

The World of Complex Surveys



Software for the Statistical
Analysis of Correlated Data



Contact

G. Gordon Brown

TEL: 919-485-5647

Email: ggbrown@rti.org

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

Complex Surveys



Software for the Statistical
Analysis of Correlated Data



Outline

Outline

- **Starting Issues**
 - **Identify Target Population**
 - **Hypotheses of Interest**
 - **Sample Design**
 - **Data Collection**
- **Data Preprocessing**
 - **Weighting**
 - **Imputation**
 - **Data Masking**
 - **Variance Units**
- **Data analysis**

$$-\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = -\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Starting Issues



Software for the Statistical
Analysis of Correlated Data



Sun Tzu on Statistics?

Victorious warriors win first and then go to war, while defeated warriors go to war first and then seek to win – Sun Tzu, ~400BC.

Figure out what you are going to do, then collect data. Don't collect data first and then try and figure out what you are going to do.

Target Population

- ▶ Entire set of individuals to which findings are to be extrapolated
- ▶ Individual members of the population whose characteristics are to be measured are called *population elements*

Hypotheses of Interest

- ▶ Determine research questions
- ▶ Create a set of hypotheses that explore research questions
 - May be simple or complex
- ▶ Determine 'best' statistical method to use to test hypotheses

Determine Sample Design

Problem: Need to sample target population efficiently.

- Provide tools for sampling elements of the frame
- ▶ **Allow access to all sub-populations of interest**
- ▶ **Minimize cost and time (effort)**
- ▶ **Maximize statistical power**

Sample Design: A problem

- ▶ **What are some difficulties to collecting data?**
- ▶ **A specific problem**
 - Drug use in the USA
 - Target population – non-institutionalized persons who are citizens or full time residents of the USA over the age of 2.
 - Research Questions –
 - Drugs use by: gender, age groups, race/ethnicity/ other SES factors.
 - Tobacco, Alcohol, Marijuana, Cocaine.

Specific problems

- ▶ **How to sample a population of interest?**
- ▶ **Is a simple random sample feasible? Why/Why not?**
- ▶ **Sufficient sample size for subgroups of interest**
 - Too small – no power or precision
 - Too big – more time and money
- ▶ **Travel costs; Time**
 - Too much travel = too much cost and time

Solutions

▶ SRS is not feasible

- No list; population too spread out; too much effort

▶ Solution: stratification

- Stratify sample by geo-region: States; groups of states; sections of states
- strata guarantee sampling of subpopulations

▶ Cluster sample within geo-region

- Select census tracts/block: then households
- Creates cluster-correlated data

Clustering and Intracluster Correlation

$$r = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Reasons for taking a clustered sample:

- ▶ **Field costs are prohibitive**
 - Widely scattered population

- ▶ **There exists no sampling frame for the population elements**
 - No list of all US residents
 - Do have a list of all states, counties, census tracts, etc

Features of Clustered Samples

- ▶ **Potential for clustermates to respond similarly**
 - intracluster correlation
- ▶ **Positive correlations or overdispersion,**
 - Increased variances estimates compared to SRS
 - Typical for most surveys
- ▶ **Failure to account for clustering usually leads to:**
 - Underestimated standard errors
 - Confidence intervals too narrow
 - Test statistics with inflated Type I error rates

Clustered Samples: The Necessary Evil

► Positives:

- Decreases costs
- ease of sampling
- Necessary due to lists

► Negatives:

- Increases standard errors
- Reduces power
- Complicates analysis

$$-\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = -\frac{1}{n} \sum (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x} \bar{y})$$

Data Collection and Data Management

- ▶ **Not as statically oriented as other tasks**
- ▶ **Starting statistician often do data management**
 - Write lots of SAS code
 - Creating variables, labels and formats
 - Cleaning data sets
 - Merging data sets
 - Creating code books
 - Creating summary information

Data Preprocessing



Software for the Statistical
Analysis of Correlated Data



Data Preprocessing

▶ Weighting

- Survey weights
- Analysis weights

▶ Imputation

- Item missing observations
- Allow all observations to be used

▶ Variance units

- Modify sample design for purposed of data analysis

▶ Data masking

- Protect confidentiality of participants

Unequal Weighting

Problem: Need unbiased estimation of population parameters

- Want statistics to reflect population of interest

▶ **Relates to *Probability* or *Population-Based Sampling***

- Every element in the target population has a known, non-zero probability of being included in the sample
- Each sample element has a sampling weight associated with their data

Sampling weight = **inverse of selection probability**

Unequal Weighting (continued)

- ▶ ***Sampling weight*** refers to the number of referenced individuals in target population.
- ▶ ***Analysis weights = Adjusted sampling weights***
 - Nonresponse adjustments
 - Post-stratification adjustments

Unequal Weighting (continued)

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

- ▶ **Downside: Weights add variability**
 - inefficiency,
 - loss of power,
 - increase in variance of sample statistics
 - wider confidence intervals.

