

ICPSR 20240

Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003 [United States]

Margarita Alegria

*Center for Multicultural Mental Health Research
at Cambridge Health Alliance*

James S. Jackson

*University of Michigan. Institute for Social
Research*

Ronald C. Kessler

*Harvard Medical School. Department of Health
Care Policy*

David Takeuchi

University of Washington

Training Materials

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

Terms of Use

The terms of use for this study can be found at:
<http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/20240/terms>

Information about Copyrighted Content

Some instruments administered as part of this study may contain in whole or substantially in part contents from copyrighted instruments. Reproductions of the instruments are provided as documentation for the analysis of the data associated with this collection. Restrictions on "fair use" apply to all copyrighted content. More information about the reproduction of copyrighted works by educators and librarians is available from the United States Copyright Office.

NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.



NIMH Collaborative Psychiatric Epidemiology Surveys

CPES Sample Designs and Analysis Weights

Steve Heeringa



Sample Design and Weighting Documentation

- www.icpsr.umich.edu/CPES
- Table of contents
 - Sample design
 - Weighting
- Detailed description of sample design and weighting procedures.
- Also, Heeringa et al. (2004). "Sample Designs and Sampling Methods for the Collaborative Psychiatric Epidemiology Studies (CPES)." *International Journal of Methods in Psychiatric Research*, 13(4), 221- 239.



CPES Estimation and Inference

- CPES is based on the integration of three nationally representative multi-stage area probability samples.
- Estimation will require weighted analysis to compensate for disproportionate sampling.
- Inference (design-based) will require the use of variance estimation procedures for complex sample designs.

C P E S Complex Sample Design

- 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)



Complex Sample Design

- “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 simple random sampling (SRS)
 - Unbiased estimates with measurable sampling error are still possible
 - Independence of observations, equal probabilities of selection may no longer hold

