Using SAS Survey Procs for BRFSS Descriptive Analyses

Instructor: Donna Brogan, Ph.D.

March 23, 2013 Saturday AM

2013 BRFSS Annual Conference

dbrogan@emory.edu

WORKSHOP OBJECTIVES

- Methods: telephone sampling & estimation of population parameters, within BRFSS context

- Use SAS survey procedures to:
    - Estimate **popn** total/prevalence/mean
        - Also for subpopulations and/or domains
    - Estimate prevalence ratio or odds ratio (2 x 2)
    - Compare domains on prevalence/mean
    - All with estimated s.e. & CI

PREREQUISITES

- Foundations of statistical inference
- Intermediate statistical methods
- Epid measures of risk & association
- SAS for data management
- SAS STAT for analyses of SRS data
  - E.g. Proc MEANS, FREQ, UNIVARIATE, GLM
- See references: slide 190

Review: Sample Survey Basics & Terminology

Why and How Conduct BRFSS Surveys?

Context for BRFSS Sample Survey

- **Define BRFSS target population**
  - **Adults** resident in LA in 2004 (3.3 million)
  - Noninstitutionalized, household popn (live in HU)
    - College dormitory? nursing home? military base?
  - Adult = **element** in popn (unit of analysis)
- **Population parameter(s) of interest**
  - LA: # or % of **adults** who are binge drinkers
  - LA: mean body mass index (BMI) for **adults**

Why Conduct BRFSS Sample Survey?

- Want to know **value** of popn parameter(s)
- Value **unknown** unless measure all elements
  - Too expensive to do census/enumeration

- Thus, **use sample survey methodology**
  - Select probability sample of adults from popn
  - Measure variables on sampled adults
  - Analyze sample data: **estimate** popn parameters

Select Probability Sample from BRFSS Target Popn. How??

- 1. Simple/stratified random sampling
- 2. Area probability sampling (APS)
- 3. Telephone sampling (RDD variations)
- 4. Address based sampling (ABS)

1. Simple/Stratified Random Sampling: Not Feasible

- Sampling frame: list of adults in target popn
  - Name & contact information each adult in popn
  - Impossible to construct sampling frame
- PSU(primary sampling unit)=adult (element)
- One stage of sampling to get to adult

2. Area Probability Sampling: Judged Too Expensive

- Multi-stage sampling to obtain sample adults

- PSU: 1 or more counties or county part
- SSU, TSU, etc.: CT, block, HU address
- Final sampling unit: adult (element)

- Used by NHANES & NHIS, but not BRFSS

3. Telephone Sampling: Used by BRFSS & Other Surveys

- **1st stage frame:** list of telephone numbers (PSUs) that link to target popn
  - Select sample of telephone numbers
- **2nd stage frame** for **sampled** phone number: list of adults associated with phone
  - Select 1 (or more) adults into sample

- Two stage sample to get to adult (element)

4. Address Based Sampling (ABS): Recent Method

- **1st stage frame**: list of HU addresses (PSUs) that link to target popn
  - Use USPS and 911 to construct frame
- **2nd stage frame**: list of adults (SSU) reside at sampled address
- Two stage sample
- ABS replace telephone sampling in U.S.?

BRFSS Methods after Obtain Telephone Sample of Adults

- Telephone interview of sampled adult
  - CDC core & modules, state specific questions
- Data entry and processing
- Weighting & survey design variables
- Other calculated variables, e.g. BMI
- Annual dataset for all states released to states & to public (on WEB)

Review of BRFSS RDD Telephone Sampling Methods

Phase 1: Mid 1980’s thru 2010

Phase 2: 2011 and Beyond

1st Stage BRFSS Sampling Frame Through 2010

- All possible **landline** phone numbers for state (PSU = phone number)
  - Computer generated by AC * prefix * xxxx
- Frame stratified by **phone density**
  - **High/low** density strata: high **oversampled**
  - Very low density numbers deleted from frame
- Frame maybe stratified by geography
  - State make inference to certain geog areas
  - AC & prefix used for geographic stratification

2nd Stage BRFSS Sampling Frame Through 2010

- **2nd stage frame:** list of adults reside at HU of given sampled **landline** phone number
  - 1 adult selected with equal prob from sampling frame of all adults in household
- SSU = adult (element)

BRFSS Sample of Adults Through 2010

- **Unequal** probability sample of adults for two reasons
  - Some PSU’s (phone numbers) oversampled based on phone density and/or geography
    - And, hence, some undersampled
  - Adults in HUs with only one adult have larger prob of being selected into sample, compared to adults who live in HU’s with 2 or more adults

Problems with BRFSS RDD Landline Sampling Methods

- 1. Survey response rate dropping over time
  - Sampled landline number: ring no answer
  - If answer, don’t reveal # adults in HU
  - If adult selected, not agree to be interviewed
  - Some demographic groups particularly low RR
- 2. Percent of adults “cell only” steadily increasing (especially young, rent, minority)
  - Landline frame: severe **undercoverage**

Why Worry About These Two BRFSS Problems?

- **Each** of the 2 problems **may** result in biased estimators of popn parameters
- Don't **know** if estimators biased, since don't know true value of popn parameter
  - But research points in direction of bias
- Low face validity or credibility of survey
  - 1. Survey response rate is 25%
  - 2. Noncoverage of "Cell only": 54% of adults 25-29, 50% of renters, 30% of adults

BRFSS Solutions to These 2 Problems: 2011 & Beyond

- New weighting method (**raking**) to adjust for unit nonresponse & do post-stratification
- New telephone sampling **frame**
  - Cell phone numbers on 1st stage sampling frame
    - An additional stratum to the landline strata
  - Interview cell sampled adult **only if** that adult is "cell only". If have landline, drop from sample.
  - Called **dual frame RDD** telephone sampling
- Note: nontelephone elements **not covered**

BRFSS Survey Design Variables Through 2010

- **_FinalWt**
  - Sampling weight variable to estimate all population parameters for adults
- **_Ststr**
  - 1st stage stratification variable for landline sampling frame (state, density, geographic)
- **_Psu** (in later years = Seqno)
  - Earlier years: cluster of phone numbers
  - Later years: phone number selected (marker)

More BRFSS Survey Design Variables Thru 2010

- Module for Sample Child
  - **_ChildWt,** _Ststr, _Psu
  - Target Popn: children reside in state in HU
  - Unit of analysis = child
- Interview items about housing unit
  - **_HouseWt**, _Ststr, _Psu
  - Target Popn: HUs in state
  - Unit of analysis = HU

BRFSS Sampling Weight Variables through 2010

- **Sum** of **_FinalWt** over r responding adults = # **adults** (noninst, HH) in state popn

- **Sum** of **_HouseWt** over r responding adults = # HUs in state (**occupied??**)

- **Sum** of **_ChildWt** over responding adults with child data = # children (noninst, HH) in state popn

Survey Design Variables: BRFSS Dual Frame 2011 +

- **_LLCPWT** adult final weight
  - Sampling weight variable to estimate all population parameters for adults
- **_Ststr**
  - 1st stage stratification variable for dual frame (state, density, geographic, landline/cell)
- **_Psu** (= Seqno)
  - Marker for phone number selected

More Survey Design Vars: BRFSS Dual Frame 2011 +

- **_CLLCPWT** child final weight
  - Sampling weight variable to estimate all population parameters for children
- Use above with **_Ststr** and **_Psu**

- Did not find HU sampling weight variable in 2011 dual frame BRFSS dataset
  - Would be complicated to calculate

BRFSS Sampling Weight Variables: 2011 onward

- **Sum** of **_CCLPWT** over r responding adults = # **adults** (noninst, HH) in state popn

- **Sum** of **_CLLCPWT** over responding adults with child data = # children (noninst, HH) in state popn

Population Parameters in BRFSS Surveys

Their Definition and Estimation

Nominal Variables & Population Parameters

- **Nominal** variables (categorical unordered)
- Binge drinking (yes=1, no=0)
  - **Population total** (# binge bingers)
  - **Population proportion** or % (% binge drinkers)

