPULATION PARAMETERS ARE OFTEN NOT IDENTICAL IN FORM TO THE POPULATION FORMULAS. IN FACT, IN SOME CASES, SUCH AS ESTIMATING A POWER SPECTRUM IN TIME SERIES ANALYSIS, USING THE POPULATION FORMULA PRODUCES A TERRIBLE ESTIMATE (IT IS NOT EVEN CONSISTENT).)

$\sigma_{\bar{X}}^2$  IS ESTIMATED (IN SRSWR) BY:

$$\hat{\sigma}_{\bar{X}}^2 = \frac{\hat{\sigma}_x^2}{n}.$$

THE QUANTITY  $\sigma_{\bar{X}} = \sqrt{v(\hat{\mu}_x)}$  IS CALLED THE STANDARD ERROR OF THE ESTIMATE,  $\hat{\mu}_x$ , DENOTED  $SE(\hat{\mu}_x)$ . (THE STANDARD ERROR OF THE ESTIMATE IS SIMPLY THE STANDARD DEVIATION OF THE ESTIMATE, BUT THE TERM "STANDARD ERROR" IS USED INSTEAD OF "STANDARD DEVIATION" WHEN REFERRING TO THE STANDARD DEVIATION OF ESTIMATES OR POPULATION PARAMETERS.)

IT IS ESTIMATED (IN SRSWR) BY  $\hat{\sigma}_{\bar{X}} = \sqrt{\hat{\sigma}_{\bar{X}}^2}$ .

### SAMPLING THEORY (CONT.): EXAMPLE (CONT.)

HENCE THE SAMPLE MEAN,  $\bar{X}$ , OF A SIMPLE RANDOM SAMPLE DRAWN WITH REPLACEMENT IS UNBIASED, AND ITS VARIANCE DECREASES BY THE FACTOR  $1/n$  AS THE SAMPLE SIZE  $n$  INCREASES.

IT CAN BE SHOWN THAT THE SAMPLE MEAN OF A SIMPLE RANDOM SAMPLE DRAWN WITH REPLACEMENT HAS THE MINIMUM VARIANCE OF ALL UNBIASED ESTIMATORS (OF THE POPULATION MEAN) THAT ARE LINEAR FUNCTIONS OF THE SAMPLE ("BEST LINEAR UNBIASED ESTIMATE," "BLUE").

NOTE: THE PRECEDING ESTIMATION FORMULAS APPLY TO SIMPLE RANDOM SAMPLING WITH REPLACEMENT. FOR OTHER METHODS OF SAMPLING, THE FORMULAS ARE DIFFERENT.

NOTE: THE STANDARD ERROR OF THE ESTIMATE IS OF INTEREST MAINLY FOR CONSTRUCTING INTERVAL ESTIMATES (TO BE EXAMINED SHORTLY).

NOTE ALSO:

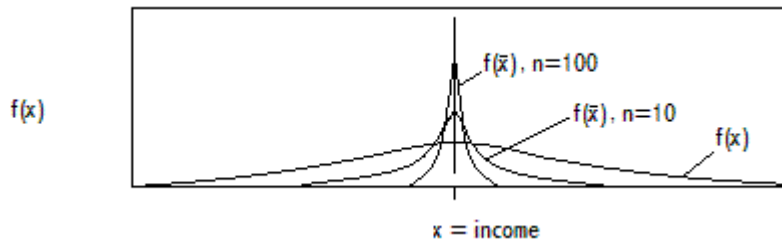
POPULATION TOTAL =  $\tau_x (= X) = N\mu_x$

SAMPLE ESTIMATE OF POPULATION TOTAL =  $\hat{\tau}_x (= \hat{X}) = N\bar{x}$

## SAMPLING THEORY (CONT.): SAMPLING DISTRIBUTION

SAMPLING DISTRIBUTION OF THE ESTIMATOR (THE PROBABILITY DISTRIBUTION OF A RANDOM VARIABLE).

CONSIDER THE HYPOTHETICAL UNIVERSE OF ALL POSSIBLE SAMPLES FOR ANY METHOD OF SAMPLING. THERE IS A NUMERICAL VALUE OF THE STATISTIC (OR ESTIMATE) FOR EVERY POSSIBLE SAMPLE. THE PROBABILITY DISTRIBUTION OF THIS STATISTIC (ESTIMATE) IS THE SAMPLING DISTRIBUTION OF THE STATISTIC.



(WEAK) LAW OF LARGE NUMBERS: AS THE SAMPLE SIZE INCREASES, THE SAMPLE MEAN OF A SIMPLE RANDOM SAMPLE DRAWN WITH REPLACEMENT BECOMES VERY CLOSE TO THE POPULATION MEAN (THE PROBABILITY THAT THE SAMPLE MEAN DIFFERS BY ANY SPECIFIED AMOUNT FROM THE POPULATION MEAN DECREASES TO ZERO AS THE SAMPLE SIZE INCREASES TO INFINITY).

CENTRAL LIMIT THEOREM: THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN (IN SIMPLE RANDOM SAMPLING WITH REPLACEMENT) APPROACHES THE NORMAL DISTRIBUTION (THE "BELL-SHAPED CURVE") AS THE SAMPLE SIZE APPROACHES INFINITY. IF  $\mu_x$  AND  $\sigma_x^2$  DENOTE THE MEAN AND VARIANCE OF THE POPULATION, THEN THE MEAN AND VARIANCE OF THE LIMITING DISTRIBUTION ARE  $\mu_x$  AND  $\sigma_x^2/n$ . WE WILL DISCUSS THE NORMAL DISTRIBUTION IN GREATER DETAIL SHORTLY.

THE TWO PRECEDING RESULTS ARE TRUE FOR ANY FINITE POPULATION (THEY HOLD FOR RANDOM SAMPLING FROM ANY PROBABILITY DENSITY WITH FINITE VARIANCE).

THESE RESULTS ARE VERY IMPORTANT IN SAMPLE SURVEY. THEY ARE THE BASIS FOR THE ESTIMATION FORMULAS THAT ARE USED TO ANALYZE THE DATA.

THEY ARE VERY USEFUL SINCE THE FORM OF THE PROBABILITY DISTRIBUTION IS USUALLY NOT KNOWN (BUT ALWAYS HAS A FINITE VARIANCE).

THEY ARE APPLICABLE FOR LARGE SAMPLE SIZES (E.G.,  $N > 30$ ).

## SAMPLING THEORY (CONT.): INTERVAL ESTIMATION

THE PRECISION OF AN ESTIMATE WILL BE DETERMINED, USING THE THEORY OF STATISTICS, BASED ON INFORMATION IN THE SAMPLE.

THE FORMULAS USED TO ESTIMATE THE POPULATION MEAN (OR OTHER CHARACTERISTIC) WILL DEPEND ON THE SAMPLE DESIGN AND SAMPLE SELECTION METHOD.

THE PRECEDING MATERIAL ON ESTIMATION HAS BEEN CONCERNED WITH POINT ESTIMATION OF POPULATION PARAMETERS. WE WILL NOW ADDRESS INTERVAL ESTIMATION.

A POINT ESTIMATE SPECIFIES A SINGLE, "LIKELY," VALUE AS THE ESTIMATE. WE KNOW SOMETHING ABOUT ITS PRECISION FROM ITS STANDARD ERROR (AND THE THEORY OF STATISTICS).

AN INTERVAL ESTIMATE OF A PARAMETER IS AN INTERVAL THAT HAS A SPECIFIED PROBABILITY OF INCLUDING THE PARAMETER (THE INTERVAL IS THE RANDOM QUANTITY, NOT THE PARAMETER).

CONFIDENCE INTERVAL. LET  $\theta$  DENOTE A POPULATION PARAMETER (E.G., THE MEAN,  $\mu$ ). LET  $T_1 = t_1(X_1, \dots, X_n)$  AND  $T_2 = t_2(X_1, \dots, X_n)$  BE TWO STATISTICS SATISFYING  $T_1 \leq T_2$  FOR WHICH  $P(T_1 < \theta < T_2) = \alpha$ , WHERE  $\alpha$  DOES NOT DEPEND ON  $\theta$ . THEN THE RANDOM INTERVAL  $(T_1, T_2)$  IS CALLED A  $100\alpha$  PERCENT CONFIDENCE INTERVAL FOR  $\theta$ .

$\alpha$  IS CALLED THE CONFIDENCE COEFFICIENT.  $T_1$  AND  $T_2$  ARE CALLED THE LOWER AND UPPER CONFIDENCE LIMITS, RESPECTIVELY.

A VALUE  $(t_1, t_2)$  OF THE RANDOM INTERVAL  $(T_1, T_2)$  IS ALSO CALLED A  $100\alpha$  PERCENT CONFIDENCE INTERVAL FOR  $\theta$ .

(SIMILAR DEFINITIONS FOR ONE-SIDED LOWER AND UPPER CONFIDENCE INTERVALS AND LIMITS.)

NOTE: CONFIDENCE INTERVALS ARE RELATED TO TESTS OF HYPOTHESIS, TO BE DISCUSSED IN DAY 2.

## SAMPLING THEORY (CONT.): CONFIDENCE INTERVALS

HOW TO DETERMINE CONFIDENCE INTERVALS: NEED TO USE INFORMATION ABOUT THE SAMPLING DISTRIBUTION OF CERTAIN STATISTICS.

EXAMPLE: FROM PROBABILITY THEORY: CHEBYCHEV INEQUALITY (FOR ANY STATISTIC FROM A FINITE POPULATION, NOT JUST  $\bar{X}$ ):

$$P(|\bar{X} - \mu_{\bar{X}}| < r\sigma_{\bar{X}}) \geq 1 - \frac{1}{r^2}$$

SO (SINCE  $\mu_{\bar{X}} = \mu_X$ )  $(\bar{X} - r\sigma_{\bar{X}}, \bar{X} + r\sigma_{\bar{X}})$  IS AN APPROXIMATE  $(1 - 1/r^2)$  PERCENT CONFIDENCE INTERVAL FOR  $\mu_X$ .

THE CHEBYCHEV-INEQUALITY METHOD OF CONSTRUCTING CONFIDENCE INTERVALS IS NOT VERY GOOD.

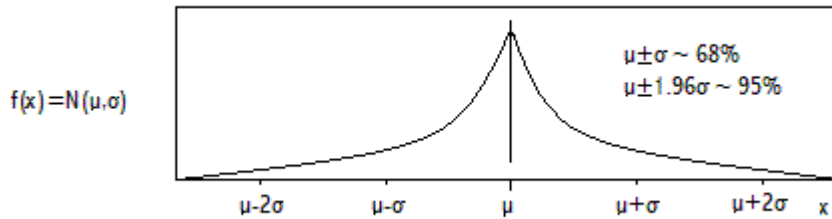
IT IS USUALLY MUCH BETTER TO OBTAIN CONFIDENCE INTERVALS FROM THE KNOWLEDGE THAT, FOR LARGE SAMPLES, THE SAMPLING DISTRIBUTION OF  $\bar{X}$  TENDS TO A NORMAL DISTRIBUTION WITH MEAN  $\mu_{\bar{X}} = \mu_X$  AND VARIANCE

$$\sigma_{\bar{X}}^2 = \sigma_X^2 / n.$$

## SAMPLING THEORY (CONT.): THE NORMAL DISTRIBUTION

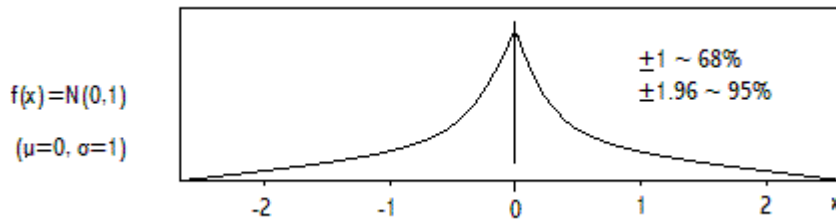
### THE NORMAL DISTRIBUTION:

PROBABILITY DENSITY FUNCTION:  $f_X(x) = f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$



THE STANDARD NORMAL DISTRIBUTION IS THE DISTRIBUTION OF  $z = \frac{x - \mu}{\sigma}$ . IT IS A

NORMAL DISTRIBUTION WITH  $\mu = 0$  AND  $\sigma = 1$ :  $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$



TABLES AVAILABLE. 95% OF THE AREA IS CONTAINED WITHIN THE INTERVAL  $\mu \pm 1.96\sigma$ . 90% OF THE AREA IS CONTAINED WITHIN THE INTERVAL  $\mu \pm 1.645\sigma$ .

## SAMPLING THEORY (CONT.): CONFIDENCE INTERVALS

CONFIDENCE INTERVAL FOR  $\bar{X}$  :

$$P(-z_{\alpha/2} < \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < z_{1-\alpha/2}) = \alpha$$

SO (REARRANGING AND SETTING  $\mu_{\bar{X}} = \mu_X$ )

$$P(\mu_X - z_{\alpha/2}\sigma_{\bar{X}} < \bar{X} < \mu_X + z_{1-\alpha/2}\sigma_{\bar{X}}) = \alpha$$

HENCE

$$P(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} < \mu_X < \bar{X} + z_{1-\alpha/2}\sigma_{\bar{X}}) = \alpha$$

HENCE

$$(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}}, \bar{X} + z_{1-\alpha/2}\sigma_{\bar{X}})$$

IS A  $100\alpha$  PERCENT CONFIDENCE INTERVAL FOR  $\mu$ , WHERE  $z_p$  DENOTES THE  $p$  PERCENTILE POINT OF THE NORMAL PROBABILITY DENSITY FUNCTION.

FOR EXAMPLE,  $z_{.025} = -1.96$  AND  $z_{.975} = 1.96$ , SO

$$(\bar{X} - 1.96\sigma_{\bar{X}}, \bar{X} + 1.96\sigma_{\bar{X}})$$

IS A 95% CONFIDENCE INTERVAL FOR  $\mu_X$ .