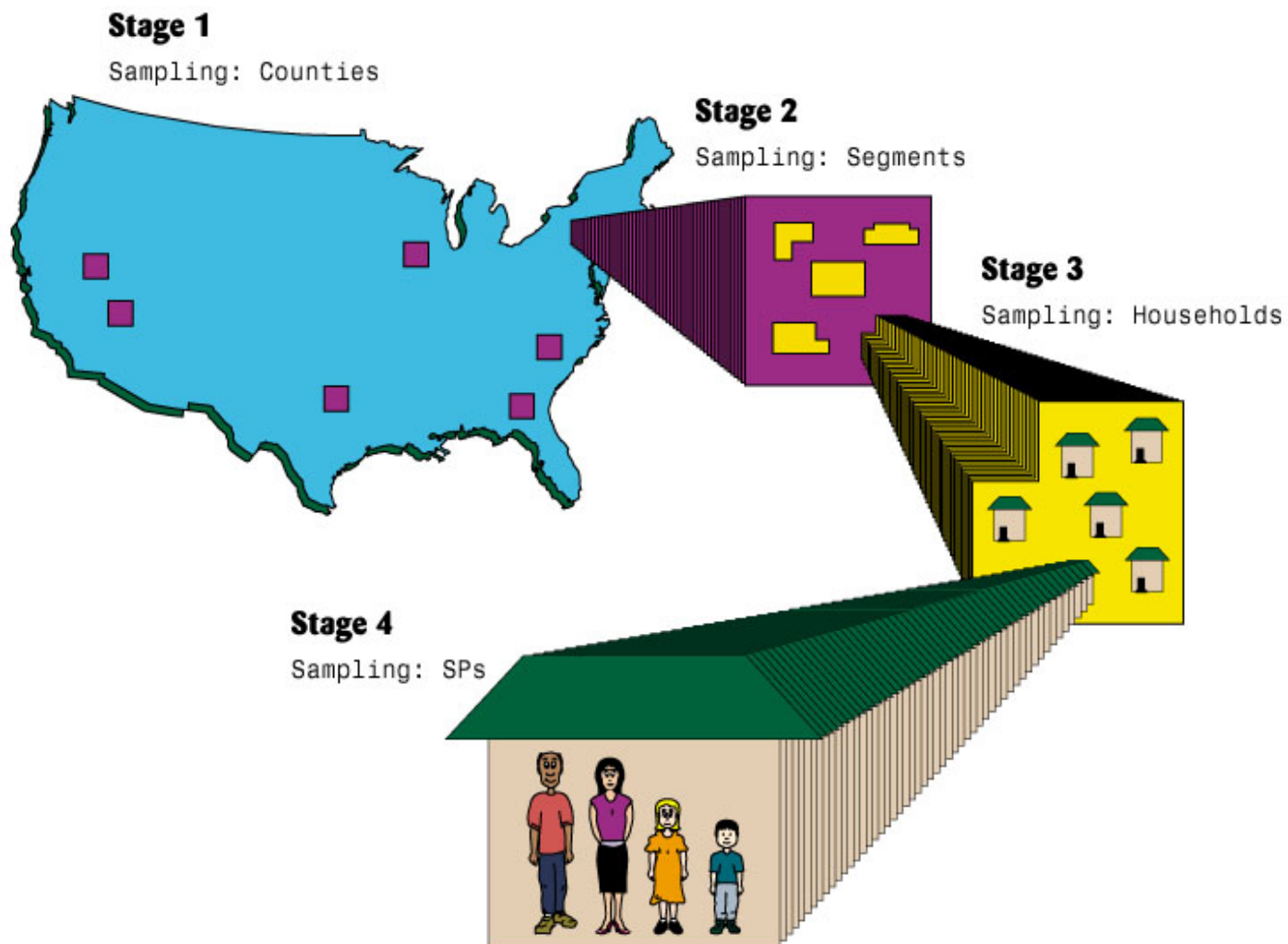


Where are Complex Sample Designs Used?

- 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.
- Multi-stage area probability sample designs for household surveys: NCS-R, NSAL, NLAAS, NHIS, NHANES, CPS, etc.



Multi-Stage Sample of U.S. Households





Consortium in Psychiatric Epidemiology Studies (CPES)

- Sampled persons are interviewed about mental health and mental health-related matters.
- Separate estimates (oversampling) required for domains defined by:
 - Age-sex groups, African Americans, Latinos, Asians and All Others.



CPES SAMPLE DESIGN (Primary stage)

- The PSUs are mostly single counties or metropolitan statistical areas (MSAs).
- Stratify PSUs by region and MSA status.
- Select PSUs by probability proportionate to estimated size (PPES) sampling (a few selected with certainty).



CPES SAMPLE DESIGN (Second Stage)

- The SSUs are area segments. Area segments are Census blocks or combinations of Census blocks with a minimum number of households, e.g. 50 for NCS-R.
- Area segments are selected with PPES- typically 6-20 area segments per PSU



CPES SAMPLE DESIGN (Third Stage)

- Before selection can occur, field staff list the housing units in the selected area segments.
- For each segment, a subsample of housing units is selected from the prepared list.



CPES SAMPLE DESIGN (Fourth Stage)

Conduct a screening interview to classify persons by domain.

- Subsample persons to give required sampling rates by domain.
 - Survey weights will reflect this subsampling step
- Conduct interviews with sample persons (SPs).



Basic Weighting Approach

- Suppose sample element i was selected with probability π_i . Then sample element i represents $(1 / \pi_i)$ elements in the population.
- That is, count the element i in the analysis by giving it a weight of $w_i = (1 / \pi_i)$.
- For example, a sample element selected with probability $1/10$ represents 10 elements in the population.



Weighting

Weighting is used to compensate for

- Unequal probabilities of selection
- Nonresponse (typically, a unit that fails to respond)
- In poststratification to adjust weighted sample distributions for certain variables (e.g., age and sex) to make them conform to the known population distribution.



Overall Weight

CPES weights incorporate simultaneously all three components, unequal probabilities of selection, nonresponse, and poststratification:

$$W = W_{sel} \cdot W_{nr} \cdot W_{pstrat}$$

- In probability samples in which the sample elements have been selected with different probabilities and each case has been assigned a weight, $W_i = 1/\pi_i$, the unbiased estimator of the population mean is:

$$\bar{y}_{weighted} = \frac{\sum_{i=1}^n W_i \cdot y_i}{\sum_{i=1}^n W_i}$$

- The simple unweighted estimator of the sample mean is a special case of the weighted estimator of the population mean with $W_i=1$ for all cases:

$$\bar{y}_{weighted} = \frac{\sum_{i=1}^n W_i \cdot y_i}{\sum_{i=1}^n W_i} = \frac{\sum_{i=1}^n 1 \cdot y_i}{\sum_{i=1}^n 1} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$



Weighted Estimates of Population Statistics

Weighted estimate of the Variance of a Random Variable: S^2

$$S_{weighted}^2 = \frac{\sum_{i=1}^n W_i (y_i - \bar{y})^2}{\sum_{i=1}^n W_i - 1}$$

Weighted Estimate of the Covariance of Variables y and x: S_{xy}

$$S_{xy,weighted} = \frac{\sum_{i=1}^n W_i (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^n W_i - 1}$$

Weighted Estimate of the
Simple Linear Regression Coefficient: β

$$\hat{\beta}_{weighted} = \frac{\sum_{i=1}^n W_i (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^n W_i (x_i - \bar{x})^2}$$

.