- Type health plan (employer, Medicaid, etc.)
  - **Population total:** (# have employer plan)
  - **Population proportion** or %: % employer plan

Ordinal Variables & Population Parameters

- **Ordinal** variables (categorical ordered)
  - Health status (excellent, good, VG, fair, poor)
  - BMI status (underweight, normal, overweight, obese, morbidly obese)
- **Population parameters**:
  - Usually as for nominal

Count Variables and Population Parameters

- **Count** variable: # ER visits past 6 mos
    - Coded 0, 1, 2, 3, etc.
- **Population total:** total number ER visits made by popn in past 6 months
- **Population mean:** mean ER visits made by popn in past 6 months (**but** many 0)
- **Population proportion** or %: % make at least 1 ER visit past 6 months

Continuous Variables and Population Parameters

- **Continuous** variables
  - Height, weight, BMI
  - # cigarettes smoked per day, among smokers
- **Population mean weight, mean BMI**
- **Subpopn mean:** mean cigs smoked per day, among smokers
- **Subpopn total:** total cigs smoked per day, among smokers

Continuous/Count Vars as Categorical or Ordinal

- BMI: low, normal, overweight, obese
- BMI: obese, not obese

- Number ER visits past 6 months
  - None, 1 or more
  - None, 1-3, 4 or more

Population Parameters: Mathematical Definition

- **Finite** target population has N elements
  - N may be large (3.3 million), but not infinite
- Let i denote element i , i = 1, 2, .., N
- Let y_i be value of variable y for element i
  - Continuous or count variable y, BMI or ER visits
  - Dichotomous variable y, e.g. disease yes/no
  - Categorical variable y, e.g. health plan type

POPULATION TOTAL Y
Continuous Variable y=BMI

- $$Y = \sum_{i=1}^{i=N} y_i$$
- Y = sum of BMI value for N popn elements

POPULATION MEAN Continuous Variable y=BMI

$$\bar{Y} = \frac{\sum_{i=1}^{i=N} y_i}{N} = \frac{Y}{N}$$

- Mean body mass index for N popn elements

Estimator of Mean BMI, Based on BRFSS Sample

$$\hat{\bar{Y}} = \frac{\sum_{k=1}^{k=r} w_k y_k}{\sum_{k=1}^{k=r} w_k} = \frac{\hat{Y}}{\hat{N}}$$

- r = # adult respondents in BRFSS dataset
- w_k = value of sampling weight variable for adult k in sample (or child k, or HU k)

POPULATION TOTAL Y
Dichotomous Var y (0,1)

- Assume y variable coded as:
  - 1=have attribute, 0 = not have attribute
- $$Y = \sum_{i=1}^{i=N} y_i$$
- Y = number of popn elements with attribute

Estimator of Popn Total Y, Dichotomous Var y (0,1)

- Assume y variable coded as:
  - 1=have attribute, 0 = not have attribute
- $$\hat{Y} = \sum_{k=1}^{k=r} w_k y_k$$

POPN MEAN (PROP) Dichotomous Var y (0,1)

$$\bar{Y} = \frac{\sum_{i=1}^{i=N} y_i}{N} = \frac{Y}{N} = P$$

- Proportion of popn elements with attribute

Estimator of Popn Proportion Dichotomous Var y (0,1)

- $$\hat{\bar{Y}} = \frac{\sum_{k=1}^{k=r} w_k y_k}{\sum_{k=1}^{k=r} w_k} = \frac{\hat{Y}}{\hat{N}} = \hat{P}$$

Terminology: Subpopulations & Domains

- **Subpopulation** (**some** elements of popn)
  - **Diabetics** only, e.g. number & % on insulin

- **Domains**—mutually exclusive/exhaustive subpopns formed by some variable
  - SEX: males & females (e.g. smoking prevalence)
  - AGEG: 3 age groups (e.g. diabetes prevalence)

Define Parameters for Subpopulations & Domains

- Form indicator variable which says if element i in popn belongs to subpopn d or domain d

- $\delta_{\mathbf{di}}$ = 1 if element i in popn belongs to subpopn or domain d
- = 0 if element i in popn does **not** belong to subpopn or domain d

Subpopn/Domain d MEAN Continuous Var y = BMI

- $$\bar{Y}_d = [\sum_{i=1}^{i=N} y_i \delta_{di}] \ / \ [\sum_{i=1}^{i=N} \delta_{di}] = \frac{Y_d}{N_d}$$

- N_d is number of popn elements in d
- Mean BMI for popn elements in d

Estimator of Mean BMI for Subpopn/Domain d

$$\hat{\bar{Y}}_d = [\sum_{k=1}^{k=r} w_k y_k \delta_{dk}] \;/\; [\sum_{k=1}^{k=r} w_k \delta_{dk}]$$

$$= \frac{\hat{Y}_d}{\hat{N}_d}$$

Subpopn/Domain d TOTAL Dichotomous Var y (0,1)

$$Y_d = \sum_{i=1}^{i=N} y_i \delta_{di}$$

- Number elements in subpopn/domain d with attribute defined by y variable (i.e. y=1)

Estimator of Subpopn or Domain d TOTAL y (0,1)

$$\hat{Y}_d = \sum_{k=1}^{k=r} w_k y_k \delta_{dk}$$

Subpopn/Domain d MEAN or Prop. Dichot Var y (0,1)

- $$\bar{Y}_d = [\sum_{i=1}^{i=N} y_i \delta_{di}] / [\sum_{i=1}^{i=N} \delta_{di}]$$
  $$= Y_d / N_d = P_d$$
- N_d is number of elements in domain d
- P_d is proportion elements in d with attribute

Estimator of Subpopn or Domain d Mean/Proportion

$$\hat{\bar{Y}}_d = [\sum_{k=1}^{k=r} w_k y_k \delta_{dk}] / [\sum_{k=1}^{k=r} w_k \delta_{dk}]$$

$$= \hat{Y}_d / \hat{N}_d = \hat{P}_d$$

- Note: y variable dichotomous (0, 1)

Relevance of Definitions for Parameters

- Recall: parameters for entire popn, for subpopn, for domains
- Helps analyst:
  - Decide what to estimate
  - Understand estimation formulas for parameters
  - Write program for sample survey software
  - Interpret computer output from survey software

VARIANCE ESTIMATION for BRFSS Surveys

Estimated Variance and Standard Error for Estimators of Popn/ Subpopn/Domain Parameters

Why need estimated S.E. for an estimator?

- Quantify sampling error (variability)
- Confidence interval on popn parameter
- Coefficient of variation for estimator
- Test hypotheses about popn parameters

- **Recall**: square root of estimated variance is estimated S.E. (standard error)

2 Factors Make Variance Estimation Nonstandard

- 1. Sampling plan is **not** SRS
- 2. Many estimators **not** linear in y or x variables, but are ratios
  - Previous slides with estimator formulas

- Often no "closed form" algebraic expression
- Thus, "approximate" estimated variance

Factor #1: NOT SRS
3 Attributes Complex Design

- A. Elements selected unequal probability
  - Easy to address
  - Do weighted analysis (see estimator formulas)
- B. Stratification in sampling plan
  - Easy to address BRFSS 1st stage stratification
  - Variance estimated within each stratum
  - Within strata estimated variances added up over strata to obtain desired estimated variance

Factor #1: NOT SRS (cont.)
3 Attributes Complex Design

- C. Elements in sample may be clustered
  - Early landline RDD sampling (Mitofsky-Waksberg) resulted in clustered adults
  - Since early 1990's list assisted landline RDD sampling (DSS, disproportionate stratified sampling) has no clustering of HUs or adults or children in BRFSS sample
  - For dual frame in 2011 +, no clustering of adults or of children