- ▶ **Variance of sample statistics is *increased* compared to SRS.**
 - IF VAR(weights) increases then VAR(statistics) increase

Weighting solutions

- ▶ **Sample weights provide ‘truth’**
- ▶ **Analysis weights estimate ‘truth’**
- ▶ **Want to minimize bias and variance of analysis weights**
 - Use Mean Square Error
- ▶ **MSE = bias² + Variance**
- ▶ **Adjustments to weights**
 - Truncation
 - Summation to control totals
 - Adjust for non-response

Imputation

$$-\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = -\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Problem: Item missing observations

- Some individuals will not answer all questions
- Not for entire observation – non-response

▶ Want complete answer for statistical purposes

- Regression covariates.

▶ Supply answers

- Hot deck
- Cold deck

▶ Multiple Imputation

- Allows determination of variance due to imputation

Data Masking/ Disclosure

Problem: Determine individuals based on answers

- Legal problems
- ▶ **Done on Public Use surveys**
- ▶ **Involves**
 - Permutation
 - Imputation
 - Resampling – 2-phase sampling
 - Removing identifying variables.

Variance Units

Problem: Complex designs = difficult numerically

▶ **Create pseudo design units**

- Pseudo strata
- Pseudo PSU (Primary Sampling Unit)

▶ **Treat sample like a With Replacement design**

- Easier to analyze

▶ **May help mask data**

- Masked Variance Units

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

Data Analysis



Software for the Statistical
Analysis of Correlated Data



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



The Survey Statistician's Problem-Variance!!!

Point estimates=**easy**

Variance estimates=**hard**

- ▶ **Need to obtain variance estimates for ALL statistics**
 - Proportions, mean, median, beta, odds ratios, etc.
- ▶ **Features of variance estimates**
 - Estimates must adjust for all features of a survey
 - Cluster, stratification, weighting, imputation, etc
 - Robust to misspecified model assumption
 - Appropriate for non-linear functions

Point Estimates

- ▶ **EASY!**
- ▶ **Similar to standard method; with weights**
- ▶ **Example: Mean**

Non-Survey

$$\bar{x} = \frac{\sum_{i=1}^n y_i}{n}$$

Survey

$$\bar{x} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$

More point estimates

- ▶ **Linear Regression model: estimate regression coefficients**

W is diagonal matrix of weights

Non-survey

$$\hat{\beta} = [\mathbf{X}'\mathbf{X}]^{-1} \mathbf{X}'\mathbf{Y}$$

Survey

$$\hat{\beta} = [\mathbf{X}'\mathbf{W}\mathbf{X}]^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y}$$

Variance

- ▶ **HARD!**
- ▶ **No simple adjustments to standard formulas**
 - Can't just multiply by weights
- ▶ **Must account for**
 - Weighting
 - Stratification
 - Clustering : intra-cluster correlation
 - Imputation
 - Without Replacement design
 - And more!

Variance: a simple example

▶ Variance a linear Regression model

Non-Survey

$$V(\hat{\beta}) = \hat{\sigma}^2 [\mathbf{X}'\mathbf{X}]^{-1}$$

Survey

$$V(\hat{\beta}) \neq \hat{\sigma}^2 [\mathbf{X}'\mathbf{W}\mathbf{X}]^{-1}$$

- ▶ First, no $\hat{\sigma}^2$ in a survey. Why?
- ▶ Second, no adjustment for design

Variance Estimation Methods

- ▶ **Delete-1 Jackknife – early 1950s**
- ▶ **Replicate Weight Jackknife – late 1950s**
- ▶ **Balanced Repeated Replication (BRR) – late 1960s**
- ▶ **Taylor Series Linearization – early 1970s**
- ▶ **Implicit Taylor Series Linearization**
 - Generalized Estimating Equations (GEE)
 - Binder (1983)
- ▶ **Survey Bootstrap – current**
 - Not available in any software, yet.

