Applied Survey Data Analysis
Design-based Inference from Complex Samples
(A brief introduction!)

5th School on Survey Sampling and Survey Methodology
Cuiaba, Brazil
October 17, 2017

Presenter: Steven G. Heeringa
University of Michigan, Ann Arbor, Michigan
sheering@umich.edu
Workshop Overview

- Overview of complex sample designs and design effects due to stratification, clustering and weighting.

- Design-based inference for complex samples.
  - Weighted estimation
  - Estimation of variance
Probability Sampling

• Probability sample design:
  – Each population element has a known, non-zero selection probability
  – Properly weighted, sample estimates are unbiased or nearly unbiased for the corresponding population statistic.
  – Variance of sample statistics can be estimated from the sample data (measurability)

• Simple random sample (SRS): A probability sample in which each element has an independent and equal chance of being selected for observation. Closest population sampling analog to independently and identically distributed (iid) data.
Complex Sample Survey Designs

• “Complex sample”:

  – A probability sample developed using sampling procedures such as stratification, clustering and weighting designed to improve statistical efficiency, reduce costs or improve precision for subgroup analyses relative to SRS

  – Unbiased estimates with measurable sampling error are still possible

  – Independence of observations, (iid), equal probabilities of selection may no longer hold
Where are Complex Sample Designs Used?

• Complex sample designs are the rule and not the exception in sample-based studies in the Social Sciences, Epidemiology, Public Health, Agriculture, Natural Resources and many other scientific fields.

• Common complex sample designs
  – Area probability samples of household populations
  – Multi-stage samples of schools, classes and students
  – Stratified samples of businesses, hospitals
  – Dual-frame and other samples in agriculture
  – RDD telephone samples
  – Natural resource samples (time and location samples)
Design Effects in Complex Sample Designs
Concept of Design Effect

\[ D^2(\hat{\theta}) = \frac{\text{Var}(\hat{\theta})_{\text{complex}}}{\text{Var}(\hat{\theta})_{\text{srs}}} = \frac{SE(\hat{\theta})^2_{\text{complex}}}{SE(\hat{\theta})^2_{\text{srs}}} \]

where:

- \( D^2(\hat{\theta}) \) = the design effect for the sample estimate, \( \hat{\theta} \);
- \( \text{Var}(\hat{\theta})_{\text{complex}} \) = the complex sample design variance of \( \hat{\theta} \); and
- \( \text{Var}(\hat{\theta})_{\text{srs}} \) = the simple random sample variance of \( \hat{\theta} \).
Design Effects for Complex Samples

\[ D^2(\hat{\theta}) \approx 1 + f(G_{\text{strat}}, L_{\text{cluster}}, L_{\text{weighting}}) \]

where:

- \( G_{\text{strat}} \) = the relative gain in precision from stratified sampling compared to SRS;
- \( L_{\text{cluster}} \) = the relative loss of precision due to clustered selection of sample elements;
- \( L_{\text{weighting}} \) = the relative loss of precision due to unequal weighting for sample elements.
General Influence of Design and Estimation Features on Variances

![Graph showing the relationship between sample size and standard error of P, with arrows indicating the effects of stratification, clustering, and weighting.]
Stratified Random Sampling Precision Gains (with Proportionate allocation)

\[ \Delta = \text{Var}(\bar{y}_{srs}) - \text{Var}(\bar{y}_{st, pr}) \]

\[ = \frac{S^2_{total}}{n} - \frac{S^2_{within}}{n} = \left[ \frac{S^2_{within} + S^2_{between}}{n} \right] - \frac{S^2_{within}}{n} \]

\[ = \frac{S^2_{between}}{n} \]

\[ = \frac{\sum_{h=1}^{H} W_h (\bar{Y}_h - \bar{Y})^2}{n} \]
Model of Design Effect for Means Under Cluster Sampling

• The design effect for the mean of a cluster sample depends on the cluster size \((B)\) and the homogeneity of the elements within the clusters, measured by the intra-class correlation, \((\rho)\):

\[
D^2(\overline{y}) = 1 + (B - 1) \rho.
\]

• For estimates of means or proportions, the intra-class correlation may be estimated by:

\[
\hat{\rho} = \frac{d^2(\overline{y}) - 1}{B - 1} = \frac{3.029 - 1}{24 - 1} = 0.088.
\]
Intra-class Correlation, $\rho$  (1)

- $\rho$ is a property of the naturally occurring clusters and of the variable under study.

- Since elements in a cluster tend to be similar to one another, $\rho$ is virtually always positive. For human populations, a positive $\rho$ may be due to:
  - *Environment*: exposure to the same climate, hazardous waste, or infectious disease.
  - *Self selection*: wealthy households tend to reside in wealthy neighborhoods and poor households in poor neighborhoods.
A Simple Model for Losses In Precision due to Weighting

$L_w = \text{Loss due to weighting} = \text{Proportionate increase in variance of an estimate of a mean.}$

$$L_w \sim CV^2(W_i) = \frac{\text{Var}(W_i)}{W^2}$$

$$= \left[ \frac{\sum_{1}^{n} W_i^2}{\left(\sum_{1}^{n} W_i \right)^2} \right] \cdot n - 1$$
Example of Weighting Loss in Disproportionate Stratified Sampling

<table>
<thead>
<tr>
<th>Stratum</th>
<th>% Hispanic Population</th>
<th>Oversampling Rate</th>
<th>Weight</th>
<th>% of Hispanic Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.2%</td>
<td>1:1</td>
<td>4</td>
<td>7%</td>
</tr>
<tr>
<td>2</td>
<td>22.8%</td>
<td>2:1</td>
<td>2</td>
<td>17%</td>
</tr>
<tr>
<td>3</td>
<td>24.1%</td>
<td>3:1</td>
<td>1.33</td>
<td>26%</td>
</tr>
<tr>
<td>4</td>
<td>33.9%</td>
<td>4:1</td>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

For \( n = 1000 \):

\[
L_W = \left[ \frac{\sum_{i=1}^{n} W_i^2}{\left( \sum_{i=1}^{n} W_i \right)^2} \right] \cdot n - 1 = \left[ \frac{2759.9}{2,148,569} \right] \cdot 1000 - 1 = .284
\]
Estimation and Inference for Complex Sample Designs
Classification of Sample Designs for Survey Data: Hansen, Madow and Tepping (1983)

<table>
<thead>
<tr>
<th>Sampling Plan</th>
<th>Method of Inference</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Design-based</td>
<td>Model-based</td>
</tr>
<tr>
<td>Probability Sample</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Model-dependent Sample</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>
Sampling Design for Population Studies

• **Sample Plan** – framework and method for selecting the observational units
  – *Probability sampling*: Hansen, Hurwitz, Madow (1953); Kish (1965)
  – *Model-assisted sampling*: Sarndal, Swensson, Wretman (1992)

• **Estimation and Inference**
  - *Design-based* (**focus** of this workshop)
  - *Model-based*
    - Bayesian Models (Little, 2003, Ch. 4 in *Analysis of Survey Data*)
Statistical Inference (Model-based)

1) Incorporates a probability model, \( f(y|x, \theta) \) for the variables.
   a) ex: \( Y(1), \ldots, Y(n) \) are independent observations from a standard normal distribution with mean \( \mu \) and variance \( \sigma^2 \).
   b) ex: coin flip – Bernoulli trial
      \[
      \begin{align*}
      P(\text{Head}) &= .50 \\
      T(\text{Tail}) &= .50
      \end{align*}
      \]

2) Goal is to estimate the parameters, \( \theta \), of \( f(y|x,\theta) \). Assumes no finite limit to the process that generates the sample \( Y_s \).

3) Estimators, tests of hypothesis, confidence intervals derived on the basis of the probability model. Extremely powerful if the probability model is correct.
   ex:
   \[
   F = \frac{GM_1M_2}{R^2} + \text{error}
   \]
“Design-Based” Inference from Sample Surveys

• Goal is inference about $\theta$, the value of the statistic in a finite population;
• Inference is based on the expected distribution of sample estimates, not on a probability model, $f(y|x,\theta)$ for Ys;
• Confidence interval approach to inference about $V$

\[
\hat{\theta} \pm t_{df,1-\alpha/2} \cdot se(\hat{\theta})
\]

Example: $\bar{y} \pm t_{x,.975} \cdot se(\bar{y})$

$30.0 \pm 1.96 \times (1.30)$
$\quad (27.45, 32.54)$

• $t_{df}(p)$ is the value of the Student – $t$ distribution with df degrees of freedom cutting off tail-probability $p/2$. Assumes normality of the sampling distribution of $\hat{\theta}$

• $se(\hat{\theta})$ the estimated standard error (square root of the variance) of the sample estimate, $\hat{\theta}$. Special computational formulas and/or methods are needed.
Sampling Distributions: SRS and Cluster Sampling

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>SRS</th>
<th>Cluster Size, b=10</th>
<th>Cluster Size, b=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>25.04614</td>
<td>24.95748</td>
<td>Mean 25.04225</td>
</tr>
<tr>
<td>SD</td>
<td>2.811056</td>
<td>4.607751</td>
<td>SD 8.485906</td>
</tr>
<tr>
<td>n=1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>24.9862</td>
<td>25.02338</td>
<td>Mean 24.97321</td>
</tr>
<tr>
<td>SD</td>
<td>2.019553</td>
<td>3.219163</td>
<td>SD 5.898481</td>
</tr>
<tr>
<td>n=5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>25.0038</td>
<td>24.99775</td>
<td>Mean 25.02288</td>
</tr>
<tr>
<td>SD</td>
<td>0.909071</td>
<td>1.438167</td>
<td>SD 2.620669</td>
</tr>
</tbody>
</table>

Mean of y
Three Elements of Design-Based Inference Using Confidence Intervals (CIs)

(1) \( \hat{\theta} \) ± (2) \( t_{df,1-\alpha/2} \cdot se(\hat{\theta}) \)

where:

\( \hat{\theta} = \) survey weighted estimate of \( \theta \);

\( t_{df,1-\alpha/2} = \) critical value from Student t with \( df \);

\( se(\hat{\theta}) = \) robust, design-corrected estimate of SE(\( \hat{\theta} \)).
Weighting and Estimation
Survey Weights

Weighting may simultaneously incorporate all three components: unequal probabilities of selection, nonresponse, and post-stratification

1) Weight for unequal probabilities of selection: $w_{sel}$;

2) Weight for sample nonresponse: $w_{nr}$;

3) Poststratification weight for population noncoverage and sampling variance reduction: $w_{ps}$.