Weighted Estimate of
Multiple Linear Regression Coefficients, $\underline{\beta}$

$$\underline{\beta} = (X'WX)^{-1} X'WY$$

where: W is an $n \times n$ diagonal matrix with weight,

$$W = \begin{bmatrix} W_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & W_n \end{bmatrix}$$

Weighted Estimation

Pseudo-Maximum Likelihood,
Weights are included in the score equations,
e.g. for Logistic Model

$$\begin{aligned} U(B) &= \frac{\delta \ln L(B)}{\delta \beta} \\ &= \sum_h \sum_{\alpha} \sum_i w_{h\alpha i} \tilde{x}'_{h\alpha i} y_{h\alpha i} \\ &\quad - \sum_h \sum_{\alpha} \sum_i w_{h\alpha i} \tilde{x}'_{h\alpha i} P_{h\alpha i}(\beta) \end{aligned}$$

Effect of Weights on Variances of Estimates of Means

L_W = Relative increase in sampling variances of estimates due to weighting

$$\sim CV^2(W_i) = \left[\frac{\sum_1^n W_i^2}{\left(\sum_1^n W_i \right)^2} \right] \cdot n - 1$$

Generally, oversampling for subgroups:

$$L_{W,SUB} < L_W$$



Example of Weighting Loss in Disproportionate Sampling

| Stratum | % Hispanic Population | Oversampling Rate | Weight | % of Hispanic Sample |
|---------|-----------------------|-------------------|--------|----------------------|
| 1 | 19.2% | 1:1 | 4 | 7% |
| 2 | 22.8% | 2:1 | 2 | 7% |
| 3 | 24.1% | 3:1 | 1.33 | 26% |
| 4 | 33.9% | 4:1 | 1 | 50% |
| | 100% | | | 100% |

Example of Weighting Loss in Disproportionate Sampling

For $n=1000$:

$$L_W = \left[\frac{\sum_1^n W_i^2}{\left(\sum_1^n W_i \right)^2} \right] \cdot n - 1$$
$$= \left[\frac{2759.9}{2,148,569} \right] \cdot 1000 - 1 = .284$$

- Analysis based on:

$$W_i = \frac{1}{\pi_i}; \quad W_i^* = a \cdot W_i; \quad W_i^{**} = \frac{W_i}{b}$$

where : a and b are constants

will yield the same results (provided the software program treats the weights correctly).

Normalizing Weights

$$W_{i,cent} = W_i \cdot \frac{n}{\sum_{i=1}^n W_i}$$

Properties of normalized weights include:

- $\sum W_{i,cent} = n$
- $\bar{W}_{cent} = 1.0$



Example: Distribution of NCS Part 2 Weights (Normalized)

| Moments | | | |
|------------------------|------------|-------------------------|------------|
| N | 5877 | Sum Weights | 5877 |
| Mean | 1.00000007 | Sum Observations | 5877.00039 |
| Std Deviation | 1.16010775 | Variance | 1.34584999 |
| Skewness | 2.8311973 | Kurtosis | 7.9882 |
| Uncorrected SS | 13785.2153 | Corrected SS | 7908.21452 |
| Coeff Variation | 116.010767 | Std Error Mean | 0.01513284 |

C P E S CPES Weighting Approach

- Separately, NCS-R, NSAL and NLAAS cases are assigned to 1 of 12 distinct race/ancestry groups
 - Vietnamese, Filipino, Chinese, All other Asian, Cuban, Puerto Rican, Mexican, All other Hispanic, Afro-Caribbean, African-American, White, All Other
 - Persons with multiple ancestry are assigned to the first applicable discrete category in this list.

C P E S CPES Weighting Approach

- Separately, NCS-R, NSAL and NLAAS cases are assigned to 1 of 11 distinct geographic domains defined based on the race/ancestry composition of the Census tract in which they resided at the time of interview.
 - Hawaii is only represented in NLAAS

C P E S CPES Weighting Approach

- The March 2002 Current Population Survey (CPS) was used to estimate the total size of the adult population in each race/ancestry stratum.
- Within each race/ancestry stratum the distribution of the population to the 11 geographic domains was estimated based on the CPES study which could provide the most precise estimates, e.g. NSAL for African Americans.

