

**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)541-7324  
[jcaldwell9@yahoo.com](mailto:jcaldwell9@yahoo.com)  
<http://www.foundationwebsite.org>

COURSE NOTES: DAY TWO

HOW TO DESIGN SURVEYS AND ANALYZE SURVEY DATA

3 April 2007

(Updated 15 April 2007, 26 March 2009, 29 May 2009, 11 January 2010)

***NO RECORDING DEVICES ALLOWED***

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

Posted at Internet

website <http://www.foundationwebsite.org/SampleSurvey3DayCourseDayTwo.pdf> .

May be copied or reposted for noncommercial use (or for evaluation by those considering attending the course), with attribution.

## 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; DISTINCTIONS BETWEEN DESCRIPTIVE AND ANALYTICAL SURVEYS

GENERAL PROCEDURE FOR DESIGNING A DESCRIPTIVE SAMPLE 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

REVIEW 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

## REVIEW OF FIRST DAY'S TOPICS

### DEFINITIONS:

SAMPLE DESIGN: SAMPLE SELECTION PROCESS; ESTIMATION PROCESS

SURVEY CONCEPTS:

ELEMENTS; INDIVIDUALS; ELEMENTARY UNITS

POPULATION: THE COLLECTION OF ELEMENTS; DEFINED BY CONTENT, UNITS, EXTENT, AND TIME

TARGET POPULATION: POPULATION OF INTEREST

SURVEY POPULATION: POPULATION SAMPLED FROM

UNIVERSE: CONCEPTUALLY INFINITE POPULATION GENERATED BY THEORETICAL MODEL OR PROCESS WHICH PRODUCES THE FINITE POPULATION

SUBCLASS: SUBPOPULATION

DOMAIN: SUBCLASS PLANNED FOR IN THE SURVEY

OBSERVATION: UNIT OBSERVED

DEFINITIONS (CONT'D):

VARIABLE: A MEASUREMENT ON AN ELEMENT ( $x_1, x_2, \dots, x_n$ )

SAMPLING UNITS: GROUPS OF ITEMS SELECTED IN SAMPLING

ELEMENT SAMPLING: SAMPLE UNIT = ELEMENT

CLUSTER SAMPLING: SAMPLE UNIT = MORE THAN ONE ELEMENT

STRATUM: SUBPOPULATION

LIST: COMPILATION OF SAMPLING UNITS

OBSERVATIONAL UNIT: UNITS FROM WHICH RESPONSES ARE OBTAINED

FRAME: LIST, OR NECESSARY PORTION

PERFECT FRAME: COMPLETE, ACCURATE, UNDUPLICATED, UP-TO-DATE

DEFINITIONS (CONT'D)

BASIC STATISTICAL CONCEPTS

POPULATION

RANDOM VARIABLE

LEVEL OF MEASUREMENT (NOMINAL, ORDINAL, INTERVAL)

PERCENTILES, MOMENTS (MEAN, STANDARD DEVIATION)

PROBABILITY DISTRIBUTION

HISTOGRAM: FREQUENCY PLOT

DISTRIBUTION PARAMETERS: VARIABLES SPECIFYING A DISTRIBUTION (E.G.,  
MEAN AND VARIANCE FOR A NORMAL DISTRIBUTION)

SAMPLE

STATISTIC

ESTIMATE, ESTIMATOR

SAMPLING DISTRIBUTION

DEFINITIONS (CONT'D)

PRECISION (RELIABILITY); BIAS (VALIDITY); ACCURACY

STANDARD ERROR (PRECISION)

COEFFICIENT OF VARIATION

MEAN SQUARE ERROR (ACCURACY) = VARIANCE + (BIAS)<sup>2</sup>

CONFIDENCE INTERVAL

NORMAL DISTRIBUTION; BINOMIAL DISTRIBUTION CORRELATION; VARIANCE

ALTERNATIVE ESTIMATORS (SIMPLE, RATIO, REGRESSION)


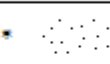
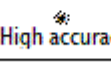
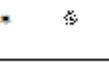
LAW OF LARGE NUMBERS; CENTRAL LIMIT THEOREM

SAMPLING DISTRIBUTIONS; BIAS, STANDARD ERROR OF THE ESTIMATE

FORMULAS FOR COMPUTING PARAMETER ESTIMATES AND PRECISION ESTIMATES

REPLACEMENT AND NONREPLACEMENT SAMPLING (BINOMIAL, HYPERGEOMETRIC)

FINITE POPULATION CORRECTION (FPC) FACTOR

		Bias	
		Low	High
Precision	Low		
	High	 High accuracy	

REVIEW OF FIRST DAY'S TOPICS (CONT'D)

TAXONOMY OF SURVEY SELECTION PROCEDURES

NONPROBABILITY PROCEDURES:

HAPHAZARD

PURPOSIVE, OR JUDGMENT ("REPRESENTATIVE" SAMPLING)

QUOTA SAMPLING (PURPOSIVE)

CAPTURE-RECAPTURE

PROBABILITY SAMPLING (EVERY UNIT HAS A KNOWN NONZERO PROBABILITY OF SELECTION) -- CAN USE STATISTICAL THEORY TO DEVELOP "GOOD" ESTIMATES AND ESTIMATE PRECISION ("RANDOM" SAMPLING; "REPRESENTATIVE" SAMPLING)

EQUAL PROBABILITIES -- AT ALL SAMPLING STAGES -- EQUAL OVERALL PROBABILITIES	UNEQUAL PROBABILITIES -- UNINTENDED (SAMPLE FRAME OR SELECTION PROBLEMS) -- OPTIMAL ALLOCATION
ELEMENT SAMPLING	CLUSTER SAMPLING -- ONE-STAGE -- SUBSAMPLING (TWO-STAGE)
UNSTRATIFIED	STRATIFIED
RANDOM SELECTION	SYSTEMATIC SELECTION
ONE-PHASE	TWO-PHASE (DOUBLE)
CROSS-SECTIONAL	LONGITUDINAL (PANEL, TIME SERIES)

## CHARACTERISTICS OF A SAMPLE DESIGN

GOAL-ORIENTED (ADDRESSES RESEARCH OBJECTIVES)

MEASURABLE PRECISION (CAN ESTIMATE STANDARD ERRORS)

FEASIBLE TO IMPLEMENT

EFFICIENT (HIGH PRECISION /COST RATIO) -- DESIGN EFFECT; "OPTIMAL" DESIGN

PRECISION (STANDARD ERROR)

BIAS

ACCURACY (TOTAL ERROR -- SAMPLING VARIABILITY + BIAS + NONSAMPLING ERRORS)

## SELECTION PROCEDURES

LIST OR FRAME

TABLE OF RANDOM NUMBERS

SYSTEMATIC SAMPLING

SAMPLING WITH OR WITHOUT REPLACEMENT

## THE ELEMENTS OF SURVEY DESIGN

SPECIFY POPULATION OF INTEREST

DEFINE ESTIMATES OF INTEREST

SPECIFY PRECISION OBJECTIVES OF THE SURVEY; RESOURCE CONSTRAINTS;  
POLITICAL CONSTRAINTS

SPECIFY OTHER VARIABLES OF INTEREST (EXPLANATORY VARIABLES,  
STRATIFICATION VARIABLES)

DEVELOP INSTRUMENTATION (DEVELOPMENT, PRETEST, PILOT TEST,  
RELIABILITY, VALIDITY)

DEVELOP SAMPLE DESIGN

DETERMINE SAMPLE SIZE AND ALLOCATION

SPECIFY SAMPLE SELECTION PROCEDURE

SPECIFY FIELD PROCEDURES

DETERMINE DATA PROCESSING PROCEDURES

DEVELOP DATA ANALYSIS PLAN

OUTLINE FINAL REPORT

## DISTINCTIONS BETWEEN DESCRIPTIVE AND ANALYTICAL SURVEYS

### DESCRIPTIVE SURVEY

- CONDITION OR STATE OF A FINITE POPULATION AT SOME POINT IN TIME
- ESTIMATION OF MEANS AND PROPORTIONS FOR THE POPULATION AND VARIOUS SUBPOPULATIONS
- ESTIMATION OF SOME BASIC RELATIONSHIPS, THROUGH CROSSTABULATIONS (BY STRATA)
- NOTE: TEST DIFFERENCES, IF MADE AT ALL, ARE MADE UNDER INFINITE POPULATION ASSUMPTION

### ANALYTICAL SURVEY

- ESTIMATION OF RELATIONSHIPS BETWEEN (DEPENDENT AND INDEPENDENT) VARIABLES, FOR A CONCEPTUALLY INFINITE POPULATION
- INFERENCES ABOUT THE PROCESS GENERATING OR ACTING ON THE POPULATION, NOT ABOUT THE POPULATION ITSELF
- MODEL BUILDING (MODEL IDENTIFICATION; ESTIMATION; TEST OF MODEL ADEQUACY; RESPECIFICATION); GENERAL LINEAR STATISTICAL MODEL (EXPERIMENTAL DESIGN, MULTIPLE REGRESSION ANALYSIS)

### DIFFERENCES

- IN DESCRIPTIVE SURVEY, WANT LARGE SAMPLE SIZES IN SUBPOPULATIONS OF INTEREST
- IN ANALYTICAL SURVEY, WANT VARIATION AND BALANCE IN VARIABLES OF INTEREST, AND ORTHOGONALITY (LOW CORRELATION) BETWEEN VARIABLES THAT ARE NOT CAUSALLY RELATED
- FPC IS IRRELEVANT IN ANALYTICAL SURVEY (INFINITE POPULATION)

A GOOD REFERENCE THAT DISCUSSES THE DISTINCTION BETWEEN DESCRIPTIVE AND ANALYTICAL SURVEYS (DESIGN-BASED, MODEL-BASED, MODEL-ASSISTED INFERENCE) IS SHARON L. LOHR'S *SAMPLING: DESIGN AND ANALYSIS* (DUXBURY PRESS, 1999).

## II. GENERAL PROCEDURE FOR DESIGNING A DESCRIPTIVE SAMPLE SURVEY (RELATE TO SLIDE "THE ELEMENTS OF SURVEY DESIGN")

### 1. SPECIFY POPULATION OF INTEREST

- TARGET POPULATION: CONTENT, UNITS, EXTENT, TIME
- UNITS OF ANALYSIS (E.G., STUDENTS, SCHOOLS, SCHOOL DISTRICTS)
- SURVEY POPULATION; RECOGNIZES PRACTICAL CONSTRAINTS
  - INACCESSABILITY OR LACK OF DATA; AVAILABILITY OF FRAME
  - EXPENSE (OF TRAVELLING, MEASURING)
  - LEGAL, POLITICAL CONSTRAINTS
  - TOTAL COST CONSTRAINTS
- WORK WITH PROGRAMMATIC PEOPLE TO DETERMINE A SURVEY POPULATION THAT WILL PERMIT MEANINGFUL RESULTS (WORTH STUDYING, ADEQUATE PRECISION)
- SUMMARIZE NATURE OF POPULATION:
  - FOR AVAILABLE VARIABLES RELATED TO THE STUDY, COMPUTE
    - MEANS
    - CROSSTABS
    - STRATUM VARIANCES
  - SAMPLING COST INFORMATION
  - CLUSTER STRUCTURE OF POPULATION (INTRACLUSTER CORRELATION COEFFICIENT)

## 2. DEFINE ESTIMATES OF INTEREST

- DEFINE VARIABLES OF INTEREST (E.G., AGE, SEX, RACE, EARNINGS, STATUS, OPINION)
- DEFINE SUBPOPULATIONS OF INTEREST (E.G. WOMEN, UNEMPLOYED, COLLEGE STUDENTS, PUBLIC SCHOOLS)
- IDENTIFY SURROGATE VARIABLES (E.G., EARNINGS VS. INCOME; BUDGET VS. EXPENDITURES)
- IDENTIFY METHODS OF OBSERVATION (RECORD SCAN, INTERVIEW, MAIL QUESTIONNAIRE)
- IDENTIFY MEASURES OF INTEREST (E.G., MEAN CHANGES IN EARNINGS, RELATIVE TO A COMPARISON GROUP)
- IDENTIFY DISTRIBUTION PARAMETERS TO BE ESTIMATED (MEAN, PERCENTILES, PERCENTAGES)

## 3. SPECIFY OBJECTIVES AND CONSTRAINTS

- PRECISION OBJECTIVES: CONFIDENCE LIMITS ON KEY ESTIMATES, FOR KEY SUBPOPULATIONS
- RESOURCE CONSTRAINTS (TIME, FUNDS, PERSONNEL)
- POLITICAL AND LEGAL CONSTRAINTS
- DATA LIMITATIONS (FRAME, MEASUREMENT)

#### 4. SPECIFY OTHER VARIABLES OF INTEREST

- STRATIFICATION VARIABLES
  - FOR PRECISION IMPROVEMENT
  - FOR COST REDUCTION
  - FOR SUBPOPULATIONS OF INTEREST
- EXPLANATORY VARIABLES
  - FOR CROSSTABS
  - FOR IMPROVED ESTIMATES (RATIO, REGRESSION)
  - FOR NONRESPONSE ANALYSIS
- SAMPLING COST DATA
- INTRASTRATUM VARIABILITY
- INTRACLUSTER CORRELATION COEFFICIENTS

## 5. DEVELOP INSTRUMENTATION

TYPE OF INSTRUMENT (MAIL, TELEPHONE, PERSONAL INTERVIEWS, DATA COLLECTION FORM)

### INSTRUMENT DESIGN

- QUESTION CONTENT
- QUESTION ORDER (RAPPORT-BUILDING FIRST, SENSITIVE LAST)
- QUESTION WORDING (NO LEADING, BIASED, DIFFICULT-TO-UNDERSTAND, DIFFICULT TO ANSWER, AMBIGUOUS, INFLAMMATORY)
- QUESTION STRUCTURE/FORMAT (OPEN OR CLOSED; NUMBER OF CATEGORIES)
- QUESTIONNAIRE LENGTH
- QUESTIONNAIRE LAYOUT (SELF-CODING, SMOOTH FLOW, CLEAR SKIP PATTERNS, RESPONSE SHEETS)
- QUESTIONNAIRE INSTRUCTIONS (PROCEDURES, PROBES)
- OPPORTUNITIES FOR INTERVIEWER COMMENT ON VALIDITY OF RESPONSE

## PRETEST

- LIMITED TRIAL (JUDGMENT SAMPLE) ON AS WIDE A VARIETY OF THE POPULATION AS POSSIBLE, TO CHECK FOR DATA AVAILABILITY, SMOOTHNESS OF FLOW, UNDERSTANDABILITY, LENGTH)