Then compute the overall weight as: $w = w_{sel} \times w_{nr} \times w_{ps}$

Examples of Survey Weighted Estimates

\[
\bar{y}_w = \frac{\sum_{i=1}^{n} W_i \cdot y_i}{\sum_{i=1}^{n} W_i} \quad \text{estimates } \bar{Y};
\]

\[
s_w^2 = \frac{\sum_{i=1}^{n} W_i \cdot (y_i - \bar{y})^2}{\sum_{i=1}^{n} W_i - 1} \quad \text{estimates } S^2 ;
\]

\[
b_{1,w} = \frac{\sum_{i=1}^{n} W_i \cdot y_i \cdot x_i}{\sum_{i=1}^{n} W_i \cdot x_i^2} \quad \text{estimates the simple linear regression coefficient, } B_1.
\]
Survey Weighted Estimation: Pseudo-Maximum Likelihood for Logistic Regression

Pseudo ln(Likelihood):

\[ Pseudo\ ln(Likelihood): = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{\alpha}^h} w_{h\alpha i} y_{h\alpha i} \cdot \ln(\pi(x_{h\alpha i})) + \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{\alpha}^h} w_{h\alpha i} (1 - y_{h\alpha i}) \cdot \ln(1 - \pi(x_{h\alpha i})) \]

where:

\[ \pi(x_i) = e^{x_iB} / (1 + e^{x_iB}), \quad \hat{\pi}(x_i) = e^{x_iB} / (1 + e^{x_iB}) \]

and \( b \) = the vector of coefficient estimates that solves:

\[ U(B) = \frac{\delta \ln L(B)}{\delta B} \bigg|_{B=b} = 0 = \]

\[ = \sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} x_{h\alpha i} y_{h\alpha i} - \sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} x_{h\alpha i} \hat{\pi}(x_{h\alpha i}) \]
Degrees of Freedom
Three Elements of Design-Based Inference (CIs)

(1) \( \hat{\theta} \pm t_{df,1-\alpha/2} \cdot se(\hat{\theta}) \)

where:

\( \hat{\theta} \) = survey weighted estimate of \( \theta \);

\( t_{df,1-\alpha/2} \) = critical value from Student t with df;

Note: df here are for the sampling distribution of estimates!

\( se(\hat{\theta}) \) = robust, design-corrected estimate of SE(\( \hat{\theta} \)).
Degrees of Freedom in Variance Estimation for Complex Sample Data

Simple Rule:

\[
\text{degrees of freedom} = \text{# of clusters} - \text{# of strata} = \sum_{h=1}^{H} a_h - H
\]

ex: Two cluster per stratum design.

\[
d.f. = 2 \cdot H - H = H
\]

See: Valliant, R. and Rust, K.F., Degrees of Freedom Approximations and Rules-of-Thumb, *Journal of Official Statistics*, Vol. 26, No. 4, 2010, pp. 585–602. (They propose a simple estimator of degrees of freedom that leads to improved confidence interval coverage relative to the simple rule above, which is currently used by most software packages.)
Degrees of Freedom in Confidence Interval Construction

\[ CI_{.95} = \hat{\theta} \pm t_{.975,df} \text{ se}(\hat{\theta}) \]

\[ t_{.975,1} = 12.706 \]
\[ t_{.975,5} = 2.5706 \]
\[ t_{.975,10} = 2.2281 \]
\[ t_{.975,20} = 2.0860 \]
\[ t_{.975,30} = 2.0423 \]
\[ t_{.975,40} = 2.0211 \]
\[ t_{.975,\infty} = 1.9600 \]

\[ Z_{.975} = 1.9600 \]
Variance of Sample Estimates
Three Elements of Design-Based Inference (CIs)

\[ \hat{\theta} \pm t_{df,1-\alpha/2} \cdot se(\hat{\theta}) \]

where:

\[ \hat{\theta} = \text{survey weighted estimate of } \theta; \]
\[ t_{df,1-\alpha/2} = \text{critical value from Student t with df;} \]
\[ se(\hat{\theta}) = \text{robust, design-corrected estimate of } SE(\hat{\theta}). \]
Complex Sample Variance Estimation

• Direct (closed form) results exist for estimating variances of descriptive estimates computed from Simple Random, Stratified Random or Equal-size Cluster Samples (which are linear statistics).

• Direct (closed form) results exist for estimating variances of linear statistics, such as totals:

\[
\theta = \sum_{i=1}^{N} a_i \cdot y_i ; \quad \hat{\theta} = \sum_{i=1}^{n} a_i \cdot y_i
\]

**Example 1, Population Total:**

\[
Y = \sum_{i=1}^{N} y_i ; \quad \hat{Y} = \sum_{i=1}^{n} w_i \cdot y_i = \sum_{i=1}^{n} y_i^*
\]

• More complicated if the estimate is a nonlinear function of sample quantities (weights and unequal size clusters make most common statistics nonlinear functions of random variables).
• In general, weighted survey estimates are nonlinear functions of linear statistics:

\[
\bar{y}_w = \frac{\sum_{h} \sum_{\alpha} \sum_{i} y_{h\alpha i} w_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i}} = \frac{x}{z}
\]

\[y_{h\alpha i} = \text{Measurement on unit } i \text{ in cluster } \alpha \text{ in stratum } h\]

\[w_{h\alpha i} = \text{Corresponding weight}\]

• Certain other statistics such as regression coefficients and correlation coefficients are also nonlinear functions of linear statistics:

\[
\hat{b} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} y_{h\alpha i} x_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} x^2_{h\alpha i}} = \frac{u}{v}
\]
Complex Sample Variance Estimation Methods

• Taylor Series approximation or linearization technique
  • Approximate nonlinear statistics as a linear function of estimates of totals, derive corresponding variance estimator
  • Leads to a specific form of the variance estimator for each statistic (default variance estimation method in SAS and Stata)

• Replication or Resampling Methods
  • Jackknife Repeated Replication (JRR)
  • Balanced Repeated Replication (BRR)
  • Bootstrap Methods
Design Variables for Variance Estimation

- General design variable inputs:
  - Stratum Code (e.g. \textit{stratum\_var}, $h = 1,\ldots,H$)
  - Cluster Code for PSUs or Elements (min. 2 / stratum)
  - Final Survey Weight ($w_i$ for each case, $i = 1,\ldots,n$)

- SAS V9.2+ (Design variables included in command code)

  \begin{verbatim}
  PROC SURVEYMEANS DATA = \textit{example};
    STRATUM \textit{stratum\_var};
    CLUSTER \textit{cluster\_var};
    WEIGHT \textit{wgt\_var};
    VAR \textit{varname};
  RUN;
  \end{verbatim}

- Stata (Global declaration of design variables using \texttt{svyset})

  \begin{verbatim}
  svyset \textit{cluster\_var} [pweight = \textit{wgt\_var}],
    strata(\textit{stratum\_var})
  \end{verbatim}
Alternative for Replication Variance Estimates

• For replicated variance estimation methods, alternative is to provide a vector of replicate weights (1 weight per replicate)
  – Eliminates the need to release sample design stratum and cluster groups, enhances disclosure protection (still need method used)
  – Enables data producers to perform nonresponse adjustment and post-stratification for each replicate

• **SAS V9.2+**
  
  PROC SURVEYMEANS DATA= VARMETHOD=JACKKNIFE;
  REPWEIGHTS=WT1-WT52; *For 52 replicate weights;
  WEIGHT wgtvar;
  VAR varname;
  RUN;

• **Stata V12.1+**

  svyset [pweight=wgtvar], jkrweight(wt*) vce(jackknife)
  svy: mean varname
Simplifying Assumptions in Practical Variance Estimation

• Primary stage units (PSUs) in multi-stage sample designs are considered to be selected with replacement from the primary stage strata. Any finite population correction for the primary stage sample is ignored. The resulting estimates of sampling variance will be slight overestimates (see Kish, 1965, 5.3B).

• Multi-stage sampling within selected PSUs results in a single ultimate cluster of observations for that PSU.