Factor # 2—Ratio Estimators
2 Approximation Methods

- **Taylor Series Linearization (TSL)**
  - In all survey software packages except WESVAR
- **Replication Techniques**
  - BRR = balanced repeated replication
  - JK = jackknife
  - Available in SUDAAN & in SAS & STATA survey procedures & in WESVAR
- BRFSS datasets are set up for using **TSL**

Taylor Series Linearization Nonlinear Estimators (e.g. Ratio)

- Expand formula for estimator as infinite series
  - Infinite series is **linear** in sample statistics
- Truncate infinite series to first few terms
- Estimate variance of truncated infinite series

Adults: Use Sample Data to Estimate Popn Total Y

- Recall--definition of popn total Y
- $$Y = \sum_{i=1}^{i=N} y_i$$ y continuous, count or discrete (0, 1)
- $$\hat{Y} = \sum_{k=1}^{k=r} w_k y_k$$ = estimator of Y

  - w_k = value of weight variable _FinalWt for respondent adult k in dataset

Rewrite equation previous slide: Estimate Popn Total Y

- r_h is # of respondent elements (adults) from stratum h (based on _Ststr)

$$\hat{Y} = \sum_{h=1}^{h=L} \sum_{k=1}^{k=r_h} w_{hk} y_{hk} = \sum_{h=1}^{h=L} \sum_{k=1}^{k=r_h} z_{hk}$$

$$z_{hk} = w_{hk} y_{hk}$$

- Statistically independent sampling across the first stage strata

Variance Estimation Within Each Stratum

- Calculate mean of the z_{hk} within stratum h

$$\bar{z}_h = \frac{1}{r_h} \sum_{k=1}^{k=r_h} z_{hk}$$

$$s_{zh}^2 = \frac{1}{(r_h - 1)} \sum_{k=1}^{k=r_h} (z_{hk} - \bar{z}_h)^2$$

Variance Estimation for $\hat{Y}$

- $$EstVar(\hat{Y}) = \sum_{h=1}^{L} r_h s_{zh}^2$$
- Estimator is on slide 57
- NOTE: **Weighted sum over strata of w/n stratum estimated variances**

Estimated Variance for Other Estimators

- Ratio estimators: need to use TSL
  - Formulas more complicated
  - But, still sum of within stratum variances
- Subpopulation or Domain Estimators
  - Easy for estimated subpopn/domain totals
  - More complicated for ratio estimators

- No more detail here—see math-stat books

BRFSS ANALYSIS

General Analytical Strategy

Prepare Dataset for Analysis

- Obtain national BRFSS dataset: WEB, other
- Subset to “state” or “states” of interest
- Subset to variables of interest
- Obtain **national** estimates from 50 + DC
  - **If** all states included questions of interest
- If analyze given module X (25 states used)
  - Inference **not national**, but **union** of 25 states

Check Coding of Variables: Recoding May Be Needed

- **_RFBING2** (binge drinking last 30 days)
  - 1=no, 2=yes, 9=dk, refuse, missing
  - Likely change 9 to . (missing) for analysis
- **_BMI4** (body mass index)
  - 0001-9988 BMI, 2672 implies 26.72
  - 9999 dk, refuse, missing
  - Change 9999 to dot, divide other values by 100

- Each adult asked above questions

Unweighted/Weighted Analyses with SAS Procs

- **Unweighted** SAS (e.g. FREQ, MEANS)
  - Results describe elements **in sample**
  - E.g., 66% **of adult respondents** are female
- **Weighted** SAS (e.g. FREQ or MEANS with **Weight** statement)
  - Point estimate is estimator of a popn paramter
  - Point estimate makes **inference** to population
    - E.g. **estimated 53%** of adults in **popn** are female
  - Will not give correct estimated s.e., CI, etc.

SAS PROCS FOR SAMPLE SURVEY DATA

General Features for Using These PROCS with BRFSS

Descriptive Survey Procs Available in SAS 9.2/9.3

- **SURVEYFREQ** (categorical data)
  - Similar to PROC FREQ, but for survey data
- **SURVEYMEANS** (continuous/categorical)
  - Similar to PROC MEANS, but for survey data
- **SURVEYREG**
  - Similar to PROC GLM, but for survey data
  - Estimate age-standardized prevalence or mean
  - Compare domains to each other
  - Macro for SurveyMeans does some of above

SAS SURVEY PROCS Describe Sample Design

- Need 3 statements below, **in general**
  - **STRATA** name(s) of 1st stage stratification variable(s)
  - **CLUSTER** name(s) of PSU variable(s)
  - **WEIGHT** name of sampling weight variable (only one variable)

BRFSS Thru 2010: Sample Design--SAS Survey Procs

- Proc Survey..... Varmethod = taylor..
- STRATA _Ststr ;
- CLUSTER _Psu ;
- WEIGHT _FinalWt ; (adult)
  - Or _ChildWt or _HouseWt
- One or more states, any ONE year
- NOT correct for >= 2 years combined

BRFSS 2011 + : Sample Design--SAS Survey Procs

- Proc Survey….. Varmethod = taylor..
- STRATA _Ststr ;
- CLUSTER _Psu ;
- WEIGHT _LLCPWT ; (adult)
  - Or _CLLCPWT for child
- One or more states, any ONE year
- NOT correct for >= 2 years combined

BRFSS Dataset for Workshop

LA 2004

la04v7.sas7bdat, n = 9064 Rs

On Workshop CD

Get BRFSS Dataset into SAS Work Directory

- SAS program **ProcFormat2013.sas** on C drive in folder Brogan/BRFSData
- Open this SAS program
- Run **“proc format”** part of program
- Choose appropriate Libname
- Read dataset into SAS Work Directory
- Run proc contents

Lecture Example 1 Nonsurvey PROCS in SAS

- Look at survey design variables
- Look at coding of some variables
- Proc Freq weighted: estimate popn parameters but no estimated s.e.
  - **Estimated** Number Binge Drinkers = 462,272 **Estimated** prev of binge drinking = 14.22%
  - **In** population of adults in LA in 2004, **IF** assume MCAR on binge drinking item nonresponse

Proc SurveyFreq

Analytical Capabilities

SurveyFreq Capabilities

- **Categorical** variables only (nominal/ordinal)
  - Tables of dimension 1, 2, 3, etc.
- Estimate popn percentage (prevalence), total
  - With estimated standard error & CI
  - With CV (coefficient of variation)
- Estimate percentages & totals for **domains**
  - With estimated SE & CI & CV (coeff of variation)

SurveyFreq Subpopulation Analyses

- No SubPopn statement in SAS survey procs
  - It **should** be available for the survey procs!

- Use indirect methods for subpopn analyses
  - These methods work in all SAS Survey Procs

SurveyFreq More Capabilities

- Estimate **association** for 2 x 2 table
  - Row = exposure, column = outcome
  - Estimate **prevalence ratio**, with CI
  - Estimate **odds ratio**, with CI
  - Estimate **prevalence difference**, with CI
  - Stratified analyses available: by a 3rd variable
- Chi-square tests for independence of 2 vars
  - Choose from 8 chi-square tests available

CV = Coefficient of Variation What is it?

- Characteristic of an estimator
- Quantifies sampling variability of estimator
  - **relative to** value of popn parameter
- Estimated CV(any estimator) =

  EstSE (estimator)/(Value of estimator)

$$\boxed{EstCV(\hat{P}) = EstS.E.(\hat{P}) / \hat{P}}$$

How use CV?