## PILOT TEST (POSSIBLY A RANDOM SAMPLE)

- TEST OF FIELD PROCEDURES
- MORE THOROUGH TEST OF INSTRUMENT
- RELIABILITY ASSESSMENT (ITEM-ITEM, ITEM-TOTAL, SPLIT-HALF CORRELATIONS)
- VALIDITY ASSESSMENT (INDEPENDENT CHECKS)
- REORDERING (FACTOR ANALYSIS)

## 6. DEVELOP SAMPLE DESIGN

- SYNTHESIZE SEVERAL DESIGN ALTERNATIVES, EMPHASIZING DIFFERENT DESIGN OBJECTIVES (E.G., ESTIMATE TOTALS VS. ESTIMATE DIFFERENCES (CONTRASTS); ESTIMATE CLUSTER MEAN VS. ELEMENT MEAN)
- DESIGN ALTERNATIVES: STRATIFIED, CLUSTERED, MULTISTAGE, TWO-PHASE
- INDICATE PRECISION, COST, OPERATIONAL PROBLEMS OF EACH DESIGN (USE ESTIMATION FORMULAS)
- ACHIEVE A CONSENSUS ON THE "BEST" OVERALL DESIGN

## 7. DETERMINE SAMPLE SIZE AND ALLOCATION

- NUMBER OF FIRST-STAGE AND SECOND-STAGE SAMPLE UNITS
- NUMBER OF FIRST-PHASE AND SECOND-PHASE UNITS
- NUMBER OF SAMPLE UNITS PER STRATUM
- PROPORTION OF PANEL TO BE REPLACED

FOR SIMPLE DESIGN WITH FEW OBJECTIVES, USE FORMULAS FOR OPTIMAL ALLOCATION (GIVEN COST, VARIANCE, INTRACLUSTER CORRELATION COEFFICIENT)

FOR COMPLEX DESIGNS WITH MANY OBJECTIVES, USE JUDGMENT TO DEVELOP SEVERAL ALTERNATIVES, PRESENT CHARACTERISTICS OF EACH

FOR DESCRIPTIVE SURVEYS, EXAMINE SAMPLE WEIGHTS (RECIPROCAL OF SELECTION PROBABILITIES -- WANT UNIFORM WEIGHTS, TO EXTENT POSSIBLE)

A MAJOR DIFFICULTY IN DETERMINING SAMPLE SIZES FOR COMPLEX SURVEYS IS THAT THE STANDARD APPROACH (OF SETTING THE DESIRED NUMERICAL VALUE OF AN ERROR BOUND EQUAL TO THE THEORETICAL (FORMULA) VALUE) INVOLVES QUANTITIES (VARIANCES) WHOSE VALUES ARE NOT KNOWN, EVEN APPROXIMATELY, PRIOR TO THE SURVEY. PRACTICAL METHODS WILL BE PRESENTED FOR DEALING WITH THIS PROBLEM.

AS MENTIONED IN DAY ONE (PAGE 58), A COMPUTER PROGRAM FOR DETERMINING SAMPLE SIZES IN SAMPLE SURVEYS (EITHER SIMPLE OR COMPLEX DESIGNS) IS POSTED

AT <http://www.foundationwebsite.org/JGCSampleSizeProgram.mdb> (A MICROSOFT ACCESS PROGRAM). IN USING THIS PROGRAM, THE "SAMPLE SIZE" IS USUALLY TAKEN TO BE THE NUMBER OF FIRST-STAGE UNITS. THE SAMPLE SIZE FOR SECOND-STAGE UNITS IS USUALLY DETERMINED BY THE VALUE OF THE INTRA-UNIT CORRELATION COEFFICIENT (WHICH VARIES ACCORDING TO THE VARIABLE OF INTEREST). (THIS WILL BE DISCUSSED LATER.) IN MOST APPLICATIONS, LITTLE IS KNOWN IN ADVANCE OF THE SURVEY ABOUT THE VALUES OF MEANS OR VARIANCES, AND THE SAMPLE SIZE IS DETERMINED ACCORDING TO EDUCATED GUESSES ABOUT CORRELATIONS AND STANDARDIZED UNITS. TO AVOID THE NECESSITY OF SPECIFYING THE VARIANCE, THE SAMPLE SIZE MAY BE DETERMINED FOR SAMPLING FOR PROPORTIONS (IN WHICH CASE THE VARIANCE IS A FUNCTION OF THE MEAN), OR BY SPECIFYING THE SIZE OF CONFIDENCE INTERVALS OR DIFFERENCES TO BE DETECTED RELATIVE TO (IN UNITS OF) THE STANDARD DEVIATION.

## 8. SPECIFY SAMPLE SELECTION PROCEDURES

- WITH OR WITHOUT REPLACEMENT (FIRST AND SECOND STAGE)
- PPS (PPMS) OR NON PPS
- SYSTEMATIC SAMPLING
- SELECTION TO ENABLE VARIANCE ESTIMATION
- CONTROLLED SELECTION

9. SPECIFY FIELD PROCEDURES

- NUMBER AND SPACING OF QUESTIONNAIRE WAVES
- LETTERS OF ENDORSEMENT
- CLEARANCES
- NATURE OF INITIAL CONTACT; CALL BACKS
- INCENTIVES FOR RESPONDENTS; FOR WORKERS; AUDITS
- FIELD EDIT, RECONTACT
- TRANSMITTAL TO CENTRAL PROCESSING FACILITY

10. SPECIFY DATA PROCESSING PROCEDURES

- LOGGING, MANUAL EDIT, CODING
- KEYING, MACHINE EDIT
- DATA BASE DESIGN (AUDIT TRAIL, UPDATE FILE)
- TREATMENT OF MISSING VALUES (CODES, IMPUTED VALUES)
- HANDLING OF NONRESPONSE SUBSAMPLE
- DATA BASE DOCUMENTATION

## 11. DEVELOP DATA ANALYSIS PLAN

### PRELIMINARY ANALYSIS

- NONRESPONSE COUNTS
- TREATMENT OF NONRESPONSE (IMPUTATION)
- FREQUENCY COUNTS, MEANS, MINIMA, MAXIMA, RANGES, STANDARD DEVIATIONS (FOR ALL VARIABLES); PEARSON CORRELATION MATRIX FOR CONTINUOUS VARIABLES
- FORM DUMMY VARIABLES FOR ALL NONORDINAL (NOMINAL) DATA, COMPUTE CRAMER CORRELATION MATRIX FOR ALL VARIABLES

### DIRECTED ANALYSIS

- COMPUTE ESTIMATES OF INTEREST (MEANS, PERCENTAGES, TOTALS, SUBPOPULATION ESTIMATES)
- COMPUTE CROSSTABS OF INTEREST, WITH TESTS OF SIGNIFICANCE
- PERFORM TESTS OF HYPOTHESES

## 12. REPORT PREPARATION

- ESTIMATES: TABLES AND CHARTS
- COMMENTARY
- GENERALIZED VARIANCES
- CHARACTERIZATION OF NONRESPONSE
- DISCUSSION OF POTENTIAL SOURCES OF BIAS (SELECTION, MEASUREMENT, NONRESPONSE, PROCESSING, ESTIMATION PROCEDURE)
- DISCUSSION OF EXTERNAL VALIDITY

### III. WHEN AND HOW TO USE SIMPLE RANDOM SAMPLING

#### 1. NATURE OF SITUATION WHICH WARRANTS USE OF SIMPLE RANDOM SAMPLING

- ESTIMATES WANTED FOR TOTAL POPULATION OR LARGE SUBPOPULATIONS
- NO MAJOR COST DIFFERENCES IN SAMPLING VARIOUS CLASSES OF SAMPLE UNITS
- RELATIVELY HOMOGENEOUS POPULATION, OR NO AUXILIARY INFORMATION
- NO COST SAVINGS IN SAMPLING "NEARBY" UNITS OR OTHER NATURAL CLUSTERS OF THE POPULATION (I.E., CLUSTER SAMPLING WOULDN'T SAVE MONEY)
- SAMPLING IS INEXPENSIVE (E,G,, COMPUTER FILES)
- SAMPLE FRAME IS AVAILABLE FOR ENTIRE POPULATION
- LIMITED ANALYSIS CAPABILITY

METHODS FOR DETERMINING SAMPLE SIZES FOR SIMPLE RANDOM SAMPLING WERE DISCUSSED IN DAY ONE OF THE COURSE.

## 2. HOW TO SELECT A SIMPLE RANDOM SAMPLE (WITH REPLACEMENT)

- REPRODUCIBLE METHODS
  - TABLE OF RANDOM NUMBERS
  - COMPUTER-GENERATED RANDOM NUMBERS (WITH SEED) ( $Nu_i$ ,  $i=1,2,\dots,N$ )
- NONREPRODUCIBLE METHODS
  - ROLLS, SPINS, SHUFFLING (WITH RESHUFFLE AFTER EACH DRAW)
  - COMPUTER-GENERATED RANDOM NUMBERS (WITHOUT SEED)
- OTHER METHODS (QUESTIONABLE)
- LAST DIGITS OF SOCIAL SECURITY NUMBER

## 3. SAMPLING WITHOUT REPLACEMENT

- FOR SMALL POPULATIONS, WILL IMPROVE PRECISION (FPC)
- COMPUTATION OF VARIANCE IS COMPLICATED

## 4. HOW TO SELECT A SIMPLE RANDOM SAMPLE WITHOUT REPLACEMENT

- REPRODUCIBLE METHODS
  - TABLE OF RANDOM NUMBERS
  - RANDOM PERMUTATIONS
  - COMPUTER-GENERATED RANDOM NUMBERS
- NONREPRODUCIBLE METHODS
  - SHUFFLING, ROLLS, SPINS
- SYSTEMATIC SAMPLING OF RANDOMLY ORDERED LIST

#### IV. WHEN AND HOW TO USE STRATIFICATION

##### 1. NATURE OF SITUATION WHICH WARRANTS USE OF STRATIFIED SAMPLING

- SUBPOPULATIONS OF INTEREST
- ADMINISTRATIVE DIFFERENCES
- COST REDUCTION
- VARIANCE REDUCTION
  - DEPENDENT VARIABLE CORRELATED WITH VARIABLE OF STRATIFICATION
  - STRATA ARE INTERNALLY HOMOGENEOUS
- HAVE AUXILIARY INFORMATION ON ALL OF POPULATION

##### 2. THE USE OF A "CERTAINTY" STRATUM

- POLITICAL REASONS
- COST REASONS
- (IN TWO-STAGE SAMPLING -- DISCUSSED LATER)
- NO PROBLEM IN ASSIGNING UNIT SELECTION PROBABILITIES BEFORE SELECTING THE SAMPLE

##### 3. ERRORS IN CLASSIFICATION

- LEAVE THE WRONG UNITS IN THE WRONG STRATA
- SLIGHT DECREASE IN EFFICIENCY
- CORRECTING THE SAMPLE CAN LEAD TO BIAS

4. HOW TO DETERMINE THE NUMBER OF STRATA AND THE STRATUM BOUNDARIES

- CASE 1: STRATIFICATION FOR SUBPOPULATIONS OF INTEREST
  - COMPARE VARIANCE OF ESTIMATED MEAN WITH STRATIFICATION TO VARIANCE WITHOUT STRATIFICATION (DESIGN EFFECT)
  - ASSESS TRADEOFF OF PRECISION OF ESTIMATES FOR TOTAL POPULATION FOR PRECISION OF ESTIMATES OF SUBPOPULATIONS
- CASE 2: STRATIFICATION FOR PRECISION IMPROVEMENT
  - NUMBER OF STRATA:
    - EMPIRICAL EVIDENCE: 2-6 CATEGORIES
  - STRATUM BOUNDARIES:
    - SUPPOSE X IS A KNOWN VARIABLE RELATED TO THE VARIABLE OF INTEREST Y.

CATEGORY	$f(x)$	$\sqrt{f(x)}$	CUM $\sqrt{f(x)}$
0-5	.4	.63	.63
5-10	.2	.45	1.08
...			
95-100	<u>.05</u>	.22	5.67
	1.00		

SET THE STRATUM BOUNDARIES TO CREATE EQUAL INTERVALS ON THE CUM $\sqrt{f(x)}$ .

## 5. POSTSTRATIFICATION

STRATIFICATION: STRATUM SIZES KNOWN IN ADVANCE OF SAMPLING, FIXED SAMPLE SIZE PER STRATUM

POSTSTRATIFICATION: STRATA DETERMINED AFTER SAMPLE IS SELECTED; STRATUM SIZES KNOWN, SAMPLE SIZE PER STRATUM IS RANDOM VARIABLE

GAINS IN PRECISION DUE TO POSTSTRATIFICATION COMPARABLE TO THOSE FROM STRATIFICATION WITH SAME EXPECTED STRATUM SAMPLE SIZE. SLIGHT LOSS IN PRECISION BECAUSE OF UNKNOWN SAMPLE SIZE.

## 6. STRATIFICATION WITH RANDOM QUOTAS

USUAL PROCEDURE: SORT UNITS INTO STRATA, DRAW SAMPLES  
EQUIVALENT PROCEDURE: SELECT SAMPLE, SORT INTO STRATA, RESELECT UNTIL STRATUM SIZES MET

## 7. STRATIFICATION TO THE LIMIT

IF SEVERAL VARIABLES OF INTEREST (SUBPOPULATIONS, EXPLANATORY VARIABLES), DESIRE TO CROSS-STRATIFY

		Public/Private			
		Urban	Rural	Urban	Rural
School Size	Very Small				
	Small				
	Medium				
	Large				

NUMBER OF CELLS MULTIPLIES QUICKLY (E.G., 5 VARIABLES OF STRATIFICATION, EACH AT 3 LEVELS, IMPLIES  $3^5 = 243$  CELLS)

UNLESS DESIRE TO APPLY SPECIAL DESIGN PROCEDURES, NEED AT LEAST TWO UNITS PER STRATUM TO ESTIMATE VARIANCE.

IF STRATIFY TO THE LIMIT OF ONE UNIT PER STRATUM, MUST "COLLAPSE" STRATA TO ESTIMATE VARIANCE (BIASED ESTIMATE)

8. MULTIPLE STRATIFICATION (DEEP STRATIFICATION, TWO-WAY STRATIFICATION, CROSS-STRATIFICATION)

STRATIFY ON VARIABLES A AND B

		B										R		
		1												
A	1	x												3
						x								
												x		
					x									3
									x					
										x				
								x						2
												x		
						x								2
											x			
			x										2	
C							x							
		2		2		2		2		2		2	n=12	

R = NUMBER OF ROWS = 6  
 C = NUMBER OF COLUMNS = 5  
 $R \times C = 6 \times 5 = 30$   
 n = SAMPLE SIZE = 12

CASE 1: STRATIFICATION ALONG MARGINALS (NOT IN CELLS)

$n > \max(R,C)$

AT LEAST 1 UNIT PER ROW  
 AT LEAST 1 UNIT PER COLUMN  
 NOT NECESSARILY 1 UNIT PER CELL  
 SPECIFY MARGINAL COUNTS, USE PROBABILITY SELECTION OF CELLS ...