IN PRACTICE, WE DO NOT KNOW THE VALUE OF  $\sigma_{\bar{X}}$ , AND WE USE THE SAMPLE ESTIMATE,  $\hat{\sigma}_{\bar{X}}$ , IN ITS PLACE. IN THIS CASE, HOWEVER, THE PROBABILITY

DISTRIBUTION OF  $\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$  IS NOT EXACTLY A STANDARD NORMAL DISTRIBUTION. IT

IS A STUDENT'S  $t$  DISTRIBUTION, WHICH IS A LITTLE "WIDER" THAN THE STANDARD NORMAL DISTRIBUTION. A REASONABLE AND CONVENIENT APPROXIMATION IS TO REPLACE THE FACTOR 1.96 BY 2, AND USE

$$(\bar{X} - 2\hat{\sigma}_{\bar{X}}, \bar{X} + 2\hat{\sigma}_{\bar{X}}) = (\bar{X} - 2SE(\bar{X}), \bar{X} + 2SE(\bar{X}))$$

AS A 95% CONFIDENCE INTERVAL.

THE QUANTITY  $2SE(\bar{X})$  IS CALLED THE "BOUND ON THE ERROR OF ESTIMATION OF  $\bar{X}$ ".

THE PRECEDING ILLUSTRATED THE CONSTRUCTION OF A CONFIDENCE INTERVAL FOR THE POPULATION MEAN,  $\mu$ , IN THE CASE OF SIMPLE RANDOM SAMPLING WITH REPLACEMENT, USING  $\bar{X}$  AS THE ESTIMATE OF THE POPULATION MEAN,  $\mu_X$ , AND  $\hat{\sigma}_{\bar{X}}$  AS ITS ESTIMATED STANDARD ERROR. IN GENERAL, FOR AN ARBITRARY SAMPLE DESIGN (AND PARAMETER TO BE ESTIMATED), WE CONSTRUCT THE CONFIDENCE INTERVAL USING THE APPROPRIATE PARAMETER ESTIMATE AND ITS ESTIMATED STANDARD ERROR.

## SAMPLING THEORY (CONT.): DIFFERENT SAMPLING METHODS

WE WILL NOW EXAMINE SEVERAL DIFFERENT TYPES OF SAMPLING USED IN SAMPLE SURVEY:

SIMPLE RANDOM SAMPLING, WITH AND WITHOUT REPLACEMENT  
STRATIFIED SAMPLING  
CLUSTER SAMPLING  
SYSTEMATIC SAMPLING  
MULTISTAGE SAMPLING  
DOUBLE SAMPLING (TWO-PHASE SAMPLING)

AND TWO ALTERNATIVE TYPES OF ESTIMATION (IN ADDITION TO THE USUAL LINEAR ESTIMATORS):

RATIO ESTIMATORS  
REGRESSION ESTIMATORS.

IN DAY ONE OF THE COURSE, WE EXAMINE THE SAMPLING METHODS (DEFINITIONS, SAMPLE SELECTION METHODS, ESTIMATION FORMULAS (POINT ESTIMATES, CONFIDENCE INTERVALS)

IN DAY TWO WE SHOW HOW TO DETERMINE WHICH ONE TO USE, AND HOW TO TAILOR IT TO THE PARTICULAR APPLICATION (I.E., DESIGN THE SURVEY).

A NOTE ON NOTATION...IT IS CUSTOMARY IN STATISTICS TO USE  $X$  TO SPECIFY AN ARBITRARY (SINGLE) RANDOM VARIABLE, AND TO USE  $X_1, X_2, \dots$  OR  $X, Y, Z, \dots$  FOR SEQUENCES OR SETS OF RANDOM VARIABLES. WHEN ONE VARIABLE DEPENDS ON ANOTHER IN SOME WAY (SUCH AS INCOME DEPENDING ON AGE OR EDUCATION), IT IS CUSTOMARY TO USE  $Y$  FOR THE "DEPENDENT" VARIABLE AND  $X$ 's FOR THE EXPLANATORY ("INDEPENDENT") VARIABLES. IN SAMPLE SURVEY, IT IS CUSTOMARY TO USE  $Y$  FOR AN ARBITRARY RANDOM VARIABLE AND FOR A DEPENDENT VARIABLE.

WHILE IT DOES NOT MATTER THEORETICALLY WHAT SYMBOL IS USED TO REPRESENT A RANDOM VARIABLE, WE SHALL DEFER TO CONVENTION AND HENCEFORTH USE  $Y$ , INSTEAD OF  $X$ , TO DENOTE AN ARBITRARY RANDOM VARIABLE. (THE CONVENTIONAL NOTATION IS GENERALLY GOOD AND HELPFUL, AND THERE IS NO GOOD REASON TO DEPART FROM IT.) WHETHER  $X$  OR  $Y$  IS USED IS NOT RELEVANT – ALL THAT MATTERS IS HOW THE RANDOM VARIABLE IS DEFINED. THE CONVENTION OF USING  $Y$  TO DENOTE A DEPENDENT VARIABLE (E.G., IN A MULTIPLE REGRESSION EQUATION) IS WELL ESTABLISHED, HOWEVER, AND THERE IS NO REASON FOR DEPARTING FROM IT.

THE NOTATION IN THIS COURSE CLOSELY FOLLOWS THE NOTATION IN *SAMPLING TECHNIQUES*, 3<sup>rd</sup> EDITION BY WILLIAM G. COCHRAN (WILEY, 1977) OR *ELEMENTARY SURVEY SAMPLING*, 2<sup>nd</sup> EDITION, BY RICHARD L. SCHEAFFER, WILLIAM MENDENHALL AND LYMAN OTT (DUXBURY PRESS, 1979). (COCHRAN HAS MORE FORMULAS, AND IS A MATHEMATICS TEXT. SCHEAFFER IS MUCH SIMPLER, AND PRESENTS JUST THE BASIC RESULTS, WITHOUT PROOF.)

SIMPLE RANDOM SAMPLING  
(FROM A FINITE POPULATION)

POPULATION SIZE =  $N$ , SAMPLE SIZE =  $n$

$$\text{POPULATION MEAN} = \mu_Y = \bar{Y} = \frac{y_1 + y_2 + \dots + y_N}{N} = \sum_{i=1}^N y_i$$

$$\text{POPULATION VARIANCE} = \sigma_Y^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu_Y)^2 = \frac{1}{N} \sum_{i=1}^N y_i^2 - \mu_Y^2$$

$$\text{ALSO } S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu_Y)^2$$

POPULATION TOTAL =  $\tau_Y (= Y) = N\mu_Y$

SAMPLE MEAN (POINT ESTIMATOR OF POP. MEAN) =  $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$  (A RANDOM VARIABLE)

OR  $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$  (A NUMBER, CALCULATED FROM A PARTICULAR SAMPLE).

SAMPLE VARIANCE =  $\hat{\sigma}_Y^2 = s_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)$  (A RANDOM VARIABLE)

OR  $\hat{\sigma}_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$  (A NUMBER, CALCULATED FROM AN ACTUAL SAMPLE).

SAMPLING WITH REPLACEMENT

SAMPLING WITHOUT REPLACEMENT

$$E(\bar{Y}) = \mu_Y$$

$$E(\bar{Y}) = \mu_Y$$

$$\text{(TRUE) VARIANCE OF } \bar{Y} = V(\bar{Y}) = \sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$$

$$\sigma_{\bar{Y}}^2 = \frac{N-n}{N} \frac{S_Y^2}{n} = \frac{N-n}{N-1} \frac{\sigma_Y^2}{n}$$

$$E(s_Y^2) = \sigma_Y^2$$

$$E(s_Y^2) = S_Y^2$$

$$\text{ESTIMATED VARIANCE OF } \bar{Y} = v(\bar{Y}) = \hat{\sigma}_{\bar{Y}}^2 = \frac{s_Y^2}{n}$$

$$\hat{\sigma}_{\bar{Y}}^2 = \frac{N-n}{N} \frac{s_Y^2}{n}$$

ESTIMATED STANDARD ERROR OF  $\bar{Y} = SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}}$

THE FACTOR  $(N-n)/N = 1 - n/N$  IS CALLED THE FINITE POPULATION CORRECTION (fpc).

IT SHOWS HOW MUCH LOWER THE VARIANCE OF THE ESTIMATE IS WITH SAMPLING WITHOUT REPLACEMENT, COMPARED TO SAMPLING WITH REPLACEMENT.

(NOTE: HERE, AND IN THE SAMPLING METHODS THAT FOLLOW, IF THE TOTAL POPULATION SIZE IS NOT KNOWN, THEN REPLACE THE  $fpc$  BY 1.)

95% CONFIDENCE INTERVAL FOR THE POPULATION MEAN:  $\bar{Y} \pm 2\hat{\sigma}_{\bar{Y}}$ .

AS NOTED EARLIER, THE TERM  $2SE(\bar{Y})$  IS CALLED THE “BOUND ON THE ERROR OF ESTIMATION OF  $\bar{Y}$ .”

SIMPLE RANDOM SAMPLING WITH BINARY DATA  
(SAME FORMULAS APPLY, BUT THEY SIMPLIFY)

EACH  $y_i = 0$  OR  $1$

$$\text{POPULATION MEAN} = \bar{Y}_N = \frac{Y_1 + Y_2 + \dots + Y_N}{N} = \frac{\text{Number of 1's in population}}{N} = P_Y$$

$$\text{POPULATION VARIANCE} = \sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu_Y)^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \mu_Y^2 = P_Y - P_Y^2 = P_Y(1 - P_Y)$$

$$\text{ALSO } S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \mu_Y)^2 = \frac{N}{N-1} P_Y(1 - P_Y)$$

$$\text{SAMPLE MEAN} = \bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n} = \frac{\text{Number of 1's in sample}}{n} = p_Y = \hat{P}_Y$$

$$\text{SAMPLE VARIANCE} = \hat{\sigma}_Y^2 = s_Y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} = \frac{np_Y(1 - p_Y)}{n-1}$$

SAMPLING WITH REPLACEMENT

SAMPLING WITHOUT REPLACEMENT

(DROP THE SUBSCRIPT Y.)

$$E(p) = P$$

$$E(p) = P$$

$$\text{(TRUE) VARIANCE OF } p = \text{var}(p) = \sigma_p^2 = \frac{P(1-P)}{n}$$

$$\sigma_p^2 = \frac{N-n}{N} \frac{S^2}{n} = \frac{N-n}{N-1} \frac{P(1-P)}{n}$$

$$E(s^2) = \sigma^2 = P(1-P)$$

$$E(s^2) = S^2 = \frac{N}{N-1} P(1-P)$$

$$\text{ESTIMATED VARIANCE OF } p = \hat{\sigma}_p^2 = \frac{s^2}{n} = \frac{p(1-p)}{n-1}$$

$$\hat{\sigma}_p^2 = \frac{N-n}{N} \frac{s^2}{n} = \frac{N-n}{N} \frac{p(1-p)}{n-1}$$

ESTIMATED STANDARD ERROR OF  $p = \hat{\sigma}_p$

95% CONFIDENCE INTERVAL FOR  $P$ :  $p \pm 2\hat{\sigma}_p$

## SELECTING RANDOM SAMPLES

RANDOMNESS IS A PROPERTY OF THE PROCESS GENERATING THE “RANDOM” NUMBERS – IT CANNOT BE PROVED FROM THE NUMBERS THEMSELVES.

### TABLES OF RANDOM NUMBERS

ENTER TABLE RANDOMLY

DOCUMENT STARTING POINT AND RECORD SELECTED NUMBERS

USE A TABLE FROM AN ACCESSIBLE (IN-PRINT) SOURCE (SO THAT YOUR SELECTION CAN BE DOCUMENTED AND VERIFIED BY OTHERS, BY ACCESSING THAT SOURCE AND SEEING THE NUMBERS YOU SELECTED).

### SYSTEMATIC SAMPLING (WILL BE TREATED IN GREATER DETAIL LATER)

SELECT EVERY  $k$ -th ITEM FROM A LIST OF THE SAMPLE UNITS (FRAME).

APPROPRIATE IF LIST IS IN RANDOM ORDER (E.G., PREPARED ARBITRARILY), BUT IT OFTEN PRODUCES INCREASES IN PRECISION IF THE LIST IS NOT IN RANDOM ORDER (E.G., A TREND IS PRESENT).

IF LIST IS NOT IN RANDOM ORDER, THEN THE SAMPLE IS NOT RANDOM, AND RESULTS MAY BE BIASED. THE GREATEST DANGER IS IF THERE IS SOME SORT OF PERIODICITY IN THE LIST.

EVERY  $k$ -th UNIT IS SELECTED FROM THE LIST,  $k=N/n$ .

IF IT IS SUSPECTED THAT THE LIST IS NOT IN RANDOM ORDER, THEN SELECT A NUMBER OF SYSTEMATIC SAMPLES (FOR EXAMPLE, TEN SYSTEMATIC SAMPLES, EACH WITH INTERVAL  $10k$ , AND EACH STARTING FROM A NEW RANDOM STARTING POINT).

COMPUTER-GENERATED (“PSEUDORANDOM”) NUMBERS. GENERATED BY MATHEMATICAL FORMULAS, STARTING WITH A “SEED”: REPRODUCIBLE, DOCUMENTABLE.

MATHEMATICAL / STATISTICAL SOFTWARE PACKAGES (E.G., *PROC SURVEYSELECT* IN SAS).

RANDOM NUMBERS GENERATED BY A HAND CALCULATOR ARE GENERALLY NOT REPRODUCIBLE. OK FOR “PERSONAL USE,” BUT NOT FOR PAID WORK FOR A CLIENT.