C P E S CPES Weighting Approach

- Separately, NCS-R, NSAL and NLAAS study-specific weights (see chart) were post-stratified to the 12 x 11 grid of population totals for race/ancestry by geographic domain.

C P E S CPES Weighting Approach

- CPES weight values for the pooled data were constructed by scaling each individual study weight by a factor equal to the proportion of the total sample in a grid cell that originated with the component study.
- The result is a CPES weight that sums to the CPS 2002 post-stratification controls in the 12 x 11 grid.

C P E S CPES Weighting Approach

- Long and Short form weight variables– NCS-R uses a two-phase sampling approach in which only a subsample of cases progress from Part 1 to Part 2.
- Therefore, there are 2 CPES weights:
 - CPESWTLG – used when one or more variables of interest are measured only in NCS-R Part 2
 - CPESWTSH – used when all variables of interest are measured in NCS-R Part 1

C P E S CPES Weighting Approach

- Weights for pooled analysis of data from only two of the three CPES data sets:
 - NCNLWTLG, NCNLWTSH: NLAAS and NCS-R
 - NCNSWTLG, NCNSWTSH: NSAL and NCS-R
 - NSNLWT : NLAAS and NSAL
- Derived as CPES weights but with final scaling step limited to proportion of sample in cell from the two studies.



Table 1: Descriptive Statistics for CPES Study Weights

| Sample | Sample | Weight | Mean | S.E. | CV |
|------------|--------|----------|---------|--------|--------|
| NCSR | Short | NCSRWTSH | 1.00 | 0.0054 | 52.28 |
| NCSR | Long | NCSRWTLG | 1.00 | 0.0127 | 95.82 |
| NLAAS | --- | NLAASWGT | 6333.6 | 86.77 | 93.42 |
| NSAL | --- | NSALWTPN | 8022.2 | 165.84 | 161.22 |
| CPES | Short | CPESWTSH | 10468.2 | 75.25 | 101.69 |
| CPES | Long | CPESWTLG | 12693.4 | 152.03 | 153.49 |
| NCSR-NLAAS | Short | ncnlwtsh | 15061.7 | 100.10 | 78.44 |
| NCSR-NLAAS | Long | ncnlwtlg | 20290.6 | 261.12 | 130.87 |
| NCSR-NSAL | Short | ncnswtsh | 13588.6 | 106.91 | 97.52 |
| NCSR-NSAL | Long | ncnswtlg | 17731.9 | 248.46 | 152.04 |
| NLAAS-NSAL | --- | nsnlwt | 19553.1 | 608.91 | 322.59 |

Table 2: Un-weighted and Weighted Prevalence of Major Depressive Disorder

| Sample | Sample size | Un-weighted Prevalence % (s.e.) | Weighted Prevalence % (s.e.) |
|---------------------------|--------------------|--|-------------------------------------|
| White-Non-Hispanic | | | |
| NCSR ^a | 6696 | 17.95(0.47) | 17.75(0.55) |
| CPES ^a | 7567 ^b | 18.00(0.44) | 17.92(0.54) |
| Black | | | |
| NCSR ^a | 1176 | 11.99(0.95) | 11.29(1.04) |
| NSAL | 3433 | 10.63(0.53) | 10.31(0.58) |
| CPES ^a | 4609 ^b | 10.98(0.46) | 10.39(0.47) |
| Hispanic | | | |
| NLAAS | 2554 | 15.66(0.72) | 13.79(0.68) |
| CPES ^a | 3615 | 15.60(0.60) | 13.71(0.63) |

a =short sample, *b*= excluding missing cases *s.e.*=standard error

Table 3: Un-weighted and Weighted Prevalence of Generalized Anxiety Disorder

| Sample | Sample size | Un-weighted Prevalence % (s.e.) | Weighted Prevalence % (s.e.) |
|---------------------------|--------------------|--|-------------------------------------|
| White-Non-Hispanic | | | |
| NCSR ^a | 6696 | 6.45(0.30) | 6.16(0.31) |
| CPES ^a | 7566 ^b | 6.25(0.28) | 6.08(0.32) |
| Black | | | |
| NCSR ^a | 1176 | 4.34(0.59) | 4.14(0.74) |
| NSAL | 3430 | 3.53(0.31) | 3.26(0.36) |
| CPES ^a | 4606 ^b | 3.73(0.28) | 3.48(0.34) |
| Hispanic | | | |
| NLAAS | 2554 | 3.48(0.36) | 2.82(0.40) |
| CPES ^a | 3615 | 3.76(0.32) | 3.03(0.34) |