CASE 2: CONTROLLED SELECTION: LATIN SQUARE DESIGN

$$n=R=C$$

		Control Classes		
		1	2	3
Strata	I	Aa	Bb	Cc
	II	Cb	Ac	Ba
	III	Bc	Ca	Ab

SIX PATTERNS OF CONTROLLED SELECTION: A, B, C, a, b,c

ASSIGN A PROBABILITY TO EACH PATTERN (E.G., 1/3), AND SELECT A PATTERN  
(NON-INDEPENDENT SAMPLING)

CASE 3: CONTROLLED SELECTION: GENERAL CASE

$$n < R, n < C$$

DEFINE A SERIES OF DESIRABLE PATTERNS SUCH THAT THE COLLECTION OF PATTERNS IN TOTO "COVER" ALL CELLS.

ASSIGN PROBABILITIES TO THE PATTERNS SUCH THAT THE EXPECTED MARGINAL TOTALS ARE AS DESIRED, AND THE PROBABILITIES OF SELECTION OF THE INDIVIDUAL ELEMENTS ARE AS UNIFORM AS POSSIBLE.

SELECT A PATTERN.

MULTIPLE STRATIFICATION: USEFUL IF SAMPLE SIZE IS SMALL (E.G., SAMPLE OF CLUSTERS), OR LARGE NUMBER OF VARIABLES OF STRATIFICATION.

#### CASE 4: OPTIMIZATION TECHNIQUES

SPECIFY A SERIES OF CONSTRAINTS, SUCH AS CONSTRAINTS ON THE PRECISION OF VARIOUS ESTIMATES, OR ON COST, OR ON STRATUM SIZES, AND DETERMINE AN ALLOCATION OVER THE STRATA THAT MINIMIZES OR MAXIMIZES A SPECIFIED OBJECTIVE FUNCTION, SUBJECT TO THESE CONSTRAINTS.

#### EXAMPLE:

##### CONSTRAINTS:

VARIANCE OF POPULATION MEAN  $\leq A1$

VARIANCE OF STRATUM MEANS  $\leq A2$

SAMPLE ALL UNITS IN TEXAS

##### OBJECTIVE:

DETERMINE SAMPLE ALLOCATION THAT MINIMIZES TOTAL COST, SUBJECT TO ABOVE CONSTRAINTS

##### PROCEDURE:

VARIOUS OPTIMIZATION METHODS, E.G., LAGRANGE MULTIPLIERS (EXAMPLE: NEYMAN ALLOCATION TO STRATA)

9. HOW TO ALLOCATE SAMPLE SIZES TO STRATA, WHEN COSTS AND VARIANCES ARE KNOWN

$$\text{COST} = C = c_o + \sum_h c_h n_h$$

OPTIMAL ALLOCATION:

$$\frac{n_h}{n} = \frac{N_h S_h / \sqrt{c_h}}{\sum N_h S_h / \sqrt{c_h}}$$

TAKE A LARGER SAMPLE IF:

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

10. HOW TO ALLOCATE SAMPLE SIZES TO STRATA, WHEN COSTS AND VARIANCES ARE UNKNOWN

- ASSUME COSTS SAME
- ASSUME VARIANCES SAME
- REVIEW VARIANCES OF ESTIMATES OF INTEREST FOR:
  - EQUAL ALLOCATION
  - PROPORTIONAL ALLOCATION (SELF-WEIGHTING)

IN A SELF-WEIGHTING DESIGN, THE SAMPLE MEAN IS AN UNBIASED ESTIMATE OF POPULATION MEAN:

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

$$\frac{n_h}{N_h} = \frac{n}{N} \text{ (THE SAME FOR ALL STRATA)}$$

$$y_{st} = \frac{\sum_i y_i}{n} \text{ (THE AVERAGE OF ALL OBSERVATIONS, REGARDLESS OF STRATUM)}$$

## 11. SELECTION OF VARIABLES OF STRATIFICATION

- COARSE DIVISIONS FOR SEVERAL VARIABLES PREFERABLE TO FINE DIVISIONS FOR A FEW VARIABLES
- NO NEED TO CROSS-CLASSIFY; COLLAPSE UNIMPORTANT CELLS ("NESTED" STRATIFICATION)
- STRATIFY ON VARIABLES UNRELATED TO EACH OTHER
- QUALITATIVE/SUBJECTIVE VARIABLES: CAN BE USED FOR STRATIFICATION~ BUT NOT IN ESTIMATION PROCEDURE
- AVOID STRATIFICATION PAST THE LIMIT OF TWO UNITS PER CELL

## V. WHEN AND HOW TO USE CLUSTER SAMPLING

### 1. NATURE OF SITUATIONS WHICH WARRANT USE OF CLUSTER SAMPLING

- NATURAL CLUSTERS OF THE POPULATION (HOUSEHOLDS, SCHOOLS, SMSAs)
- CONFINING SAMPLING TO NEARBY UNITS PRODUCES LARGE COST SAVINGS
- NO FRAME AVAILABLE FOR ALL UNITS, BUT COULD CONSTRUCT FRAME FOR A FEW CLUSTERS
- ELEMENTS WITHIN CLUSTERS ARE NOT HIGHLY "SIMILAR" WITH RESPECT TO VARIABLES OF INTEREST.

### 2. THE "CLUSTER EFFECT"

ELEMENTS WITHIN CLUSTER ARE USUALLY POSITIVELY CORRELATED  
POSITIVE INTRACLUSTER CORRELATION COEFFICIENT ( $\rho$ ) REDUCES  
PRECISION OVER THAT OF A SIMPLE RANDOM SAMPLE OF THE SAME SIZE

$$S_c^2 = S^2(1 + (m-1)\rho)$$

COST SAVING OFFSETS LOSS IN PRECISION

(NOTE: IF  $\rho$  IS NEGATIVE), CLUSTERING IMPROVES PRECISION)

### 3. DETERMINING SAMPLE SIZES IN CLUSTER SAMPLING (EQUAL-SIZE CLUSTERS)

COST FUNCTION:

$$C = c_1 Mn + c_2 \sqrt{n}$$

WHERE

M = CLUSTER SIZE

n = SAMPLE SIZE

VARIANCE FUNCTION:

$$V(\bar{y}) = \frac{S^2 - (M - 1)AM^{g-1}}{n}$$

WHERE

$$S_w^2 = AM^g \quad (g > 0)$$

IS THE WITHIN-CLUSTER VARIANCE.

WE COULD SOLVE FOR THE VALUE OF M (E.G., TO MINIMIZE THE VARIANCE, SUBJECT TO A FIXED COST), BUT WILL NOT (SEE COCHRAN'S BOOK FOR THIS). THE ABOVE VARIANCE FUNCTION IS RARELY EVER KNOWN. THIS FORMULATION WAS DEVELOPED SIMPLY TO PROVIDE GUIDANCE IN DETERMINING THIS CLUSTER SIZE. IN FACT, IN MOST SOCIOECONOMIC SURVEYS, THE CLUSTER SIZE IS NOT DETERMINED BY THE SURVEY DESIGNER, BUT IS TAKEN AS AN EXISTING ADMINISTRATIVE UNIT (E.G., CENSUS ENUMERATION AREA, DISTRICT, VILLAGE, SCHOOL). BASED ON THIS MODEL, THE FOLLOWING RECOMMENDATIONS ARE MADE.

THE OPTIMAL CLUSTER SIZE (M) BECOMES SMALLER WHEN:

- LENGTH OF INTERVIEW INCREASES
- TRAVEL BECOMES CHEAPER
- THE ELEMENTS BECOME MORE DENSE (CLUSTERS CLOSER TOGETHER)
- THE TOTAL AMOUNT OF MONEY (C) INCREASES
- THE CLUSTERS ARE INTERNALLY HOMOGENEOUS (THE WITHIN-CLUSTER VARIANCE IS SMALL RELATIVE TO THE POPULATION VARIANCE)

ALTHOUGH THE ABOVE FORMULATION IS NOT USUALLY USED TO DETERMINE M, IT CAN BE USED TO DETERMINE THE VALUE OF n (THE NUMBER OF CLUSTERS OF SIZE M TO SELECT. THE VALUE OF n IS THE SOLUTION OF THE FOLLOWING EQUATION:

$$2c_1 M \sqrt{n} / c_2 = \sqrt{(1 + 4Cc_1 M / c_2^2)} - 1$$

4. VARIABLE-SIZE CLUSTERS: SAMPLING WITH PROBABILITIES PROPORTIONAL TO SIZE (PPS) (SAMPLING WITH REPLACEMENT)

- SIMPLE RANDOM SAMPLE OF CLUSTERS YIELDS ESTIMATE OF POOR PRECISION IF CLUSTERS VARY IN SIZE
- LARGE INCREASE IN PRECISION IF SELECT CLUSTERS WITH PROBABILITIES PROPORTIONAL TO SIZE
- FORMULAS FOR VARIANCES ARE SIMPLER

(NOTE: PPS IS NOT SELF-WEIGHTING)

5. VARIABLE-SIZE CLUSTERS: SAMPLING WITH PROBABILITIES PROPORTIONAL TO A MEASURE OF SIZE (PPMS)

IF WE DON'T KNOW CLUSTER SIZE EXACTLY, SET PROBABILITIES PROPORTIONAL TO A MEASURE OF SIZE

$$\hat{Y}_{PPES} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i}$$

## 6. VARIABLE-SIZE CLUSTERS: PPMS, WITHOUT REPLACEMENT

IN STRATIFIED CLUSTER SAMPLING, THERE MAY BE A SMALL NUMBER OF LARGE CLUSTERS IN A STRATUM, SO THAT FPC IS NOT NEGLIGIBLE

IF WE USE SAMPLING WITHOUT REPLACEMENT, FORMULAS FOR VARIANCE BECOME VERY COMPLICATED

USE THE RAO-HARTLEY-COCHRAN TECHNIQUE TO ENABLE VARIANCE ESTIMATION:

- ASSIGN UNITS TO GROUPS AT RANDOM, MAKING THE NUMBER OF UNITS PER GROUP AS NEARLY EQUAL AS POSSIBLE
- SELECT 1 UNIT FROM EACH GROUP, PPS

THE ESTIMATOR

$$\hat{Y}_{RHC} = \sum_{g=1}^n Z_g \frac{y_g}{z_g}$$

IS UNBIASED. ITS ESTIMATED VARIANCE IS:

$$v(\hat{Y}_{RHC}) = \frac{N^2 + k(n-k) - Nn}{N^2(n-1) - k(n-k)} \sum Z_g \left( \frac{y_g}{z_g} - \hat{Y}_{RHC} \right)^2$$

WHERE

$$N = Qn + k, \quad k < n$$

WHERE THERE ARE  $k$  GROUPS OF SIZE  $Q+1$  AND  $n-k$  GROUPS OF SIZE  $Q$ .

## 7. STRATIFICATION OF CLUSTERS

AS AN ALTERNATIVE TO PPS, WE CAN STRATIFY CLUSTERS BY SIZE, AND SELECT WITH EQUAL PROBABILITIES FROM WITHIN STRATA.

## 8. STRATIFICATION OF CLUSTERS: THE USE OF A CERTAINTY STRATUM

IN PPS SELECTION, THE USUAL PROCEDURE IS AS FOLLOWS:

<u>UNIT</u>	<u>SIZE</u> $M_i$	<u>CUM. SIZE</u> $\Sigma M_i$	<u>ASSIGNED</u> <u>RANGE</u>	<u>RANDOM</u> <u>NUMBER</u>
1	4	4	1-4	
2	6	10	5-10	x
...				
N	3	372	369-372	x

WITH REPLACEMENT: NO PROBLEM

WITHOUT REPLACEMENT: IF SOME CLUSTERS ARE LARGE, OR THE NUMBER OF CLUSTERS IS SMALL IT MAY NOT BE POSSIBLE TO IMPLEMENT PPS.

SOLUTION: DEFINE SKIP INTERVAL,  $k$ . DEFINE THE CRITICAL SIZE AS

$$s = \frac{2}{3}k$$

PLACE ALL CLUSTERS OF SIZE GREATER THAN  $s$  IN A CERTAINTY STRATUM. APPLY PPS TO NONCERTAINTY STRATUM. USE PROPORTIONAL SAMPLING FROM CERTAINTY CLUSTERS.

## 9. GENERAL RECOMMENDATIONS REGARDING CLUSTER SAMPLING

- PPS OR STRATIFY BY SIZE (AND SELECT WITH EQUAL PROBABILITIES)
- CERTAINTY STRATUM FOR LARGE CLUSTERS
- SEVERAL ESTIMATES OF INTEREST (CHECK TEXTBOOK FOR FORMULAS)
- BE CAREFUL – IF CLUSTERS ARE HIGHLY INTERNALLY HOMOGENEOUS, THE "EFFECTIVE" SAMPLE SIZE EQUALS NUMBER OF CLUSTERS, NOT TOTAL NUMBER OF UNITS
- USUAL CROSSTABS, TESTS OF SIGNIFICANCE NOT APPLICABLE

IN CLUSTER SAMPLING, A KEY PARAMETER IS THE INTRACLUSTER CORRELATION COEFFICIENT ("ICC"). IT IS THE CORRELATION COEFFICIENT BETWEEN PAIRS OF UNITS WITHIN THE SAME CLUSTER, DEFINED AS

$$\rho = E(y_{ij} - \mu)(y_{ik} - \mu) / E(y_{ij} - \mu)^2$$

WHERE THE NUMERATOR IS AVERAGED OVER ALL DISTINCT PAIRS OF ELEMENTS (SUBUNITS) WITHIN THE SAME CLUSTER, AND THE DENOMINATOR IS AVERAGED OVER ALL ELEMENTS. THE ICC IS A MEASURE OF THE INTERNAL HOMOGENEITY OF CLUSTERS.

IN CLUSTER SAMPLING THERE ARE TWO MEANS OF INTEREST – THE MEAN PER UNIT (CLUSTER) AND THE MEAN PER ELEMENT. THE MEAN PER UNIT IS THE MEAN OF THE CLUSTER TOTALS,  $\Sigma y_i / N$ , AND THE MEAN PER ELEMENT IS THE MEAN OF THE ELEMENTS,  $\mu = \Sigma y_i / NM$ .