### REASON FOR DOCUMENTATION:

- REVIEW OF WORK (TO ENSURE / ESTABLISH CORRECTNESS OF SAMPLING PROCEDURES)
- LEGAL TESTIMONY (TO PROVE CORRECTNESS IN A COURT OF LAW)

ESTIMATION OF SAMPLE SIZE  
SIMPLE RANDOM SAMPLING WITH REPLACEMENT (“SRSWR”)

A RECOMMENDED SAMPLE SIZE MAY BE DETERMINED BY SPECIFYING A LIMIT ON THE STANDARD ERROR OF THE ESTIMATE (OR THE SIZE OF A 95% CONFIDENCE INTERVAL), AND SOLVING FOR  $n$ .

EXAMPLE:

SUPPOSE THAT WE WANT A 95% CONFIDENCE INTERVAL FOR THE POPULATION MEAN TO BE OF SIZE  $\bar{Y} \pm E$ .

THEN, SINCE THE FORMULA FOR A 95% CONFIDENCE INTERVAL (IN SRSWR) IS

$\bar{Y} \pm 2\sigma_{\bar{x}} = \bar{Y} \pm 2\frac{\sigma_Y}{\sqrt{n}}$ , WE SET  $E = 2\frac{\sigma_Y}{\sqrt{n}}$  AND SOLVE FOR  $n$ :

$$n = \left( \frac{2\sigma_Y}{E} \right)^2$$

TO USE THIS FORMULA, WE TO SPECIFY  $E$ , AND WE NEED AN ESTIMATE OF THE STANDARD DEVIATION,  $\sigma_Y$ .

THIS IS OBTAINED FROM PREVIOUS SURVEYS, REPORTS, OR JUDGMENT (E.G., IF WE JUDGE THAT MOST OF THE POPULATION COVERS A RANGE OF 200,000, THEN WE COULD ESTIMATE  $\sigma_Y = 200,000/4 = 50,000$ ).

FOR EXAMPLE, IF  $\sigma_Y = 50,000$  AND  $E=5,000$ , THEN  $n = 400$ .

THE PRECEDING METHOD FOR DETERMINING SAMPLE SIZE DOES NOT TAKE COST (BUDGET RESTRICTIONS) INTO ACCOUNT. IT IS APPROPRIATE FOR SRSWR OR FOR SRSWOR IF  $N$  IS LARGE.

NOTE THAT IN DETERMINING SAMPLE SIZES, WE USE THE FORMULAS FOR THE POPULATION (TRUE) VALUES OF THE VARIANCE OR STANDARD ERROR OF THE ESTIMATE – WE DO NOT USE THE SAMPLE FORMULAS SINCE WE ARE NOT USING SAMPLE DATA (WE DO NOT YET HAVE A SAMPLE!).

GENERAL NOTE ON SAMPLE SIZE DETERMINATION FOR DESCRIPTIVE SURVEYS (NOT JUST SIMPLE RANDOM SAMPLING)

A FREE COMPUTER PROGRAM FOR DETERMINING SAMPLE SIZES, BOTH FOR SIMPLE RANDOM SAMPLING AND FOR MORE COMPLEX DESIGNS, IS AVAILABLE FROM BRIXTON HEALTH (A PUBLIC HEALTH AND EPIDEMIOLOGY CONSULTANCY IN LIVERPOOL, ENGLAND, UK) AT <http://www.brixtonhealth.com/samplexs.html> .

THE COMPUTER PROGRAM USED BY THE AUTHOR TO DETERMINE SAMPLE SIZES FOR SURVEYS IS POSTED AT <http://www.foundationwebsite.org/JGCSampleSizeProgram.mdb> (A MICROSOFT ACCESS PROGRAM). IT IS DESIGNED PRIMARILY FOR USE IN DETERMINING SAMPLE SIZES FOR ANALYTICAL SURVEYS. FOR DESCRIPTIVE SURVEYS, THE USUAL APPROACH TO

SAMPLE SIZE DETERMINATION IS TO SPECIFY A DESIRED LEVEL OF PRECISION (FOR AN ESTIMATE OF INTEREST) AND TO DETERMINE THE SAMPLE SIZE THAT PRODUCES THAT LEVEL OF PRECISION. FOR ANALYTICAL SURVEYS, THE USUAL APPROACH IS TO SPECIFY A DESIRED LEVEL OF POWER FOR A SPECIFIED TEST OF HYPOTHESIS (E.G., ABOUT THE SIZE OF A “DOUBLE DIFFERENCE” ESTIMATE), AND TO DETERMINE THE SAMPLE SIZE THAT PRODUCES THAT LEVEL OF POWER.

THERE ARE MANY OTHER SOURCES OF INFORMATION ABOUT SAMPLE SURVEY DESIGN AND SAMPLING ON THE INTERNET WORLD WIDE WEB.

ESTIMATION OF SAMPLE SIZE  
SIMPLE RANDOM SAMPLING WITH REPLACEMENT ("SRSWR")

IN SAMPLING FOR PROPORTIONS,  $\sigma = \sqrt{P(1-P)}$  (DROPPING THE SUBSCRIPT X), SO

$$n = \left(\frac{2}{E}\right)^2 P(1-P).$$

TO USE THIS FORMULA, WE NEED TO SPECIFY THE PROPORTION,  $P$ .

SINCE  $P(1-P)$  ASSUMES ITS MAXIMUM VALUE FOR  $P=.5$ , THE MAXIMUM SIZE FOR A 95%

CONFIDENCE INTERVAL,  $p \pm 2 \frac{\sqrt{P(1-P)}}{\sqrt{n}}$ , IS  $p \pm \frac{2(.5)}{\sqrt{n}}$ , AND SO, SETTING

$E = \frac{2(.5)}{\sqrt{n}} = \frac{1}{\sqrt{n}}$  AND SOLVING FOR  $n$ , WE OBTAIN THE REQUIRED SAMPLE SIZE AS:

$$n = \frac{1}{E^2}.$$

FOR EXAMPLE, IF  $E=.05$ , THEN  $n = 400$ . IF  $E = .03$  THEN  $n = 1,111$ . (NOTE: MANY TELEVISION OPINION POLLS HAVE  $n = 1,000$ , AND HAVE AN ERROR OF ESTIMATION OF ABOUT .03.)

ESTIMATION OF THE POPULATION TOTAL  
(SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT)

THE POPULATION MEAN IS  $\mu_Y = Y / N$ .

THE POPULATION TOTAL IS  $\tau_Y = N\mu_Y$ .

AN ESTIMATOR OF THE POPULATION TOTAL IS

$$\hat{\tau}_Y = N\bar{Y}.$$

THE ESTIMATED VARIANCE OF  $\hat{\tau}_Y$  IS

$$v(\hat{\tau}) = \sigma_{\hat{\tau}}^2 = N^2 \sigma_{\bar{y}}^2 = N^2 \frac{N-n}{N} \frac{s_Y^2}{n}.$$

THE "BOUND ON THE ERROR OF ESTIMATION" IS  $2\sigma_{\hat{\tau}}$ .

FOR SIMPLE RANDOM SAMPLING WITH REPLACEMENT, SIMPLY DROP THE FINITE POPULATION CORRECTION (*fpc*),  $(N - n) / N = 1 - n / N$ .

SAMPLING VARIANCE OF OTHER STATISTICS (SRSWR)  
(MEDIAN, PERCENTILES, STANDARD DEVIATION, COEFFICIENT OF VARIATION)

IN THE PRECEDING, WE HAVE GIVEN FORMULAS FOR THE VARIANCES (TRUE AND ESTIMATED) OF THE SAMPLE MEAN,  $\bar{Y}$ , AND THE ESTIMATED POPULATION TOTAL,  $\hat{Y}$ .

HERE ARE STANDARD ERRORS FOR SOME OTHER STATISTICS:

IF THE PARENT POPULATION IS NORMAL, THE STANDARD ERROR OF THE SAMPLE MEDIAN IS  $1.25 \frac{\sigma}{\sqrt{n}}$

IF THE PARENT POPULATION IS NORMAL, THE STANDARD ERROR OF THE SAMPLE COEFFICIENT OF VARIATION IS  $\frac{CV}{\sqrt{2n}}$ , WHERE CV DENOTES THE POPULATION COEFFICIENT OF VARIATION.

IF THE PARENT POPULATION IS NORMAL, THE STANDARD ERROR OF THE SAMPLE STANDARD DEVIATION IS  $\frac{\sigma}{\sqrt{2n}}$ .

## SUMMARY OF MAIN RESULTS

THE GOAL IS ESTIMATION OF FINITE-POPULATION PARAMETERS: MEAN ( $\mu_y$  OR  $\bar{Y}_N$ ), TOTAL ( $\tau_y$  OR  $Y$ ). THE FOLLOWING ARE THE MAIN FORMULAS INVOLVED:

FORMULAS FOR ESTIMATORS OF THE POPULATION PARAMETERS: SAMPLE MEAN ( $\hat{\mu}_y$  OR  $\bar{y}$ ), ESTIMATED POPULATION TOTAL ( $\hat{\tau}_y$  OR  $\hat{Y}$ ).

(FROM THIS POINT ON, WE WILL USE  $\tau_y$  RATHER THAN  $Y$ , TO DENOTE THE POPULATION TOTAL, TO AVOID CONFUSION WITH THE SYMBOL  $Y$  USED TO DENOTE A RANDOM VARIABLE.)

FORMULAS FOR THE TRUE VALUES OF THE VARIANCES OR STANDARD DEVIATIONS (STANDARD ERRORS) OF THE ESTIMATORS. (THESE ARE USED IN THE ESTIMATION OF THE SAMPLE SIZES REQUIRED TO ACHIEVE SPECIFIED LEVELS OF PRECISION.)

FORMULAS FOR ESTIMATING THE STANDARD ERRORS OF THE SAMPLE ESTIMATES: STANDARD ERROR OF  $\hat{\mu}_y$ , STANDARD ERROR OF  $\hat{\tau}_y$ . (THESE ARE USED TO INDICATE THE LEVEL OF PRECISION OF THE SAMPLE ESTIMATES.)

BOUND ON THE ERROR OF ESTIMATION (TWICE THE ESTIMATED STANDARD ERROR OF THE ESTIMATE)

95% CONFIDENCE INTERVAL: THE ESTIMATE  $\pm$  TWICE THE ESTIMATED STANDARD ERROR OF THE ESTIMATE

THE ABOVE SUMMARY WILL APPLY NOT JUST TO SIMPLE RANDOM SAMPLING (WITH OR WITHOUT REPLACEMENT), BUT TO MANY OF THE OTHER TYPES OF SAMPLING TO BE DISCUSSED.

SO, IN GENERAL, ALL WE REALLY NEED TO KNOW ABOUT EACH SURVEY DESIGN IS WHAT IS THE FORMULA FOR A GOOD ESTIMATE (OF A SPECIFIED POPULATION PARAMETER, SUCH AS THE POPULATION MEAN OR TOTAL), AND THE FORMULA FOR THE STANDARD ERROR OF THE ESTIMATE. FROM THIS WE CAN STATE THE "BOUND ON THE ERROR OF ESTIMATION" (TWICE THE STANDARD ERROR OF THE ESTIMATE) AND A 95% CONFIDENCE INTERVAL.

## USING SUPPLEMENTARY (AUXILIARY) INFORMATION TO ASSIST SURVEY DESIGN

IF NOTHING IS KNOWN ABOUT THE POPULATION IN ADVANCE OF SAMPLING (EXCEPT FOR A LIST OF SAMPLING UNITS), THEN SIMPLE RANDOM SAMPLING IS ALL THAT CAN BE DONE.

INFORMATION KNOWN ABOUT THE POPULATION PRIOR TO SAMPLING CAN ENABLE THE CONSTRUCTION OF AN IMPROVED (MORE EFFICIENT) SAMPLE DESIGN (HIGHER PRECISION FOR THE SAME SAMPLING EFFORT (SMALLER SAMPLE, LOWER COST), OR ACHIEVEMENT OF A SPECIFIED LEVEL OF PRECISION FOR LESS SAMPLING EFFORT).

THIS INFORMATION MAY BE ABOUT THE PRIMARY VARIABLE(S) OF INTEREST, OR ABOUT VARIABLES RELATED TO THEM (E.G., IN AN INCOME SURVEY, MAY KNOW LAST YEAR'S INCOME, OR VARIABLES RELATED TO INCOME, SUCH AS QUALITY OF NEIGHBORHOOD, OR AGE, OR EDUCATION).

THIS INFORMATION MAY BE QUALITATIVE OR QUANTITATIVE, BUT IT MUST BE DETERMINABLE FOR EACH SAMPLE UNIT IN THE FRAME.

### QUALITATIVE (NOMINAL):

RICH OR POOR (NEIGHBORHOODS, SOIL REGIONS)  
ADVANCED OR RETARDED (ECONOMIC REGIONS)  
MORE OR LESS DENSELY POPULATED  
URBAN OR RURAL  
RESIDENCE, ETHNICITY

### QUANTITATIVE (ORDINAL, INTERVAL):

INCOME DATA WHEN SURVEYING EXPENDITURE DATA  
HEIGHT WHEN ESTIMATING WEIGHT  
AGE WHEN ESTIMATING BLOOD PRESSURE  
SEX WHEN ESTIMATING MARKET PREFERENCES  
EDUCATIONAL LEVEL OR NATIONALITY WHEN SURVEYING ATTITUDES  
POLITICAL AFFILIATION WHEN SURVEYING POLITICAL OPINIONS  
SAMPLING COSTS (URBAN, RURAL) WHEN SURVEYING SCHOOLS

THIS INFORMATION WILL ENABLE US TO CONSTRUCT A VARIETY OF SURVEY DESIGNS THAT ARE MORE EFFICIENT THAN SIMPLE RANDOM SAMPLING:

STRATIFIED SAMPLING  
CLUSTER SAMPLING  
MULTISTAGE SAMPLING  
DOUBLE SAMPLING (TWO-PHASE SAMPLING)

## STRATIFIED RANDOM SAMPLING

DEFINITION: A STRATIFIED RANDOM SAMPLE IS ONE OBTAINED BY SEPARATING THE SAMPLE UNITS (POPULATION ELEMENTS) INTO NONOVERLAPPING GROUPS, CALLED STRATA, AND SELECTING A SIMPLE RANDOM SAMPLE FROM EACH STRATUM.