a =short sample, b= excluding missing cases s.e.=standard error

Table 4: Un-weighted and Weighted Prevalence of Post Traumatic Stress Disorder

| Sample | Sample size | Un-weighted Prevalence % (s.e.) | Weighted Prevalence % (s.e.) |
|---------------------------|--------------------|--|-------------------------------------|
| White-Non-Hispanic | | | |
| NCSR ^a | 4180 | 10.26(0.47) | 6.84(0.54) |
| CPES ^b | 4180 ^b | 10.26(0.47) | 6.92(0.53) |
| Black | | | |
| NCSR ^a | 679 | 12.37(1.26) | 7.4(0.86) |
| NSAL | 3419 ^c | 9.15(0.49) | 9.10(0.54) |
| CPES ^a | 4098 ^c | 9.69(0.46) | 8.76(0.49) |
| Hispanic | | | |
| NLAAS | 2554 | 5.29(0.44) | 4.42(0.45) |
| CPES ^a | 3259 ^c | 6.17(0.42) | 4.74(0.39) |

a = long sample, b = only assessed in NCS-R, c = excluding missing cases s.e.=standard error

Analysis of Complex Survey Data and Survival Analysis

- This session will cover an overview of issues in analysis of complex sample data and illustrate use of some of these concepts using CPES data
- Analysis topics demonstrated with complex sample data adjustments during the CPES training
 - Survival Analysis-covered in detail during this session
 - Descriptive** will be covered in detail in a later session
 - Regression** will be covered in detail in a later session
- Pat Berglund-Senior Research Associate-Institute for Social Research/Harvard Medical School
- June 18, 2008 8:30-10am

CPES Overview

- This section will focus on the CPES surveys
 - Overall sample design
 - CPES instruments (NCS-R, NSAL, NLAAS)
 - Weights
 - Complex sample features and variables

Key WebSites

- Along with the CPES website as the starting point each project has their own site
 - NCSR: www.hcp.med.harvard.edu/ncs
 - NSAL: <http://www.rcgd.isr.umich.edu/prba/nsal>
 - NLAAS: <http://www.multiculturalmentalhealth.org/nlaas.asp>
- Other useful sites for analysis of complex survey data:
 - SAS: www.sas.com
 - Stata: www.stata.com
 - Sudaan software: www.rti.org/SUDAAN/
 - SPSS software: www.spss.com
 - Iveware software: www.isr.umich.edu/src/smp/ive
 - R software: <http://www.r-project.org/>
 - Mplus software: <http://www.statmodel.com/>
 - Wesvar software: <http://www.westat.com/wesvar/>

Overview of Public Release Dataset

- All raw and selected diagnostic, demographic, sample design and weight variables are included in one file, overall n= 20113
 - NCS-R has 2 sections
 - n=9282 for the entire part 1 sample with a sub-sample of respondents that completed part 2 of the instrument, n=5692 part 2 respondents
- Raw variables-includes all variables that could be released while keeping disclosure issues in mind
- Diagnostic variables-includes selected diagnostic variables such as ICD and DSM disorders along with age of onset, age of recency, lifetime, 12 month disorders, and 30 day disorders

Overview of Public Release Dataset

- Demographic variables-includes selected demographic and design variables
- The dataset can be downloaded in various formats such as SAS transport, ASCII with SAS, SPSS, or Stata setup statements
- Associated documentation tools
 - online and Adobe/pdf format codebook
 - Adobe/pdf format of the instrument
 - SDA analysis system
- Other tools are related literature links and background information on the study and analysis tips including sample programs

Survey Instruments

- NSAL and NLAAS have one section in their respective instrument while the NCS-R has 2 sections

- The instruments can be viewed from the CPES website
 - Under “Interactive Documentation” and then “Sample Design”
 - The key point here is that each survey is a “complex” survey based on stratified/multi-stage/probability samples
 - These samples should be analyzed with techniques that take the complex sample design into account