AN APPROXIMATE EXPRESSION FOR THE VARIANCE OF THE SAMPLE MEAN PER ELEMENT IS

$$\text{var (sample mean per element)} = [(1 - f) S^2 / n] (1 + (M - 1) \rho)$$

THE FIRST TERM, IN BRACKETS, IS THE VARIANCE OF THE SAMPLE MEAN IN SIMPLE RANDOM SAMPLING (SRS). THE SECOND TERM,  $1 + (M - 1) \rho$ , IS A FACTOR THAT SHOWS HOW MUCH THE VARIANCE OF THE SAMPLE MEAN CHANGES IN CLUSTER SAMPLING, FROM THAT FOR SRS. THE TERM  $1 + (M - 1) \rho$  IS KISH'S "DESIGN EFFECT," OR "DEFF" FOR SAMPLING CLUSTERS OF SIZE M.

THIS FACTOR,  $\text{DEFF} = 1 + (m - 1)\rho$ , INDICATES HOW MUCH THE VARIANCE OF THE SAMPLE MEAN DIFFERS IN CLUSTER SAMPLING FROM THE VARIANCE IN SIMPLE RANDOM SAMPLING. SINCE THE ESTIMATED SAMPLE SIZE FOR SRS IS PROPORTIONAL TO THE VARIANCE, DEFF ALSO INDICATES HOW MUCH THE SAMPLE SIZE OF CLUSTERS MUST BE INCREASED TO ACHIEVE THE SAME PRECISION AS A SRS OF THE SAME SIZE. *THIS IS THE PRINCIPAL WAY IN WHICH THE SAMPLE SIZE IS ESTIMATED FOR CLUSTER SAMPLING (I.E., ESTIMATE THE SAMPLE SIZE FOR SRS AND MULTIPLY BY THE DEFF.* SINCE M IS TYPICALLY GIVEN, ALL THAT IS NEEDED TO KNOW DEFF IS THE VALUE OF  $\rho$ . IF CLUSTERS ARE HIGHLY INTERNALLY HOMOGENEOUS, USE A LARGE VALUE OF  $\rho$ , SUCH AS .5 TO .9. IF THE ELEMENTS OF A CLUSTER VARY ABOUT AS MUCH AS IN THE GENERAL POPULATION, USE A SMALL VALUE OF  $\rho$ , SUCH AS 0

- .3. NOTE THAT THE VALUE OF  $\rho$  GENERALLY DECREASES AS THE CLUSTER SIZE (M) INCREASES. NOTE ALSO THAT THE VALUE OF  $\rho$  DIFFERS FOR EVERY VARIABLE OF INTEREST (I.E., FOR WHICH THE MEAN IS TO BE ESTIMATED). USE THE VALUE OF  $\rho$  CORRESPONDING TO THE MOST IMPORTANT VARIABLES OF INTEREST.

FOR MOST APPLICATIONS,  $\rho$  IS POSITIVE, SO THE FACTOR IS POSITIVE. (IT CAN BE NEGATIVE ONLY IF M IS VERY SMALL (FOR EXAMPLE, GENDER IN TWO-PERSON HOUSEHOLDS).) IF  $S^2$  DENOTES THE POPULATION VARIANCE,  $S_1^2$  DENOTES THE VARIANCE AMONG CLUSTER MEANS, AND  $S_2^2$  DENOTES THE WITHIN-CLUSTER VARIANCE, THEN THE FOLLOWING RELATIONSHIPS ARE APPROXIMATE:

$$\rho = (MS_1^2 - S^2) / [(M - 1)S^2]; S_2^2 = S^2(1 - \rho); (1 - \rho) / \rho = S_2^2 / (S_1^2 - S_2^2/M).$$

IF M IS LARGE, THE FOLLOWING RELATIONSHIPS ARE APPROXIMATE:

$$S^2 = S_1^2 + S_2^2; S_1^2 = \rho S^2; S_2^2 = (1 - \rho) S^2.$$

## VI. WHEN AND HOW TO USE SYSTEMATIC SAMPLING

### 1. REASONS FOR USING SYSTEMATIC SAMPLING

- TO SELECT A SIMPLE RANDOM SAMPLE (FROM A RANDOMLY ORDERED LIST)
- FOR VARIANCE REDUCTION (SAMPLING FROM TREND DATA)
- TO SELECT A SAMPLE QUICKLY, WHEN IT DOESN'T MATTER IF WE CAN COMPUTE THE VARIANCE (E.G., SECOND-STAGE SAMPLING WHEN THE SAMPLING FRACTION ( $n/N$ ) OF THE FIRST-STAGE UNITS IS SMALL)
- TO SELECT A SAMPLE QUICKLY IN GENERAL (BEST TO SELECT SEVERAL REPLICATED SAMPLES IN ORDER TO BE ABLE TO COMPUTE THE VARIANCE)

### 2. NATURE OF SITUATION WHICH WARRANTS USE OF SYSTEMATIC SAMPLING

- MANUAL RECORD SYSTEM (PHYSICAL LIST, CARD FILES)
- FILES IN RANDOM ORDER
- LIMITED TIME OR RESOURCES FOR SELECTING SAMPLE
- NO PERIODICITIES SUSPECTED IN DATA

### 3. HOW TO SELECT A SYSTEMATIC SAMPLE

- EVERY k-th, WITH RANDOM START
- EVERY k-th, WITH SEVERAL RANDOM STARTS
- INTEGER SAMPLING INTERVAL (k)
- NON-INTEGERS SAMPLING INTERVAL

## VII. WHEN AND HOW TO USE MULTISTAGE SAMPLING (TWO-STAGE)

### 1. NATURE OF SITUATION WHICH WARRANTS USE OF A MULTISTAGE DESIGN

AS FOR CLUSTER SAMPLING -- EXCEPT THAT IT IS IMPRACTICAL OR INEFFICIENT TO SAMPLE ALL OF THE CLUSTER (LARGE CLUSTER SIZE, LARGE INTRACLUSTER CORRELATION COEFFICIENT)

EXAMPLE: SCHOOLS; SMSAs; HOSPITALS; CLINICS

## 2. ESTIMATED MEAN IN TWO-STAGE SAMPLING (EQUAL-SIZED PRIMARY UNITS, WITHOUT REPLACEMENT)

### NOTATION AND FORMULAS:

$y_{ij}$  = VALUE FOR THE j-th ELEMENT IN THE i-th PRIMARY UNIT

$$\bar{y}_i = \frac{\sum_{j=1}^m y_{ij}}{m} = \text{SAMPLE MEAN PER ELEMENT IN THE } i\text{-th PRIMARY UNIT}$$

$$\bar{\bar{y}} = \frac{\sum_{i=1}^n \bar{y}_i}{n} = \text{OVERALL SAMPLE MEAN PER ELEMENT}$$

$S_1^2$  = VARIANCE AMONG PRIMARY UNIT MEANS

$S_2^2$  = VARIANCE AMONG ELEMENTS WITHIN PRIMARY UNITS

$\bar{\bar{y}}$  IS AN UNBIASED ESTIMATE OF THE POPULATION MEAN PER ELEMENT

$$V(\bar{\bar{y}}) = \frac{N-n}{N} \frac{S_1^2}{n} + \frac{M-m}{M} \frac{S_2^2}{mn}$$

$$v(\bar{\bar{y}}) = \frac{1-f_1}{n} \hat{S}_1^2 + \frac{f_1(1-f_2)}{mn} \hat{S}_2^2$$

NOTE: IF  $f_1 = n/N$  IS SMALL, WE CAN USE SYSTEMATIC SAMPLING IN THE SECOND-STAGE UNITS, SINCE WE DON'T NEED TO BE ABLE TO ESTIMATE  $S_2^2$ .

### 3. OPTIMAL SAMPLING AND SUBSAMPLING FRACTIONS (EQUAL-SIZED PRIMARY UNITS)

$$C = c_1n + c_2nm$$

$$V(\bar{y}) = \frac{1}{n}(S_1^2 - \frac{S_2^2}{M}) + \frac{1}{mn}S_2^2 - \frac{1}{N}S_1^2$$

THE OPTIMAL SECOND-STAGE SAMPLE SIZE IS

$$m_{opt} = \frac{S^2}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{\frac{c_1}{c_2}}$$

THE VALUE OF  $n$  IS DETERMINED BY SETTING EITHER THE COST OR THE VARIANCE, AND SOLVING THE RESPECTIVE EQUATION FOR  $n$ .

IF  $\rho$  DENOTES THE INTRACLUSTER CORRELATION COEFFICIENT (DEFINED EARLIER, IN THE CLUSTER-SAMPLING SECTION), THE FOLLOWING APPROXIMATION HOLDS, FOR  $\rho$  NOT EQUAL TO ZERO:

$$m_{opt} = \frac{S^2}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{\frac{c_1}{c_2}} \approx \sqrt{\frac{c^1}{c^2} \frac{1-\rho}{\rho}}$$

NOTE: IF  $\rho$  IS SMALL,  $m_{opt}$  IS LARGE; IF  $\rho$  IS LARGE,  $m_{opt}$  IS SMALL.

THE COMMENTS MADE ABOUT  $\rho$  IN CLUSTER SAMPLING GENERALLY APPLY TO MULTISTAGE SAMPLING.

THE EXPRESSION INVOLVING  $\rho$  IS USED BECAUSE THE VARIANCES ARE USUALLY NOT KNOWN, BUT  $\rho$  CAN BE APPROXIMATED. MOST SURVEYS ARE CONCERNED WITH COLLECTING DATA ON A NUMBER OF VARIABLES, AND THE VALUE OF  $\rho$  IS DIFFERENT FOR EACH OF THEM. THE FIRST-STAGE SAMPLE UNITS (PRIMARY SAMPLING UNITS, OR PSUs) FOR A SURVEY ARE OFTEN DETERMINED BY ADMINISTRATIVE CONVENIENCE, E.G., AS CENSUS ENUMERATION AREAS, VILLAGES, OR DISTRICTS, FOR WHICH DATA EXIST TO FACILITATE SAMPLE DESIGN. THE OPTIMAL WITHIN-PSU SAMPLE SIZE IS DETERMINED BY CONSIDERING HOW HOMOGENEOUS THE PSUs ARE, ON AVERAGE, OVER ALL IMPORTANT VARIABLES. FOR EXAMPLE, IF THE PSUs ARE HIGHLY INTERNALLY HOMOGENEOUS, USE A LARGE VALUE FOR  $\rho$  (E.G.,  $\rho = .5$  TO  $.9$ ), AND IF THE PSUs ARE NOT HIGHLY INTERNALLY HOMOGENEOUS, USE A SMALL VALUE FOR  $\rho$  (E.G.,  $\rho = 0$  TO  $.3$ ). IN MANY SOCIO-ECONOMIC SURVEYS, THE VALUE OF  $m$  IS IN THE RANGE 15-30 (E.G., A SAMPLE OF 15 HOUSEHOLDS IS SELECTED FROM EACH VILLAGE (PSU)). IN MANY APPLICATION, THE OPTIMUM IS "FLAT" (I.E., IS NOT HIGHLY SENSITIVE TO CHANGES IN THE VALUE OF  $\rho$ .)

ONCE THE VALUE OF  $m$  (WITHIN-FIRST-STAGE-UNIT SAMPLE SIZE) IS DETERMINED, IT REMAINS TO DETERMINE THE VALUE NUMBER,  $n$ , OF UNITS TO

SELECT. THE PROCEDURE FOR DOING THIS IS THE SAME AS FOR CLUSTER SAMPLING. THE VARIANCE EXPRESSION PRESENTED ABOVE IS NOT USEFUL FOR ESTIMATING THE SAMPLE SIZE, SINCE THE VARIANCES ARE TYPICALLY NOT KNOWN PRIOR TO THE SURVEY.

IN TERMS OF  $\rho$ , AN APPROXIMATE EXPRESSION FOR THE VARIANCE, FOR N AND M LARGE, IS

$$V(\bar{y}) \approx \frac{1}{nm} S^2 [1 + (m-1)\rho]$$

WHERE  $S^2$  DENOTES THE POPULATION VARIANCE. AS IN THE CASE OF CLUSTER SAMPLING, THE TERM PRECEDING THE TERM IN BRACKETS IS THE VARIANCE FOR THE SAMPLE MEAN OF A SIMPLE RANDOM SAMPLE. THE TERM IN BRACKETS,  $1 + (m-1)\rho$ , IS KISH'S DEFF. AS IN THE CASE OF CLUSTER SAMPLING, THE NUMBER OF FIRST-STAGE UNITS TO SELECT IS OBTAINED BY DETERMINING THE SAMPLE SIZE (FOR A SPECIFIED LEVEL OF PRECISION) FOR SIMPLE RANDOM SAMPLING AND MULTIPLYING BY THE VALUE OF DEFF.

#### 4. UNEQUAL-SIZED PRIMARY UNITS: SAMPLING WITH EQUAL PROBABILITIES

IF STRATIFY PRIMARY UNITS BY SIZE, USE SELECTION WITH EQUAL PROBABILITIES

UNBIASED ESTIMATE:

$$\bar{Y}_u = \frac{N}{nM_0} \sum_{i=1}^n M_i y_i = \frac{1}{n\bar{M}} \sum_{i=1}^n M_i y_i$$

THIS ESTIMATE IS SELF-WEIGHTING IF  $f_{2i} = \frac{m_i}{M_i} = \text{CONSTANT} = f_2$

#### 5. UNEQUAL-SIZED PRIMARY UNITS: PPS OR PPES SAMPLING

IF PRIMARY UNITS ARE NOT STRATIFIED BY SIZE, USE PPS OR PPES SAMPLING

UNBIASED ESTIMATE FOR PPES:

$$Y_{PPES} = \frac{1}{nM_0} \sum_{i=1}^n \frac{M_i \bar{y}_i}{z_i}$$

THIS ESTIMATOR IS SELF-WEIGHTING IF  $\frac{M_i}{z_i m_i} = \text{CONSTANT} = \frac{n}{f_0}$ ,

I.E., IF  $m_i = \frac{M_i}{z_i} \times \text{CONSTANT}$  THEN

$$\bar{Y}_{PPES} = \sum_i \sum_j y_{ij} / f_0 M_0$$

FOR PPS,  $z_i = \frac{M_i}{M_0}$ ; UNBIASED ESTIMATE IS:

$$\bar{Y}_{PPS} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

THIS ESTIMATOR IS SELF-WEIGHTING IF  $m_i = \text{CONSTANT} = m$ :

$$\bar{Y}_{PPS} = \sum_i \sum_j y_{ij} / nm$$

**6. UNEQUAL-SIZED PRIMARY UNITS: OPTIMAL SAMPLING AND SUBSAMPLING FRACTIONS, SAMPLING WITH EQUAL PROBABILITIES**

$$\text{COST} = C = c_1 n + c_2 \sum_{i=1}^n m_i + c_3 \sum_{i=1}^n M_i$$

(THE LAST TERM IS THE LISTING COST).