- Decide if estimator variability too high

- NCHS guideline
  - Do not report value of any estimator if its estimated CV exceeds 0.30 (i.e. 30%)

- Some follow NCHS guideline, some not

Lecture Example 2
SurveyFreq

Population (adult) analysis:

Prevalence of Binge Drinking

Number of Binge Drinkers

LecEx 2A SurveyFreq Default output

- proc surveyfreq data = La04 **varmethod = taylor** ;
- **strata _ststr / List** ;
- **cluster _psu** ;
- **weight _finalwt** ;

- **tables _rfbing2** ; /* default printout */

LecEx 2B, SurveyFreq Add output options

- proc surveyfreq data = LA04 ;
- strata _ststr ; /* drop List option */
- cluster _psu **;**
- weight _finalwt ;

- **tables _rfbing2 / cl clwt cv cvwt ;**

DDF for Sample Survey

Denominator degrees of freedom

- DDF = number of PSUs in sample less number of 1st stage strata in sample design

- DDF for BRFSS LA 2004 dataset:
  - Each R in dataset is a PSU, hence 9064 PSUs
  - 18 1st stage (PSU) strata: 2 density by 9 regions
  - Thus, BRFSS DDF = 9064 – 18 = 9046

How Does SAS Use DDF in Its Calculations?

- Construct confidence intervals
  - Obtains critical value for CI, e.g. 95%, by going to Student t-distribution with degrees of freedom = ddf
- Conduct statistical tests of significance to test null hypotheses
- DDF for BRFSS survey typically thousands
- DDF for other surveys, e.g. APS, typically much smaller

What is Item Nonresponse?

- Obsn in dataset supposed to have value for a given variable, but does not
- Alcohol questions asked of all adults, so all obsns should have value for _RfBing2
- However, 179 obsns coded 9 (changed to dot) for _RfBing2
- They cannot be in analysis in LecEx02

Item Nonresponse: Default Method SAS survey procs

- SAS survey procs assume **MCAR**
  - **Missing completely at random**
- MCAR = those not respond to item like those who do respond to item, on average
- If assume MCAR, point estimate of mean, prevalence, etc. makes inference to popn
- SAS deletes from analysis any obsns with missing data for analysis variable(s)

Item Nonresponse: Other Method SAS survey procs

- Add **NOMCAR** to **PROC** statement
  - Does not make MCAR assumption
- **Subpopn** defined as adults in popn who would answer item(s), if asked
- SAS does correct subpopn analysis
- Point estimate makes inference to **subpopn** rather than to entire popn
- This method is default in SUDAAN

LecEx 2C, SAS With NOMCAR Option

- proc surveyfreq data = La04 **NoMcar** ;
- strata _ststr ;
- Cluster _psu
- weight _finalwt ;
- **tables _rfbing2 / cl clwt cv cvwt ;**
- Some estimated standard errors & CIs differ slightly from LecEx 2B (SurveyFreq)

What Should I Use in SAS? Default MCAR or NOMCAR

- Only s.e. impacted, not point estimate
- Most people use MCAR without realizing it
- NOMCAR requires stated results as:
  - "in subpopn of those who would respond to.."
- I generally use NOMCAR because…
  - Is SUDAAN default
  - Estimated s.e.'s often slightly larger
  - Infer to entire popn if further assume MCAR

How SurveyFreq Estimates Popn Total

- _RFBING2 coded as 1=no, 2=yes
- How estimate total number binge drinkers?
- SAS forms indicator variable y for binge drinker
  - y =1 if _RFBING2 = 2 (i.e. drinker)
  - y =0 if _RFBING2 = 1 (i.e. ≠ 2 and ≠ ., not drinker)

$$\hat{Y} = \sum_{k=1}^{k=8885} w_k y_k = \textbf{estimated \# binge drinkers}$$

How SurveyFreq Computes CI on Popn Total

- Symmetrical CI around point estimate

$$CI = \hat{Y} \pm [EstS.E.(\hat{Y})] * t_{ddf,1-\alpha/2}$$

- t = critical value from Student t distbn
  - Cuts off area (1-α/2) to left of critical value
  - Degrees of freedom = ddf = denominator degrees of freedom for the survey

How SurveyFreq Estimates Popn Percent

- First estimate **proportion** who binge drink

$$\hat{P} = \frac{\hat{Y}}{\hat{N}} = [\sum_{k=1}^{k=8885} w_k y_k / \sum_{k=1}^{k=8885} w_k]$$

= estimated proportion who binge drink

- Multiply estimated proportion by 100

How SurveyFreq Computes CI on Popn Percentage

- **By default:** Wald confidence interval, symmetrical around point estimate

$$CI = EstPopn\% \pm (EstS.E.) * t_{ddf,1-\alpha/2}$$

- t = critical value from Student t distbn
- **Other options in SAS 9.3**
  - CL (type=logit), SUDAAN default CI method for percentages

Ex 2 (SAS): Results with 2 Item Nonresponse Methods

Estimates	Default MCAR	Use NOMCAR
Binge Prev %	14.22	14.22
SE binge prev%	0.5301…..	0.5301…..
CI binge prev%	(13.18, 15.26)	(13.18, 15.26)
# binge drinkers	462,272	462,272
SE # drinkers	18029	18050
CI # drinkers	426930,497613	426890,497654

Estimate # Drinkers when Item Nonresponse

- Estimated # drinkers: **462,272**
  - Slight underestimate since 179 not respond

- Revised estimate for total, assume MCAR
  - (.142167) * (3322812) = **472,394**
- Approx estimated S.E. for revised total
  - (3322812) * EstSE (est prev .142167)

Lecture Example 3 SurveyFreq

Domain Analysis

Domains: males and females

Dependent Var: Binge Drinking

Lecture Example 3 SurveyFreq

- Estimate binge drinking prevalence, by sex
- Define 2 domains of interest:
    - Males and females
    - Use variable SEX to define the two domains
    - NOTE: no missing data on variable SEX

- Each domain, estimate #/% who binge drink

LecEx 3A—SAS, 2 way table, default output

- proc surveyfreq data = La04 NoMcar ;
- strata _ststr ;
- cluster _psu ;
- weight _finalwt ;
- **tables sex * _rfbing2 / Row ;**
  - /* Sex is row variable, & it defines domains. Binge is column variable. Ask for **row** percents on **tables** statement. */

LecEx 3B—SAS Optional output & suppress output

- proc surveyfreq data = La04 NoMcar ;
- strata _ststr ; cluster _psu ;
- weight _finalwt ;

- **tables sex * _rfbing2 / Row CL clwt cv cvwt nocellpercent ;**

How SURVEYFREQ Estimates Popn Total for Males

- How estimate total number male binge drinkers?
- SAS forms indicator variable y for binge drinking
  - y =1 if _RFBING2 = 2 (binge drinker)
  - y =0 if _RFBING2 = 1 (not binge drinker)
- SAS forms indicator variable for male

$$\delta_{mk} = 1 \textbf{ if sample element k is male}$$

$$\delta_{mk} = 0 \textbf{ if sample element k is not male}$$

How SURVEYFREQ Estimates Popn Total for Males

- Estimated number of male binge drinkers is:

$$\hat{Y}_m = \sum_{k=1}^{k=8885} w_k \delta_{mk} y_k$$

How SURVEYFREQ Estimates Popn Percent for Males

- Among males, estimated proportion who are binge drinkers is:

$$\hat{P}_m = \sum_{k=1}^{k=8885} w_k \delta_{mk} y_k \,/ \sum_{k=1}^{k=8885} w_k \delta_{mk}$$

- Multiply estimated proportion by 100

How Compare Domains? SurveyFreq

Example:

Compare Males to Females on Binge Drinking

Compare 2 Domains on Binge Drinking

- **Testing hypothesis approach**
  - Several chi-square tests for survey data
  - Null: 2 variables (sex & binge) independent
- **Estimation approach for 2 x 2 table**
  - Strength of association between 2 variables
  - Prevalence ratio (PR) & odds ratio (OR)
  - Prevalence difference (PD)

SurveyFreq expects 2 x 2 table set up as follows for OR

- **Row** Variable is **Exposure**
  - Lower code(row 1)=Exposed, Not Exposed(row2)
- **Column** Variable is **Disease**
  - Lower code(col 1)= Disease, No Disease (col 2)
- If your variables **not** coded this way,
  - Recode variables
  - Reinterpret output to what you want
  - Perhaps can use ORDER = …. option on PROC for SurveyFreq