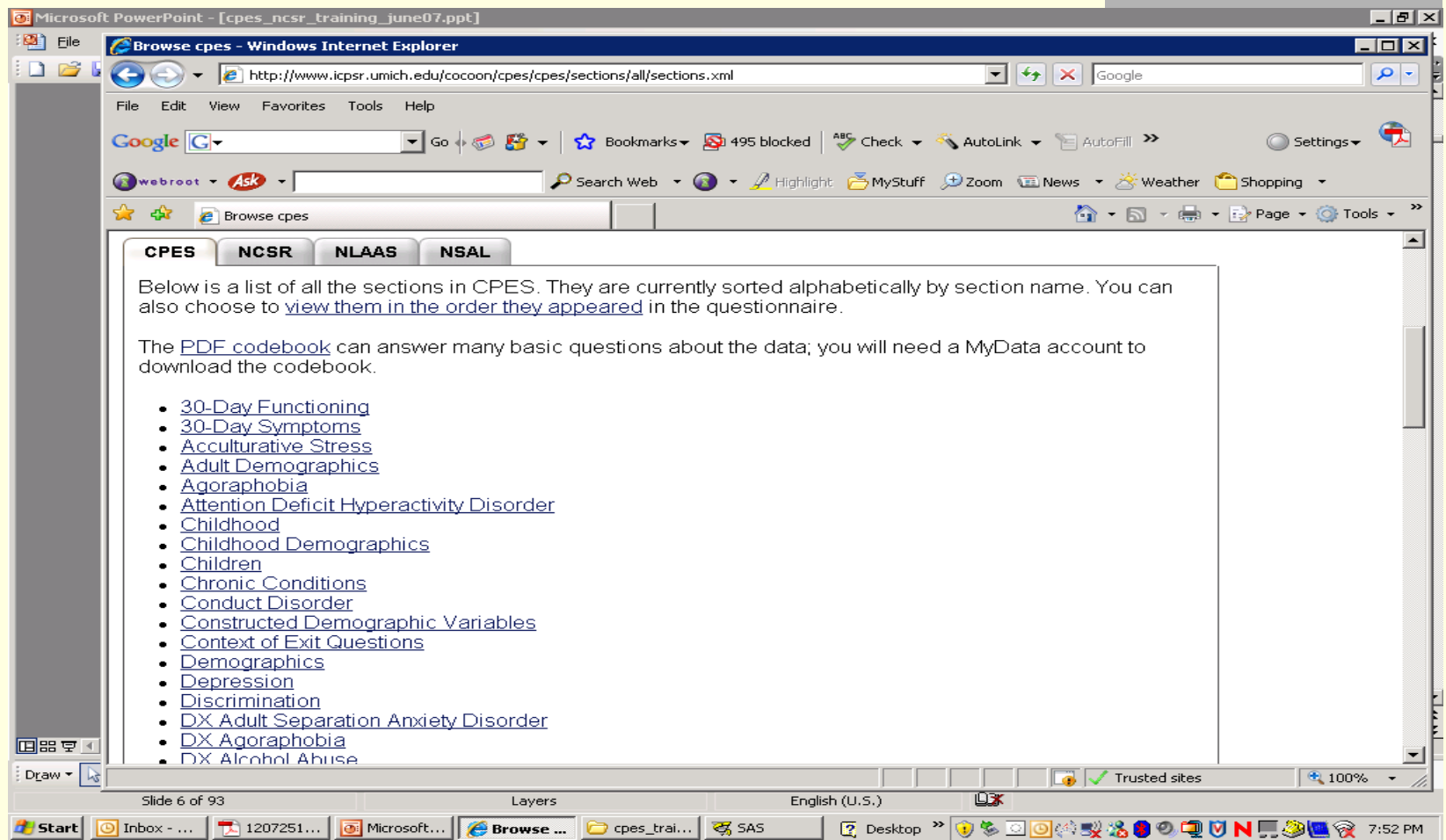
Sample Design Information

The screenshot shows a Microsoft PowerPoint presentation titled "cpes_ncsr_training_june07.ppt". The current slide is titled "Using CPES" and is the 5th of 92 slides. The slide content includes a list of links for "Using CPES":

- [Introduction](#)
 - [National Comorbidity Survey Replication \(NCS-R\) \(NCS-R Specific Documentation\)](#)
 - [National Survey of American Life \(NSAL\)](#)
 - [National Latino and Asian American Study \(NLAAS\)](#)
- [Sample Design](#)
 - [National Comorbidity Study Replication \(NCS-R\) sample design](#)
 - [National Study of American Life \(NSAL\) sample design](#)
 - [National Latino and Asian American Study \(NLAAS\) Sample Design](#)
- [Questionnaire Development](#)
- [Survey Management](#)
- [Data Collection](#)

The background of the slide features a banner for "Collaborative Psychiatric Epidemiology Surveys" with the "CPES" logo and navigation tabs: Background, Using CPES, Interactive Documentation, Download Data, Publications, and Online Analysis. A navigation bar at the bottom of the banner includes links: home - related sites - search - contact us - help - MyData options...