THE STRATA ARE DEFINED ON THE BASIS OF AUXILIARY INFORMATION.

### REASONS (INDICATIONS) FOR STRATIFICATION:

1. POPULATION ELEMENTS ARE MORE HOMOGENEOUS (LESS VARIABLE) WITHIN STRATA THAN IN THE GENERAL POPULATION (WITH RESPECT TO THE VARIABLES OF INTEREST).
2. COST OF SAMPLING MAY BE LOWER IN SOME STRATA (ADMINISTRATIVE CONVENIENCE).
3. ESTIMATES OF POPULATION PARAMETERS CAN BE READILY OBTAINED FOR EACH STRATUM.

### EXAMPLE 1:

IN A COUNTY CONTAINING FIVE VOTING DISTRICTS, WE WISH TO ESTIMATE THE PROPORTION OF REGISTERED VOTERS WHO FAVOR A PARTICULAR ELECTION CANDIDATE. WE WANT AN OVERALL ESTIMATE FOR THE COUNTY AND ESTIMATES FOR EACH VOTING DISTRICT. A LIST OF REGISTERED VOTERS IS AVAILABLE. A SAMPLE OF 100 VOTERS IS SELECTED FROM EACH DISTRICT.

THE DATA ARE COLLECTED AND ANALYZED (USING FORMULAS TO BE PRESENTED), AND THE PROPORTIONS ARE ESTIMATED FOR THE COUNTY AND EACH DISTRICT.

### EXAMPLE 2:

IT IS DESIRED TO ESTIMATE THE TOTAL NUMBER OF COMPUTERS IN ALL SCHOOLS IN ZAMBIA. IT IS KNOWN THAT MOST RURAL SCHOOLS AND MOST SMALL SCHOOLS DO NOT HAVE COMPUTERS. ALSO, IT IS KNOWN THAT MANY PRIVATE SCHOOLS DO HAVE COMPUTERS. IT IS ALSO KNOWN THAT MOST LARGE SCHOOLS ARE IN URBAN AREAS. RECENT DATA ON SCHOOL SIZE AND OWNERSHIP ARE KNOWN FROM LAST YEAR'S ANNUAL SCHOOL CENSUS.

IN THIS CASE, USE OF A STRATIFIED SAMPLE DESIGN, WITH STRATIFICATION BY SCHOOL SIZE, URBAN/RURAL STATUS, AND OWNERSHIP STATUS, WOULD PROBABLY BE A GOOD CHOICE. (NOTE: FEW PRACTICAL SAMPLE SURVEYS COLLECT DATA ON ONLY ONE VARIABLE. WHILE THIS DESIGN MAY BE GOOD FOR ESTIMATING THE TOTAL NUMBER OF COMPUTERS, IT WOULD NOT NECESSARILY BE THE BEST FOR ESTIMATING SOME OTHER PARAMETER, SUCH AS THE TYPE OF WATER SOURCE FOR THE SCHOOL (NONE, WELL, PUMP, MUNICIPAL WATER). ALL OF THE SURVEY OBJECTIVES AND CONSTRAINTS MUST BE CONSIDERED TOGETHER IN DESIGNING THE SURVEY.

STRATIFIED RANDOM SAMPLING  
NOTATION / ESTIMATION FORMULAS

NUMBER OF STRATA:  $L$

WILL USE THE SUFFIX / SUBSCRIPT  $h$  TO DENOTE AN ARBITRARY STRATUM, AND SUFFIX / SUBSCRIPT  $i$  TO DENOTE AN ARBITRARY UNIT WITHIN A STRATUM.

NUMBER OF UNITS (IN STRATUM  $h$ ) = "SIZE" OF STRATUM  $h$ :  $N_h$

NUMBER OF UNITS IN SAMPLE (IN STRATUM  $h$ ):  $n_h$

VALUE OF THE  $i$ -th UNIT:  $y_{hi}$

STRATUM WEIGHT:  $W_h = \frac{N_h}{N}$

SAMPLING FRACTION:  $f_h = \frac{n_h}{N_h}$

TRUE MEAN:  $\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$

SAMPLE MEAN:  $\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$  ( $= \hat{Y}_h$ )

TRUE VARIANCE:  $S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1}$

(WILL USE  $S^2$  RATHER THAN  $\sigma^2$ , SINCE (1) IT IS CUSTOMARY IN SAMPLE SURVEY; AND (2) THE FORMULAS ARE A LITTLE SIMPLER.)

SAMPLE VARIANCE:  $s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$

TOTAL POPULATION SIZE:  $N = N_1 + N_2 + \dots + N_k$

TOTAL SAMPLE SIZE:  $n = n_1 + n_2 + \dots + n_k$

THE POPULATION TOTAL, MEAN AND VARIANCE ARE DEFINED THE SAME AS BEFORE.

STRATIFIED RANDOM SAMPLING  
ESTIMATION FORMULAS

$$E(\bar{y}_h) = \bar{Y}_h$$

$$V(\bar{y}_h) = \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

$$E(s_h^2) = S_h^2$$

ESTIMATE OF THE POPULATION MEAN (st STANDS FOR “STRATIFIED”):

$$\bar{y}_{st} = \frac{\sum_{h=1}^L N_h \bar{y}_h}{N} = \sum_{h=1}^L W_h \bar{y}_h$$

THIS IS NOT, IN GENERAL, THE SAME AS THE SAMPLE MEAN, WHICH IS:

$$\bar{y} = \frac{\sum_{h=1}^L n_h \bar{y}_h}{n}$$

THE ESTIMATE  $\bar{y}_{st}$  IS EQUAL TO  $\bar{y}$  ONLY IF:

$$\frac{n_h}{n} = \frac{N_h}{N} \text{ OR } \frac{n_h}{N_h} = \frac{n}{N} \text{ OR } f_h = f$$

I.E., IF THE STRATIFICATION INVOLVES A PROPORTIONAL ALLOCATION OF THE SAMPLE TO THE STRATA (SAMPLE SIZES PROPORTIONAL TO STRATUM SIZES), THEN THE SAMPLE IS SAID TO BE “SELF-WEIGHTING.”

STRATIFIED RANDOM SAMPLING  
ESTIMATION FORMULAS / MAJOR RESULTS

MAJOR RESULTS:

THE ESTIMATE  $\bar{y}_{st}$  IS AN UNBIASED ESTIMATE OF THE POPULATION MEAN,  $\bar{Y}$ , I.E.,

$$E(\bar{y}_{st}) = \bar{Y}$$

THE TRUE VARIANCE OF  $\bar{y}_{st}$  IS:

$$V(\bar{y}_{st}) = \frac{1}{N^2} = \sum_{h=1}^L \frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h}$$

AN UNBIASED ESTIMATE OF THE VARIANCE OF  $\bar{y}_{st}$  IS:

$$v(\bar{y}_{st}) = s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{s_h^2}{n_h} = \sum_{h=1}^L \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^L \frac{W_h s_h^2}{N}$$

APPROXIMATE 95% CONFIDENCE LIMITS ARE HENCE AS FOLLOWS:

FOR THE POPULATION MEAN:

$$\bar{y}_{st} \pm 2s(\bar{y}_{st})$$

FOR THE POPULATION TOTAL:

$$N\bar{y}_{st} \pm 2Ns(\bar{y}_{st})$$

STRATIFIED RANDOM SAMPLING  
ALLOCATION OF SAMPLE TO STRATA

PROPORTIONAL ALLOCATION (SELF-WEIGHTING):

$$n_h = \frac{nN_h}{N}$$

OPTIMAL (MINIMUM VARIANCE), IF COST OF SAMPLING IS THE SAME IN ALL STRATA, BUT THE STRATUM VARIANCES MAY DIFFER:

$$n_h = \frac{nN_h S_h}{\sum_{h=1}^L N_h S_h}$$

THIS IS CALLED THE "NEYMAN" ALLOCATION.

OPTIMAL (MINIMUM VARIANCE), WITH SAMPLING COST FUNCTION ("LINEAR" COST FUNCTION):

$$\text{COST} = C = c_0 + \sum_{h=1}^L c_h n_h$$

$$n_h = n \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^L (W_h S_h / \sqrt{c_h})} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^L (N_h S_h / \sqrt{c_h})}$$

GUIDELINES: TAKE A LARGER SAMPLE IN A STRATUM IF

1. THE STRATUM IS LARGER
2. THE STRATUM IS MORE VARIABLE INTERNALLY
3. SAMPLING IS CHEAPER IN THE STRATUM

THE MORE WE KNOW ABOUT THE VARIABLE OF INTEREST ( $y$ ), THE BETTER JOB WE CAN DO OF STRATIFICATION. IN SOME CASES, IT MAY BE WORTHWHILE TO CONDUCT A PRELIMINARY SAMPLE TO OBTAIN SOME INFORMATION THAT WOULD ASSIST STRATIFICATION.

IN SAMPLING FOR PROPORTIONS, THE FORMULAS BECOME A LITTLE SIMPLER – SEE A TEXT ON SAMPLE SURVEY FOR THE FORMULAS IN THAT CASE.

SYSTEMATIC SAMPLING: WHEN THE UNITS ARE ARRANGED IN DESCENDING OR ASCENDING ORDER, THEN SYSTEMATIC SAMPLING HAS A SIMILAR EFFECT AS STRATIFICATION. (WE ADDRESS SYSTEMATIC SAMPLING LATER.)

## GAINS IN PRECISION FROM STRATIFICATION

THE GAIN IN PRECISION FROM STRATIFICATION OVER SIMPLE RANDOM SAMPLING DEPENDS MAINLY (APART FROM COST CONSIDERATIONS) ON HOW MUCH LESS THE VARIATION WITHIN STRATA IS COMPARED TO THE VARIATION OVER THE GENERAL POPULATION.

### A SIMPLE MODEL.

CONSIDER THE CASE IN WHICH THE OBSERVED RANDOM VARIABLE,  $X$ , IS THE SUM OF TWO INDEPENDENT RANDOM VARIABLES, A "STRATUM" COMPONENT,  $X_S$ , AND A "WITHIN-STRATUM" COMPONENT,  $X_W$ :

$$X = X_S + X_W$$

SUPPOSE THAT  $E(X_S) = \mu_S$ ,  $E(X_W) = 0$ ,  $V(X_S) = \sigma_S^2$  AND  $V(X_W) = \sigma_W^2$ .

THEN  $V(X) = V(X_S) + V(X_W)$ , OR  $\sigma^2 = \sigma_S^2 + \sigma_W^2$ .

SUPPOSE THAT THERE ARE  $L$  STRATA AND THAT AN EQUAL NUMBER OF UNITS,  $n_s$ , IS SELECTED FROM EACH STRATUM. THE TOTAL SAMPLE SIZE IS  $n = Ln_s$ .

THE ESTIMATOR OF THE POPULATION MEAN IS:

$$\hat{\mu}_X = \frac{\sum_{h=1}^L \hat{\mu}_h}{L}$$

WHERE

$$\hat{\mu}_h = \frac{\sum_{i=1}^{n_s} x_{hi}}{n_s}.$$

ITS VARIANCE IS:

$$v(\hat{\mu}_X) = \frac{\sigma_W^2}{Ln_s}.$$

THE VARIANCE OF A SIMPLE RANDOM SAMPLE (WITH REPLACEMENT) OF SIZE  $n = Ln_s$  IS:

$$v(\mu_X) = \frac{\sigma^2}{Ln_s}.$$

SO THE RATIO OF THE VARIANCES OF STRATIFIED RANDOM SAMPLING TO SIMPLE RANDOM SAMPLING IN THIS CASE IS  $\sigma_W^2/\sigma^2$ .

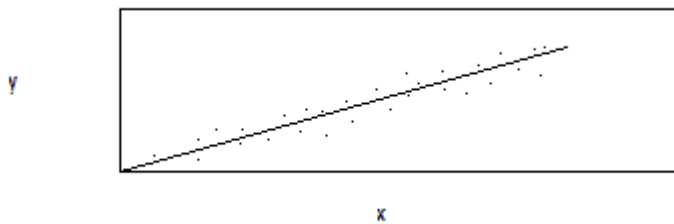
## ALTERNATIVE ESTIMATION TECHNIQUES: RATIO AND REGRESSION ESTIMATORS

### RATIO ESTIMATORS (IN SIMPLE RANDOM SAMPLING)

VARIABLE OF PRIMARY INTEREST (“RESPONSE” VARIABLE):  $Y$

SUPPOSE THAT THERE IS ANOTHER (“AUXILIARY”) VARIATE,  $X$ , CORRELATED WITH  $Y$ , AND KNOWN FOR EACH UNIT OF THE SAMPLE, AND FOR WHICH WE KNOW THE POPULATION TOTAL,  $\tau_X$ .

SUPPOSE FURTHER THAT THE RELATIONSHIP BETWEEN THE RESPONSE VARIABLE AND THE AUXILIARY VARIABLE IS LINEAR THROUGH THE ORIGIN:



THEN A RATIO ESTIMATOR IS A GOOD CHOICE. IT IS, IN FACT, THE BEST CHOICE IF THE VARIANCE OF THE RESPONSE VARIABLE,  $Y$ , ABOUT THE LINE IS PROPORTIONAL TO  $X$ .