2 x 2 Table expected by SurveyFreq

	Disease Yes = 1	Disease No = 2	COLUMN TOTAL
Expose Yes = 1	$\hat{N}_{11} = A$	$\hat{N}_{12} = B$	$\hat{N}_{1+} = A + B$
Expose No= 2	$\hat{N}_{21} = C$	$\hat{N}_{22} = D$	$\hat{N}_{2+} = C + D$
ROW TOTAL	$\hat{N}_{+1} =$ $A + C$	$\hat{N}_{+2} =$ $B + D$	$\hat{N}_{++} = A + B$ $+ C + D$

Odds Ratio Calculation by SurveyFreq

- For row 1 (exposed) estimates ODDS of being in column 1 (outcome of interest)
- For row 2 (nonexposed) estimates ODDS of being in column 1
- Takes ratio (exposed to nonexposed) of the 2 estimated ODDS
- Familiar formula, BUT table has estimated population totals, NOT sample size

Odds Ratio Calculation in SurveyFreq

OR

$$EstOR = \frac{\hat{N}_{11} \big/ \hat{N}_{12}}{\hat{N}_{21} \big/ \hat{N}_{22}} = \frac{\hat{N}_{11}\hat{N}_{22}}{\hat{N}_{12}\hat{N}_{21}} = \frac{AD}{BC}$$

Odds Ratio Calculation if Variables Coded Differently

- **Both** variables reverse coded from what software expects: get OR you want
- **One** variable reverse coded: get **inverse** of OR you want
  - Take reciprocal of estimated odds ratio and reciprocal of lower/upper limits of confidence interval in order to get the OR that you want

Prev Ratio Calculation by SurveyFreq

- For column (disease) variable, you define if column 1 or 2 is outcome of interest
- For each row, software estimates prevalence of being in specified column
- SurveyFreq takes ratio of two estimated prevalences, with row1 in numerator & row 2 in denominator (no choice)

“Prevalence Ratio” col 1 SurveyFreq

PR1

$$EstPR1 = \frac{\hat{N}_{11} \Big/ \hat{N}_{1+}}{\hat{N}_{21} \Big/ \hat{N}_{2+}} = \frac{A/(A+B)}{C/(C+D)}$$

“Prevalence Ratio” col 2 SurveyFreq

PR2

$$EstPR2 = \frac{\hat{N}_{12} \big/ \hat{N}_{1+}}{\hat{N}_{22} \big/ \hat{N}_{2+}} = \frac{B/(A+B)}{D/(C+D)}$$

Prevalence Difference Calculation by SurveyFreq

- For column (disease) variable, you define if column 1 or 2 is outcome of interest
- For each row, software estimates prevalence of being in specified column
- Software subtracts row2 prevalence from row1 prevalence (no choice)

PrevDiff Calculation (col 1) by SurveyFreq

$$Row1prev = \hat{N}_{11} / \hat{N}_{1+} = A/(A+B)$$

$$Row2prev = \hat{N}_{21} / \hat{N}_{2+} = C/(C+D)$$

$$Totalprev = \hat{N}_{+1} / \hat{N}_{++} = \frac{(A+C)}{(A+B+C+D)}$$

$$prevdiff = \hat{N}_{11} / \hat{N}_{1+} - \hat{N}_{21} / \hat{N}_{2+}$$

PrevDiff Calculation (col 2) by SurveyFreq

$$Row1prev = \hat{N}_{12} / \hat{N}_{1+} = B/(A+B)$$

$$Row2prev = \hat{N}_{22} / \hat{N}_{2+} = D/(C+D)$$

$$Totalprev = \hat{N}_{+2} / \hat{N}_{++} = \frac{(B+D)}{(A+B+C+D)}$$

$$prevdiff = \hat{N}_{12} / \hat{N}_{1+} - \hat{N}_{22} / \hat{N}_{2+}$$

SurveyFreq Syntax for Odds Ratio, Prev Ratio, PrevDiff

- Request options on **Tables** statement
- **Reminder**: **only** for **2 x 2** table
- **OR** odds ratio, column 1 & column 2 prevalence ratio (“relative risk”)
- **RISK** prevalence (risk) for row 1, row 2, & union, prev difference (row 1 – row 2), for **each** of the 2 columns
- **RISK1** or **RISK2** RISK (above), but only for chosen column

Lecture Example 7 SurveyFreq

Odds ratio

Prevalence Ratio

Prevalence Difference

Sex and Binge Drinking

LecEx 7A SurveyFreq OR & RISK _RFbing2

- proc surveyfreq data = La04 NoMcar .. ;
- strata _ststr ; cluster _psu ;
- weight _finalwt ;
- **tables sex * _Rfbing2 / row**
  or risk nocellpercent ;
- Note: _rfbing2 **not** coded as SAS expects, i.e. column 2 is outcome of interest

LecEx 7B SurveyFreq OR & Risk1 Binger

- proc SurveyFreq data = La04 NoMcar….. ;
- strata _ststr ; cluster _psu ;
- weight _finalwt ;
- **tables sex * binger / row**
  or risk1 nocellpercent ;
- Note: binger is coded as SAS expects, i.e. column 1 is outcome of interest, use Risk1

LecEx 7C SurveyFreq OR & Risk1 3 variables

- proc SurveyFreq data = La04 NoMcar ... ;
- strata _ststr ; cluster _psu ;
- weight _finalwt ;
- **tables _age3r * sex * binger / row or risk1 nocellpercent ;**
- Note: “stratified” (by age) analysis of 2 x 2 tables (sex * binger)

Prev Ratio, Odds Ratio, Prev Diff: Use which one?

- Each assesses relationship between 2 variables
- DB personal preference: prev ratio over odds ratio
  - Estimate prevalence ratio directly, survey design
  - Don't need to use OR as "pretend" risk ratio, as is done in case-control studies (no other choice)
- Rare outcome (disease): OR ≅ PR
- Common outcome: OR maybe lot larger than PR
  - Estimated OR = 3.96 and PR =3.29 for binge (M to F)
- May want OR if planning logistic regression
- Lots of discussion on this topic in epid literature

Subpopulation Analyses in SAS Survey Procedures

No Subpopulation Statement available yet in SAS Survey Procedures

Example A: Analysis of a Subpopulation

- Subpopulation = diagnosed diabetics
  - **Diabetes**: 1=yes, 2=no, . = no answer
- Variable of interest **Insulin**:
  - For diabetics: 1=yes, 2=no, .= no answer
  - **All others: insulin value is blank, . or .S**
  - DB coding preference: . versus .S
- Subpopn parameters to estimate:
  Among **diabetic adults**, % & # take insulin

Example B: Analysis of a Subpopulation

- Subpopulation = diagnosed diabetics
  - **Diabetes**: 1=yes, 2=no, . = no answer
- Variable of interest **BMI**:
  - For diabetics: BMI = some value, or .(dot)
  - **All others** have value of BMI also, or .(dot)
- Subpopn parameter to estimate: Among **diabetic adults**, mean BMI

Theory of Subpopulation Analyses

- Earlier formulas calculate **point estimates**: use **entire sample** with indicator variable to “zero out” obsns not in subpopulation
- For estimated standard error, also use **entire sample.** Obsns in dataset who do **not** belong to subpopn **contribute** to calculation of estimated s.e.
- Domain analyses: examples of subpopns

Subpopulation Analysis in SAS Survey Procedures

- No subpopulation statement in SAS
  - Option in SUDAAN, STATA, SPSS & WesVar
- SAS knows **how** to conduct subpop analyses
  - **Does so** for NoMcar & for domain analyses
- But, not let **you** define your own subpop
- Default & "workaround methods" suggested by SAS for **your** subpop analyses **may** be cumbersome &/or underestimate s.e.