CPES Sections



NCS-R Instrument - Sections and Flow

The NCS-R instrument is divided into 2 parts

- Part 1 includes sections 1-14 with an additional demographic section for those that do not go on to complete Part 2
- Part 2 includes detailed questions about additional disorders such as gambling disorder, childhood disorders such as conduct disorder and ADD, social networks, family history/risk factors and other detailed sections such as finances
- At the end of the Pharmacoepidemiology section, a series of questions directing flow into Part 2 of the survey are included, other key flow questions are included in the Screener section and these are referenced in the Pharmacoepi section as well

Flow into Part 2 of Interview

- Questions at the end of the pharmacoepi section channel respondents to various sections of the questionnaire
- These questions are related to the screening questions of the Screener section (see instrument instructions)
- Overall strategy of using the Screener section is detailed in the paper “The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)”,
RONALD C. KESSLER, T. BEDIRHAN ÜSTÜN

Rationale for 2 Parts to NCS-R Instrument

- Interview Length and Analysis content (see Kessler Design and Field paper)

“Part II included assessments of risk factors, consequences, services, and other correlates of the core disorders. Part II also included assessments of additional disorders that were either of secondary importance or that were very timeconsuming to assess.

Part II was administered only to 5,692 of the 9,282 NCS-R respondents, over-sampling those with clinically significant psychopathology. All respondents who did not receive Part II were administered a brief demographic battery and were then either terminated or sampled in their appropriate proportions into sub-sampled interview sections that are described below.”

CPES Sample Design

- The CPES is a combination of 3 surveys and thus each sample design should be considered when using the combined data set
 - What are elements of design for each survey and how was this accounted for in the CPES combined file?
 - Implications for analysis when using the combined file?
 - Implications for analysis when using each separate file?
 - Our focus is on the CPES analyzed as a combined data set

NCSR Sample Design

- Nationally representative multi-stage clustered area probability sample of households. Interviewed people in the age range 18 years and older, rather than in the NCS-1 age range of 15-54. The exclusion of the 15-17 age range was dictated by carrying out a separate NCS Adolescent survey of 10,000 respondents in the age range 13-17. The inclusion of the age range 55 years and older was based on the desire to study the entire adult age range. Part II was administered only to 5,692 of the 9,282 Part I respondents, including all Part I respondents with a lifetime disorder plus a probability subsample of other respondents.
- See Kessler, Ronald C.; Berglund, P.; Chiu, W.T.; Demler, O.; Heeringa, S.; Hiripi, E.; Jin, R.; Pennell, B.E.; Walters, E.E.; Zaslavsky, A.; Zheng, H., "The US National Comorbidity Survey Replication (NCS-R): Design and field procedures." *International Journal of Methods in Psychiatric Research*. 2004, 13, (2), 69 - 92.

NSAL Sample Design

- The NSAL survey populations included all US adults in the three target groups who were age 18 and older and resided in households located in the coterminous 48 states. The African-American survey population included only Black adults who did not identify ancestral ties in the Caribbean. The Afro-Caribbean survey population was limited to Black adults who self-identified as being of Caribbean ancestry. The White survey population included all Caucasian adults except persons of self-reported Hispanic ancestry. Institutionalized persons including individuals in prisons, jails, nursing homes, and long-term medical or dependent care facilities were excluded from the study population. Military personnel living in civilian housing were eligible for the study but residents of housing located on a military base or military reservation were excluded. The NSAL survey populations were restricted to adults who were able to complete the interview in English.
- See CPES website for more details.

NLAAS Sample Design

- The National Latino and Asian American Study (NLAAS) is a nationally representative community household survey that estimates the prevalence of mental disorders and rates of mental health service utilization by Latinos and Asian Americans in the United States. The central aims of the NLAAS were threefold. First, to describe the lifetime and 12-month prevalence of psychiatric disorders and the rates of mental health services use for Latino and Asian American populations using nationwide representative samples of these groups. Second, to assess the associations among social position, environmental context, and psychosocial factors with the prevalence of psychiatric disorders and utilization rates of mental health services. Third, to compare the lifetime and 12-month prevalence of psychiatric disorders, and utilization of mental health services of Latinos and Asian Americans with national representative samples of non-Latino whites (drawn from the National Comorbidity Study-Replication (NCS-R) and African Americans (drawn from the National Survey of American Life (NSAL)).
- See CPES website for more details.