THE RATIO ESTIMATE OF  $\tau_Y$ , THE POPULATION TOTAL (FOR  $Y$ ) IS:

$$\hat{\tau}_{YR} = \frac{n\bar{y}}{n\bar{x}} \tau_X = \frac{\bar{y}}{\bar{x}} \tau_X$$

(THE PRECEDING IS SHEAFFER'S NOTATION. IN COCHRAN'S NOTATION, USING  $X$  AND  $Y$  TO DENOTE THE POPULATION TOTALS, THIS FORMULA IS:

$$\hat{Y}_R = \frac{y}{x} X = \frac{\bar{y}}{\bar{x}} X$$

WHERE  $x$  AND  $y$  DENOTE THE SAMPLE TOTALS OF THE  $x_i$  AND  $y_i$ , RESPECTIVELY, AND  $X$  DENOTES THE POPULATION TOTAL FOR THE  $x_i$ . THIS NOTATION IS CONFUSING, SINCE  $X$  REFERS BOTH TO A RANDOM VARIABLE AND A POPULATION TOTAL (OF THE  $x_i$ ), AND  $x$ , WHICH WOULD NORMALLY REFER TO A SPECIFIC VALUE OF THE  $X$  RANDOM VARIABLE, INSTEAD REFERS TO THE SAMPLE TOTAL OF THE  $x_i$ .)

THE RATIO ESTIMATE OF  $\mu_Y$ , THE POPULATION MEAN (FOR  $Y$ ) IS:

$$\hat{\mu}_{YR} = \frac{n\bar{y}}{n\bar{x}} \mu_X = \frac{\bar{y}}{\bar{x}} \mu_X$$

(OR, IN COCHRAN'S NOTATION:

$$\hat{Y}_R = \frac{y}{x} \bar{X} = \frac{\bar{y}}{\bar{x}} \bar{X} ).$$

THE RATIO ESTIMATE IS BIASED, BUT THE BIAS IS NEGLIGIBLE IN LARGE SAMPLES. IT IS CONSISTENT, I.E., ITS AVERAGE TENDS TO THE TRUE VALUE AS THE SAMPLE SIZE INCREASES.

NOTE: FROM THIS POINT ON IN THIS COURSE, WE WILL NOT PRESENT FORMULAS FOR THE TRUE VARIANCES OF ALL OF THE SAMPLE ESTIMATES DISCUSSED. THE TRUE VARIANCE IS NEEDED TO DETERMINE SAMPLE SIZE, BUT IT IS NOT USED IN THE ANALYSIS OF THE SAMPLE DATA. THE FORMULAS FOR THE ESTIMATED VARIANCES (BASED ON THE SAMPLE DATA) WILL ALWAYS BE PRESENTED (SINCE THEY ARE ALWAYS NEEDED IN THE ANALYSIS OF THE SAMPLE DATA), BUT THE TRUE FORMULAS WILL BE PRESENTED ONLY WHEN ADDRESSING THE PROBLEM OF DETERMINING SAMPLE SIZE (IN ADVANCE OF THE SURVEY). STANDARD REFERENCE TEXTS MAY BE CONSULTED FOR THE FORMULAS FOR THE TRUE VARIANCES, IN THOSE CASES IN WHICH THEY ARE NOT PRESENTED HERE.

SOME DISCUSSION OF DETERMINING SAMPLE SIZES WAS PRESENTED EARLIER, IN THE CASE OF SIMPLE RANDOM SAMPLING. DETERMINING SAMPLE SIZES FOR OTHER SAMPLE DESIGNS IS ADDRESSED IN DAY TWO OF THE COURSE.

## RATIO ESTIMATORS IN SIMPLE RANDOM SAMPLING

### SUMMARY OF RESULTS

ESTIMATOR OF THE POPULATION RATIO,  $R$ :

$$r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

ESTIMATED VARIANCE OF  $r$ :

$$v(r) = \frac{N-n}{nN} \frac{1}{\mu_x^2} \frac{\sum_{i=1}^n (y_i - rx_i)^2}{n-1}$$

BOUND ON THE ERROR OF ESTIMATION OF  $r$ :

$$2\sqrt{v(r)} = 2\sqrt{\frac{N-n}{nN} \frac{1}{\mu_x^2} \frac{\sum_{i=1}^n (y_i - rx_i)^2}{n-1}}$$

IF THE POPULATION MEAN FOR  $X$ ,  $\mu_x$ , IS UNKNOWN, USE THE SAMPLE ESTIMATE,  $\bar{X}^2$ , TO APPROXIMATE  $\mu_x^2$ .

RATIO ESTIMATOR OF THE POPULATION TOTAL:

$$\hat{t}_{RY} = r\tau_X$$

ESTIMATED VARIANCE OF  $\hat{t}_{RY}$ :

$$v(\hat{t}_{RY}) = \tau_X^2 v(r).$$

NOTE THAT IT IS NECESSARY TO KNOW  $\tau_X$ , THE POPULATION TOTAL FOR  $X$ , IN ORDER TO ESTIMATE  $\tau_Y$  BY THE RATIO ESTIMATION METHOD.

AS USUAL, AN APPROXIMATE 95% CONFIDENCE INTERVAL FOR A POPULATION PARAMETER IS GIVEN BY:

PARAMETER ESTIMATE PLUS/MINUS 2 (ESTIMATED STANDARD ERROR OF THE PARAMETER ESTIMATE),

WHERE THE ESTIMATED STANDARD ERROR OF THE PARAMETER ESTIMATE IS THE SQUARE ROOT OF ITS ESTIMATED VARIANCE. IN THE PRECEDING CASE:

$$\hat{\tau}_{RY} \pm 2\sqrt{v(\hat{\tau}_{RY})}$$

RATIO ESTIMATOR OF THE POPULATION MEAN:

$$\hat{\mu}_{RY} = r\mu_X$$

ESTIMATED VARIANCE OF  $\hat{\tau}_{RY}$  :

$$v(\hat{\tau}_{RY}) = \mu_X^2 v(r).$$

NOTE THAT IT IS NECESSARY TO KNOW  $\mu_X$ , THE POPULATION MEAN FOR X, IN ORDER TO ESTIMATE  $\mu_Y$  BY THE RATIO ESTIMATION METHOD.

## RATIO ESTIMATORS IN STRATIFIED RANDOM SAMPLING

TWO APPROACHES:

1. SEPARATE RATIO ESTIMATE. CONSTRUCT A SEPARATE RATIO ESTIMATE FOR EACH STRATUM, AND THEN FORM A WEIGHTED AVERAGE OF THESE SEPARATE ESTIMATES AS A SINGLE ESTIMATE OF THE POPULATION RATIO.
2. COMBINED RATIO ESTIMATE. ESTIMATE THE MEANS FOR Y AND X USING THE FORMULA FOR STRATIFIED RANDOM SAMPLING, AND THEN USE THE RATIO OF THESE OVERALL MEANS AS A RATIO ESTIMATOR.

WHICH METHOD IS PREFERRED DEPENDS ON THE NATURE OF THE POPULATION AND THE DESIGN. IF THE RATIO VARIES FROM STRATUM TO STRATUM, THE SEPARATE ESTIMATE IS USUALLY BETTER (MORE PRECISE). IF THE SAMPLE SIZE IS SMALL IN EACH STRATUM, THE COMBINED RATIO ESTIMATE IS USUALLY BETTER.

THE FORMULAS FOR RATIO ESTIMATORS IN STRATIFIED RANDOM SAMPLING ARE SOMEWHAT COMPLICATED. REFER TO COCHRAN, SAMPLING TECHNIQUES FOR THE FORMULAS. THERE ARE TECHNIQUES AVAILABLE TO REDUCE OR REMOVE THE BIAS, AND TO IMPROVE THE VARIANCE OF THE ESTIMATE.

PRODUCT ESTIMATORS:

IF X AND Y TAKE ONLY POSITIVE VALUES AND THE CORRELATION IS NEGATIVE, THEN A RATIO ESTIMATE IS NOT APPROPRIATE, BUT A SIMILAR ESTIMATE, CALLED A PRODUCT ESTIMATOR, IS INDICATED:

$$\hat{\mu}_{yp} = \frac{\bar{x}}{\mu_x} \bar{y}, \quad \hat{t}_{yp} = N \frac{\bar{x}}{\mu_x} \bar{y},$$

(OR, IN COCHRAN'S NOTATION:

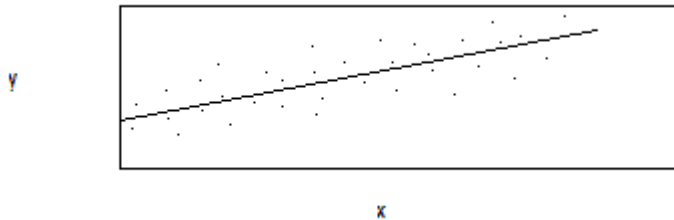
$$\hat{Y}_p = \frac{\bar{x}}{\bar{X}} \bar{y}, \quad \hat{Y}_p = N \frac{\bar{x}}{\bar{X}} \bar{y}).$$

REGRESSION ESTIMATORS IN SIMPLE RANDOM SAMPLING

VARIABLE OF PRIMARY INTEREST (“RESPONSE” VARIABLE): Y

SUPPOSE THAT THERE IS ANOTHER (“AUXILIARY”) VARIATE, X, CORRELATED WITH Y, AND KNOWN FOR EACH UNIT OF THE SAMPLE, AND FOR WHICH WE KNOW THE POPULATION TOTAL,  $\tau_x$ .

SUPPOSE FURTHER THAT THE RELATIONSHIP BETWEEN THE RESPONSE VARIABLE AND THE AUXILIARY VARIABLE IS LINEAR, *BUT NOT NECESSARILY THROUGH THE ORIGIN*, AS WAS ASSUMED IN THE CASE OF RATIO ESTIMATION:



THEN USE OF A LINEAR REGRESSION ESTIMATOR IS APPROPRIATE.

FOR EXAMPLE, MAY KNOW LAST YEAR’S SCHOOL BUDGET FOR EACH SCHOOL IN THE COUNTRY (FROM AN ANNUAL SCHOOL CENSUS), AND WANT TO OBTAIN A PRELIMINARY ESTIMATE THIS YEAR’S BUDGET FROM A SAMPLE OF SCHOOLS.

(LINEAR) REGRESSION ESTIMATOR OF A POPULATION MEAN,  $\mu_Y$ :

$$\hat{\mu}_{YL} = \bar{y} + b(\mu_x - \bar{x})$$

WHERE

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

ESTIMATED VARIANCE OF  $\hat{\mu}_{YL}$ :

$$v(\hat{\mu}_{YL}) = \frac{N-n}{Nn} \frac{1}{n-2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\hat{\mu}_{YL})}$ .

FOR AN ESTIMATE OF THE POPULATION TOTAL (FOR Y), USE  $\hat{\tau}_{YL} = N\hat{\mu}_{YL}$

CLUSTER SAMPLING  
(SINGLE-STAGE CLUSTER SAMPLING)

A CLUSTER SAMPLE IS A SIMPLE RANDOM SAMPLE IN WHICH EACH SAMPLING UNIT IS A COLLECTION, OR CLUSTER, OF ELEMENTS. IN THIS CASE THE SAMPLING UNITS ARE THE CLUSTERS, AND THE ELEMENTS WITHIN THE UNITS ARE CALLED SUBUNITS.

CLUSTER SAMPLING IS FAR MORE COST-EFFECTIVE THAN SIMPLE RANDOM SAMPLING OR STRATIFIED SAMPLING, IF

1. THE COST OF OBTAINING A FRAME THAT LISTS ALL OF THE POPULATION ELEMENTS IS HIGH; OR
2. THE COST OF OBTAINING OBSERVATIONS INCREASES SUBSTANTIALLY AS THE DISTANCE SEPARATING THE SAMPLE ELEMENTS INCREASES.

EXAMPLES:

IN MANY COUNTRIES THERE ARE NO COMPLETE, UP-TO-DATE LISTS OF HOUSEHOLDS OR FARMS, AND THE COST OF CONSTRUCTING A FRAME OF ALL UNITS OF THE POPULATION WOULD BE PROHIBITIVE. IT IS MUCH CHEAPER TO DIVIDE THE COUNTRY INTO GEOGRAPHIC AREAS (I.E., CONSTRUCT AN AREA FRAME), SELECT A RANDOM SAMPLE OF GEOGRAPHIC AREAS, AND OBSERVE ALL OF THE ELEMENTS (HOUSEHOLDS, FARMS) WITHIN EACH SELECTED AREA.

SUPPOSE THAT IN A CITY, CITY BLOCKS CONTAIN AN AVERAGE OF 20 HOUSEHOLDS EACH. INTERVIEWING ALL HOUSEHOLDS IN A SAMPLE OF 50 BLOCKS WILL COST SUBSTANTIALLY LESS THAN INTERVIEWING A SIMPLE RANDOM SAMPLE OF 1,000 HOUSEHOLDS. ALSO, A FRAME OF CITY BLOCKS MAY BE READILY AVAILABLE, WHEREAS A FRAME OF HOUSEHOLDS MAY NOT.

OTHER EXAMPLES:

- SELECT A SAMPLE OF STUDENTS OR TEACHERS BY MEANS OF A CLUSTER SAMPLE OF SCHOOLS
- SELECT A SAMPLE OF PATIENTS OR DOCTORS BY SELECTING A SAMPLE OF HOSPITALS OR NURSING HOMES
- SELECT A SAMPLE OF PRODUCTION UNITS BY SELECTING A CLUSTER SAMPLE OF BOXES OF UNITS COMING OFF A PRODUCTION LINE

IN CLUSTER SAMPLING, IT IS DESIRED THAT CLUSTERS BE INTERNALLY HETEROGENEOUS (WITH RESPECT TO THE CHARACTERISTICS BEING MEASURED). IF ALL OF THE ELEMENTS WITHIN A CLUSTER ARE VERY SIMILAR THEN RELATIVELY LITTLE INFORMATION IS PROVIDED COMPARED TO A SIMPLE RANDOM SAMPLE OF THE SAME SIZE. THIS IS THE OPPOSITE OF STRATIFIED SAMPLING, WHERE IT IS DESIRED TO HAVE THE STRATA AS INTERNALLY HOMOGENEOUS AS POSSIBLE.