DB WorkAround Method for Subpop Analyses in SAS

- **Always** use **NoMCAR** on PROC statement
- For obsns **not** in subpop, code value of dependent variable = dot (e.g. . or .x)
- For obsns where DK if in subpopn due to item nonresponse, code dep var = . or .x
- Yields standard subpopulation analysis
  - SAS output agrees with SUDAAN with SUBPOPN

Lecture Example 8 SurveyFreq

Subpopulation Analysis of Diagnosed Diabetics

LecEx 8A
Check coding of variables

- **Proc Freq data = La04 ;**
- **TABLES diabetes * insulin /**
  list missing;
- Diabetes= 1=yes (840)
  - Insulin: 1=yes (217), 2=no (622), .=miss (1)
- Diabetes =2= no (8206), Insulin = .
- Diabetes = . = dk (18), Insulin = .

LecEx 8B: Estimate Prevalence of Diabetes

- Proc SurveyFreq data = La04 NoMcar ..;
- Strata _ststr ;
- Cluster _psu ;
- Weight _finalwt ;

- **TABLES diabetes / CL CLwt ;**

LecEx 8C: % and # of Diabetics Take Insulin

- Proc SurveyFreq data = La04 **NoMCAR** nosummary ;
- Strata _ststr ; Cluster _psu ;
- Weight _finalwt ;
- **TABLES insulin / cl clwt ;**
- DB work-around method: subpop analysis
  - Variable Insulin coded dot: obsns not in subpop

LecEx 8D. Among Diabetics, % and # Take Insulin, by Sex

- Proc SurveyFreq Data=La04 **nomcar**
- Strata _ststr ; Cluster _psu ;
- Weight _finalwt ;
- **TABLES sex * insulin / row CL nocellpercent risk1 OR ;**
- DB workaround method. Note that value of variable Insulin is dot for all obsns not in subpop

Proc SurveyMeans

Analytical Capabilities

SurveyMeans Basic Capabilities

- **Continuous/count** variables (BMI, ER visits)
  - Estimate Mean & Total with s.e., CI, CV
  - Estimate Percentiles
- **Categorical** variables (binge, marital status)
  - Estimate Percentage/proportion & Total with s.e., CI, CV
- Above for entire popn, domains, subpop
  - Need workaround method for subpopn analysis

SAS SurveyMeans Additional Capabilities

- Estimate population parameters that are ratios (used infrequently, but can be useful)

- One-sided confidence intervals
  - <u (-∞ is lower limit); >s (+∞ is upper limit)

- Compare domains to each other
  - **Only** in SurveyMeans **macro** available on WEB

SurveyMeans Syntax for BRFSS Survey, 1 year

- Proc SurveyMeans data = .. **options** ;
- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;
- **Var** _bmir _bmi4cat _RfBing2 **;**
- **Class** _bmi4cat _RfBing2 ;
  - **Class** statement identifies vars on **Var** statement analyzed as categorical; other vars on **Var** statement analyzed as continuous

SurveyMeans Keywords DOMAIN statement

- **Domain** Sex Race4 Age3r ;
  - Identifies domains for analysis
  - Variables on **VAR** statement analyzed for each level of each **DOMAIN** variable
  - Correct subpop analyses done by SAS here
- **BY** statement: do not use, use **DOMAIN**
  - Because standard error estimated correctly with DOMAIN statement & **not** with By

Some Options on PROC SurveyMeans Statement

- **ALL** (outputs all statistics)
- **NOBS MEAN STDERR CLM**
  - Above 4 are default for means/proportions
- **CV NMISS** (# obsns missing in analysis)
- **SUM** (estimated total for y variable)
- **STD** (estimated s.e. of estimated total)
- **CLSUM** (CI on total—2 sided)
- **CVSUM** (estimated CV of estimated total)

Lecture Example 9 SurveyMeans

Continuous and Categorical Dependent Variables

Lecture Example 9
LecEx 9A

- Estimate mean BMI: _Bmir
- Estimate binge drink prev (distribution):
  - _RfBing2 or Binger or Binge01
- 9A, check variables for coding/missing
  - Proc freq ; tables _rfbing2 ; 179 missing
  - Proc univariate ; var _bmir   ; 497 missing, also min = 6.68, max = 99.98 (OUTLIERS?)
- Note: I analyze _bmir values as real

LecEx 9B SurveyMeans Default

- Proc SurveyMeans data=La04 NoMcar ;
- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;
- **Var _Bmir Binge01 _RfBing2 ;**
- **Class _RfBing2 ;**
- /*default: get nobs, mean, stderr, clm */

LecEx 9C _Bmir with Options, SurveyMeans

- Proc SurveyMeans data = La04 **nobs nmiss mean stderr cv clm min max range lclm uclm df** NoMcar ;
- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;
- **Var _Bmir ;**

LecEx 9C Binge01 SurveyMeans, Options

- Proc SurveyMeans data = La04
  nobs nmiss mean stderr cv clm
  lclm uclm sum std clsum cvsum
  lclsum uclsum df NoMcar ;
- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;
- **VAR Binge01 ;**

LecEx 9D Percentiles SurveyMeans _bmir

- Proc SurveyMeans data = La04 NoMcar **quartiles percentile=(42 64) ;**

- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;

- **Var _Bmir ;**

Lecture Example 10 SurveyMeans

Domain Analyses (Sex) for BMI and Binge Drinking

LecEx 10. Sex Domains: SurveyMeans

- Proc SurveyMeans data=La04 NoMcar ;
- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;
- **Var _Bmir _RfBing2 ;**
- **Class _RfBing2 ;**
- **Domain Sex ;**

Do Males/Females Differ on Binge? BMI? SurveyMeans

- Cannot answer using SurveyMeans
  - Unless use SurveyMeans **macro** on WEB
- For binge drinking, use SURVEYFREQ
  - TABLES sex * _rfbing2 / chisq ;
  - Use prev ratio, prev difference, odds ratio (?)
- For mean BMI, can use SURVEYREG
  - Dependent = _BMIR,   Independent = SEX
  - Test regression coefficient for SEX
  - Not illustrated here

SAS MACRO %SMSUB

- http://support.sas.com/kb/25/033.html
- Supplements SURVEYMEANS calculations
- Contrasts for means, totals, & ratios
- Real SUBPOP statement
- Ratio estimates for subgroups
- Subgroup & overall estimates in 1 table

Lecture Example 11 SurveyMeans

Domains Formed by Cross-Classification of Two Variables

LecEx 11. Mean _Bmir SurveyMeans

- Proc SurveyMeans data=La04 NoMcar ;
- Strata _Ststr ; Cluster _Psu ;
- Weight _FinalWt ;

- **Var _Bmir ;**
- **Domain race4 sex sex * race4 ;**

Estimated Mean BMI, by RaceEth & Sex, LA, 2004

Race/Eth	Male	Female
W_NH	27.7	26.1
B_NH	27.9	29.1
HISPANIC	27.1	27.2
OTH_NH	28.4	26.6

Lecture Example 12 SurveyMeans

Subpopulation Analysis:

Same Procedure as Discussed Earlier

Subpopulation Analyses: Adult Diagnosed Diabetics

- Estimate percentage on insulin (diabetics)
  - INSULIN: missing value for all nondiabetics

- Estimate mean BMI for diabetics only
  - _BMIR—nondiabetics have value for variable

LecEx12B. Insulin among Diabetics. SurveyMeans

- Proc SurveyMeans **NoMcar** …. ;
- Strata _STSTR ; Cluster _PSU ;
- Weight _FinalWt ;
- VAR Insulin ;
- CLASS Insulin ;
- DB work-around method for subpopn
- INSULIN coded dot for all nondiabetics

LecEx 12C. Mean BMI among Diabetics. SurveyMeans

- DB method for subpopn
- Recode _bmir to dot if obsn is **not** a diagnosed diabetic; new dataset bmi_diab

- Proc Surveymeans **NoMcar data = bmi_diab ...**
- Strata ... ; Cluster .... ; Weight ... ;
- Var _bmir ;