• Assume: single-stage selection of ultimate clusters, with replacement, where all elements in a given ultimate cluster are sampled. Greatly simplifies variance estimation formulae, and variance in ultimate clusters from PSUs within a stratum is the dominant source of variance in sample estimates.
Taylor Series Linearization: Function of 2 variables

\[ f(x, z) \approx f(x_o, z_o) + (x - x_o) \left( \frac{\partial f}{\partial x} \right)_{x=x_o, z=z_o} + (z - z_o) \left( \frac{\partial f}{\partial z} \right)_{x=x_o, z=z_o} \]

\[ \text{var}(f(x, z)) \approx A^2 \text{var}(x) + B^2 \text{var}(z) + 2AB \text{cov}(x, z) \]
Taylor Series Linearization: Function of 2 or more variables

\[ f(x, z) = R = \bar{y}_w = \frac{\sum_h \sum_\alpha \sum_i y_{h\alpha i} w_{h\alpha i}}{\sum_h \sum_\alpha \sum_i w_{h\alpha i}} = \frac{x}{z} \]

\[ A = \frac{\partial f}{\partial x} = \frac{1}{z_o} \]

\[ B = \frac{\partial f}{\partial z} = -\frac{x_o}{z_o^2} \]

\[ \text{Var}\left(\frac{x}{z}\right) \approx \text{Var}(x) + R^2 \times \text{Var}(z) - 2 \times R \times \text{Cov}(x, z) \]

\[ R = \frac{x_o}{z_o} \]

where: \( x_o, z_o \) are the values of \( x, z \) computed from the sample.
Simple Data Example  (4 Strata, 2 PSUs/Stratum)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>PSU (Cluster)</th>
<th>Case</th>
<th>( y_i )</th>
<th>( w_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.58</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.48</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>.42</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.57</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.46</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>.50</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.21</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>.44</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.43</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>.64</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.55</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.50</td>
<td>2</td>
</tr>
</tbody>
</table>
Taylor Series Approach: Variance Estimate for Mean of \( y \) in Data Example

\[
\bar{y}_{w,TSL} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} y_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i}} = \frac{\sum_{h} \sum_{\alpha} \sum_{i} x_{h\alpha i}}{\sum_{h} \sum_{\alpha} \sum_{i} z_{h\alpha i}} = \frac{x}{z} = \frac{11.37}{24} = 0.47375
\]

\[z_0 = \sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i} = 24\]

\[\text{var}(x) = 0.9777; \quad \text{var}(z) = 6.0000; \quad \text{cov}(x, z) = 2.4000\]

\[\text{var}(\bar{y}_{w,TSL}) \approx \frac{\text{var}(x) + (\bar{y}_{w,TSL})^2 \cdot \text{var}(z) - 2 \cdot \bar{y}_{w,TSL} \cdot \text{cov}(x, z)}{v_0^2}\]

\[= 0.9777 + 0.4737^2 \cdot 6.0000 - 2 \cdot 0.4737 \cdot 2.4000
\]

\[= \frac{0.00008731}{24^2} = 0.009343\]

\[se(\bar{y}_{w,TSL}) = 0.009343\]
Alternatives to Taylor Series Linearization

- The linearization technique is useful if the estimate can be expressed as a function of sample totals; readily available in SAS and Stata
- Linearization requires analytic manipulations, computation of derivatives
- Not directly suitable for percentiles such as the median or functions of percentiles (non-smooth functions)
- Replication techniques are useful for almost any type of estimate
Common Steps in Replicated Methods

1. Create replicates **using sampled cases** (unique to each method). Software uses sampling error codes provided in the data set.
   - **JRR replicate sample**: leave out one PSU, and re-weight cases in “deletion stratum”.
   - **BRR replicate sample**: given two PSUs per stratum, leave out one PSU from each stratum (based on a “balanced” Hadamard matrix to eliminate covariances when estimating variances), and re-weight cases (see ASDA, Chapter 3).

2. Create revised weight for replicate sample.

3. Compute weighted estimates of population statistic of interest for each replicate using replicate weight.

4. Apply replicated variance estimation formula to derive standard errors.

5. Construct confidence intervals (or hypothesis tests) based on estimated statistics, standard errors, and correct degrees of freedom (generally use a – H approximation).
Replication-based Variance Estimation: Options in SAS and Stata

- **Jackknife Repeated Replication (JRR)**
  - SAS: PROC SURVEY VARMETHOD = JK;
  - Stata: Use `svy, vce(jackknife):` commands

- **Balanced Repeated Replication (BRR)**
  - SAS: PROC SURVEY VARMETHOD = BRR;
  - Stata: Use `svy brr, hadamard(...)`: commands
  - Requires two cluster/PSU codes per stratum!

- **Bootstrapping**
  - SAS: Currently no option
  - Stata: see Kolenikov (2010, Stata Journal).
  - Example:
    - `. svyset [pw=finalwgt], vce(bootstrap) bsrweight(bw*)`
    - `. svy: logistic highbp height weight age female`
Replicate Weighting

• JRR, BRR, and bootstrap programs adjust the values of the sample selection weights to create “replicate weights” using the above procedures. Default - no adjustment to original survey weights.

• If nonresponse adjustment and post-stratification adjustments are included in the final survey weight, survey statisticians advocate that these be recomputed for each sample replicate.

• WesVar PC (Westat, Inc.) assists in developing replicate weight adjustments that reflect nonresponse and post-stratification.

• **Replicate weights** are added to data set:

\[ w=[w(1), ..., w(K)] \]
JRR: Constructing Replicates, Replicate Weights

- There are several approaches to constructing the JRR replicates. The Delete One method that is often the default in survey analysis software is illustrated here.

- For the Delete One JRR variance estimator, each stratum will contribute \( a(h) \) replicates, where for each replicate one of the clusters in a single stratum is deleted.

- In our data example, \( H=4 \) and \( a(h)=2 \) for \( h=1,\ldots,4 \), so the total number of required JRR replicates for the Delete One JRR is:

\[
\sum_{h=1}^{4} (a_h) = 8
\]
**JRR: Constructing Replicates, Replicate Weights**

- Suppose there are $H$ strata with $a(h)$ clusters. $H=4$, $a(h)=2$ in example data set.

- A JRR replicate is constructed by deleting one PSU from one stratum. The first replicate leaves out the 1st PSU in Stratum 1.

- The replicate weight for this first replicate multiplies the weights for remaining cases in the “deletion stratum” by a factor of $a(h)/[a(h)-1]$. This equals 2 in our example. Replicate 1 weight values remain unchanged for cases in all other strata.
## JRR Replicate 1: Data Example

<table>
<thead>
<tr>
<th>Stratum</th>
<th>PSU (Cluster)</th>
<th>Case</th>
<th>$y_i$</th>
<th>$w_{i,\text{rep}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>.42</td>
<td>2x2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>.57</td>
<td>2x2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>.46</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>.50</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>.21</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>.44</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>.43</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>.64</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>.55</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>.50</td>
<td>2</td>
</tr>
</tbody>
</table>
### JRR Replicate 2 (of 8): Data Example

<table>
<thead>
<tr>
<th>Stratum</th>
<th>PSU (Cluster)</th>
<th>Case</th>
<th>$y_i$</th>
<th>$w_{i,rep}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.58</td>
<td>1x2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.48</td>
<td>2x2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.46</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>.50</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.21</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>.44</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.43</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>.64</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.55</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.50</td>
<td>2</td>
</tr>
</tbody>
</table>
JRR Variance Estimation: Delete 1

- If stratum and cluster codes are available in the data set:

\[
\text{var}_{JRR}(\hat{q}) = \sum_{h=1}^{H} \frac{a_h - 1}{a_h} \sum_{\alpha=1}^{a_h} (\hat{q}_{(h\alpha)} - \hat{q})^2
\]

where

\(\hat{q}_{(h\alpha)}\) = weighted estimate of Q for replicate where cluster \(\alpha\) in stratum \(h\) has been deleted;

\(df = a - H = \sum_{h=1}^{H} a_h - H\)

- If only \(a\) replicate weights are available in the data set:

\[
\text{var}_{JRR}(\hat{q}) = \frac{a - 1}{a} \sum_{k=1}^{a} (\hat{q}_k - \hat{q})^2
\]

\(df = a - 1\)
JRR: Estimating the Sampling Variance

\[
\text{var}_{JRR\_DelOne} (\hat{q}) = 0.5 \cdot \sum_{k=1}^{8} (\hat{q}_k - \hat{q})^2
\]

\[
= 0.5 \cdot \sum_{k=1}^{8} (\bar{y}_k - \bar{y})^2
\]

\[
= .0000801449
\]

\[
se_{JRR} (\bar{y}) = \sqrt{.0000801449} = .008952
\]

\[
CI (\bar{y}) = 0.47375 \pm t_{1-\alpha/2, 4} \cdot .008952
\]

\[
= 0.47375 \pm 2.7764 \cdot (.008952) = (0.44890, 0.49860)
\]
Half Sample Replicates

• Assume a paired selection design (2 PSUs per stratum design)
• A half sample is defined by choosing one PSU from each stratum.
• A complement of a half sample is made up of all those PSUs not in the half sample. A complement is also a half sample.
• There are $2^H$ possible half samples and their complements. We only need H half samples for variance estimation. (Why?)
# Half-Sample Replication

Consider the following choice of 4 half samples:

<table>
<thead>
<tr>
<th>Halfsample</th>
<th>Stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

+ First element (PSU) is selected
- Second element (PSU) is selected
Balanced Repeated Replication (BRR)

• Square arrays of + and - that define balanced half samples are called Hadamard matrices. Plackett and Burman (Biometrika, 1946) have tabulated these matrices for H=4,8,12,16,…, 200 (all multiples of 4)

• Computer algorithms are available for creating these matrices (Stata, WesVar PC).

• What if the number of strata is not multiple of 4? In this case only partial balance can be achieved.

• For H=3 drop one column from the matrix for H=4. For H=5 choose the matrix with H=8 and drop 3 strata (columns).
BRR: Replicates, Replicate Weights

• Suppose there are H strata with a(h) clusters. H=4, a(h)=2 in example data set

• A BRR replicate is constructed by deleting one PSU from each stratum according to the pattern specified in the Hadamard matrix. Example: the first half-sample replicate leaves out the 2\textsuperscript{nd} PSU in Strata 1,2,3 and the 1\textsuperscript{st} PSU in Stratum 4.

• The replicate weight for this first replicate multiplies the weights for remaining cases in the half-sample by a factor of 2.
BRR Replicate 1: Data Example

<table>
<thead>
<tr>
<th>Stratum</th>
<th>PSU (Cluster)</th>
<th>Case</th>
<th>$y_i$</th>
<th>$w_{i,rep}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.58</td>
<td>1x2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.48</td>
<td>2x2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1x2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.46</td>
<td>2x2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>.39</td>
<td>1x2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.47</td>
<td>2x2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>.47</td>
<td>2x2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.50</td>
<td>2x2</td>
</tr>
</tbody>
</table>
BRR: Constructing Replicates, Replicate Weights(2)

• Example: the second replicate leaves out the 2\textsuperscript{nd} PSU in Stratum 1 and the 1\textsuperscript{st} PSU in Strata 2,3,4.