FROM THE EXPRESSIONS FOR THE VARIANCE, WE CAN SHOW:

$$m_{opt} = \frac{S^2}{\sqrt{S_1^2 - S_2^2 / \bar{m}}} \sqrt{\frac{c_1}{c_2}}$$

IF  $\rho_{\bar{m}}$  DENOTES THE AVERAGE VALUE OF THE INTRA-UNIT CORRELATION COEFFICIENT (DEFINED AS IN THE SECTION ON CLUSTER SAMPLING), THEN THE FOLLOWING APPROXIMATION HOLDS, FOR  $\rho_{\bar{m}}$  NOT EQUAL TO ZERO.

$$m_{opt} = \frac{S^2}{\sqrt{S_1^2 - S_2^2 / \bar{m}}} \sqrt{\frac{c_1}{c_2}} \approx \sqrt{\frac{c^1}{c^2} \frac{1 - \rho_{\bar{m}}}{\rho_{\bar{m}}}}$$

**7. UNEQUAL-SIZED PRIMARY UNITS: OPTIMAL SAMPLING AND SUBSAMPLING FRACTIONS, WITH PPES SAMPLING**

COST = AS ABOVE

$$z_i \propto M_i \sqrt{\frac{\rho_{\bar{m}} S_d^2}{c_1 + c_3 M_i}}$$

WHERE  $S_d^2$  IS THE VARIANCE AMONG ALL UNITS IN THE POPULATION.

- IF  $c_3 M_i$  IS SMALL AND  $\rho_{\bar{m}}$  IS CONSTANT, PPS IS BEST ( $z_i \sim M_i$ )
- IF  $c_3 M_i$  IS SMALL AND  $\rho_{\bar{m}}$  DECREASES WITH INCREASING  $M_i$ , THEN OPTIMAL PROBABILITIES LIE BETWEEN  $z_i \sim M_i$  AND  $z_i \sim \sqrt{M_i}$ .
- IF  $c_3 M_i$  IS LARGE,  $z_i$  IS BETWEEN  $\sqrt{M_i}$  AND A CONSTANT (EQUAL PROBABILITIES)
- IF  $c_1$  AND  $c_3 M_i$  ARE COMPARABLE,  $z_i \sim \sqrt{M_i}$  IS REASONABLE.

## 8. STRATIFICATION

- AS IN CLUSTER SAMPLING, IF THERE ARE SOME VERY LARGE UNITS, PLACE THEM IN A CERTAINTY STRATUM
- FOR SELF-WEIGHTING ESTIMATES, USE PROPORTIONAL SAMPLING FROM THE UNITS IN THE CERTAINTY STRATUM ( $\rho$  WILL LIKELY BE LOW SINCE THE UNITS ARE LARGE)
- FOR SELF-WEIGHTING ESTIMATE, USE PPS SAMPLING FROM THE UNITS IN THE NONCERTAINTY STRATUM, WITH SELECTION OF AN EQUAL NUMBER OF ELEMENTS FROM EACH

## 9. SELECTION WITH UNEQUAL PROBABILITIES WITHOUT REPLACEMENT

- FOR FEW PRIMARY UNITS, USE REPLACEMENT SAMPLING
- TO ESTIMATE VARIANCE, USE RAO-HARTLEY-COCHRAN TECHNIQUE:
  - ASSIGN UNITS TO GROUPS AT RANDOM, WITH NUMBER OF UNITS IN A GROUP AS NEARLY EQUAL AS POSSIBLE
  - SELECT ONE UNIT PER GROUP, PPS
  - UNBIASED ESTIMATE AVAILABLE FOR VARIANCE

## 10. GENERAL RECOMMENDATIONS REGARDING TWO-STAGE DESIGNS

IF MUST LIST WHOLE POPULATION TO IMPLEMENT PPS, FORGET IT (USE PPES OR EQUAL PROBABILITIES)

FORM CERTAINTY STRATUM OF LARGEST CLUSTERS (USE 2/3 k RULE, AS IN CLUSTER SAMPLING)

USE PROPORTIONAL SAMPLING FROM CERTAINTY CLUSTERS

FOR NONCERTAINTY STRATUM, USE EITHER:

- PPS, PLUS EQUAL NUMBER OF ELEMENTS PER UNIT; OR
- EQUAL PROBABILITIES, WITH PROPORTIONAL SAMPLING IN EACH UNIT

IF USE REPLACEMENT SAMPLING FOR PRIMARY UNITS, USE RHC METHOD OF SELECTION

IF  $f_1$  IS SMALL, CAN USE SYSTEMATIC SAMPLING IN SECOND-STAGE SELECTION (DON'T NEED ESTIMATE OF WITHIN-CLUSTER VARIANCE)

CONSIDER ALTERNATIVE ESTIMATES (UNBIASED, SAMPLE MEAN, RATIO-TO-SIZE)

USE SELF-WEIGHTING DESIGN WITHIN STRATA, OVERALL UNLESS LOSE TOO MUCH PRECISION

DETERMINING THE FIRST-STAGE-UNIT SAMPLE SIZE BY ESTIMATING THE SAMPLE SIZE FOR SIMPLE RANDOM SAMPLING AND MULTIPLYING THIS VALUE BY THE VALUE OF THE DESIGN EFFECT,  $DEFF$ , CORRESPONDING TO THE VALUE OF THE INTRA-UNIT CORRELATION COEFFICIENT,  $\rho$ , FOR THE MORE IMPORTANT SURVEY VARIABLES (FOR WHICH THE MEAN IS TO BE ESTIMATED).

## VIII. WHEN AND HOW TO USE DOUBLE SAMPLING

### 1. NATURE OF SITUATION WHICH WARRANTS USE OF DOUBLE SAMPLING

- NEED INFORMATION ON COSTS, VARIANCES, AUXILIARY VARIABLE (STRATIFICATION)
- SCREENING

### 2. DETERMINATION OF SAMPLE SIZE IN DOUBLE SAMPLING

$n'$  = FIRST PHASE SAMPLE SIZE

$n$  = SECOND PHASE SAMPLE SIZE

$$\text{COST} = C = nc_n = n'c_{n'}$$

OPTIMAL ALLOCATION:

$$\frac{n'}{n} = \sqrt{\frac{c_n V_{n'}}{c_{n'} V_n}}$$

WHERE

$V_n$  = WITHIN-STRATUM VARIANCE

$V_{n'}$  = BETWEEN-STRATUM VARIANCE

NOTES:

IF  $V_{n'}/V_n$  IS VERY LARGE, STRATIFICATION IS EFFECTIVE, AND IT PAYS TO HAVE A LARGE FIRST-PHASE SAMPLE, I.E.,  $n'/n$  IS LARGE.

IF  $c_n/c_{n'}$  IS VERY LARGE, THE FIRST-PHASE SAMPLE IS INEXPENSIVE, AND SO  $n'/n$  IS LARGE.

## IX. HOW TO RESOLVE CONFLICTING/MULTIPLE SURVEY DESIGN OBJECTIVES

- DETERMINE OPTIMAL ALLOCATIONS FOR VARIOUS OBJECTIVES, AND SELECT A GOOD COMPROMISE DESIGN.
- PLACE CONSTRAINTS ON THE VARIANCES OF KEY ESTIMATES, USE OPTIMIZATION METHODS TO DETERMINE AN ALLOCATION THAT SATISFIES THE CONSTRAINTS (E.G., NEYMAN ALLOCATION TO STRATA).
- MINIMIZE A LINEAR COMBINATION OF THE VARIANCES OF SEVERAL KEY ESTIMATES (NOT RECOMMENDED. IT IS DIFFICULT, AND MAY FAIL TO ADDRESS IMPORTANT CONSTRAINTS.)
- BEST APPROACH IS ITERATIVE: EXAMINE ALTERNATIVES. SEEK A DESIGN THAT SATISFIES ALL IMPORTANT CONSTRAINTS, EVEN THOUGH IT MAY NOT BE "OPTIMAL" WITH RESPECT TO A SINGLE OBJECTIVE FUNCTION.

## PART TWO: HOW TO DESIGN ANALYTICAL SURVEYS

### I. REVIEW OF REGRESSION ANALYSIS (GENERAL LINEAR STATISTICAL MODEL, INCLUDING EXPERIMENTAL-DESIGN AND QUASI-EXPERIMENTAL DESIGN MODELS)

REFERENCE: SHARON L. LOHR, *SAMPLING: DESIGN AND ANALYSIS* (DUXBURY PRESS, 1999)

#### 1. LINEAR REGRESSION MODEL (UNIVARIATE)

$$y_i | x_{1i}, x_{2i}, \dots, x_{mi} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_m x_{mi} + e_i$$

WHERE

$y_i$  = DEPENDENT (RESPONSE) VARIABLE

$x_{ij}$  = INDEPENDENT (EXPLANATORY) VARIABLE

$b_i$  = REGRESSION COEFFICIENT (PARAMETER)

$e_i$  = ERROR TERM (MEAN = 0, VARIANCE =  $\sigma^2$ )

(NOTE: IT IS COMMON TO WRITE REGRESSION EQUATIONS USING LOWER-CASE SYMBOLS FOR THE RANDOM VARIABLES, AND TO USE UPPER-CASE SYMBOLS FOR "CROSSPRODUCTS" MATRICES.)

ASSUMPTIONS:

THE  $e_i$ 's ARE UNCORRELATED WITH EACH OTHER AND WITH THE  $x$ 's, AND HAVE ZERO MEAN AND THE SAME VARIANCE.

**OPTIONAL:** IN MATRIX NOTATION:

$$\underline{y} = X' \underline{b} + \underline{e}$$

WHERE

$X$  = OBSERVATION MATRIX (DESIGN MATRIX)

$\underline{y}$  = VECTOR OF OBSERVED  $y$ 's

$\underline{b}$  = VECTOR OF PARAMETERS

$\underline{e}$  = ERROR VECTOR

LEAST SQUARES ESTIMATE OF THE b's:

$$\underline{b} = (XX') * X \underline{y}$$

WHERE

$XX'$  = CROSSPRODUCTS MATRIX

$(XX')^*$  = GENERALIZED (CONDITIONAL) INVERSE OF  $XX'$

NOTE: THE MODEL IS LINEAR IN THE PARAMETERS, NOT IN THE  $x$ 's.

SOME OF THE  $x$ 's MAY BE NONLINEAR, E.G.,

$$y_i | x = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_1x_2 + e_i$$

DISCRETE INDEPENDENT VARIABLES (BINARY VARIABLES, DICHOTOMOUS VARIABLES, INDICATOR VARIABLES, DUMMY VARIABLES, CATEGORICAL VARIABLES, NOMINAL VARIABLES, QUALITATIVE VARIABLES)

$X_7 = 0$  FOR MALE, 1 FOR FEMALE

$X_8 = 0$  FOR WHITE, 1 FOR BLACK, 2 FOR NATIVE AMERICAN, 3 FOR HISPANIC, 4 FOR ASIAN, 5 FOR OTHER

$X_9 = -2$  FOR VERY DISSATISFIED,  $-1$  FOR DISSATISFIED,  $0$  FOR NEITHER SATISFIED NOR DISSATISFIED,  $1$  FOR SATISFIED,  $2$  VERY SATISFIED

CAN INCLUDE  $X_7$  AND  $X_9$  IN THE REGRESSION MODEL, BUT NOT  $X_8$ , SINCE IT IS NOT ORDERED. DEFINE DUMMY VARIABLES:

$X_{10} = 0$  IF  $X_8=0$ ,  $1$  IF  $X_8 \neq 0$

$X_{11} = 0$  IF  $X_8=1$ ,  $1$  IF  $X_8 \neq 1$

$X_{12} = 0$  IF  $X_8=2$ ,  $1$  IF  $X_8 \neq 2$

$X_{13} = 0$  IF  $X_8=3$ ,  $1$  IF  $X_8 \neq 3$

$X_{14} = 0$  IF  $X_8=4$ ,  $1$  IF  $X_8 \neq 4$

$X_{15} = 0$  IF  $X_8=5$ ,  $1$  IF  $X_8 \neq 5$

MAY INCLUDE ANY FIVE OF THESE SIX DUMMY VARIABLES IN THE MODEL. SHOULD NOT INCLUDE ALL SIX, BECAUSE THEY SUM TO 1 (A PERSON MUST BE OF ONE RACE), AND THIS LINEAR DEPENDENCY WOULD CAUSE THE MODEL TO BE INDETERMINANT (IF NO LINEAR DEPENDENCIES, THEN CAN INVERT THE CROSSPRODUCTS MATRIX).

## BINARY DEPENDENT VARIABLE

IF THE DEPENDENT VARIABLE IS BINARY, THEN THE ERROR TERMS WILL NOT ALL HAVE THE SAME VARIANCES. TWO PROBLEMS ARISE:

1. HOMOSCEDASTICITY (EQUAL-VARIANCE) ASSUMPTION IS VIOLATED

USE GENERALIZED LEAST SQUARES (GAUSS-MARKOV) ESTIMATION

PERFORM REGRESSION ITERATIVELY: E.G., USE FIRST ESTIMATE FOR  $E(y)=p$ , AND SET VARIANCE EQUAL TO  $p(1-p)$

2.  $E(y)$  IS A PROBABILITY, BUT THE PREDICTIONS WILL FALL BEYOND THE  $(0, 1)$  INTERVAL

APPLY A LOGISTIC TRANSFORMATION (SUFFICIENT STATISTIC)

$$y_i^* = \ln\left(\frac{y_i}{1 - y_i}\right)$$

AND USE A REGRESSION MODEL FOR THE LOGISTIC VARIABLE:

$$y_i^* = b_0 + b_1x_1 + \dots$$

## INTERACTION TERMS; POOLED VS. SEPARATE REGRESSION MODELS

### SEPARATE REGRESSION EQUATIONS:

$$y_i = \begin{cases} b_0 + b_1x_1 + b_2x_2 + e & \text{if } x_3 = 0 \\ b_0^* + b_1^*x_1 + b_2^*x_2 + e & \text{if } x_3 = 1 \end{cases}$$

### COMBINED (POOLED) REGRESSION EQUATIONS:

$$y_i = b_0 + b_1'x_1 + b_2'x_2 + b_3'x_1x_3 + b_4'x_1x_4 + e$$

USE SEPARATE REGRESSION EQUATIONS IF A NEW VARIABLE IS INTERACTING WITH A LARGE NUMBER OF THE OLD VARIABLES

## LINEAR DEPENDENCY

$$\text{cov}(x_1, x_2) = E(x_1 - x_1)(x_2 - x_2) = E(x_1 x_2) - E(x_1)E(x_2)$$

$$\rho = \text{corr}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sqrt{\text{var}(x_1) \text{var}(x_2)}} \quad (0 \leq \rho \leq 1)$$

IF THERE IS A LINEAR DEPENDENCY AMONG THE  $x$ 's, E.G.,

$$x_1 + x_2 + x_3 = 1$$

THEN CAN'T USE A REGRESSION PACKAGE THAT CALCULATES ONLY MATRIX INVERSES (NOT GENERALIZED INVERSES).