CLUSTER SAMPLING  
NOTATION AND ESTIMATION FORMULAS

NOTATION (WE DROP THE SUBSCRIPT Y FROM THE POPULATION PARAMETERS, SINCE Y IS THE ONLY RANDOM VARIABLE OF CONCERN HERE:

$N$  = NUMBER OF CLUSTERS IN THE POPULATION

$n$  = NUMBER OF CLUSTERS SELECTED, USING SIMPLE RANDOM SAMPLING

$m_i$  = NUMBER OF ELEMENTS IN THE  $i$ -th CLUSTER

$$\bar{m} = \frac{1}{n} \sum_{i=1}^n m_i = \text{AVERAGE CLUSTER SIZE FOR THE SAMPLE}$$

$$M = \sum_{i=1}^n m_i = \text{NUMBER OF ELEMENTS IN THE POPULATION}$$

$$\bar{M} = \frac{M}{N} = \text{AVERAGE CLUSTER SIZE FOR THE POPULATION}$$

$y_i$  = TOTAL OF THE VALUES FOR ALL OBSERVATIONS IN THE  $i$ -th CLUSTER

ESTIMATION FORMULAS:

NOTE: THERE ARE ALTERNATIVE ESTIMATORS FOR THE VARIANCE IN CLUSTER SAMPLING, AND SOME OF THEM ARE COMPLICATED (AND INVOLVE ADVANCED STATISTICAL TECHNIQUES SUCH AS ANALYSIS OF VARIANCE). THE FORMULAS PRESENTED HERE FOLLOW SCHAEFFER, AND ARE THE LEAST COMPLICATED. THEY ARE BIASED, BUT THE BIAS IS LOW FOR LARGE  $n$  AND DISAPPEARS IF THE CLUSTER SIZES ( $m_i$ ) ARE ALL EQUAL.

ESTIMATOR OF THE POPULATION MEAN,  $\mu$ :

$$\hat{\mu}_{YC} = \bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$$

ESTIMATED VARIANCE OF  $\bar{y}$ :

$$v(\bar{y}) = \frac{N-n}{Nn\bar{M}^2} \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\bar{y})}$

IF  $\bar{M}$  IS UNKNOWN, REPLACE IT BY THE ESTIMATE  $\bar{m}$ .

ESTIMATOR OF THE POPULATION TOTAL,  $\tau$ , WHEN  $M$  (THE TOTAL NUMBER OF ELEMENTS IN THE POPULATION) IS KNOWN. SINCE  $\tau = M\mu$ ,

$$\hat{\tau}_{YC1} = M\bar{y} = M \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$$

ESTIMATED VARIANCE OF  $M\bar{y}$ :

$$v(M\bar{y}) = M^2 v(\bar{y}) = N^2 \frac{N-n}{Nn} \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(M\bar{y})}$

IN ORDER TO USE THE ESTIMATOR  $M\bar{y}$  TO ESTIMATE THE POPULATION TOTAL, IT IS, OF COURSE, NECESSARY TO KNOW THE VALUE OF  $M$  (THE TOTAL NUMBER OF ELEMENTS IN THE POPULATION). IN SITUATIONS IN WHICH CLUSTER SAMPLING IS APPROPRIATE,  $M$  MAY NOT BE KNOWN.

AN ESTIMATOR OF THE POPULATION TOTAL,  $\tau$ , THAT DOES NOT REQUIRE KNOWLEDGE OF  $M$  IS GIVEN BELOW.

DEFINE

$$\bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_i = \text{AVERAGE OF THE CLUSTER TOTALS FOR THE } n \text{ SAMPLE CLUSTERS}$$

THEN

$$\hat{\tau}_{YC2} = N\bar{y}_t = \frac{N}{n} \sum_{i=1}^n y_i$$

ESTIMATED VARIANCE OF  $N\bar{y}_t$ :

$$v(N\bar{y}_t) = N^2 v(\bar{y}_t) = N^2 \frac{N-n}{Nn} \frac{\sum_{i=1}^n (y_i - \bar{y}_t)^2}{n-1}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(N\bar{y}_t)}$

CLUSTER SAMPLING MAY BE COMBINED WITH STRATIFIED SAMPLING, WHERE A CLUSTER SAMPLE IS SELECTED FROM EACH STRATUM. THE ESTIMATION FORMULAS ARE SOMEWHAT COMPLICATED, AND THE STUDENT IS REFERRED TO A REFERENCE TEXT ON SAMPLE SURVEY, SUCH AS COCHRAN, *SAMPLING TECHNIQUES*.

## SYSTEMATIC SAMPLING

NUMBER THE  $N$  UNITS OF THE FRAME (POPULATION) FROM 1 TO  $N$  IN SOME ORDER (E.G., AN EXISTING LIST, OR A CARD INDEX).

TO SELECT A SAMPLE OF  $n$  UNITS, TAKE A UNIT AT RANDOM FROM THE FIRST  $k$  UNITS AND EVERY  $k$ -th UNIT THEREAFTER, WHERE  $k = N/n$ . THIS TYPE OF SYSTEMATIC SAMPLE IS CALLED AN "EVERY  $k$ -th" OR "ONE-IN- $k$ " SYSTEMATIC SAMPLE.

ADVANTAGES OF SYSTEMATIC SAMPLING:

1. IT IS EASIER TO PERFORM, AND HENCE LESS PRONE TO ERRORS, THAN SIMPLE RANDOM SAMPLING.
2. IT IS OFTEN MORE PRECISE THAN SIMPLE RANDOM SAMPLING (SINCE IT IN EFFECT STRATIFIES THE POPULATION INTO  $n$  STRATA – THE FIRST  $k$  UNITS, THE SECOND  $k$  UNITS, AND SO ON).

EXAMPLE: SELECT A SAMPLE OF  $n$  SHOPPERS ON A STREET CORNER. DO NOT KNOW THE TOTAL POPULATION SIZE ( $N$ ), AND CANNOT LIST THE POPULATION. COULD SELECT EVERY 20-th SHOPPER UNTIL HAVE OBTAINED A SAMPLE OF SIZE  $n = 50$ .

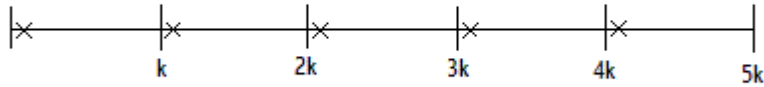
SYSTEMATIC SAMPLING IS MORE PRECISE THAN SIMPLE RANDOM SAMPLING IF THE VARIATION AMONG UNITS IN THE SAME SYSTEMATIC SAMPLE IS GREATER THAN THE VARIATION AMONG THE WHOLE POPULATION. (IF UNITS WITHIN THE SAME SYSTEMATIC SAMPLE ARE SIMILAR TO EACH OTHER, THEN THEY DO NOT PROVIDE AS MUCH INFORMATION AS A SIMPLE RANDOM SAMPLE.)

IF THE LIST IS IN RANDOM ORDER, THEN SYSTEMATIC SAMPLING PRODUCES THE SAME RESULTS AS SIMPLE RANDOM SAMPLING (WITHOUT REPLACEMENT).

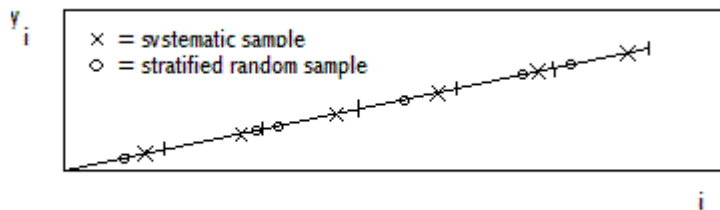
IF THE LIST IS NOT IN RANDOM ORDER, SYSTEMATIC SAMPLING MAY BE MORE PRECISE THAN SIMPLE RANDOM SAMPLING, OR LESS PRECISE. IN THIS CASE, IT IS NOT POSSIBLE TO ESTIMATE THE VARIANCE OF THE ESTIMATE FROM THE SAMPLE. IN ORDER TO BE ABLE TO ESTIMATE THE VARIANCE, IT IS NECESSARY TO TAKE TWO OR MORE SYSTEMATIC SAMPLES, I.E, SELECT TWO OR MORE STARTING POINTS AT RANDOM, AND TAKE A SYSTEMATIC SAMPLE STARTING FROM EACH STARTING POINT. IF  $m$  (E.G., 10) SYSTEMATIC SAMPLES ARE SELECTED, THEN EACH ONE IS OF SIZE  $n/m$ , WHERE  $n$  DENOTES THE TOTAL SAMPLE SIZE DESIRED (AND THE SKIP INTERVAL IS  $m$  TIMES AS LARGE AS IT WOULD HAVE BEEN HAD A SINGLE SYSTEMATIC SAMPLE BEEN SELECTED). THE PROCESS OF SELECTING SEVERAL SYSTEMATIC SAMPLES (FROM RANDOMLY SELECTED STARTING POINTS) IS CALLED "REPEATED SYSTEMATIC SAMPLING".

THE PRECISION OF A SYSTEMATIC SAMPLE DEPENDS ON HOW THE POPULATION IS ORDERED. IF THE UNITS OF THE POPULATION ARE ORDERED, THEN SYSTEMATIC SAMPLING IS MORE PRECISE THAN SIMPLE RANDOM SAMPLING. IF NEARBY UNITS ARE SIMILAR (HIGHLY CORRELATED), THEN SYSTEMATIC SAMPLING IS USUALLY MORE PRECISE THAN SIMPLE RANDOM SAMPLING. IF SOME SORT OF PERIODIC VARIATION IS PRESENT IN THE POPULATION, SYSTEMATIC SAMPLING CAN PRODUCE VERY POOR RESULTS IF ONLY A SINGLE SYSTEMATIC SAMPLE IS SELECTED.

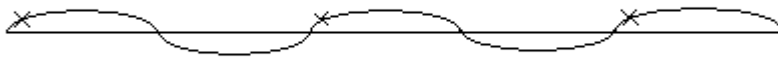
## SCHEMATIC REPRESENTATION OF SYSTEMATIC SAMPLING



### EXAMPLE 1: SYSTEMATIC SAMPLING IN A POPULATION WITH A LINEAR TREND

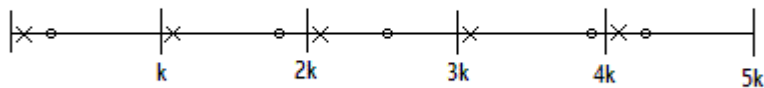


### EXAMPLE 2: SYSTEMATIC SAMPLING IN A POPULATION WITH PERIODIC VARIATION



SYSTEMATIC SAMPLING IS EQUIVALENT TO STRATIFIED SAMPLING IN THE CASE IN WHICH A SINGLE ITEM IS SELECTED FROM EACH STRATUM (OR  $m$  ITEMS ARE SELECTED FROM EACH STRATUM, IF  $m$  SYSTEMATIC SAMPLES ARE SELECTED).

x = systematic sample  
o = stratified random sample



SYSTEMATIC SAMPLING IS EQUIVALENT TO CLUSTER SAMPLING IN THE CASE IN WHICH A SINGLE CLUSTER IS SELECTED (OR  $m$  CLUSTERS ARE SELECTED, IF  $m$  SYSTEMATIC SAMPLES ARE SELECTED).

Sample Number					
1	2	...	i	...	k
$y_1$	$y_2$		$y_i$		$y_k$
$y_{k+1}$	$y_{k+2}$		$y_{k+i}$		$y_{2k}$
...					
$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	...	$y_{(n-1)k+i}$	...	$y_{nk}$

ESTIMATION FORMULAS FOR SYSTEMATIC SAMPLING, IN THE CASE OF A RANDOMLY ORDERED LIST: THE SAME AS FOR SIMPLE RANDOM SAMPLING.

$$\hat{\mu} = \bar{y}_{SY} = \frac{\sum_{i=1}^n y_i}{n}$$

ESTIMATED VARIANCE OF  $\bar{y}_{SY}$ :

$$v(\bar{y}_{SY}) = \frac{N-n}{N} \frac{s^2}{n}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\bar{y}_{SY})}$

IF  $N$  IS UNKNOWN, REPLACE THE  $fpc$ ,  $(N-n)/N$  BY 1.

ESTIMATION OF THE POPULATION TOTAL WHEN  $N$  (THE POPULATION SIZE) IS KNOWN (WHICH IS OFTEN NOT THE CASE IN SYSTEMATIC SAMPLING):

$$\hat{\tau} = N\bar{y}_{SY}$$

ESTIMATED VARIANCE OF  $\hat{\tau}$ :

$$v(N\bar{y}_{SY}) = N^2 v(\bar{y}_{SY})$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(N\bar{y}_{SY})}$

### ESTIMATION FORMULAS FOR REPEATED SYSTEMATIC SAMPLING

$n$  = TOTAL SAMPLE SIZE

$m$  = NUMBER OF SYSTEMATIC SAMPLES, EACH OF SIZE  $n/m$  ( $n$  IS ASSUMED TO BE AN INTEGRAL MULTIPLE OF  $m$ )

$k'$  = SKIP INTERVAL FOR EACH SYSTEMATIC SAMPLE =  $mN/n$

$$\bar{y}_i = \frac{\sum_{j=1}^{n/m} y_{ij}}{n/m} = \text{SAMPLE MEAN OF THE } i\text{-th SYSTEMATIC SAMPLE}$$

ESTIMATOR OF THE POPULATION MEAN  $\mu$  USING  $m$  ONE-IN- $k'$  SYSTEMATIC SAMPLES:

$$\hat{\mu} = \sum_{i=1}^m \frac{\bar{y}_i}{m}$$

ESTIMATED VARIANCE OF  $\hat{\mu}$  :

$$v(\hat{\mu}) = \frac{N-n}{N} \frac{\sum_{i=1}^m (\bar{y}_i - \hat{\mu})^2}{m(m-1)}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\hat{\mu})}$  .