• Again, the replicate weight for this second replicate multiplies the weights for remaining cases in the half-sample by a factor of 2.

• Four half-sample replicates are created based on the deletion pattern in the Hadamard matrix
## BRR Replicate 2: Data Example

<table>
<thead>
<tr>
<th>Stratum</th>
<th>PSU (Cluster)</th>
<th>Case</th>
<th>$y_i$</th>
<th>$w_{i,2,\text{brr}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.58</td>
<td>1x2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>.48</td>
<td>2x2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>.50</td>
<td>2x2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.21</td>
<td>1x2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>.44</td>
<td>1x2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.43</td>
<td>1x2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>.47</td>
<td>2x2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>.50</td>
<td>2x2</td>
</tr>
</tbody>
</table>
BRR: Constructing Estimates

Four replicates are created and the weighted estimate:

\[ \hat{q}_r = \bar{y}_r = \frac{\sum_{i \in \text{rep}} y_i \cdot w_{i,\text{rep}}}{\sum_{i \in \text{rep}} w_{i,\text{rep}}} \]

is computed for each of \( r=1,\ldots,4 \) replicates

\[ \hat{q}_1 = 0.4708; \quad \hat{q}_2 = 0.4633; \quad \hat{q}_3 = 0.4614; \quad \hat{q}_4 = 0.4692; \]

\[ \hat{q} = \frac{\sum_{i=1}^{n} y_i \cdot w_i}{\sum_{i=1}^{n} w_i} \]

the full sample estimate is also computed.

\[ \hat{q} = 0.4737 \]
BRR: Estimating the Sampling Variance

$$\text{var}_{BRR}(\hat{q}) = \frac{1}{c} \sum_{r=1}^{4} (\hat{q}_r - \hat{q})^2$$  
  
[Formula v2 for half samples]

$$= \frac{1}{c} \sum_{r=1}^{4} (\bar{y}_r - \bar{y})^2$$

$$= .00007248$$

$$\text{se}_{BRR}(\bar{y}) = \sqrt{.00007248} = .008515$$

$$\text{CI}(\bar{y}) = 0.47375 \pm t_{1-\alpha/2,4} \cdot .008515$$

$$= 0.47375 \pm 2.7764 \cdot (.008515) = (0.45011, 0.49739)$$

Recall that JRR resulted in a 95% CI of (0.4493, 0.4981), resulting in extremely similar inferences.
Rao-Wu Rescaling Bootstrap: Replicate Formation

• b=1,…, B bootstrap replicates formed by sampling with replacement (SWR) \( m_h \) PSUs from each of the h=1,…H primary stage strata.

• Recommendation (Rust and Rao, 1996) is to set \( m_h = a_h - 1 \) (one less than the number of sample PSUs in stratum h)

• For \( m_h = a_h - 1 \), the bootstrap weight for the selected replicate is:
  \[
  w_{h\alpha i}^{(b)} = w_{h\alpha i} \cdot \frac{a_h}{(a_h - 1)} \cdot r_{h\alpha}^{(b)},
  \]

  where : \( r_{h\alpha}^{(b)} = \) the count of times PSU \( \alpha \)
  is selected in replicate b.
Rao-Wu Rescaling Bootstrap: Variance Estimation

- Variance estimate for \( b=1, \ldots, B \) bootstrap estimates is the Monte Carlo approximation:

\[
\text{var}_{\text{Boot}}(\hat{q}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{q}_b - \hat{q})^2
\]

- For large numbers of bootstrap replicates, i.e. \( b>>100 \) a histogram of the bootstrap simulates the sampling distribution of the estimator
  - May be used to examine asymptotic normality assumption
  - May be used to derive asymmetric CIs for population values
Empirical Comparison of Methods

• Data Set: Health and Retirement Survey: HRS Wave 1 (1992)

• Software/Variance estimation methods
  – SRS
  – WesVar BRR and JRR with Replicate Weighting for Nonresponse and Poststratification
  – SAS TSL, BRR, JRR using only full sample weight and program creation of replicates
Comparison of Variance Estimation Methods
HRS (1992) Means, Full Adult Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Mean (Wgt)</th>
<th>SRS (SE)</th>
<th>W-BRR (SE)</th>
<th>W-JKK (SE)</th>
<th>S-TSL (SE)</th>
<th>S-BRR (SE)</th>
<th>S-JKK (SE)</th>
<th>R-BS (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of School</td>
<td>9759</td>
<td>12.31</td>
<td>0.031</td>
<td>0.078</td>
<td>0.077</td>
<td>0.077</td>
<td>0.077</td>
<td>0.077</td>
<td>0.076</td>
</tr>
<tr>
<td>Body Weight (lbs)</td>
<td>9759</td>
<td>172.35</td>
<td>0.374</td>
<td>0.447</td>
<td>0.432</td>
<td>0.435</td>
<td>0.434</td>
<td>0.435</td>
<td>0.404</td>
</tr>
<tr>
<td>Memory Recall (Words in Test 10)</td>
<td>9085</td>
<td>7.57</td>
<td>0.028</td>
<td>0.064</td>
<td>0.064</td>
<td>0.066</td>
<td>0.066</td>
<td>0.066</td>
<td>0.067</td>
</tr>
</tbody>
</table>
Comparison of Variance Estimation Methods

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Coef Est (Wgt)</th>
<th>SRS (SE)</th>
<th>W-BRR (SE)</th>
<th>W-JKK (SE)</th>
<th>S-TSL (SE)</th>
<th>S-BRR (SE)</th>
<th>S-JKK (SE)</th>
<th>R-BS (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>4.88</td>
<td>0.120</td>
<td>0.170</td>
<td>0.170</td>
<td>0.166</td>
<td>0.170</td>
<td>0.166</td>
<td>0.178</td>
</tr>
<tr>
<td>Male</td>
<td>-0.93</td>
<td>0.052</td>
<td>0.063</td>
<td>0.065</td>
<td>0.064</td>
<td>0.064</td>
<td>0.064</td>
<td>0.061</td>
</tr>
<tr>
<td>Black</td>
<td>-1.24</td>
<td>0.084</td>
<td>0.089</td>
<td>0.091</td>
<td>0.093</td>
<td>0.094</td>
<td>0.093</td>
<td>0.096</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.51</td>
<td>0.107</td>
<td>0.140</td>
<td>0.143</td>
<td>0.141</td>
<td>0.141</td>
<td>0.141</td>
<td>0.139</td>
</tr>
<tr>
<td>SCHLYRS</td>
<td>0.27</td>
<td>0.009</td>
<td>0.011</td>
<td>0.012</td>
<td>0.011</td>
<td>0.011</td>
<td>0.011</td>
<td>0.012</td>
</tr>
</tbody>
</table>
Software Packages for Complex Samples

• **SAS V9.3+** PROC SURVEYxxxx commands

• **STATA v13+** svy procedures (www.stata.com)

• **SUDAAN V11+** (www.rti.org/sudaan)
  - Stand-alone or SAS-callable
  - Large set of PROCS

• **IBM SPSS® Statistics**: Complex Samples Module
  – Add-on module, requires SPSS base to run
  – Needs to be purchased separately from SPSS base
  – Visit http://www.spss.com/ for more information
Free-Ware Software Packages
(see ASDA website for Example Apps)

• WesVar PC®
  – Free stand-alone software developed by WESTAT corporation
  – Excellent graphical user interface
  – Visit http://www.westat.com/ for more information

• IVEware: SAS-based (macros) or stand-alone
  – Descriptive and regression modeling
  – Incorporates multiple imputation capability using the sequential regression or chained equations method
  – SASMOD, SYNTHESIZE, COMBINE functions
  – http://www.iveware.org
Additional Software Packages

• R Survey Package ([http://cran.r-project.org](http://cran.r-project.org))
  – Free Survey package under Packages menu (Windows) or Packages and Data (Mac)
  – svydesign(\textit{arguments}), svyrepdesign() to specify design and weights
  – svymean(), svytotal(), svyplot()
  – svyratio(), svyglm(), svyolr(), svyloglin(), svycoxph()

• Others: Mplus (for modeling)
Applied Survey Data Analysis

Site Overview

This site contains information about the text "Applied Survey Data Analysis" (first and second editions) including author biographies, links to public release data sets and related sites, code and output for analysis examples replicated in current software packages, and information about new publications of interest to survey data analysis. Other features include a FAQ log and links to other software and statistical sites. We plan to intermittently update this site with news about ongoing statistical and software advances in the field of analysis of survey data.

Special Notes from Authors

ASDA-Second Edition is Available as of June 28, 2017!

Project Overview

Applied Survey Data Analysis is the product born of many years of teaching applied survey data analysis classes and practical experience analyzing survey data. We have taught various versions of this course in the ISR/SRC Summer Institute Program, as part of University of Michigan/CSCAR, and within the Survey Methodology Program at University of Michigan and University of Maryland. Our goal has been to integrate teaching materials and practical analysis knowledge into a textbook geared to a level accessible for graduate students and working analysts who may have varying levels of statistical and analytic expertise. We intend to update the materials on this website as statistical and software improvements emerge with the goal of assisting analyst and researchers performing survey data analysis.

Information About Authors

Patricia A. Berglund is a Senior Research Associate in the Survey Methodology Program at the Institute for Social Research. She has extensive experience in the use of computing systems for data management and complex sample survey data analysis. She works on research projects in youth substance abuse, adult mental health, and survey methodology using data from Army STARRS, Monitoring the Future, the National Comorbidity Surveys, World Mental Health Surveys, Collaborative Psychiatric Epidemiology Surveys, and various other national and international surveys. In addition, she is involved in development, implementation, and teaching of analysis courses and computer training programs at the Survey Research Center-Institute of Social Research. She also lectures in the SAS® Business Intelligence Knowledge Series. mailto:berg@umich.edu
ASDA Website: Example Code and Results

- Updated review of software packages for the analysis of complex sample survey data.