SOLUTION: REMOVE LINEAR DEPENDENCIES (BY CHANGING THE VARIABLES INCLUDED IN THE MODEL).

## MULTICOLLINEARITY

IF  $\text{corr}(x_1, x_2)$  IS HIGH, THEN  $\text{corr}(b_1, b_2)$  IS HIGH, AND THE STANDARD ERROR OF THE ESTIMATES  $b_1$  AND  $b_2$  WILL BE LARGE .

SOLUTIONS:

- REMOVE ONE OF THE  $x$ 's
- TRANSFORM E.G.,  $x_3 = x_1 - x_2$
- ORTHOGONALIZE THROUGH CONTROL OF  $x$ 's (PRIOR TO DATA COLLECTION)
- ORTHOGONALIZE THROUGH FACTOR ANALYSIS OR PRINCIPAL COMPONENTS ANALYSIS (LOSE NATURAL VARIABLES)

## CATEGORICAL REPRESENTATIONS OF CONTINUOUS VARIABLES

$$x_1 = \text{AGE} \quad (0 \leq x_1 \leq 120)$$

OR

$$x_1 = \begin{cases} 0 & \text{for } 0 \leq x_1 \leq 17 \\ 1 & \text{for } 18 \leq x_1 \leq 65 \\ 2 & \text{for } x_1 > 65 \end{cases}$$

ADVANTAGES OF CATEGORICAL REPRESENTATION: SURVEY DESIGN (STRATIFICATION)

## OTHER REGRESSION TOPICS

- MULTIVARIATE (EACH OBSERVATION CONSISTS OF A VECTOR OF OBSERVED  $y$ 's)
- SIMULTANEOUS EQUATION MODELS (CAN'T USE ORDINARY LEAST SQUARES ESTIMATION PROCEDURE)
- AUTOCORRELATED RESIDUALS (USE GENERALIZED LEAST SQUARES, NONLINEAR REGRESSION)
- PATH ANALYSIS
- RIDGE REGRESSION
- STEPWISE REGRESSION
- HALF-NORMAL PLOTS
- WEIGHTED REGRESSION
- REGRESSION OVER FINITE POPULATION
- VARIANCE-STABILIZING TRANSFORMATIONS
- COMPUTATIONAL CONSIDERATIONS (STORED CROSSPRODUCTS MATRIX)
- ANALYSIS OF RESIDUALS
- ERRORS IN OBSERVED VARIABLES ( $x$ 's)
- MISSPECIFIED MODEL
- TESTS OF SIGNIFICANCE OF REGRESSION COEFFICIENTS
- MULTIPLE CORRELATION COEFFICIENT
- PRINCIPLE OF CONDITIONAL ERROR

## II. GENERAL PROCEDURES FOR DESIGNING AN ANALYTICAL SURVEY

### 1. SAMPLE SURVEY DESIGN FOR ANALYSIS

WANT:

- HIGH PRECISION ON ESTIMATES OF  $b$ 's
- LOW CORRELATION AMONG ESTIMATES OF  $b$ 's

IMPLIES NEED FOR:

- GOOD "BALANCE" (SPREAD) ON  $x$ 's
- HIGH DEGREE OF ORTHOGONALITY (LOW CORRELATION) AMONG  $x$ 's

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

### 2. PROBLEMS IN DESIGN OF AN ANALYTICAL SURVEY

WANT:

- BALANCE ON A LARGE NUMBER OF VARIABLES
- ORTHOGONALITY (OR LOW CORRELATION) AMONG LARGE NUMBER OF VARIABLES

IN ADDITION, CONSTRAINTS ON:

- COST
- PRECISION (SAMPLE SIZE)
- STRATIFICATION AND OTHER DESIGN CONSTRAINTS (E.G., CLUSTERS)

NOTE: ATTENTION CENTERS ON INDEPENDENT VARIABLES

- IN DESCRIPTIVE SURVEY, TRY TO STRATIFY ON PREMEASURE OF DEPENDENT VARIABLES
- IN ANALYTICAL SURVEY, TRY TO STRATIFY ON PREMEASURE OF INDEPENDENT VARIABLES

#### NOTES ON DETERMINING SAMPLE SIZE IN ANALYTICAL SURVEYS:

REGRESSION COEFFICIENTS ARE SIMILAR TO DIFFERENCES. DETERMINE THE SAMPLE SIZE REQUIRED TO ACHIEVE A SPECIFIED LEVEL OF PRECISION FOR A DIFFERENCE (OR POWER FOR A TEST INVOLVING A DIFFERENCE), USING THE SAMPLE-SIZE PROGRAM CITED EARLIER (POSTED AT INTERNET LOCATION <http://www.foundationwebsite.org/JGCSampleSizeProgram.mdb> ).

NOTE THAT THE STANDARD METHOD OF ESTIMATING SAMPLE SIZE FOR ANALYTICAL SURVEYS IS TO DETERMINE THE SAMPLE SIZE THAT PROVIDES A SPECIFIED LEVEL OF POWER FOR TESTS OF HYPOTHESIS ABOUT DIFFERENCES (E.G., ABOUT THE SIZE OF A DOUBLE-DIFFERENCE ESTIMATE OF PROGRAM IMPACT), RATHER THAN THE APPROACH (FOR DESCRIPTIVE SURVEYS) OF DETERMINING THE SAMPLE SIZE THAT PROVIDES A SPECIFIED LEVEL OF PRECISION FOR AN ESTIMATE (SUCH AS A MEAN OR TOTAL). THAT APPROACH ("POWER ANALYSIS") IS NOT DESCRIBED IN THIS COURSE, BUT IS DESCRIBED IN THE CITED PROGRAM, AND ALSO IN THE ARTICLE "SAMPLE

SURVEY DESIGN FOR EVALUATION,” POSTED  
AT <http://www.foundationwebsite.org/SampleSurveyDesignForEvaluation.pdf> .

PROGRAM EVALUATIONS OFTEN INVOLVE THE USE OF ANALYTICAL SURVEYS TO DETERMINE DOUBLE-DIFFERENCE (“DIFFERENCE-IN-DIFFERENCE”) ESTIMATES OF PROGRAM IMPACT, IN A PRETEST-POSTTEST-WITH-COMPARISON-GROUP DESIGN. (IN STATISTICAL TERMINOLOGY, THE DOUBLE-DIFFERENCE ESTIMATE IS THE INTERACTION EFFECT OF TREATMENT AND TIME.) THE CITED PROGRAM CAN DETERMINE SAMPLE SIZES EXPLICITLY FOR THIS TYPE OF DESIGN. (IT USES THE DEFF TO REPRESENT THE DESIGN EFFECT OF ALL FACTORS OTHER THAN THE PRETEST-POSTTEST-WITH-COMPARISON-GROUP STRUCTURE (SUCH AS STRATIFICATION AND MULTISTAGE SAMPLING).)

IF THE DESIGN IS HIGHLY STRUCTURED, THE FORM OF THE IMPACT ESTIMATE WILL BE SIMPLE (E.G., A DOUBLE DIFFERENCE)

IF THE DESIGN IS NOT HIGHLY STRUCTURED, A “GENERALIZED REGRESSION” (“GREG”) ESTIMATOR WILL BE USED (SEE SHARON L. LOHR, *SAMPLING: DESIGN AND ANALYSIS* (DUXBURY PRESS, 1999) FOR DISCUSSION), BY SUBSTITUTING MEAN VALUES OF THE EXPLANATORY VARIABLES IN A REGRESSION MODEL THAT RELATES IMPACT TO EXPLANATORY VARIABLES.

THE PROBLEM OF DETERMINING SAMPLE SIZES FOR COMPLEX SURVEYS IS NOT SIMPLE OR EASY, AND THE PROBLEM OF DETERMINING SAMPLE SIZES FOR ANALYTICAL SURVEYS IS EVEN MORE COMPLICATED AND DIFFICULT. REFER TO THE ARTICLE “SAMPLE SURVEY DESIGN FOR EVALUATION” FOR MORE INFORMATION ON THIS TOPIC,  
AT <http://www.foundationwebsite.org/SampleSurveyDesignForEvaluation.pdf> . THE SAMPLE-SIZE ESTIMATION SHOULD BE MATCHED TO THE SURVEY DESIGN, EITHER EXPLICITLY (AS IN THE CASE OF THE PRETEST-POSTTEST-WITH-COMPARISON-GROUP DESIGN TREATED IN THE CITED SAMPLE-SIZE PROGRAM) OR VIA THE DESIGN EFFECT, DEFF (OR A COMBINATION OF BOTH).

SIMPLE RANDOM SAMPLING IS NOT EFFICIENT FOR ESTIMATING REGRESSION-MODEL PARAMETERS OR DIFFERENCES. WITH A SIMPLE RANDOM SAMPLE, THE VARIANCE OF AN ESTIMATED DIFFERENCE BETWEEN TWO SAMPLE HALVES IS FOUR TIMES THE VARIANCE OF THE ESTIMATED MEAN, AND THE VARIANCE OF A DOUBLE DIFFERENCE INVOLVING FOUR EQUAL GROUPS IS 16 TIMES AS LARGE. THESE NUMBERS LEAD TO VERY LARGE SAMPLE SIZES. THE SAMPLE SIZE IS REDUCED TO REASONABLE LEVELS BY TWO METHODS: (1) INTRODUCTION OF CORRELATIONS INTO THE SAMPLE BY MEANS OF MATCHING (PAIRED COMPARISONS BETWEEN TREATMENT AND CONTROL UNITS) AND PANEL SAMPLING (REINTERVIEW OF THE SAME UNITS AT TIME 2); AND (2) THE USE OF POWER ANALYSIS TO DETERMINE SAMPLE SIZES (I.E., DETERMINING THE SAMPLE SIZE THAT DETECTS A SPECIFIED DIFFERENCE IN IMPACT WITH A SPECIFIED PROBABILITY (POWER)). THESE METHODS ARE DISCUSSED IN THE CITED ARTICLE.

### 3. TWO CONCEPTUAL APPROACHES TO DESIGN OF ANALYTICAL SURVEYS

#### OBJECTIVE-FUNCTION APPROACH (AS FOR THE NEYMAN ALLOCATION TO STRATA):

O.F. = LINEAR COMBINATION OF VARIANCES OF ESTIMATES OF INTEREST

SELECT DESIGN TO MINIMIZE OBJECTIVE FUNCTION

#### CONSTRAINT-DRIVEN APPROACH:

SELECT DESIGN TO MEET A VARIETY OF CONSTRAINTS (CONSTRAINED OPTIMIZATION WITH NUMEROUS CONSTRAINTS)

OBJECTIVE FUNCTION APPROACH HAS NOT PROVED TO BE PRODUCTIVE

(CAN'T DETERMINE SUITABLE SCALAR UTILITY FUNCTION)

CONSTRAINT-DRIVEN APPROACH IS FEASIBLE

#### 4. METHODS FOR DESIGNING ANALYTICAL SURVEYS

FOR "SMALL PROBLEMS" (FEW EXPLANATORY VARIABLES, SUCH AS AN EXPERIMENTAL DESIGN):

1. CROSS STRATIFICATION
2. MARGINAL STRATIFICATION
3. CONTROLLED SELECTION (GOODMAN, KISH)
4. EXPERIMENTAL DESIGN AND QUASI-EXPERIMENTAL DESIGNS

FOR LARGE PROBLEMS (MANY EXPLANATORY VARIABLES, SUCH AS A QUASI-EXPERIMENTAL DESIGN WITH MANY COVARIATES):

5. A GENERAL METHODOLOGY FOR DESIGNING ANALYTICAL SURVEYS (SETTING OF VARIABLE SELECTION PROBABILITIES TO OBTAIN DESIRED EXPECTED STRATUM ALLOCATIONS)

### III. ILLUSTRATION OF METHODS FOR THE DESIGN OF ANALYTICAL SURVEYS

#### 1. CROSS STRATIFICATION

##### CASE 1. CELL-BY-CELL STRATIFICATION (CROSS-STRATIFICATION)

SAMPLE SIZE EXCEEDS NUMBER OF CELLS

n1	n2			

ORDINARY STRATIFIED DESIGN

##### CASE 2. MARGINAL STRATIFICATION

NUMBER OF CELLS EXCEEDS SAMPLE SIZE; SAMPLE SIZE EXCEEDS EVERY NUMBER OF MARGINAL CATEGORIES

	1			2			
	1	2	3	1	2	3	
1							4
2							2
3							4
	2	1	2	2	1	2	n=10

R = 6  
C = 3  
n = 10  
(R x C = 18)

ROW AND COLUMN TOTALS ARE FIXED

CELLS ARE RANDOMLY SELECTED TO SATISFY THESE CONSTRAINTS

PROCEDURE:

x										
	x									
						x				
		x								
									x	
			x							
x										
						x				
								x		
				x						
2	1	2		2	1	2				10

RANDOM PERMUTATION OF n=10

x	x	x		x		4
		x			x	2
x			xx		x	4
2	1	2	2	1	2	10

PROBLEM: LOW ORTHOGONALITY. PROCEDURE BETTER SUITED FOR DESCRIPTIVE SURVEYS, WHERE ORTHOGONALITY IS NOT AN ISSUE.

## 2. CONTROLLED SELECTION (CONTROLS BEYOND STRATIFICATION)


$3 \times 3 \times 2 \times 2 \times 2 = 72$  CELLS

SAMPLE SIZE (E.G., SMSAs) = 40

SOME CELLS HAVE NO POPULATION MEMBERS

PLACE FIVE SAMPLE PATTERNS OF 40 ON THE GRID, SO THAT THE MARGINAL AND CROSSTAB CONSTRAINTS ARE GENERALLY SATISFIED (BALANCE + ORTHOGONALITY), SO THAT ALL NONEMPTY CELLS ARE COVERED AT LEAST ONCE (TO ALLOW FOR ESTIMATION OF MEANS AND TOTALS), SO THAT THE PROBABILITY OF SELECTION OF INDIVIDUAL ELEMENTS IS AS EVEN AS POSSIBLE (TO ALLOW FOR PRECISE ESTIMATION OF MEANS AND TOTALS), WITH THE MORE DESIRABLE PATTERNS HAVING HIGHER PROBABILITIES OF SELECTION, SELECT ONE OF THE FIVE SAMPLES (PROBABILITY SAMPLE).