LecEx 12D. Mean BMI among Diabetics. SurveyMeans

- Another method for subpop analysis

- Proc Surveymeans **NoMcar data = La04 …**
- Strata … ; Cluster …. ; Weight … ;
- **Var _bmir ;**
- **Domain Diabetes ;**
- Get twice the output that you want

Compare Domains to Each Other

Categorical Variables Only

Chi-Square Tests on

Two Way Tables, R x C

Chi-Square Tests-Survey Data R x C Table

- Are 2 categorical vars related (associated)?
  - Males/females same prevalence binge drinking?
    - 2x2: also prev difference, prev ratio, odds ratio
  - Three age domains same prevalence?
  - Four race/eth domains same BMI cat distbn?
- Null Hypothesis:
  - Two variables are statistically independent
- Alternate Hypothesis
  - Two variables not statistically independent

SurveyFreq: 4 Types Chi-Square Tests, all Pearson

- **Pearson** type test (based on proportions)
  - Observed minus expected number of elements in a cell—**weighted** of course
- **WCHISQ** request gives 2 tests (W = Wald)
  - Unadjusted F Wald, adjusted F Wald
  - Unadjusted = adjusted for 2 x 2 table
- **CHISQ** Rao-Scott Pearson modification
- **CHISQ1** Minor variation on **CHISQ**

SurveyFreq: 4 More Chi-Square Tests

- Loglinear test (based on log odds ratios)
  - **WLLCHISQ** request gives 2 tests (W = Wald)
  - Unadjusted F Wald, adjusted F Wald
    - Unadjusted = adjusted for a 2 x 2 table

- Likelihood ratio type test (ratio obs/exp)
  - **LRCHISQ** Rao-Scott LR modification
  - **LRCHISQ1** minor variation on LRCHISQ

8 (or 6) Chi Square Tests! Which one(s) to use?

- SAS manual--discussion & references
  - Several anticonservative **if** table sparse **& if** survey DDF small wrt (R-1)(C-1)
- STATA manual recommendation
  - Always use Rao-Scott Pearson (CHISQ option in SURVEYFREQ)
- BRFSS surveys—typically very large ddf
  - So no worry about small survey DDF

Lecture Example 4 (2 x 2) SurveyFreq Chi-Square

- Proc SurfeyFreq data = La04 NoMcar ;
- strata _ststr ; cluster _psu ;
- weight _finalwt ;
- **TABLES sex * _rfbing2 / ROW CL chisq chisq1 lrchisq lrchisq1 wchisq wllchisq nocellpercent ;**
- Everything after slash mark is an option
- Request 6 chi-square tests, as illustration

Interpretation of Significant Chi-Square Tests (2 x 2)

- CHISQ, CHISQ1, LRCHISQ, LRCHISQ1, WCHISQ
  - Prevalence of binge drinking not equal for males & females in popn: males higher
- WLLCHISQ
  - Odds of binge drinking not equal for males & females in popn: males higher

Lecture Example 5 (3 x 2) SurveyFreq Chi-Square

- Proc SurveyFreq NoMcar … ;
- Strata _ststr ; Cluster _psu ;
- Weight _finalwt ;
- **TABLES age3r * _rfbing2 / ROW chisq chisq1 lrchisq lrchisq1 wchisq wllchisq nocellpercent ;**
- Since no CL option, no cell percent output

Interpretation of Significant Chi-Square Tests (3 x 2)

- CHISQ, CHISQ1, LRCHISQ, LRCHISQ1, WCHISQ
  - Prevalence of binge drinking not equal for 3 age domains in popn
- WLLCHISQ
  - Odds of binge drinking not equal for 3 age domains in popn
- Tests **not** say **how** age domains differ on prevalence or odds

Lecture Example 6
3 way table in SURVEYFREQ

- Proc SurveyFreq data = La04 NoMcar …. ;
- Strata _ststr ; Cluster _psu ; Weight…;
- **TABLES age3r * sex * _rfbing2 / ROW chisq nocellpercent ;**
- Analysis: for **each** level of age3r,
  - Prevalence of binge drinking, by sex
  - Chi-square test of sex and binge drinking

Interpretation of Significant CHISQ Tests in Example 6

- For each age domain, males/females in the population differ on binge drinking prev: males higher

- Estimated binge drinking prevalences

Age 18-34:	34% M	12% F
Age 25-54:	21% M	7% F
Age 55+:	10% M	2% F

LecEx13. SurveyFreq Binge Prevalence by Race/Eth

- Proc SurveyFreq data = La04 NoMcar;
- Strata _ststr ; Cluster _psu ;
- Weight _finalwt ;

- Tables race4 * binge01 / row **chisq** CL nowt ;

Example 13 Results
Estimated Binge Prevalence

- **WNH 15.7% Hisp 23.7%**
- **BNH 10.4% OtherNH 12.4%**
- Rao-Scott chi-square test: p < .0001
  - All 4 domains not have same prevalence
  - SurveyFreq: not indicate which domains differ
- SurveyMeans: no option compare domains
  - Except if use SAS MACRO %SMSUB
- Can compare domains with SurveyReg

Compare Domains to Each Other on Mean or Prevalence

Can Use SAS SurveyReg
With Contrast and Estimate

Some Characteristics of SAS SURVEYREG

- **Linear regression**
  - Dependent variable continuous (usually)
  - Independent vars—continuous/categorical
- Similar to nonsurvey PROC GLM
  - Can use **Contrast** & **Estimate** statements
- Wald F test used to test default null hypotheses & those from **Contrast** or **Estimate** requests (sometimes is t-test)

Use SURVEYREG to Compare Domains

- Fit a “cell mean” model (no intercept)
  - Dependent variable: continuous (e.g. BMI) or dichotomous coded 1,0 (e.g. BINGE01)
  - Independent variable: domain variable
- Vector of regression coefficients is domain means or proportions
- Contrast: form linear combinations of regression coeffs want to estimate or test

What Is A Linear Contrast? Quick Review: BMI / Sex

- Define a vector of domain (sex) means
- $\begin{vmatrix} \bar{Y}_M \\ \bar{Y}_F \end{vmatrix}$ mean BMI
- Define row vector of constants (linear contrast) $\begin{vmatrix} 1 & -1 \end{vmatrix}$

Linear Contrast BMI/Sex

- Take product of two vectors (row x column)

$$\begin{vmatrix} 1 & -1 \end{vmatrix} \quad \begin{vmatrix} \bar{Y}_M \\ \bar{Y}_F \end{vmatrix} \quad = \bar{Y}_M - \bar{Y}_F$$

- Want to estimate or test domain differences
- Tell SurveyReg cell mean model, dependent var (BMI), ind. variable (sex), & linear contrast

Another Linear Contrast Example: BMI/Race

- Define a vector of domain (race) means-BMI
- $$\begin{vmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \overline{Y}_3 \\ \overline{Y}_4 \end{vmatrix}$$
- Define row vector of constants (linear contrast) $$\begin{vmatrix} 1 & 0 & -1 & 0 \end{vmatrix}$$

Another Linear Contrast Example: BMI/Race

- Multiply 2 vectors together (row x column)

$$\begin{vmatrix} 1 & 0 & -1 & 0 \end{vmatrix} \begin{vmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \overline{Y}_3 \\ \overline{Y}_4 \end{vmatrix} \qquad = \overline{Y}_1 - \overline{Y}_3$$

- Want to estimate or test domain differences
- Tell SurveyReg cell mean model, dependent var (BMI), ind. variable (race), & linear contrast

Another Linear Contrast Example: Binge/Race

- Define a vector of domain (race) props
- $\left|\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{matrix}\right|$ proportion binge drink
- Define row vector of constants (linear contrast) $\left|\begin{matrix} 1 & 0 & -1 & 0 \end{matrix}\right|$

Another Linear Contrast Example: Binge/Race

- Multiply 2 vectors together (row x column)

$$\begin{vmatrix} 1 & 0 & -1 & 0 \end{vmatrix} \begin{vmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{vmatrix} = P_1 - P_3$$