**Summary**


Designed for readers working in a wide array of disciplines who use survey data in their work, this book continues recent versions of real-world survey data sets. Although the authors continue to use Stata for most examples in the software code for replicating the examples on the book's updated Web site.

**Links to Data Sets for First and Second Editions**

**National Comorbidity Survey-Replication (Collaborative Psychiatric Epidemiology Surveys)**

http://www.icpsr.umich.edu/NCS

http://www.hcpr.med.harvard.edu/ncs (for NCS-R specific information)

**National Health and Nutrition Examination Survey (National Center for Health Statistics)**

http://www.cdc.gov/nchs/

**Health and Retirement Survey (Institute for Social Research-University of Michigan)**

http://hrsonline.isr.umich.edu

**European Social Survey (ESS)**

http://www.europeansocialsurvey.org/

**United States Census Bureau**

http://www.census.gov/
THANK YOU!
Applied Survey Data Analysis: Analytic Methods and Software for Descriptive Analysis of Survey Data

5th School on Survey Sampling and Survey Methodology
Cuiaba, Brazil
October 18, 2017

Presenter: Steven G. Heeringa
University of Michigan, Ann Arbor, Michigan
sheering@umich.edu
Workshop Overview

• Overview of software for analysis of complex sample survey data

• Design-based estimation for univariate statistics

• Simple bivariate analysis of complex sample survey data
Software Packages with Programs Available for Analysis of Complex Sample Survey Data

**SAS**: The “SURVEY” procedures
- PROC SURVEYMEANS
- PROC SURVEYREG
- PROC SURVEYFREQ (Version 9+)
- PROC SURVEYLOGISTIC (Version 9+)
- PROC SURVEYPHREG

**Stata**: The “svy” commands
- svyset: define design variables
- svydes: describe survey design
- svy: mean
- svy: regress
- many additional “svy” commands
Software Available, cont’d

**IVEware**: SAS-based (macros) or stand-alone
- %DESCRIBE (descriptive statistics)
- %REGRESS (regression modeling)
- %SASMOD (additional SAS procs)

**SUDAAN**: SAS-based or stand-alone procedures
- DESCRIPT
- CROSSTAB
- REGRESS
- LOGISTIC
- many others...for more info, visit http://www.rti.org/sudaan/
Free-Ware Software Packages
(see ASDA website for Example Apps)

• **WesVar PC®**
  – Free stand-alone software developed by WESTAT corporation
  – Excellent graphical user interface

• **IVEware: SAS-based (macros) or stand-alone**
  – Descriptive and regression modeling
  – Incorporates multiple imputation capability using the sequential regression or chained equations method
  – SASMOD, SYNTHESIZE, COMBINE functions
  – [http://www.iveware.org](http://www.iveware.org)
Additional Software Packages

• R Survey Package (http://cran.r-project.org)
  – Free Survey package under Packages menu
    (Windows) or Packages and Data (Mac)
  – svydesign(*arguments*), svyrepdesign() to
    specify design and weights
  – svymean(), svytotal(), svyplot()
  – svyratio(),
    svyglm(),svyolr(),svyloglin(),svycoxph()

• Others: Mplus (for modeling)
DESCRIPTIVE ANALYSIS FOR CONTINUOUS VARIABLES
Descriptive Analysis of Complex Sample Data

- Estimation of Population Totals
- Estimation of Population Means and Proportions
- Estimation of Population Ratios
- Graphical Representation of Estimated Population Distributions
- Estimation of Population Percentiles
- Subpopulation Estimations
- Estimation of Functions of Descriptive Statistics
  - Differences of means for two subpopulations
  - Differences of means over time (longitudinal)
Descriptive Statistics: Totals

\[ \hat{Y}_w = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} = \hat{Y}. \]

\[ \hat{Y}_w = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} = \hat{M} \leq \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} \cdot 1 = \hat{N} \]

for \( Y=\{1 \text{ if member of subpopulation, } 0 \text{ otherwise} \} \)
### Descriptive Statistics: Totals (Stata)

**Total Persons w/ Lifetime MDE by Marital Status (NCS-R)**

```stata
use "s701/ncsr_analysis_examples_c1_c10_2011.dta", clear
* Generate population weights, given that weights are scaled to sum to 1
  . gen popweight = ncsrwtsh * 209128094 / 9282
  . svyset seclustr [pweight = popweight], strata(sestrat)
  . svy: total mde, over(mar3cat)
  . mat list e(b)
  . estat effects
```