### 3. CONTROLLED SELECTION – SIMPLER EXAMPLE

	1			2		
	1	2	3	1	2	3
1	*	o	xo	x	o	x
2	xo	x	*	o	x	*
3	xo	o	x	xo	o	x

TWO PATTERNS, DESIGNATED BY x AND o. ASTERISK (\*) DENOTES NO POPULATION.

PATTERN x (MORE DESIRABLE): Prob = .6  
PATTERN o (LESS DESIRABLE):  $\frac{\text{Prob} = .4}{\text{Prob} = 1.0}$

ALL CELLS COVERED (EVERY POPULATION ELEMENT HAS A NONZERO PROBABILITY OF SELECTION, AND THE PROBABILITIES OF SELECTION ARE KNOWN – THE BASIC REQUIREMENT FOR A PROBABILITY SAMPLE).

NOTE: THIS EXAMPLE DOESN'T CONSIDER THE SIZE OF THE POPULATION IN EACH CELL, I.E., ADJUSTMENT OF THE PATTERNS TO MAINTAIN REASONABLY BALANCED PROBABILITIES.

## CONTROLLED SELECTION -- ADVANTAGES AND DISADVANTAGES

### ADVANTAGES:

- THE SAMPLE IS GUARANTEED TO BE "GOOD" FROM AN ANALYTICAL VIEWPOINT (GOOD BALANCE, HIGH LEVEL OF ORTHOGONALITY)

### DISADVANTAGES:

- DIFFICULT TO APPLY (DETERMINING "PREFERRED" PATTERNS, THAT MEET CONTROLS)
- CORRELATED SAMPLE - MUST USE RESAMPLING METHODS TO COMPUTE VARIANCE ESTIMATES

#### 4. EXPERIMENTAL DESIGNS AND QUASI-EXPERIMENTAL DESIGNS

SIMPLE "BLOCK" DESIGNS (ONE TREATMENT VARIABLE)

FACTORIAL DESIGNS (SEVERAL TREATMENT VARIABLES)

QUASI-EXPERIMENTAL DESIGNS (E.G., PRETEST / POSTTEST / COMPARISON-GROUP DESIGN WITH NUMEROUS COVARIATES)

#### 5. A GENERAL METHODOLOGY FOR DESIGNING ANALYTICAL SAMPLE SURVEYS, WHEN THE NUMBER OF EXPLANATORY VARIABLES IS LARGE: EXPECTED MARGINAL STRATIFICATION USING VARIABLE SELECTION PROBABILITIES – OVERVIEW

1. ESTIMATE THE TOTAL SAMPLE SIZE, BASED ON POWER CALCULATIONS FOR AN ASSUMED DESIGN (E.G., A PRETEST / POSTTEST / COMPARISON-GROUP DESIGN). (USE THE SAMPLE-SIZE PROGRAM CITED EARLIER; THE SAMPLE SIZE USUALLY REFERS TO THE NUMBER OF FIRST-STAGE UNITS. THE SAMPLE SIZE FOR SUBUNITS WITHIN THE FIRST-STAGE UNITS DEPENDS ON THE INTRA-UNIT CORRELATION COEFFICIENT; IT IS OFTEN 15-30 FOR SURVEYS OF HOUSEHOLDS WITHIN CENSUS ENUMERATION UNITS OR VILLAGES.)
2. SPECIFY A DESIRED MARGINAL STRATIFICATION FOR ALL INDEPENDENT VARIABLES AND ANTICIPATED INTERACTIONS
3. SET THE UNIT SELECTION PROBABILITIES SUCH THAT THE EXPECTED STRATUM SAMPLE SIZES ARE CLOSE TO THE DESIRED VALUES.
4. DO MATCHING, AS REQUIRED, TO IMPROVE THE PRECISION OF ESTIMATES OF DIFFERENCES (AND INCLUDE THE CORRELATIONS IN THE SAMPLE-SIZE ESTIMATE).
5. SELECT A SAMPLE ACCORDING TO THE VARIABLE SELECTION PROBABILITIES (WHEN MATCHING IS USED, INCLUDE THE MATCHING ITEM(S) IN THE SAMPLE, FOR EACH UNIT SELECTED).
6. SINCE THE DESIGN IS COMPLEX, USE RESAMPLING METHODS TO CALCULATE VARIANCES OF ESTIMATES.

4. A GENERAL METHODOLOGY FOR DESIGNING ANALYTICAL SURVEYS (EXPECTED MARGINAL STRATIFICATION USING VARIABLE SELECTION PROBABILITIES) – BRIEF SUMMARY (OMITS CONSIDERATION OF MATCHING)

NOTE: THERE IS NO STANDARD REFERENCE TEXT FOR THE FOLLOWING MATERIAL ON THE DESIGN OF ANALYTICAL SURVEYS. THIS MATERIAL WAS DEVELOPED BY THE COURSE DEVELOPER (J. G. CALDWELL) DURING THE COURSE OF HIS CONSULTING IN THIS FIELD (IN THE 1970s). FOR DETAILS, REFER TO THE ARTICLE, “SAMPLE SURVEY DESIGN FOR EVALUATION,” AT <http://www.foundationwebsite.org/SampleSurveyDesignForEvaluation.pdf> .

(1) FOR EACH SUBSTANTIVE ISSUE TO BE ADDRESSED IN THE SURVEY ANALYSIS, IDENTIFY ALL DEPENDENT AND INDEPENDENT VARIABLES.

- IN LARGE SURVEYS, MAY BE SEVERAL HUNDRED VARIABLES
- SEPARATE DEPENDENT AND INDEPENDENT VARIABLES
- SKETCH ANALYTICAL MODELS FOR EACH DEPENDENT VARIABLE (FUNCTIONAL FORM NOT IMPORTANT)

(2) FOR EACH UNIT OF THE SAMPLE FRAME, IDENTIFY PREMEASURES OF THE MAJOR INDEPENDENT VARIABLES OF INTEREST. (THE FOLLOWING STEPS REFER ONLY TO THESE VARIABLES.)

- MAY BE MANY VARIABLES (E.G., FROM PREVIOUS SURVEYS, FROM GOVERNMENT REPORTING SYSTEMS, FROM GEOGRAPHIC INFORMATION SYSTEMS)
- TRY TO INCLUDE AS MANY IMPORTANT CONCEPTS AS POSSIBLE
- EVEN A CRUDE MEASURE IS BETTER THAN NO MEASURE

(3) CONVERT ALL CONTINUOUS VARIABLES AND ORDINAL VARIABLES TO ORDERED CATEGORICAL VARIABLES. COMPUTE MARGINAL FREQUENCY COUNTS OF ALL VARIABLES. RANK ALL INDEPENDENT VARIABLES IN ORDER OF IMPORTANCE.

- TWO OR THREE CATEGORIES PER VARIABLE
- USE NATURAL BREAK POINTS OR QUANTILES (PERCENTILES)
- COMBINE SIMILAR CATEGORIES INTO SINGLE CATEGORY
- COMPUTE MARGINAL FREQUENCY COUNTS FOR ALL VARIABLES
- RANK ALL VARIABLES IN ORDER OF IMPORTANCE (WILL HAVE TO MAKE COMPROMISES LATER)
- MOTIVATION: WANT TO ASSURE BALANCE ON ALL IMPORTANT VARIABLES

(4) CALCULATE CRAMER (NONPARAMETRIC) CORRELATION MATRIX OF ALL (CATEGORICAL) VARIABLES. IDENTIFY SETS OF CORRELATED VARIABLES (“EYEBALL,” NOT FACTOR ANALYSIS).

- MOTIVATION: WANT TO ASSURE HIGH LEVEL OF ORTHOGONALITY FOR VARIABLES IN A REGRESSION MODEL
- USE “EYEBALL” TO IDENTIFY SETS OF HIGHLY CORRELATED VARIABLES, NOT FACTOR ANALYSIS OR PRINCIPAL COMPONENTS ANALYSIS
- MOTIVATION: WITHIN EACH SET, NEED TO DECIDE WHETHER VARIABLES SHOULD BE ORTHOGONALIZED

(5) FOR EACH SET OF CORRELATED VARIABLES, DETERMINE WHETHER THEY REPRESENT A SIMILAR OR DIFFERENT CONCEPT, FROM THE POINT OF VIEW OF SUBSTANTIVE THEORY. FOR IMPORTANT VARIABLE PAIRS THAT ARE NOT CAUSALLY RELATED BUT HIGHLY CORRELATED, DEFINE PRODUCT VARIABLES (OR A CROSS-CLASSIFICATION). COMBINE OR DROP SIMILAR VARIABLES.

- IN REGRESSION ANALYSIS, WOULD COMBINE OR DROP SIMILAR VARIABLES. DON'T WANT TO ORTHOGONALIZE THESE VARIABLES. WANT TO COMBINE OR DROP THEM BEFORE CONSTRUCTING THE DESIGN, RATHER THAN LATER.
- IN REGRESSION ANALYSIS, WANT NON-CAUSALLY-RELATED VARIABLES TO HAVE LOW CORRELATIONS. WANT TO ORTHOGONALIZE THESE VARIABLES (TO THE EXTENT POSSIBLE).
- AFTER COMBINING OR DROPPING VARIABLES, WILL HAVE A MUCH SMALLER SET, E.G., 20 OR 30.

(6) FOR EACH OF THE RESULTANT VARIABLES (OR CROSS-CLASSIFICATION), SPECIFY A DESIRED ALLOCATION OF THE SAMPLE TO THE CATEGORIES (E.G., 50-50, 40-10-40).

- DECIDE ON DESIRED SAMPLE SIZE
- LOOK AT ACTUAL FREQUENCY COUNTS TO SEE WHAT ALLOCATIONS ARE POSSIBLE.

(7) SET THE UNIT SELECTION PROBABILITIES SUCH THAT THE EXPECTED SAMPLE SIZE FOR EACH CATEGORY IS CLOSE TO THE DESIRED SAMPLE SIZE FOR THE CATEGORY.

- CONSTRAINTS WILL GENERALLY BE INCONSISTENT
- USE IMPORTANCE ORDERING OF DEPENDENT VARIABLES TO RELAX CONSTRAINTS UNTIL FEASIBLE SOLUTION IS FOUND
- FOR LARGE PROBLEMS, USE COMPUTER OPTIMIZATION MODEL TO FIND FEASIBLE SOLUTION.
- FOR SMALL PROBLEMS, CAN DETERMINE A GOOD ALLOCATION BY HAND
- IN MOST APPLICATIONS, THERE WILL BE THE POSSIBILITY OF NUMEROUS EMPTY CELLS – NOT A PROBLEM.
- IN VIEW OF DESIRE TO ESTIMATE MEANS AND TOTALS, CONSTRAIN SOLUTION TO REPRESENT EACH CELL WITH SOME NONZERO PROBABILITY (SOME OF THE CATEGORIES MAY HAVE HAD ZERO DESIRE SAMPLE SIZES – THIS CHANGES THAT).

(8) SELECT A PROBABILITY SAMPLE FROM THE DESIGN

- SINCE THE DESIGN IS COMPLEX, USE RESAMPLING METHODS TO ESTIMATE VARIANCES

## PART III. HOW TO ANALYZE SURVEY DATA

### I. STANDARD ESTIMATION PROCEDURES FOR DESCRIPTIVE SURVEYS

#### A. SUMMARY OF PROCEDURES

##### 1. PRELIMINARY ANALYSIS (LARGE DATA SETS)

DO THE FOLLOWING FOR THE ENTIRE SAMPLE AND FOR SUBPOPULATIONS OF INTEREST (E.G., BY STRATUM). THIS IS A STANDARD NONRESPONSE ANALYSIS AND "NONPARAMETRIC" SUMMARY OF THE SAMPLE DATA. (A "FIRST-CUT" LOOK AT THE SAMPLE DATA MAY IGNORE COMPLEXITIES OF THE SAMPLE DESIGN (AND TESTS OF SIGNIFICANCE), AND SIMPLY PRESENT ESTIMATES OF CHARACTERISTICS OF THE SAMPLE.)

##### UNIVARIATE ANALYSIS:

- UNIT (QUESTIONNAIRE) NONRESPONSE. CALCULATE PROPORTION OF SAMPLE UNITS RESPONDING. CATEGORIZE BY REASON. NOTE VARIABLES CORRELATED WITH NONRESPONSE.
- ITEM NONRESPONSE. DEFINE A NO-RESPONSE INDICATOR VARIABLE FOR EACH VARIABLE. NOTE VARIABLES HAVING HIGH LEVEL OF NONRESPONSE, AND VARIABLES CORRELATED WITH NONRESPONSE.
- FOR INTERVAL-SCALE VARIABLES, COMPUTE MEAN OR PERCENTAGE, MEDIAN, MIN, MAX, RANGE, VARIANCE, STANDARD DEVIATION, HISTOGRAM.
- FOR CATEGORICAL VARIABLES COMPUTE MIN, MAX, RANGE AND HISTOGRAM.
- FOR GEOGRAPHIC DATA, PLOT ALL SAMPLE UNITS ON A MAP.

##### BIVARIATE ANALYSIS:

- RECODE ALL INTERVAL-SCALE VARIABLES BY QUINTILE
- RECODE ALL CATEGORICAL VARIABLES HAVING MORE THAN 10 CATEGORIES INTO TEN OR LESS CATEGORIES (PREFERABLY 5 OR LESS). DEFINE INDICATOR VARIABLES FOR HIGH-INTEREST CATEGORIES.
- CALCULATE CRAMER COEFFICIENT OF ASSOCIATION FOR ALL VARIABLES (I.E., EACH VARIABLE WITH EACH OTHER). NOTE STRONG CORRELATIONS.
- CONSTRUCT TABLES OR GRAPHS (SCATTER PLOTS, LOESS / LOWESS (LOCALLY WEIGHTED SCATTERPLOT SMOOTHING) CURVES) TO ILLUSTRATE STRONG RELATIONSHIPS.