SUDAAN

- ▶ **SUDAAN specializes in complex sample surveys, repeated measures, and cluster-correlated data.**
- ▶ **SUDAAN correctly accounts for**
 - Stratification
 - Clustering
 - Multi-stage sampling
 - Unequal weighting
 - Without replacement sampling
 - Other sample design features

Other Software

- ▶ Other standard statistical packages (e.g., SAS[®], SPSS[®]) do not uniformly address the survey data problems in all analytical procedures
- ▶ **SAS** now has some survey procedures that incorporate correlated data methods (SURVEYMEANS, SURVEYREG, SURVEYFREQ, SURVEYLOGISTIC)
- ▶ **Stata** has methods for analyzing complex survey data
- ▶ SPSS has a new module, Complex Samples

Real Data: Example using WIC data

WIC Mothers and Infants:

Two-stage clustered design

Strata = Region

PSU = WIC local agencies

Population Size: Approx 506,000 WIC participants

Sample Size = 953 WIC participants

Outcome of Interest: Breastfeeding initiation

Estimate: Percentage of mothers who breastfed their infant

Comparison Domains:

Race groups (white vs. non-whites)

Is there a difference in Breast feeding initiation between White and non-White women?

- ▶ Use CROSSTAB procedure in SUDAAN
- ▶ 2x2 Table of mother's race vs breastfeeding initiation
- ▶ Use Chi-sq test – based on weighted sample size

BFEED	BF-YES	BF-NO
X	522	431
RACE		
White	249	231
480	51.04	48.96
Non-White	273	200
473	56.40	43.60

Consequences of Not Fully Accounting for Complex Design on Chi-sq test

$$-\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = -\frac{1}{n} \sum (x_i - \bar{x})y_i$$

Method	% Breastfeed (SE): White	% Breastfeed (SE): Non-White	Chi-Square	P-value
Weights as Counts	51.04 (0.10)	56.40 (0.10)	1462.5	0.001
Weights Normalized (Sum to Sample Size)	51.04 (2.28)	56.40 (2.28)	2.75	0.097
Weighted, Accounting for Complex Design	51.04 (3.18)	56.40 (4.99)	0.94	0.343

The Final Word

- ▶ **Research takes careful planning**
- ▶ **Surveys are complicated**
- ▶ **Complex surveys are challenging to analysis**
 - **VARIANCE!!**
 - Weights, Clustering, Stratification, etc
- ▶ **Standard assumption no longer valid**
- ▶ **Standard statistical procedures/analysis are not valid**
- ▶ **Use survey appropriate software**
 - SUDAAN, SAS survey procedures, etc...

Questions/Comments

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

The RTI Statistician



Software for the Statistical
Analysis of Correlated Data



Statistics at RTI

▶ Research Areas

- SUDAAN – analysis of complex data
- Small area estimation – another survey feature
- Variance estimate
- Weight adjustment – GEM

▶ Survey Methodology

- Large surveys – National Survey on Drug Use and Health
- Smaller surveys – O*NET, NSCAW, NFLIS

▶ Data Management

▶ Sampling/Sample Design

Variety of Statistical Work

► Topical areas

- Epidemiology
- Environment
- Genomics
- Education
- Multi-site studies
- Medical

Key Skills to Being a Good Statistician

▶ Effective Communication

- Writing ability
- Oral presentation
- Communicate up and down the ladder

▶ Work well with others

- Be a good teammate
- Be able to direct your teammates
- Be able to advise your superiors (They may not be statisticians)

Skills (continued)

▶ Have a good concept of statistic

- But, become a specialist/expert in area of your choice
- You can't know everything
 - But, be aware of as much as you can
 - Never hesitate to ask questions!

▶ Programming skills

- Learn SAS
- Learn SAS
- Learn SAS

Parting Thoughts

- ▶ **Statistics is everywhere!**
 - See JSM as an example
 - Also a good place to look for jobs 😊
- ▶ **Statistician typically work with others**
 - Pure statisticians are rare
 - Find your field of interest
 - Pursue this with reckless abandon

The End



Software for the Statistical
Analysis of Correlated Data