<table>
<thead>
<tr>
<th>Subpop</th>
<th>n</th>
<th>Estimated Total Lifetime MDE</th>
<th>Estimated Standard Error</th>
<th>95% CI</th>
<th>(d^2(\hat{Y}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married</td>
<td>5322</td>
<td>20,304,190</td>
<td>1,584,109</td>
<td>(17,199,395, 23,408,986)</td>
<td>6.07</td>
</tr>
<tr>
<td>Sep./Wid./Div.</td>
<td>2017</td>
<td>10,360,671</td>
<td>702,601</td>
<td>(8,983,558, 11,737,783)</td>
<td>2.22</td>
</tr>
<tr>
<td>Never Married</td>
<td>1943</td>
<td>9,427,345</td>
<td>773,137</td>
<td>(7,912,024, 10,942,667)</td>
<td>2.95</td>
</tr>
</tbody>
</table>
```
Descriptive Statistics: Totals (SAS V9.2+)
Total Persons w/ Lifetime MDE by Marital Status (NCS-R)

data ncsr;
    set S701_analysis_ex_c1_c10_2011;
    ncsrwtsel_pop = ncsrwtsel * 209128094 / 9282;
run;

proc surveymeans data=ncsr nobsum df stderr clsum;
    strata sestrat;
    cluster seclustr;
    weight ncsrwtsel_pop;
    var mde;
    domain mar3cat;
run;

(Output not shown here.)
Descriptive Statistics: Means and Proportions

Means: \[ \bar{y}_w = \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{\alpha h}} w_{\alpha hi} y_{\alpha hi}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{\alpha h}} w_{\alpha hi}} = \frac{\hat{Y}}{\hat{N}} \]

Proportions: \[ \bar{y}_w = \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{\alpha h}} w_{\alpha hi} y_{\alpha hi}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{\alpha h}} w_{\alpha hi}} = \frac{\hat{M}}{\hat{N}} = p \]

for \( y = \{0, 1\} \)
Programs for Estimating Descriptive Statistics and their Standard Errors

- **SAS**: PROC SURVEYMEANS (continuous / binary variables), PROC SURVEYFREQ (categorical variables)

- **Stata**: svy: mean (continuous / binary variables), svy: prop, svy: tab (categorical variables)
Descriptive Statistics: Means and Proportions
Mean total household assets (HRS 2006).

**data hrs;** set s701.hrs_analysis_ex_c1_c9_2011; **run;**
**proc surveymeans data=hrs nobls df mean stderr clm;**
  strata stratum; cluster secu; weight kwgthh;
  domain kfinr; /* want kfinr = 1 for financial reporters of households */
  var h8atota;
**run;**

**Stata:**
. svyset secu [pweight = kwgthh], strata(stratum)
. svy, subpop(if kfinr == 1): mean h8atota
. estat effects

<table>
<thead>
<tr>
<th>n</th>
<th>df</th>
<th>$\bar{y}_w$</th>
<th>$se(\bar{y}_w)$</th>
<th>$CI_{.95}(\bar{y}_w)$</th>
<th>$d^2(\bar{y}_w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11,942</td>
<td>56</td>
<td>$527,313$</td>
<td>$28,012$</td>
<td>($471,196, $583,429)$</td>
<td>$1.52$</td>
</tr>
</tbody>
</table>
Descriptive Statistics: Ratios of Two Variables

\[ \hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{\alpha_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{\alpha_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} x_{h\alpha i}} \]

\[ \text{var}(\hat{R}) = \text{var}(\hat{X}) + \hat{R}^2 \cdot \text{var}(\hat{Y}) - 2 \cdot \hat{R} \cdot \text{cov}(\hat{Y}, \hat{X}) \neq \frac{\hat{X}^2}{\hat{Y}^2} \]
Descriptive Statistics: Ratios of Two Variables.
Ratio of HD Chol/Tot Chol (NHANES 2005-06)

- `gen age18p = 1 if age >= 18 and age != .`
- `replace age18p = 0 if age < 18`
- `svyset sdmvpsu [pweight=wtmec2yr], strata(sdmvstra)`
- `svy, subpop(age18p): ratio (lbdhdd/lbxtc)`

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$df$</td>
<td>$\hat{R}$</td>
<td>$se(\hat{R})$</td>
<td>$CI_{.95}(\hat{R})$</td>
</tr>
<tr>
<td>4996</td>
<td>15</td>
<td>0.275</td>
<td>0.002</td>
<td>(0.271, 0.280)</td>
</tr>
</tbody>
</table>
Descriptive Estimation: Don’t Forget Graphics!

- Graphics are a powerful tool in the effect descriptive presentation of survey results

- Important to remember that graphics should reflect survey weighting to provide an unbiased portrayal of population distributions

- Software systems and even programs within single systems may require special procedures to ensure graphics reflect population estimates.

- STATA enables inclusion of survey weights as PWEIGHTS in many graphics programs.
Descriptive Estimation: Histogram
Total cholesterol (NHANES 2005-06)

- generate int_wtmec2yr = int(wtmec2yr)
- histogram lbxtc if age18p [fweight=int_wtmec2yr]
Boxplot: Total Cholesterol by Gender (NHANES 2005-06)

- graph box lbxtc [pweight=wtmec2yr] if age18p==1, by(female)
Estimation of Percentiles

• Analysts are often interested in estimating more than just the mean to describe the distribution of a variable in a population
• Example: 95th percentile of PSA levels in a national sample of men over age 40
• Example: Median total assets (a very skewed variable) for rural homes
• Example: Quartiles of systolic blood pressure for African-American females
The Ungrouped Estimation Method

• First, generate a weighted estimate of the CDF for a variable:

\[
F(x) = \frac{\sum_{i=1}^{N} I(y_i \leq x)}{N}
\]

• A weighted estimator of this CDF can be written as follows:

\[
\hat{F}(x) = \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_\alpha} w_{h\alpha i} I(y_{h\alpha i} \leq x)}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_\alpha} w_{h\alpha i}}
\]
Software for Design-based Estimation of Percentiles

- SUDAAN: Taylor Series Linearization (TSL)
- WesVar PC: Balanced Repeated Replication (recommended)
- Procedures enabling appropriate computation of standard errors and confidence intervals (using BRR) are not readily implemented in the current versions of Stata and SPSS
- See Section 5.3.5 of the course text for examples using SUDAAN and WesVar
proc descript;
nest stratum secu;
weight kwgthh;
subpopn finr = 1;
var h8atota;
percentiles 25 75 / median;
setenv decwidth = 1;
run;

<table>
<thead>
<tr>
<th>Percentile</th>
<th>SUDAAN (TSL)</th>
<th>WesVar PC (BRR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{Q}_p$</td>
<td>$se(\hat{Q}_p)$</td>
</tr>
<tr>
<td>$Q_{25}$</td>
<td>$39,852$</td>
<td>$3,167$</td>
</tr>
<tr>
<td>$Q_{50}$ (Median)</td>
<td>$183,309$</td>
<td>$10,233$</td>
</tr>
<tr>
<td>$Q_{75}$</td>
<td>$495,931$</td>
<td>$17,993$</td>
</tr>
</tbody>
</table>
SAS Software for Design-based Estimation of Percentiles and Inference

• **SAS V9.2+:**
  
  PROC SURVEYMEANS DATA = … Q1 MEDIAN Q3;
  OR:
  PROC SURVEYMEANS DATA= … VARMETHOD=BRR PERCENTILES=(25 50 90…);
  – Variance estimates using replicate weights (BRR recommended) or linearization both possible
Estimation of Percentiles: SAS V9.2+
Total Household Assets (HRS 2006)

**proc surveymeans** data = hrs  <varmethod=brr> percentiles = (25 50 75) ;
  strata stratum ;
  cluster secu ;
  weight kwgthh ;
  where kfinr = 1 ;  /* domain not available; SAS still complains that where is wrong...*/
  var h8atota ; run;

---

**Data Summary**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Strata</td>
<td>56</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>112</td>
</tr>
<tr>
<td><strong>Number of Observations</strong></td>
<td><strong>12558</strong></td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>11942</td>
</tr>
<tr>
<td>Number of Obs with Nonpositive Weights</td>
<td>616</td>
</tr>
<tr>
<td>Sum of Weights</td>
<td>53853171</td>
</tr>
</tbody>
</table>

**Quantiles**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>Percentile</th>
<th>Estimate</th>
<th>Std Error</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>H8ATOTA</td>
<td>Total Assets: HH</td>
<td>25% Q1</td>
<td>39853</td>
<td>3258.139382</td>
<td>33326.094 46379.769</td>
</tr>
<tr>
<td></td>
<td>Total Assets: HH</td>
<td>50% Median</td>
<td>183309</td>
<td>9977.330641</td>
<td>163322.041 203296.031</td>
</tr>
<tr>
<td></td>
<td>Total Assets: HH</td>
<td>75% Q3</td>
<td>495931</td>
<td>17394</td>
<td>461086.518 530776.236</td>
</tr>
</tbody>
</table>
Subpopulation Analysis

• Examples of Subpopulations:
  – Men, women
  – Age groups
  – Disease groups (asthma sufferers)
  – Voters
  – Home Owners
  – Persons with income >$20,000
“Unconditional” vs. “Conditional” Analysis

• Cochran (1977), West et al. (2008) – “Conditional” analysis “conditions” on observed subpopulation sizes as though they were fixed. Results from using “if”, “by”, or “where” statements when analyzing the data.

• Conditional analysis is OK for simple random samples…

• …but not necessarily for stratified samples:
  – Distribution of subpopulation cases, \( m(h) \), to strata \( h=1,…H \) is a random variable. Rarely fixed.
  – Correct if the subpopulation of interest is used to define explicit strata, e.g., Census Region in a multi-stage national sample of U.S. households, or gender in a stratified (by sex) sample of men and women in a University student body.
“Unconditional” vs. “Conditional” Analysis

• “Unconditional” analysis treats stratum subpopulation sizes as a random variable, \( m(h), h=1, \ldots, H \)
  • Variability of subpopulation across strata and to clusters within strata (including \( m(hi)=0 \)) must be reflected in the variance estimation.

• Stata subpop() and SAS DOMAIN keywords/options ensure that this variability is reflected in standard errors of estimates.
Unconditional Analyses in Stata

- over(varname) option for command
  - varname is a categorical variable
  - Analyses (correct) will be replicated for each level of the categorical varname

- subpop(varname) for command
  - varname is a user-generated 0,1 variable where 1 indicates membership in subclass
  - applies to all svy procedures
Unconditional Analyses in Stata

• The Stata subpop( ) feature always produces the correct result.

• Stata examines the design distribution (by strata and clusters) of the subpopulation. If a stratum contains 0 cases, that stratum and its clusters do not contribute to degrees of freedom, i.e., $DF = \# \text{clusters} - \# \text{of strata}$. 
Functions of Estimates: Stata svy: mean
Mean Household Assets by Education (2006 HRS)

• `gen finr = 1`
• `replace finr = 0 if kfinr != 1`
• `svyset secu [pweight = kwgthh], strata(stratum)`
• `svy, subpop(finr): mean h8atota, over(edcat)`

<table>
<thead>
<tr>
<th>Education of Head</th>
<th>Stata Label</th>
<th>$\bar{y}_w$</th>
<th>$se(\bar{y}_w)$</th>
<th>$CI_{.95}(\bar{y}_w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11 yrs</td>
<td>1</td>
<td>$178,386$</td>
<td>$24,561$</td>
<td>$(129,184, 227,588)$</td>
</tr>
<tr>
<td>12 yrs</td>
<td>2</td>
<td>$328,392$</td>
<td>$17,082$</td>
<td>$(294,171, 362,613)$</td>
</tr>
<tr>
<td>13-15 yrs</td>
<td>3</td>
<td>$455,457$</td>
<td>$27,000$</td>
<td>$(401,369, 509,545)$</td>
</tr>
<tr>
<td>16+ yrs</td>
<td>4</td>
<td>$1,107,204$</td>
<td>$102,113$</td>
<td>$(902,646, 1,311,762)$</td>
</tr>
</tbody>
</table>
Unconditional Analyses in SAS

- PROC SURVEYMEANS / SURVEYREG / SURVEYLOGISTIC;
  domain varname;
  – keyword will produce analysis for each level of the categorical variable varname

- PROC SURVEYFREQ;
  tables domain*var1*var2;
  – Produces separate table of var1*var2 for each level of the categorical variable (e.g. gender) that is first listed in the statement.
SAS Notes

• PROC SURVEYREG, PROC SURVEYLOGISTIC: *domain* statement

• No built in feature for subpopulation analysis in these procedures, prior to Version 9.2.

• Alternatives:
  – Subsetting “if” (NOT recommended; SAS will give you a warning message if used)
SAS Code for Subpopulation Estimates Example (Output not shown).

```sas
proc format;
value edf 1='0-11 Years' 2='12 Years' 3='13-15 Years' 4='16+ Years';

title "Analysis Example 5.12: Proportions in SubGroups: HRS";
proc surveymeans data=hrs mean clm;
strata stratum;
cluster secu;
weight kwgthh;
domain kfinr*edcat;
var h8atota;
format edcat edf.;
run;
```
SAS Notes

• SAS counts all strata and clusters toward estimation of degrees of freedom, even under domain xxx; statement.

• “trick” may lead to overestimation of degrees of freedom. Strata with 0 subpopulation cases will be counted toward df, e.g., always df=32 for NHANES II. Will yield narrower CIs than subpopulation analysis in other software systems such as STATA or R.
Sampling Errors for Functions of Estimates:
Differences of Subpopulation Means

\[ \text{var} \left( \sum_{j=1}^{J} a_j \hat{\theta}_j \right) = \sum_{j=1}^{J} a_j^2 \text{var}(\hat{\theta}_j) + 2 \cdot \sum_{j=1}^{J-1} \sum_{k>j}^{K} a_j a_k \cdot \text{cov}(\hat{\theta}_j, \hat{\theta}_k) \]

where: \( a_j, a_k \) are any chosen constants.

Example:

\[ \text{var}(\bar{y}_{\text{sub1}} - \bar{y}_{\text{sub2}}) = \text{var}(\bar{y}_{\text{sub1}}) + \text{var}(\bar{y}_{\text{sub2}}) - 2\text{cov}(\bar{y}_{\text{sub1}}, \bar{y}_{\text{sub2}}) \]

where: \( \bar{y}_{\text{sub1}}, \bar{y}_{\text{sub2}} \) are estimates of the mean of \( y \) for two subclasses.
Sampling Errors for Functions of Estimates: Complex Sample Designs

• Sampling errors for statistics that are functions of survey estimates require computations of variances and covariances of estimates. These should be stored by the software in a variance/covariance matrix.

• In complex sample designs, the covariance terms for subpopulation estimates are often positive due to the clustering (non-independence) in the design.

• This is true even when the subpopulations are distinct (male and female). The covariance will be zero if the subpopulations are defined by distinct design strata (e.g., Northeast vs. South Region).
Stata Example: Compute Subpopulation Means

```
. svy, vce(linearized): mean numadl, over(arthrtis)

1: arthrtis = 1
2: arthrtis = 2

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>numadl</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.9487063</td>
<td>.0356663</td>
<td>.8780096 1.019403</td>
</tr>
<tr>
<td>2</td>
<td>.3940189</td>
<td>.0279537</td>
<td>.3386098 .449428</td>
</tr>
</tbody>
</table>
```
**Stata Example (2):** Compute the standard error for the difference of the subpopulation means.