## 2. PLANNED ANALYSIS

- FOR TOTAL POPULATION AND SUBPOPULATIONS OF INTEREST, SUMMARIZE NONRESPONSE PATTERNS
- ADJUST SAMPLE WEIGHTS FOR NONRESPONSE
- CALCULATE ESTIMATES INTEREST (USE APPROPRIATE FORMULAS OR PROCEDURES)
- CROSSTABS OF INTEREST (BY STRATUM OR WEIGHTED)
- COMPUTE EXPANSION ESTIMATES
- CONDUCT TESTS OF SIGNIFICANCE

## 3. SPECIAL ANALYSES

- ALTERNATIVE ESTIMATORS (RATIO, REGRESSION)
- POSTSTRATIFICATION
- ADDITIONAL SUBPOPULATIONS
- REDEFINE NONRESPONSE STRATA, RECOMPUTE ESTIMATES
- RESAMPLING ESTIMATES OF VARIANCES
- TABLES OF GENERALIZED VARIANCES
- $\chi^2$  (CHI-SQUARED) TESTS FOR CLUSTER-SAMPLE CROSSTABS

## B. STANDARD ESTIMATION PROCEDURES FOR DESCRIPTIVE STATISTICS – DETAILS

### ESTIMATION

EXPANSION MULTIPLIERS FOR MEANS, TOTALS AND OTHER STATISTICS FOR SIMPLE RANDOM SAMPLING AND STRATIFIED SAMPLING, PPS SAMPLING.

POST-STRATIFICATION FOR INCREASING EFFICIENCY.

STRATIFIED SAMPLING:

#### ESTIMATED TOTALS:

$$\hat{Y}_i = \sum_{h=1}^L \frac{N_h}{n_h} \sum_{j=1}^{n_{hi}} y_{hij}$$

#### ESTIMATED MEANS:

$$\hat{\bar{Y}}_i = \frac{Y_i}{N_i} = \frac{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{j=1}^{n_{hi}} y_{hij}}{\sum_{h=1}^L \frac{N_h}{N_h} n_{hi}}$$

### ALTERNATIVE ESTIMATES USING AUXILIARY INFORMATION

- RATIO OR REGRESSION ESTIMATES
- COMBINED RATIO ESTIMATES

### VARIANCE ESTIMATION

- FOR SIMPLE DESIGNS, EXACT FORMULAS ARE AVAILABLE AND EASY TO COMPUTE

## VARIOUS SHORT-CUT METHODS FOR ESTIMATING VARIANCES IN COMPLEX SURVEYS

### 1. THE METHOD OF RANDOM GROUPS (REPLICATION) (INTERPENETRATING SAMPLES)

TWO OR MORE SAMPLES FROM THE SAME POPULATION. THESE SAMPLES MAY BE INDEPENDENT OR NOT. THE SAMPLING VARIANCE IS COMPUTED FROM THESE TWO OR MORE ESTIMATES

### 2. RESAMPLING METHODS: THE JACKKNIFE, BOOTSTRAP AND BALANCED REPEATED REPLICATION (BRR)

E.G., THE METHOD OF BALANCED HALF-SAMPLES:  
FOR K STRATA THERE ARE  $2^K$  REPLICATES, A SAMPLE OF L BALANCED REPLICATES SELECTED WITHOUT REPLACEMENT,  
 $L < 2^K$ .

### 4, GENERALIZED VARIANCE FUNCTIONS (GAT CURVES)

THE VARIANCE OR RELATIVE VARIANCE OF THE ESTIMATE IS ALGEBRAICALLY RELATED TO THE VALUE OF THE ESTIMATE.

SIMPLE ALGEBRAIC MODELS SUCH AS THE FOLLOWING ARE USED.

$$V^2 = \alpha + \frac{\beta}{x}$$

### 5. LINEARIZATION (TAYLOR-SERIES APPROXIMATION)

EXACT ALGEBRAIC EXPRESSIONS FOR THE SAMPLING VARIANCES FOR NONLINEAR ESTIMATORS (RATIO ESTIMATOR, REGRESSION ESTIMATOR, ETC.) ARE USUALLY NOT AVAILABLE. LINEAR FUNCTIONS OF OBSERVATIONS ARE EMPLOYED TO ESTIMATE APPROXIMATELY THE SAMPLING VARIANCE.

## DOMAINS OF STUDY (SUBPOPULATIONS)

NOTATION (FROM COCHRAN, *SAMPLING TECHNIQUES*):

POPULATION SIZE  $N$ , SAMPLE SIZE  $n$ .

THE  $j$ -th DOMAIN CONTAINS  $N_j$  UNITS.

SAMPLE OF SIZE  $n_j$ , OBSERVATIONS  $y_{jk}$ ,  $k=1,2,\dots,n_j$

SIMPLE RANDOM SAMPLING

ESTIMATED MEAN (FOR DOMAIN  $j$ ):

$$\bar{y}_j = \sum_{k=1}^{n_j} \frac{y_{jk}}{n_j}$$

ESTIMATED STANDARD ERROR OF THE ESTIMATED MEAN:

$$\frac{s_j}{\sqrt{n_j}} \sqrt{1 - \frac{n_j}{N_j}} \quad (\text{USE } \frac{n_j}{N_j} = \frac{n}{N} \text{ if } N_j \text{ IS NOT KNOWN})$$

WHERE

$$s_j^2 = \sum_{k=1}^{n_j} \frac{(y_{jk} - \bar{y}_j)^2}{n_j - 1}$$

ESTIMATED TOTAL (FOR DOMAIN  $j$ ):

$$\hat{Y}_j = \frac{N}{n} \sum_{k=1}^{n_j} y_{jk}$$

( $N_j \bar{y}_j$  USUALLY WON'T WORK SINCE  $N_j$  IS USUALLY NOT KNOWN.)

ESTIMATED STANDARD ERROR OF THE ESTIMATED TOTAL:

$$\frac{Ns'}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

WHERE

$$s'^2 = \frac{1}{n-1} \left( \sum_{k=1}^{n_j} y_{jk}^2 - \frac{(\sum_{k=1}^{n_j} y_{jk})^2}{n} \right)$$

NOTE THAT IN THE ESTIMATED VARIANCES FOR THE TOTAL, THE DIVISORS INVOLVE  $n$ , NOT  $n_j$ . THIS IS BECAUSE THE EXPANSION FACTOR IS  $N/n$ , NOT  $N/n_j$ . HENCE, THE SAMPLE IS REGARDED AS A SAMPLE OF SIZE  $n$  IN WHICH  $n_j$  UNITS (OF THE  $j$ -th DOMAIN HAVE VALUES  $y_{jk}$  AND  $n - n_j$  HAVE VALUE ZERO.

## C. ALTERNATIVE ESTIMATION PROCEDURES

### 1. SYNTHETIC ESTIMATES

UNBIASED ESTIMATE FROM A LARGE AREA

WANT ESTIMATE FOR A SMALL AREA

USE LARGE-AREA INFORMATION TO OBTAIN SMALL-AREA ESTIMATE

CASE 1. USE MEAN OF LARGE AREA FOR SMALL AREA

CASE 2. ESTIMATE, FROM LARGE AREA, THE VALUE FOR VARIOUS SUBPOPULATIONS (STRATA), FOR WHICH COUNTS ARE KNOWN FOR THE SMALL AREA. FORM A WEIGHTED AVERAGE FOR THE SMALL AREA.

#### EXAMPLE:

FOR THE LARGE AREA, IT IS KNOWN THAT:

WHITES:      MEAN = 7  
BLACKS:      MEAN = 9

FOR THE SMALL AREA, IT IS KNOWN THAT:

WHITES = 60% (OF THE POPULATION OF THE SMALL AREA)  
BLACKS = 40%.

THE SYNTHETIC ESTIMATE OF THE MEAN FOR THE SMALL AREA IS THE WEIGHTED AVERAGE OF THE MEANS FROM THE LARGE AREA, USING THE POPULATION PROPORTIONS FOR THE SMALL AREA:  $.6(7) + .4(9) = 7.8$

## 2. RAKING

FROM A NEW SURVEY, ESTIMATE A CROSSTABULATION.

MAY KNOW MARGINAL TOTALS FROM A BETTER SOURCE (E.G., A CENSUS).

“RAKING” IS A PROCEDURE FOR ADJUSTING THE CROSSTAB TO MATCH THE MARGINALS (LEAST SQUARES PROCEDURE)


ITERATIVE PROCEDURE; SLOW IN MULTIVARIATE CASE.

## II. STANDARD ESTIMATION PROCEDURES FOR ANALYTICAL SURVEYS

### 1. PRELIMINARY ANALYSIS -- SAME AS FOR DESCRIPTIVE SURVEY

### 2. PLANNED ANALYSIS

- MODEL SPECIFICATION
  - IF WELL-SPECIFIED, DON'T NEED TO USE WEIGHTS
  - IF NOT SURE -- USE WEIGHTS
- CHOOSE BETWEEN SEPARATE VS, POOLED REGRESSIONS (WITH INTERACTION TERMS)
- DRAW SUBSAMPLES FOR PRELIMINARY ANALYSIS
- IMPUTE MISSING VALUES
  - BY REGRESSION
  - BY "HOT DECK" (CONDITIONED ON APPROPRIATE VARIABLES)
- REVISE WEIGHTS TO ACCOUNT FOR NONRESPONSE
- TRANSFORM VARIABLES (E.G., LOGISTIC MODEL)
- STORE CROSSPRODUCTS MATRIX
- RUN PRELIMINARY REGRESSIONS (ON STORED MATRIX)
- REDUCE MULTICOLLINEARITY (LOOK AT CORRELATIONS BETWEEN ESTIMATED COEFFICIENTS)
  - DROP VARIABLES
  - TRANSFORM VARIABLES
  - USE FACTOR ANALYSIS AS LAST RESORT
- ELIMINATE INSIGNIFICANT VARIABLES BY DROPPING OR COMBINING

### 3. REVISE MODEL

- TESTS OF MODEL ADEQUACY
  - PLOTS OF RESIDUALS
  - REGRESSIONS ON DIFFERENT SUBPOPULATIONS
  - SUBSTANTIVE CHECKS (ON COEFFICIENTS)
  - IF MODEL USED FOR PREDICTION, CHECK ACCURACY WITH HISTORICAL DATA
- RESPECIFICATION OF MODEL / REESTIMATION OF COEFFICIENTS
- RUN MODEL WITH AND WITHOUT WEIGHTS, NOTE DIFFERENCES

### 4. TESTS OF SIGNIFICANCE

- USE ORDINARY LEAST SQUARES (OLS) FOR BULK OF ANALYSIS
  - INFINITE POPULATION
  - CORRECTLY SPECIFIED MODEL
- FOR ROUGH TESTS OF SIGNIFICANCE, USE "EFFECTIVE" SAMPLE SIZE (TAKING INTO ACCOUNT DESIGN EFFECT)
- USE RESAMPLING TO ESTIMATE VARIANCES
- USE PRINCIPLE OF CONDITIONAL ERROR TO TEST HYPOTHESES ABOUT VARIABLES (WITH JUDGMENT)
- GENERAL LINEAR MODEL CAN HANDLE INTRACLUSTER CORRELATIONS
- CHI-SQUARED FOR CROSSTABS IN CLUSTER SAMPLE
- CAUTION: CAN'T MAKE CAUSAL INFERENCES (GUIDANCE FOR EXPERIMENTAL DESIGN)

III. COMPUTER PROGRAMS FOR ANALYSIS OF SURVEY DATA:  
OUTLINE OF TOPICS FOR THIRD DAY

1. COMPUTER PROGRAMS FOR ANALYSIS OF SURVEY DATA

- SURVEY DATA ENTRY PROGRAMS: EPI-INFO (CDC), CSPro (CENSUS)
- STATISTICAL ANALYSIS PROGRAM PACKAGES: STATA, SPSS, SAS, S, S-PLUS, R, ZELIG, STATISTICA
- MANY STATISTICAL-PACKAGE PROGRAMS ARE DESIGNED PRIMARILY FOR DESCRIPTIVE SURVEYS (FINITE POPULATION CORRECTION) OR SIMPLE RANDOM SAMPLING (E.G., MODEL-BASED APPROACH TO REGRESSION ANALYSIS), NOT FOR DESIGN-BASED ANALYSIS OF COMPLEX SURVEY DATA
- CAN USE WEIGHTS TO OBTAIN ESTIMATES, BUT NOT TESTS OF SIGNIFICANCE
- FOR DESCRIPTIVE SURVEYS, CAN USE THESE PROGRAMS FOR ESTIMATION OF MEANS, AND CROSSTABULATION FOR NONCLUSTER SAMPLING WITHIN A STRATUM
- FOR ANALYTICAL SURVEYS, CAN USE THESE PROGRAMS FOR ESTIMATION OF REGRESSION COEFFICIENTS IN NONCLUSTER SAMPLING. FOR CLUSTER SAMPLING, USE "EFFECTIVE" SAMPLE SIZE (VIA DEFF) FOR ROUGH TESTS OF SIGNIFICANCE, OR GENERALIZED LEAST SQUARES, OR SELECT SINGLE SAMPLE UNIT FROM EACH CLUSTER.
- SOME COMPUTER PROGRAMS PERFORM ESTIMATION TAKING INTO ACCOUNT THE SURVEY SAMPLE STRUCTURE, AND INCLUDE RESAMPLING METHODS FOR VARIANCE ESTIMATION
- EXAMPLES OF COMPUTER-PROGRAM PACKAGES DESIGNED TO ANALYZE SURVEY DATA FROM THE DESIGN-BASED APPROACH (FINITE-POPULATION APPROACH) INCLUDE: SUDAAN, OSIRIS, PC CARP, Wes VarPC, CENVAR, CLUSTERS, AND VPLX. (SEE SHARON L. LOHR'S *SAMPLING: DESIGN AND ANALYSIS* (DUXBURY PRESS, 1999), PP.313-315 FOR DISCUSSION OF STATISTICAL SOFTWARE.)

DAY 3: SPECIAL TOPICS / PRACTICAL PROBLEMS IN SURVEY DESIGN

SURVEY DESIGN FOR MONITORING AND EVALUATION  
INSTRUMENTATION, DATA COLLECTION AND SURVEY FIELD PROCEDURES  
PREPARATION OF OMB CLEARANCE FORMS  
LONGITUDINAL SURVEYS  
SAMPLE FRAME PROBLEMS  
SAMPLING FOR RARE ELEMENTS  
TREATMENT OF NONRESPONSE  
NONSAMPLING ERRORS (RELIABILITY/VALIDITY)  
RANDOMIZED RESPONSE  
RANDOM DIGIT DIALING  
MAJOR NATIONAL AND INTERNATIONAL SURVEYS  
QUESTIONS AND ANSWERS

SEE STEVEN K. THOMPSON'S *SAMPLING* (2<sup>ND</sup> EDITION) FOR DISCUSSION OF OTHER DESIGN TOPICS (E.G., NETWORK SAMPLING, CAPTURE-RECAPTURE, LINE-INTERCEPT SAMPLING, SPATIAL SAMPLING / KRIGING, ADAPTIVE SAMPLING)