- Want to estimate or test domain differences
- Tell SurveyReg cell mean model, dependent var (binge01), ind. variable (race), & linear contrast

Lecture Example 14A SURVEYREG: BMI & Race/Eth

- **Proc surveyreg data =**
- **Strata _ststr ; Cluster _psu ;**
- **Weight _Finalwt ;**
- **CLASS Race4 ;** /* precede model */
- **Model _bmir = Race4 / NOINT Solution CLparm ;**
  - No intercept in model (cell mean model)

Cell Mean Model _bmir and Race4

- Vector of popn regression coeffs

$$\begin{vmatrix}\overline{Y}_1\\ \overline{Y}_2\\ \overline{Y}_3\\ \overline{Y}_4\end{vmatrix}$$

- 1st regr coeff is WNH mean BMI, 2nd is BNH, 3rd is Hispanic, 4th is OtherNH

SURVEYREG Contrast/Estimate Statements

- **CONTRAST statement**
  - Tests null hypothesis: popn value of specified contrast equals zero
- **ESTIMATE statement**
  - Estimates popn value of specified contrast
  - With estimated standard error & CI (option)
- Statements used here as in PROC GLM
  - GLM is only for SRS

Lecture Ex 14B (slide edit) Add statements to Ex 14A

- CONTRAST 'BNH minus WNH'
  RACE4 -1 1 0 0 ;

$$-\overline{Y}_1 + \overline{Y}_2$$

- ESTIMATE 'BNH minus WNH'
  RACE4 -1 1 0 0 ;

$$-\overline{Y}_1 + \overline{Y}_2$$

Lecture Ex 14B (slide edit) Add statements to Ex 14A

- CONTRAST ‘Hispanic minus WNH’
  RACE4 -1 0 1 0 ;

$$-\bar{Y}_1 + \bar{Y}_3$$

- ESTIMATE ‘Hispanic minus WNH’
  RACE4 -1 0 1 0 ;

$$-\bar{Y}_1 + \bar{Y}_3$$

Lecture Ex 14B (slide edit) Add statements to Ex 14A

- CONTRAST ‘BNH minus Hispanic’
  RACE4 0 1 -1 0 ;

$$\boxed{\overline{Y}_2 - \overline{Y}_3}$$

- ESTIMATE ‘BNH minus Hispanic’
  RACE4 0 1 -1 0 ;

$$\boxed{\overline{Y}_2 - \overline{Y}_3}$$

Conclusions Regarding Race/Eth and Mean BMI

- For population of noninstitutionalized adults resident in LA in 2004 (who would agree to report height & weight, if asked):
- 1. BNHs have higher mean BMI than WNHs
- 2. No evidence to question assumption that Hispanics & WNHs have same mean BMI
- 3. BNHs have higher mean BMI than Hispanics

Compare 4 Race/Ethnicity Domains on Binge Prevalence

- In previous LecEx 14, use binge01 as dependent variable instead of _bmir.
- Cell mean model will estimate binge prevalence for each race/ethnicity domain
- Compare domains to each other with Contrast or Estimate

Lecture Example 14C
Use SURVEYREG

- **Proc surveyreg data =**
- **Strata _ststr ; Cluster _psu ;**
- **Weight _Finalwt ;**
- **CLASS Race4 ;** /* precede model */
- **Model binge01 = Race4 / NOINT Solution CLparm ;**
  - No intercept in model (cell mean model)

Cell Mean Model Binge01 and Race4

- Vector of popn regression coeffs

$$\left|\begin{matrix}P_1\\P_2\\P_3\\P_4\end{matrix}\right|$$

- 1st regr coeff is WNH prev, 2nd is BNH prev, 3rd is Hispanic prev, 4th is OtherNH prev

Lecture Ex 14D
Add statements to Ex 14C

- CONTRAST ‘WNH minus BNH’
  RACE4 1 -1 0 0 ;

$$\boxed{P_1 - P_2}$$

- ESTIMATE ‘WNH minus BNH’
  RACE4 1 -1 0 0 ;

$$\boxed{P_1 - P_2}$$

Lecture Ex 14D
Add statements to Ex 14C

- CONTRAST ‘Hispanic minus WNH’
  RACE4 -1 0 1 0 ;

$$\boxed{-P_1 + P_3}$$

- ESTIMATE ‘Hispanic minus WNH’
  RACE4 -1 0 1 0 ;

$$\boxed{-P_1 + P_3}$$

Lecture Ex 14D (slide edit) Add statements to Ex 14C

- CONTRAST ‘Hispanic minus BNH’
  RACE4 0 -1 1 0 ;

$$\boxed{-P_2 + P_3}$$

- ESTIMATE ‘Hispanic minus BNH’
  RACE4 0 -1 1 0 ;

$$\boxed{-P_2 + P_3}$$

Conclusions Regarding Race/Eth & Binge Drink Prev

- For population of noninstitutionalized adults resident in LA in 2004 (who would agree to provide alcohol consumption info, if asked):
- 1. BNHs have lower binge prev than WNHs
- 2. WNHs vs. Hispanics: p = .0549

  Estimated diff = .0804, est se = .0419
- 3. BNHs have lower binge prev than Hispanics

REFERENCES

References on Sample Survey Design and Analysis

Recommended Books: Surveys & Their Analysis

- Heeringa, Steven, BT West, PA Berglund. Applied Survey Data Analysis, Chapman & Hall/CRC, Boca Raton, FL, 2010. Excellent. $84 list.
- Groves, Robert et al, Survey Methodology, 2nd edn., John Wiley, 2009, paper, $85 list.
  - Introduction/overview of all aspects of surveys
- Korn, Edward & Barry Graubard, Analysis of Health Surveys, John Wiley, 1999. $165 list.
  - Strategies for survey data analysis, math-stat useful

Recommended Books: Sampling Methods & Analysis

- Lee, Enu Sul & Robert Forthofer. Analyzing Complex Survey Data, 2nd edn, 2006, Sage Publs.
  - Short, concepts oriented, condensed Korn/Graubard
- Lohr, Sharon. Sampling: Design and Analysis. 2010, Brooks/Cole, Cengage Learning.
  - Applied introduction to sampling (algebra)
  - Clear explanations and real-life examples
- Cochran, William G. Sampling Techniques: 3rd Edition. 1977, John Wiley. Math-stat.

Some Useful WEB Sites

- http://www.amstat.org/sections/srms
  - ASA, Survey Research Methods Section
  - What Is A Survey? booklets excellent
- http://www.hcp.med.harvard.edu/statistics/survey-soft/ Software for survey data
- http://www.aapor.org . Go to Resources & Education, then Researchers, then: Best Practices, Standard Definitions Response Rate (2011), Poll/Survey FAQ. Excellent discussions.

Special Issues of Public Opinion Quarterly

- Vol. 70, No. 5, 2006. “Special Issue: Nonresponse Bias in Household Surveys”
- Vol. 71, No. 5, 2007. “Special Issue: Cell Phone Numbers & Telephone Surveying in U.S.
- Vol. 74, No.5, 2010. “Special Issue: Total Survey Error”
- http://www.oxfordjournals.org/our_journals/poq/collectionspage.html PH Survey Methods

Some Survey Research Journals

- Survey Methods: Insights from the Field. http://surveyinsights.org/ (electronic)
- Journal of Survey Statistics & Methodology. http://www.oxfordjournals.org/our_journals/jssam/
- Survey Methodology. http://www.statcan.gc.ca/ads-annonces/12-001-x/index-eng.htm

Lab Exercises
See MS-Word documents

- Estimate # diabetics & diabetes prevalence
  - Then by sex, by age, by race/eth, race/eth * sex
- Compare males/females on diabetes via prevalence ratio, risk difference, odds ratio
  - Now do comparison within each level of race/eth
- For subpopulation of diagnosed diabetics:
  - Estimate mean age 1st told diabetic
  - Estimate # take diab pills & prevalence diab pills