```
. lincom [numadl]1-[numadl]2
```

( 1)  [numadl]1 - [numadl]2 = 0

|            | Coef. | Std. Err. |    t |    P>|t|     | [95% Conf. Interval] |
|------------|-------|-----------|------|---------|---------------------|
| (1)        | **.5546874** | **.0408345** | **13.58** | **0.000** | **.4737463** - **.6356285** |

**Note:** Difference in two estimates from previous slide!
Stata Example: Display Variance/Covariance Matrix for the Two Subpopulation Means

. estat vce

Covariance matrix of coefficients of mean model

<table>
<thead>
<tr>
<th>numadl</th>
</tr>
</thead>
<tbody>
<tr>
<td>e(V)</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>numadl</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>
Difference of Means: Stata lincom post-estimation
Total Assets by Education Level of Head: 2006 HRS

Command immediately following svy: mean command from Slide 15:
lincom [h8atota]1 - [h8atota]4

<table>
<thead>
<tr>
<th>Education of Head</th>
<th>$\bar{y}<em>{0-11} - \bar{y}</em>{16+}$</th>
<th>$se(\bar{y}<em>{0-11} - \bar{y}</em>{16+})$</th>
<th>$CI_{.95}(\bar{y}<em>{0-11} - \bar{y}</em>{16+})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11 vs. 16+</td>
<td>-$928,818$</td>
<td>$108,250$</td>
<td>(-$1,145,669, -$711,967)</td>
</tr>
</tbody>
</table>

Note: Difference in two estimates.
Stata Example: Display Variance-Covariance Matrix for the Subpopulation Means

. estat vce

Covariance matrix of coefficients of mean model

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.032 × 10^8</td>
<td>0.714 × 10^8</td>
<td>-1.794 × 10^8</td>
<td>-3.438 × 10^8</td>
</tr>
<tr>
<td>2</td>
<td>2.918 × 10^8</td>
<td>-0.126 × 10^8</td>
<td>1.209 × 10^8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>7.290 × 10^8</td>
<td>0.679 × 10^8</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1.043 × 10^{10}</td>
</tr>
</tbody>
</table>
Difference of Means: Hand Computation
Total Assets by Education Level of Head: 2006 HRS

\[ se(\bar{y}_{0-11} - \bar{y}_{16+}) = \sqrt{\text{var}(\bar{y}_{0-11}) + \text{var}(\bar{y}_{16+}) - 2 \cdot \text{cov}(\bar{y}_{0-11}, \bar{y}_{16+})} \]

\[ = \sqrt{[6.032 + 104.3 - 2 \cdot (-3.438)] \times 10^8} \]

\[ = $108,250 \]
Example: SAS Difference in Means using PROC SURVEYMEANS and PROC SURVEYREG

• Use of PROC SURVEYMEANS to calculate means/se

• Subsequent use of PROC SURVEYREG with a LSMEANS / diff statement for test if means are significantly different
proc surveymeans data=one;
strata sestrat ; cluster seclustr ;
weight ncsrwtlg ;
var bmi ;
domain ed4cat;
run ;

The SURVEYMEANS Procedure

Data Summary

<table>
<thead>
<tr>
<th>Number of Strata</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Clusters</td>
<td>84</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>9282</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>5692</td>
</tr>
<tr>
<td>Number of Obs with Nonpositive Weights</td>
<td>3590</td>
</tr>
<tr>
<td>Sum of Weights</td>
<td>5692.00048</td>
</tr>
</tbody>
</table>

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>N</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>Body Mass Index</td>
<td>5099</td>
<td>27.160539</td>
<td>0.129700</td>
<td>26.8987938 27.4222842</td>
</tr>
</tbody>
</table>

Domain Analysis: 1=0-11 Years 2=12 Years 3=13-15 Years 4=16+ Years

<table>
<thead>
<tr>
<th>1-0.11 Years</th>
<th>2-12 Years</th>
<th>3=13-15 Years</th>
<th>4=16+ Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Label</td>
<td>N</td>
<td>Mean</td>
</tr>
<tr>
<td>bmi</td>
<td>Body Mass Index</td>
<td>767</td>
<td>27.371506</td>
</tr>
<tr>
<td>bmi</td>
<td>Body Mass Index</td>
<td>1521</td>
<td>27.705417</td>
</tr>
<tr>
<td>bmi</td>
<td>Body Mass Index</td>
<td>1528</td>
<td>27.161737</td>
</tr>
<tr>
<td>bmi</td>
<td>Body Mass Index</td>
<td>1283</td>
<td>26.265901</td>
</tr>
</tbody>
</table>
PROC SURVEYREG Output