ESTIMATOR OF THE POPULATION TOTAL,  $\tau$  , USING  $m$  ONE-IN- $k'$  SYSTEMATIC SAMPLES:

$$\hat{\tau} = N\hat{\mu}$$

ESTIMATED VARIANCE OF  $\hat{\tau}$  :

$$v(\hat{\tau}) = N^2v(\hat{\mu})$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\hat{\tau})}$

## TWO-STAGE CLUSTER SAMPLING (SUBSAMPLING); MULTISTAGE SAMPLING)

A TWO-STAGE CLUSTER SAMPLE IS A SAMPLE OBTAINED BY SELECTING A SAMPLE OF CLUSTERS AND THEN SELECTING A SAMPLE OF ELEMENTS FROM EACH SAMPLED CLUSTER. SAMPLING MAY BE EXTENDED TO MORE THAN TWO STAGES, IN WHICH CASE IT IS CALLED MULTISTAGE SAMPLING.

THE CLUSTERS ARE CALLED THE PRIMARY UNITS AND THE ELEMENTS SELECTED FROM CLUSTERS ARE CALLED SECONDARY UNITS.

THE MOTIVATION FOR USING TWO-STAGE CLUSTER SAMPLING IS THAT THE UNITS WITHIN A CLUSTER MAY BE SIMILAR, SO THAT IT IS INEFFICIENT TO OBSERVE ALL OF THEM.

CLUSTERS ARE USUALLY FORMED FROM UNITS THAT ARE GEOGRAPHICALLY PROXIMATE OR EASY TO ADMINISTER.

THE SURVEY DESIGN OBJECTIVE IN TWO-STAGE CLUSTER SAMPLING IS TO BALANCE THE NUMBER OF FIRST-STAGE AND SECOND-STAGE UNITS TO ACHIEVE MAXIMUM PRECISION FOR A FIXED COST. THIS PROBLEM WILL BE ADDRESSED IN DAY 2.

### NOTATION:

$\mu$  = POPULATION MEAN

$\tau$  = POPULATION TOTAL

$N$  = NUMBER OF CLUSTERS IN THE POPULATION

$n$  = NUMBER OF CLUSTERS SELECTED, USING SIMPLE RANDOM SAMPLING

$M_i$  = NUMBER OF ELEMENTS IN THE  $i$ -th CLUSTER

$m_i$  = NUMBER OF ELEMENTS SELECTED IN A SIMPLE RANDOM SAMPLE FROM THE  $i$ -th CLUSTER

$M = \sum_{i=1}^n m_i$  = NUMBER OF ELEMENTS IN THE POPULATION

$\bar{M} = \frac{M}{N}$  = AVERAGE CLUSTER SIZE FOR THE POPULATION

$y_{ij}$  = VALUE OF THE  $j$ -th SAMPLE ELEMENT FROM THE  $i$ -th SAMPLE CLUSTER

$$\bar{y}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$$

## TWO-STAGE CLUSTER SAMPLING: ESTIMATION FORMULAS

### ESTIMATION FORMULAS:

UNBIASED ESTIMATOR OF THE POPULATION MEAN,  $\mu$ :

$$\hat{\mu} = \frac{N}{M} \frac{\sum_{i=1}^n M_i \bar{y}_i}{n}$$

ESTIMATED VARIANCE OF  $\hat{\mu}$  :

$$v(\hat{\mu}) = \frac{N-n}{N} \frac{1}{n\bar{M}^2} s_b^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^n M_i^2 \frac{M_i - m_i}{M_i} \frac{s_i^2}{m_i}$$

WHERE

$$s_b^2 = \frac{\sum_{i=1}^n (M_i \bar{y}_i - \bar{M} \hat{\mu})^2}{n-1}$$

AND

$$s_i^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\hat{\mu})}$

ESTIMATION OF THE POPULATION TOTAL,  $\tau$  :

$$\hat{\tau} = M\hat{\mu} = N \frac{\sum_{i=1}^n M_i \bar{y}_i}{n}$$

ESTIMATED VARIANCE OF  $\hat{\tau}$  :

$$v(\hat{\tau}) = M^2 v(\hat{\mu}) = \frac{N-n}{N} \frac{N^2}{n} s_b^2 + \frac{N}{n} \sum_{i=1}^n M_i^2 \frac{M_i - m_i}{M_i} \frac{s_i^2}{m_i}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{M^2 v(\hat{\mu})}$

THE ESTIMATOR  $\hat{\mu}$  GIVEN ABOVE REQUIRES A KNOWLEDGE OF  $M$ , THE TOTAL NUMBER OF ELEMENTS IN THE POPULATION. IN CLUSTER SAMPLING, THIS IS OFTEN

UNKNOWN. IN THIS CASE, IT IS ESTIMATED AS THE AVERAGE CLUSTER SIZE MULTIPLIED BY THE NUMBER OF CLUSTERS IN THE POPULATION,  $N$ :

$$N \frac{\sum_{i=1}^n M_i}{n}$$

IF WE REPLACE  $M$  IN THE FORMULA FOR  $\hat{\mu}$ , THEN WE OBTAIN A RATIO ESTIMATOR, SINCE BOTH THE NUMERATOR AND DENOMINATOR ARE RANDOM VARIABLES. THE ESTIMATION FORMULAS FOR RATIO ESTIMATORS THEN APPLY.

RATIO ESTIMATOR OF THE POPULATION MEAN,  $\mu$ :

$$\hat{\mu}_R = \frac{\sum_{i=1}^n M_i \bar{y}_i}{\sum_{i=1}^n M_i}$$

ESTIMATED VARIANCE OF  $\hat{\mu}_R$ :

$$v(\hat{\mu}_R) = \frac{N-n}{N} \frac{1}{n\bar{M}^2} s_R^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^n M_i^2 \frac{M_i - m_i}{M_i} \frac{s_i^2}{m_i}$$

WHERE

$$s_R^2 = \frac{\sum_{i=1}^n M_i^2 (\bar{y}_i - \hat{\mu}_R)^2}{n-1}$$

AND

$$s_i^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

BOUND ON THE ERROR OF ESTIMATION:  $2\sqrt{v(\hat{\mu}_R)}$

THE ESTIMATOR  $\hat{\mu}_R$  IS BIASED, BUT THE BIAS IS LOW FOR LARGE  $n$ .

DOUBLE SAMPLING  
(TWO-PHASE SAMPLING)

SURVEY DESIGNS CAN BE SUBSTANTIALLY IMPROVED (REDUCED SAMPLE SIZE, REDUCED COST, INCREASED PRECISION) WITH KNOWLEDGE OF:

AN AUXILIARY VARIATE  
VARIANCES (IN THE POPULATION, WITHIN AND BETWEEN STRATA, WITHIN AND BETWEEN CLUSTERS)

SUCH DATA MAY BE COLLECTED IN A PRELIMINARY SAMPLE. SUCH A PROCEDURE IS CALLED DOUBLE SAMPLING, OR TWO-PHASE SAMPLING.

EXAMPLES:

1. TO ENABLE STRATIFICATION. WOULD LIKE TO STRATIFY ON AGE AND SEX, BUT DO NOT HAVE AGE AND SEX DATA. IT MAY BE FEASIBLE TO CONDUCT A LARGE PRELIMINARY SURVEY (E.G., USING SYSTEMATIC SAMPLING), RECORD AGE AND SEX FOR EACH SAMPLED PERSON, AND CONDUCT A SECOND SURVEY, STRATIFIED BY AGE AND SEX, FOR INTERVIEW.
2. TO ENABLE ANALYTICAL COMPARISONS. WOULD LIKE TO MAKE COMPARISONS BETWEEN SUBGROUPS OF THE POPULATION (E.G., MALE VS. FEMALE, WHITE VS. BLACK, EMPLOYED VS. UNEMPLOYED).
3. TO ENABLE RATIO OF REGRESSION ESTIMATION. CONDUCT A LARGE SURVEY TO RECORD  $Y$ , AND A MUCH SMALLER SURVEY TO RECORD ONE OR MORE VARIATES,  $X_1, X_2, \dots$  RELATED TO IT.
4. REPEATED SAMPLING OF THE SAME POPULATION (LONGITUDINAL, OR PANEL, SURVEYS). ESTIMATION OBJECTIVES:
  - a. ESTIMATE CHANGE IN MEAN FROM ONE TIME TO ANOTHER (BEST TO RETAIN SAME SAMPLE FOR ALL SAMPLING OCCASIONS).
  - b. ESTIMATE THE MEAN OVER ALL TIME (BEST TO DRAW A NEW SAMPLE ON EACH OCCASION)
  - c. ESTIMATE THE MEAN FOR THE MOST RECENT TIME (SAME PRECISION FROM KEEPING THE SAME SAMPLE FOR ALL OCCASIONS OR BY CHANGING IT ON EVERY OCCASION. REPLACEMENT OF PART OF THE SAMPLE ON EACH OCCASION MAY BE A BETTER ALTERNATIVE).

OPTIMAL ALLOCATION (IN CASE ONE, TO ENABLE STRATIFICATION): CHOOSE THE PRELIMINARY-SURVEY SAMPLE SIZE AND THE SECOND-SURVEY SAMPLE SIZE TO MINIMIZE THE VARIANCE OF THE STRATIFIED ESTIMATE, SUBJECT TO SPECIFIED TOTAL COST.

FORMULAS WILL NOT BE PRESENTED FOR DOUBLE SAMPLING. CONSULT COCHRAN, *SAMPLING TECHNIQUES*, FOR DISCUSSION AND FORMULAS.

## DAY 2: HOW TO DESIGN SURVEYS AND ANALYZE SURVEY DATA

### PART ONE: HOW TO DESIGN DESCRIPTIVE SURVEYS

OVERVIEW OF SECOND DAY'S COURSE CONTENT

REVIEW OF FIRST DAY'S TOPICS

THE ELEMENTS OF SURVEY DESIGN

DISTINCTION BETWEEN DESCRIPTIVE AND ANALYTICAL SURVEYS

GENERAL PROCEDURE FOR DESIGNING A DESCRIPTIVE SURVEY

WHEN AND HOW TO USE SIMPLE RANDOM SAMPLING

WHEN AND HOW TO USE STRATIFICATION

WHEN AND HOW TO USE A CLUSTERED DESIGN

WHEN AND HOW TO USE SYSTEMATIC SAMPLING

WHEN AND HOW TO USE A MULTISTAGE DESIGN

WHEN AND HOW TO USE DOUBLE SAMPLING

HOW TO RESOLVE CONFLICTING / MULTIPLE SURVEY DESIGN OBJECTIVES

### PART TWO: HOW TO DESIGN ANALYTICAL SURVEYS

SURVEY OF REGRESSION ANALYSIS

GENERAL PROCEDURE FOR DESIGNING AN ANALYTICAL SURVEY

HOW TO USE MULTIPLE STRATIFICATION FOR AN ANALYTICAL DESIGN

HOW TO USE CONTROLLED SELECTION FOR AN ANALYTICAL DESIGN

### PART THREE: HOW TO ANALYZE SURVEY DATA

STANDARD ESTIMATION PROCEDURES FOR DESCRIPTIVE SURVEYS

STANDARD ESTIMATION PROCEDURES FOR ANALYTICAL SURVEYS

COMPUTER PROGRAMS FOR ANALYSIS OF SURVEY DATA

OUTLINE OF TOPICS FOR THIRD DAY

Reference List  
Sample Survey Design and Analysis

There are thousands of textbooks on statistics, and many on sample survey. Below are some from my personal library. They are somewhat old, but the basic theory has not changed. For recent texts, check a university bookstore or Internet book vendors (Amazon, Barnes & Noble).

Of the following list, I would recommend Mood's book for an introduction to mathematical statistics, almost any introductory book for an elementary introduction to statistics, Scheaffer's book for an elementary introduction to sample survey, Cochran's book for a detailed mathematical discussion of sample survey, and Kish's and Des Raj's books for a somewhat less mathematical discussion.

General statistics (undergraduate-level mathematical statistics)

Mood, Alexander M., Franklin Graybill and Duane C. Boes, *Introduction to the Theory of Statistics*, 3<sup>rd</sup> edition, McGraw Hill, 1974

Snedecor, George W. and William G. Cochran, *Statistical Methods*, 8<sup>th</sup> edition, Iowa State University Press, 1989

Brunk, H. D., *An Introduction to Mathematical Statistics*, Ginn and Company, 1960

Fraser, D. A. S., *Statistics: An Introduction*, Wiley, 1958

Dixon, Wilfred J. and Frank J. Massey, Jr., *Introduction to Statistical Analysis*, 2<sup>nd</sup> edition, McGraw Hill, 1957

General statistics (less mathematical)

Crow, Edwin L., Frances A. Davis and Margaret W. Maxfield, *Statistics Manual: With Examples Taken from Ordnance Development*, Dover Publications, 1960

Downie, N. M. and R. W. Heath, *Basic Statistical Methods*, 4<sup>th</sup> edition, Harper & Row, 1974

Survey sampling, less mathematical

Scheaffer, Richard L., William Mendenhall and Lyman Ott, *Elementary Survey Sampling*, 2<sup>nd</sup> edition, Duxbury Press, 1979 (6<sup>th</sup> edition 2005).