```
proc surveyreg data=one;
strata sestrat ; cluster seclustr ;
weight ncsrwtlg ;
class ed4cat ;
model bmi=ed4cat / solution ;
lsmeans ed4cat / diff;
run ;
```
Example: Estimating a Difference of Means over Time
Total Assets: HRS 2004 vs. 2006

gen weight = jwgthh
replace weight = kwgthh if year == 2006
gen finr04 = 1 if (year==2004 & jfinr==1)
gen finr06 = 1 if (year==2006 & kfinr==1)
gen finr0406 = 1 if finr04==1 | finr06==1
svyset secu [pweight = weight], strata(stratum)
svy, subpop(finr0406): mean totassets, over(year)

<table>
<thead>
<tr>
<th>Contrast</th>
<th>( y_{2004} - y_{2006} )</th>
<th>se(( y_{2004} - y_{2006} ))</th>
<th>CI_{.95}(( y_{2004} - y_{2006} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 vs. 2006</td>
<td>-$115,526</td>
<td>$20,025</td>
<td>(-$155,642, -$75,411)</td>
</tr>
</tbody>
</table>

Contrast

\[
\text{ Contrast } \quad y_{2004} - y_{2006} \\
\]

\[
\text{ se}(y_{2004} - y_{2006}) \\
\]

\[
\text{ CI}_{.95}(y_{2004} - y_{2006}) \\
\]
SIMPLE BIVARIATE ANALYSIS OF COMPLEX SAMPLE DATA
NCS-R Example: 2 x 2 Table

- Example: Suppose that we are interested in studying the association between gender and a lifetime major depression episode (1=yes, 0=no).
### NCS-R: MDE x Gender

Weighted Estimates of Total Proportions

<table>
<thead>
<tr>
<th></th>
<th>MDE=NO</th>
<th>MDE=YES</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0)</td>
<td>(1)</td>
<td></td>
</tr>
<tr>
<td><strong>MEN</strong> (A)</td>
<td>$p_{A0}=0.406$</td>
<td>$p_{A1}=0.072$</td>
<td>$p_{A+}=0.478$</td>
</tr>
<tr>
<td><strong>WOMEN</strong> (B)</td>
<td>$p_{B0}=0.402$</td>
<td>$p_{B1}=0.120$</td>
<td>$p_{B+}=0.522$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$P_{+0}=0.808$</td>
<td>$P_{+1}=0.192$</td>
<td>$p_{++}=1.000$</td>
</tr>
</tbody>
</table>
Proportions of Interest

Total proportion: \( p_{A0} = \frac{\hat{N}_{A0}}{\hat{N}_{++}} \)

Row proportion: \( p_{0|A} = \frac{\hat{N}_{A0}}{\hat{N}_{A+}} \)

Column proportion: \( p_{A|0} = \frac{\hat{N}_{A0}}{\hat{N}_{+0}} \)
Estimation of Total and Row Proportions

svyset seclustr [pweight = ncsrwtsh], strata(sestrat)
svy: tab sex mde, se ci deff
svy: tab sex mde, row se ci deff

<table>
<thead>
<tr>
<th>Description</th>
<th>Parameter</th>
<th>Estimated Proportion</th>
<th>Linearized SE</th>
<th>95% CI</th>
<th>Design Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Proportions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, no MDE</td>
<td>$\pi_{A0}$</td>
<td>0.406</td>
<td>0.007</td>
<td>(0.393, 0.421)</td>
<td>1.87</td>
</tr>
<tr>
<td>Male, MDE</td>
<td>$\pi_{A1}$</td>
<td>0.072</td>
<td>0.003</td>
<td>(0.066, 0.080)</td>
<td>1.64</td>
</tr>
<tr>
<td>Female, no MDE</td>
<td>$\pi_{B0}$</td>
<td>0.402</td>
<td>0.005</td>
<td>(0.391, 0.413)</td>
<td>1.11</td>
</tr>
<tr>
<td>Female, MDE</td>
<td>$\pi_{B1}$</td>
<td>0.120</td>
<td>0.003</td>
<td>(0.114, 0.126)</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Row Proportions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No MDE</td>
<td>Male</td>
<td>$\pi_{0</td>
<td>A}$</td>
<td>0.849</td>
<td>0.008</td>
</tr>
<tr>
<td>MDE</td>
<td>Male</td>
<td>$\pi_{1</td>
<td>A}$</td>
<td>0.151</td>
<td>0.008</td>
</tr>
<tr>
<td>No MDE</td>
<td>Female</td>
<td>$\pi_{0</td>
<td>B}$</td>
<td>0.770</td>
<td>0.006</td>
</tr>
<tr>
<td>MDE</td>
<td>Female</td>
<td>$\pi_{1</td>
<td>B}$</td>
<td>0.230</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Estimation of Total and Row Proportions (SAS)

proc surveyfreq data = ncsr;
  weight ncsrwtsr;
  stratum sestrat;
  cluster seclustr;
  tables sexf*mde / row cl deff;
run;
Confidence Intervals for Proportions:
Proportion 18+ with Irregular Heartbeat
[NHANES (2005-06) – Recode not shown]

Stata Code:
svyset sdmvpsu [pweight = wtmec2yr], strata(sdmvstra)
svy, subpop(age18p): tab irregular, se ci col deff
svy, subpop(age18p): proportion irregular
svy, subpop(age18p): mean irregular

<table>
<thead>
<tr>
<th>Variable</th>
<th>$n$</th>
<th>Design $df$</th>
<th>Estimated Proportion</th>
<th>Linearized SE</th>
<th>95% CI</th>
<th>Design Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irregular</td>
<td>5,121</td>
<td>15</td>
<td>0.030</td>
<td>0.007</td>
<td>(0.018, 0.048)</td>
<td>10.35</td>
</tr>
</tbody>
</table>

Stata svy: tab, CI based on logit transform technique

| Irregular | 5,121 | 15          | 0.030                | 0.007         | (0.015, 0.044) | 10.35         |

Stata svy: prop, CI computed using the standard symmetric interval

| Irregular | 5,121 | 15          | 0.030                | 0.007         | (0.015, 0.044) | 10.35         |

Stata svy: mean, CI computed using the standard symmetric interval
• The logistic or polytomous regression model framework can be used analyze categorical data when one variable can be designated as the dependent variable and other variables as explanatory or predictor variables.

• Many times we are interested only in investigating associations between a set of categorical variables.

• A common approach is to compute a chi-square statistic under the null hypothesis, and fail to reject the null hypothesis if this chi-square statistic is within the range of values that would be expected.
Chi-square Tests: Pearson and Likelihood Ratio

• The chi-square test statistic is a measure of distance between the expected and observed counts in cells.

• Pearson: \( X^2_{Pearson} = n_{++} \cdot \sum_r \sum_c (p_{rc} - \hat{\pi}_{rc})^2 / \hat{\pi}_{rc} \)

• Likelihood Ratio: \( G^2 = 2 \cdot n_{++} \cdot \sum_r \sum_c p_{rc} \times \ln \left( \frac{p_{rc}}{\hat{\pi}_{rc}} \right) \)

• If the observed data are consistent with the null hypothesis, then we would expect this distance measure (the test statistic) to be small.
Adjusting for Design Effects

• How to incorporate the design features into this analysis? Since ignoring the design features (weighting) can introduce bias in the estimated proportions that goes into making the Chi-square statistics, the above analysis is not valid.

• Two approaches:
  – Fellegi (JASA, 1980) corrects the SRS Chi-square statistic.
Rao-Scott Method

• Use the weighted proportions in the construction of Chi-square statistics. Develop an approximate F-reference distribution for determining a p-value.

• This is analytically complicated and requires computation of generalized design effects. These are the eigenvalues of the “design effect” matrix for the proportions used to compute the Chi-square statistic.

• Most software packages enabling analysis of complex sample survey data have implemented this approach.
Design-adjusted $X^2$ Test Statistics

• First order correction (SAS system default):

$$X^2_{R-S} = \frac{X^2_{Pearson}}{GDEFF},$$

$$G^2_{R-S} = \frac{G^2}{GDEFF}$$

• $GDEFF$: the mean of the eigenvalues, or the “average design effect” (see ASDA, p. 166)
Design-adjusted $X^2$ Test Statistics

- Second order correction
- Stata system default, SAS V9.3 option:

\[
X^2_{R-S} = X^2_{Pearson} / GDEFF / (1 + a^2), \]
\[
G^2_{R-S} = G^2 / GDEFF / (1 + a^2)
\]

where:

\( a = \text{coefficient of variation of eigenvalues of } D \text{ matrix.} \)
Rao-Scott Chi-Square
(First order correction as in SAS default).

\[ R - S \ Chi - Square : \]

\[ Q_{rs} = Q_{Pearson} / GDEFF \]
\[ \sim \chi^2_{(R-1)(C-1)} \text{ under } H_0 \]

or:

\[ F = Q_{rs} / (R - 1)(C - 1) \]
\[ \sim F_{(R-1)(C-1),(R-1)(C-1)s} \text{ under } H_0 \]

where: \( s = \# \text{clusters} - \# \text{strata in complex design.} \)
Alcohol Dependence vs. Education Level for Young Adults Age 18-28. Source: NCS-R.

In Stata:
svyset seclustr [pweight = ncsrwtsh], strata(sestrat)
svy, subpop(if 18<=age<29): tab ed4cat ald, row se ci deff

In SAS V9.2+:
title "Analysis Example : Proportions by Subgroups: NCSR" ;
proc surveyfreq data=ncsr ;
   strata sestrat ;
   cluster seclustr ;
   weight ncsrwtsh;
   tables age29*ed4cat*ald / row deff chisq ;
* tables age29*ed4cat*ald / row deff chisq(secondorder) ; *(OPTIONAL);
   format ed4cat edf. ald aldf. ; run ;
Table 6.6: Design-adjusted Analysis of Alcohol Dependence vs. Education Level for Young Adults Age 18-28. Source: NCS-R.

<table>
<thead>
<tr>
<th>Education Level (Grades)</th>
<th>Alcohol Dependence (ALD) Row Percentages</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 = No</td>
<td>1 = Yes</td>
<td>Total</td>
</tr>
<tr>
<td>0-11</td>
<td></td>
<td>0.909 (0.029)</td>
<td>0.091 (0.029)</td>
<td>1.000</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>0.951 (0.014)</td>
<td>0.049 (0.014)</td>
<td>1.000</td>
</tr>
<tr>
<td>13-15</td>
<td></td>
<td>0.951 (0.010)</td>
<td>0.049 (0.010)</td>
<td>1.000</td>
</tr>
<tr>
<td>16+</td>
<td></td>
<td>0.931 (0.014)</td>
<td>0.069 (0.014)</td>
<td>1.000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0.940 (0.009)</td>
<td>0.060 (0.009)</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Tests of Independence (Note the difference!)

<table>
<thead>
<tr>
<th>Unadjusted $X^2$</th>
<th>$P \left( \chi^2_{(3)} &gt; X^2_{Pearson} \right)$</th>
<th>Rao-Scott $F$</th>
<th>$P \left( F_{2.75, 115.53} &gt; F_{R-S} \right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X^2_{Pearson} = 27.21$</td>
<td>$p &lt; 0.0001$</td>
<td>$F_{R-S,Pearson} = 1.64$</td>
<td>$p = 0.18$</td>
</tr>
</tbody>
</table>

Parameters of the Rao-Scott Design-Adjusted Test

| $n_{18-29} = 1275$ | Design $df = 42$ | $GDEFF = 6.62$ | $a = 0.56$ |
Applied Survey Data Analysis

Project Overview
Information about Authors
Links to Data Sets
Links to Additional Sites
Survey Data Analysis Publications
Professional Reviews
Frequently Asked Questions
Supplemental Code

Site Overview

This site contains information about the text "Applied Survey Data Analysis", (first and second editions) including author biographies, links to public release data sets and related sites, code and output for analysis examples replicated in current software packages, and information about new publications of interest to survey data analysts. Other features include a FAQ, log and links to other software and statistical sites. We plan to intermittently update this site with news about ongoing statistical and software advances in the field of analysis of survey data.

Special Notes from Authors

ASDA-Second Edition is Available as of June 28, 2017!

Project Overview

Applied Survey Data Analysis is the product born of many years of teaching applied survey data analysis classes and practical experience analyzing survey data. We have taught various versions of this course in the ISR/JRC Summer Institute Program, as part of University of Michigan/ESCAR, and within the Survey Methodology Program at University of Michigan and University of Maryland. Our goal has been to integrate teaching materials and practical analysis knowledge into a textbook geared to a level accessible for graduate students and working analysts who may have varying levels of statistical and analytic expertise. We intend to update the materials on this website as statistical and software improvements emerge with the goal of assisting analysts and researchers performing survey data analysis.

Information About Authors

Patricia A. Berglund is a Senior Research Associate in the Survey Methodology Program at the Institute for Social Research. She has extensive experience in the use of computing systems for data management and complex sample survey data analysis. She works on research projects in youth substance abuse, adult mental health, and survey methodology using data from Army STARSS, Monitoring the Future, the National Comorbidity Surveys, World Mental Health Surveys, Collaborative Psychiatric Epidemiology Surveys, and various other national and international surveys. In addition, she is involved in development, implementation, and teaching of analysis courses and computer training programs at the Survey Research Center-Institute for Social Research. She also lectures in the SAS® Institute-Business Knowledge Series. 

Email: pberg@umich.edu
ASDA Website: Example Code and Results

- Updated review of software packages for the analysis of complex sample survey data.

**Summary**


Designed for readers working in a wide array of disciplines who use survey data in their work, this book continues t methods of survey data analysis. An example-driven guide to the applied statistical analysis and interpretation of su recent versions of real-world survey data sets. Although the authors continue to use Stata for most examples in the software code for replicating the examples on the book's updated Web site.

**Links to Data Sets for First and Second Editions**

**National Comorbidity Survey-Replication (Collaborative Psychiatric Epidemiology Surveys)**

http://www.nci.cancer.gov/views (for online documentation tools and data download)

http://www.hcp.med.harvard.edu/nccs (for NCS-R specific information)

**National Health and Nutrition Examination Survey (National Center for Health Statistics)**

http://www.cdc.gov/nchs/

**Health and Retirement Survey (Institute for Social Research-University of Michigan)**

http://hrsponline.isr.umich.edu

**European Social Survey (ESS)**

http://www.europeansocialsurvey.org/

**United States Census Bureau**

http://www.census.gov/
THANK YOU!