Babbie, Earl, *Survey Research Methods*, Wadsworth Publishing Company, 1973

Rubin, Allen and Earl Babbie, *Research Methods for Social Work*, 3<sup>rd</sup> edition, Brooks / Cole Publishing Company, 1997

Des Raj, *The Design of Sample Surveys*, McGraw-Hill, Inc., 1972

Kish, Leslie, *Survey Sampling*, John Wiley & Sons, 1965

Survey sampling, mathematical

Cochran, William G., *Sampling Techniques*, 3<sup>rd</sup> edition, John Wiley & Sons, Inc., 1977

Des Raj, *Sampling Theory*, McGraw-Hill, Inc., 1968

Hansen, Morris H., William N. Hurwitz and William G. Madow, *Sample Survey Methods and Theory*, volumes 1 and 2, John Wiley & Sons, Inc., 1953

Survey sampling, additional references (mix of mathematical and less mathematical)

Deming, William Edwards, *Some Theory of Sampling*, Dover Publications, 1950

Deming, W. Edwards, *Sample Design in Business Research*, John Wiley & Sons, 1960

Johnson, N.L. and Smith, H. Jr., *New Developments in Survey Sampling*, John Wiley & Sons, Inc., 1969

Murthy, M.N., *Sampling, Theory and Methods*, Statistical Publishing Society, 1967

Stephan, F.F., and McCarthy, P.J., *Sampling Opinions - An Analysis of Survey Procedures*, John Wiley & Sons, Inc., 1958

Sukhatme, P.V., and Sukhatme, B.V., *Sampling Theory of Surveys with Applications*, P.V. Sukhatme and B.V. Sukhatme, 1970

Williams, Bill, *A Sampler on Sampling*, John Wiley & Sons, 1978

Yates, Frank, *Sampling Theory of Censuses and Surveys*, 3<sup>rd</sup> edition, Charles Griffin and Company, 1960.

Slonin, Morris James, *Sampling* (original title: *Sampling in a Nutshell*), Simon and Schuster, 1960

Singh, Daroga and F. S. Chaudhary, *Theory and Analysis of Sample Survey Designs*, John Wiley & Sons, 1986

Rubin, Donald B., *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, 1987

Little, Roderick J. A. and Donald B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, 1987

Groves, Robert M., Paul P. Biemer, Lars E. Lyberg, James T. Massey, William L. Nicholls II, Joseph Waksberg (editors), *Telephone Survey Methodology*, John Wiley & Sons, 1988

Ghosh, M. and G. Meeden, *Bayesian Methods for Finite Population Sampling*, Chapman & Hall, 1997

Casley, D. J. and D. A. Lury, *Data Collection in Developing Countries*, 2<sup>nd</sup> edition, Clarendon Press, 1987 (very little on sample survey, most about data collection in general)

Iarossi, Giuseppe, *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*, The World Bank, 2006

References that discuss techniques for analytical surveys (model-based approach) and resampling

Rao, J. N. K. and D. R. Bellhouse, "History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis," *Survey Methodology*, June 1990, Statistics Canada

Risto Lehtonen and Erikki Pahkinen, *Practical Methods for Design and Analysis of Complex Surveys*, 2<sup>nd</sup> edition, Wiley, 2004

Thompson, Steven K., *Sampling*, 2<sup>nd</sup> edition, Wiley, 2002

Shao, Jun and Dongsheng Tu, *The Jackknife and Bootstrap*, Springer, 1995

Efron, B. and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993

Wolter, Kirk M., *Introduction to Variance Estimation*, Springer, 1985

Cohen, Jacob, *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, 1969. (Discusses the determination of sample size by specification of the power of tests of hypothesis, rather than the precision of estimates.)

References on Experimental Design and Quasi-experimental Design

Cochran, William G. and Gertrude M. Cox, *Experimental Designs*, 2<sup>nd</sup> edition, Wiley, 1950, 1957

Campbell, Donald T. and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, 1966. Reprinted from Handbook of Research on Teaching, N. L. Gage (editor), Rand McNally, 1963.

Cook, Thomas D. and Donald T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* Houghton Mifflin, 1979

Résumé of Course Developer: Joseph George Caldwell, Ph.D.  
Consultant in Statistics, Economics, Operations Research and Computer Science

Education...

Ph.D., Statistics, University of North Carolina at Chapel Hill, 1966  
B.S., Mathematics, Carnegie-Mellon University, 1962

Consultant...

to US government agencies, state governments, corporations, and foreign governments

Director/Supervisor of projects in the areas of...

- o sample survey design of major national surveys and statistical reporting systems
- o statistical experimental design and data analysis (SPSS, SAS)
- o computer models and information systems design (C, Xbase, Oracle SQL, MS Access)
- o expert systems / geographic information systems (ArcView)
- o systems and software engineering (C, Visual Basic, FORTRAN, DOD-STD-2167A, ISO12207)
- o operations research / management science and statistics in industrial and defense applications
- o monitoring and evaluation, planning and policy analysis of government programs in health, education, human services, urban problems, rural development, agriculture, tax policy analysis, and public finance
- o game theory (zero-sum and non-zero-sum, constrained games, ill-conditioned problems; computer solutions of complex games)
- o international development in the Philippines, Haiti, Egypt, Bangladesh, Ghana, Malawi, Botswana, Zambia and Timor-Leste

Manager of contract research firm (seven years); successful bidder on numerous technical contracts, including four Small Business Innovation Research (SBIR) contracts. Director of more than twenty projects for US government and other clients.

Professor of Statistics at the University of Arizona, Tucson, Arizona

Developer of technical seminars and computer program packages in sample survey design, forecasting, demographic projection, and geographic information systems

Languages: Native in English; working knowledge of Spanish, French; limited Portuguese, German, Arabic

Summary of Experience. Dr. Caldwell's professional career in research and research management has centered on the use of modern analysis techniques to solve practical problems in government, commercial, industrial, and military applications. He has directed major technical projects; developed technical training seminars; accomplished significant research results in statistics; developed statistical, demographic, and geographic-information-system computer program packages; designed statistical reporting and management information systems; and served as professor of statistics, consultant, and manager of a contract research firm.

Contact information:

Permanent address: 503 Chastine Drive, Spartanburg, SC 29301-5977 USA. Tel. 1-(864)439-2772, e-mail [jcaldwell9@yahoo.com](mailto:jcaldwell9@yahoo.com)

## CAPABILITIES AND EXPERIENCE IN STATISTICS

Education. Dr. Caldwell holds a PhD degree in mathematical statistics from the University of North Carolina at Chapel Hill. In his graduate studies, he specialized in the theory of experimental design and algebraic coding theory. His doctoral dissertation advisor was Prof. R. C. Bose, regarded as the "father" of the mathematical theory of experimental design, and developer of the Bose-Chaudhuri-Hocquenghem (BCH) codes, the best known class of codes for correcting random errors in noisy communication channels. In his doctoral dissertation, Dr. Caldwell developed the best-known class of codes for correcting additive and synchronization errors in noisy communication channels.

Experience. Dr. Caldwell has over thirty years' experience as a consultant and teacher of statistics. He has provided statistical consultation in a wide variety of fields, including sample survey design and analysis; statistical analysis of data; time series analysis and forecasting; simulation and modeling of industrial and military systems; test and evaluation of communications systems; industrial quality control; process control and product improvement; and planning, policy analysis, and program evaluation in health, education, social services, and economic development.

Experience in Monitoring and Evaluation. An area of specialization in which he has applied statistical methodology is monitoring and evaluation. He developed survey designs for a number of monitoring systems and program evaluation studies in the US and foreign countries. In the US, he directed a number of national projects in program monitoring and evaluation, including the Vocational Rehabilitation Evaluation Standards Study for the US Rehabilitation Services Administration; Social Services Effectiveness Evaluation for West Virginia; the Day Care Cost Benefit Study for the US Department of Health and Human Services; Cost-Benefit Analysis of National Institute for Alcohol Abuse and Alcoholism Treatment Centers; Medicaid Standards Impact Assessment. He developed the sampling plans for several national state/federal social and economic programs, including the Sampling Manual for Utilization Review of Medicaid; the Sampling Manual for Social Services (Title XX) Reporting Requirements; and the Sampling Manual for Office of Child Support Enforcement Reporting Requirements. He developed the survey design for the Department of Housing and Urban Development Housing Market Practices Survey; the Research Design for the Urban Arterials Section of the Highway Capacity Manual; and the survey design for the Elementary and Secondary School Civil Rights Survey.

Overseas, he served as Project Director and Chief of Party for the Economic and Social Impact Analysis / Women in Development Project in the Philippines. This project provided consulting in research design (experimental design, quasi-experimental design, survey design, survey instrument design) for a broad range of development projects (health, nutrition, and family planning; education; integrated agricultural production and marketing, aquaculture production, and agro-reforestation; integrated area development; feeder roads; ports; local water systems; electrification; small-scale industries, and tourism). He served as Manager of Monitoring and Evaluation for the Local Development II – Provincial Project in Egypt. This project was the largest USAID-funded local-level rural development project in the world. On this project, which involved the funding of 16,000 local-level projects, a sample survey design was constructed to enable assessment of program impact based on a sample of about 800 projects. The projects included potable water, waste water, roads, buildings, rolling stock, environment, and information systems.

Teaching. Dr. Caldwell served as an adjunct professor of statistics at the University of Arizona. He taught the graduate course, Sampling Theory and Methods, and the undergraduate course,

Statistical Methods in Management (for all students of business, public administration, and management information systems).

Technical Training. Dr. Caldwell developed and marketed the technical seminar, Sample Survey Design and Analysis. This popular three-day course has been given on an advertised basis, and also on an in-house basis at the US Bureau of Labor Statistics.

Research in Statistical Methodology. Dr. Caldwell served as a consultant to the US Department of Education's National Center for Education Statistics, on the Statistical Analysis Group in Education (SAGE) program. In this work, he developed a new approach to the treatment of nonresponse in longitudinal surveys. For the US Office of Naval Research, he directed the project, "Fast Algorithms for Estimation, Prediction and Control." This project was concerned with the development of an estimation methodology that could be used as an alternative to the conventional least-squares procedure, in ill-conditioned estimation problems (singularity, missing values).

Statistical Software Development / Time Series Analysis. Dr. Caldwell developed the first commercially available computer program package for implementation of the Box-Jenkins time series methodology. The Box-Jenkins (autoregressive integrated moving average) models are useful in system identification problems, such as forecasting, control, and linear predictive coding of speech.

Sample Survey Design. Dr. Caldwell developed the design for many important national sample surveys and statistical reporting systems. He specializes in the development of analytical survey designs to collect data for model development, and has developed new techniques for handling nonresponse in longitudinal surveys. Surveys he designed include:

- o Ghana Trade and Investment Program Survey
- o Malawi Annual Primary School Enrollment Survey
- o National survey of local development projects in Egypt
- o National Center for Health Services Research (NCHSR) Hospital Cost Data Study
- o Professional Standards Review Organization (PSRO) Data Base Development Study
- o Study of the Impact of National Health Insurance on Bureau of Community Health Service Users
- o 1976 Survey of Institutionalized Persons
- o Housing and Urban Development (HUD) Housing Market Practices Survey
- o Research Design for the Urban Arterials Section of the Highway Capacity Manual
- o Elementary and Secondary School Civil Rights Survey

Statistical Program Monitoring Systems. He developed the sampling manuals for the following state-federal reporting systems:

- o Sampling Manual for Utilization Review of Medicaid
- o Sampling Manual for Social Services Reporting Requirements (Title XX)
- o Sampling Manual for Office of Child Support Enforcement Reporting Requirements

Management Information Systems. He developed the Personnel Management Information System (PMIS) for the civil service of the Government of Malawi and the Education Management Information System (EMIS) for the Government of Zambia.

Experimental Design and Quality Control. He developed statistical experimental designs for test and evaluation, simulation model run-sets, chemical and physical experimentation, and industrial quality control applications.

Data Analysis. He has applied statistical software to analyze sample survey data, including the Urban Institute's Study of Salaries in Academia, surveys to collect price data for commodities in Haiti, and surveys of the implementation, operational, and service-delivery status of local development projects in Egypt. He is an expert in the analysis of time series data, and has analyzed data collected in accordance with statistical experimental designs. He has applied the full range of statistical analysis procedures, including time series analysis, multiple regression analysis, multivariate analysis of variance, components-of-variance analysis, factor analysis, and nonparametric analysis.

He is expert in the use of modern commercial statistical analysis software (e.g., SPSS, SAS) and the use of related microcomputer software (e.g., Microsoft Access database management system).

Positions.

Consultant, 1974-present (various organizations, including the Academy for Educational Development, Wachovia Bank, Chemonics International, Bank of Botswana, United Nations)  
President and Manager, Vista Research Corporation, Tucson and Sierra Vista, AZ, 1988-91  
Professor of Statistics, University of Arizona, Tucson, AZ, 1982-86  
Director of Research and Development and Principal Scientist of US Army Electronic Proving Ground's Electromagnetic Environmental Test Facility, Bell Technical Operations, Tucson and Sierra Vista, AZ, 1982-86, 1986-88  
Principal Engineer, SINGER Systems and Software Engineering, Tucson, AZ, 1986  
President and Manager, Vista Research Corporation, Alexandria, VA, and Tucson, AZ, 1977-81  
Vice President, JWK International Corporation, Annandale, VA, 1974-76  
Principal, Planning Research Corporation, McLean, VA, 1972-74  
Member of the Technical Staff, Lambda Corporation / General Research Corporation, McLean, VA, 1967-72  
Senior Operations Research Analyst, Deering Milliken Research Corporation, Spartanburg, SC, 1966-67  
Operations Research Analyst, Research Triangle Institute, Research Triangle Park, NC, 1964-66

jgcsamp20070403.doc