Considerations for Analysis of the CPES

Analysis Considerations

■ Weights

- Weights and the sample design must be considered when analyzing the CPES data, either as a combined file or as a subset of the 3 surveys
- Weights are included for a number of combinations of the surveys and we focus on use of the full CPES data set

■ Sample Design

- Due the clustering of the CPES combined sample designs, variance estimation from standard software procedures is incorrect since it assumes a simple random sample
- In general, variance/SE's and hypothesis tests will be under estimated due to clustering and stratification along with weighting

CPES Weights

- Each of the 3 data sets are weighted to adjust for differential probabilities of selection of respondents within households and differential non-response as well as to adjust for residual differences between the sample and the United States population.
- An overall CPES weight is available and should be used for combined analysis
- An additional weight was developed for the NSAL and NLAAS to match the NCSR Part 2 weight construction
- It is essential that the analyst account for the complex sample design and use the correct weights in every analysis
- See CPES website for details on each survey's weight construction

Why Use Weights?

- Weighting is used to compensate for:
 - Unequal probabilities of selection
 - Nonresponse (typically, a unit that fails to respond)
 - In poststratification to adjust weighted sample distributions for certain variables (e.g., age and sex) to make them conform to the known population distribution.
 - It is used to improve the accuracy (minimize bias) of sample estimates and to compensate for noncoverage and nonresponse

Basic Weighting Approach

- Suppose sample element i was selected with probability p_i . Then sample element i represents $(1 / p_i)$ elements in the population.
- That is, count the element i in the analysis by giving it a weight of
- $w_i = (1 / p_i)$.
- For example, a sample element selected with probability $1/10$ represents 10 elements in the population.

Overall Weight

- Weighting may incorporate simultaneously all three components, unequal probabilities of selection, nonresponse, and poststratification:
- Weight for unequal probabilities of selection: w_1 ;
- Weight for sample nonresponse: w_2 ;
- Poststratification weight for population noncoverage and sampling variance reduction: w_3 .

Then compute the overall weight as:

$$W = w_1 \times w_2 \times w_3$$

- From Heeringa - "Analysis of Complex Sample Survey Data" course, ISR SRC-Summer Institute

Bias Example

- Use of final weights is important to obtain correct, unbiased prevalences, as an example I present a table that outlines the effects of not using weights for Mexico, one of the countries in the WMH Initiative

Mexico

Table 1: Sociodemographic distribution of the Mexico sample compared to population¹

| | P1 Unweighted | P2 Unweighted | P1 Weighted | P2 Weighted | Census |
|--------------|---------------|---------------|-------------|-------------|--------|
| Sex | | | | | |
| Male | 39.4 | 36.1 | 47.6 | 47.7 | 47.7 |
| Female | 60.6 | 63.9 | 52.4 | 52.3 | 52.3 |
| Age | | | | | |
| 18-24 | 21.4 | 24.3 | 24.5 | 25.4 | 24.7 |
| 25-29 | 14.2 | 13.6 | 15.5 | 15.9 | 15.6 |
| 30-34 | 14.0 | 12.2 | 13.6 | 12.5 | 13.6 |
| 35-39 | 13.7 | 13.1 | 12.3 | 11.5 | 12.1 |
| 40-44 | 11.0 | 10.3 | 10.1 | 10.6 | 9.9 |
| 45-49 | 8.0 | 8.3 | 7.7 | 7.8 | 7.8 |
| 50-54 | 6.5 | 6.9 | 6.3 | 6.3 | 6.4 |
| 55-59 | 4.8 | 5.6 | 4.7 | 4.9 | 4.9 |
| 60-65 | 6.4 | 5.7 | 5.3 | 5.2 | 5.1 |
| Region | | | | | |
| Metropolitan | 31.5 | 32.6 | 28.0 | 27.6 | 27.6 |
| Northwest | 9.8 | 10.9 | 7.9 | 8.0 | 8.0 |
| North | 14.5 | 13.7 | 15.3 | 15.1 | 15.1 |
| Central West | 12.7 | 13.4 | 12.3 | 12.4 | 12.4 |
| Central East | 15.1 | 14.9 | 17.1 | 17.5 | 17.5 |
| South East | 16.4 | 14.6 | 19.4 | 19.3 | 19.3 |

¹Presents only the sociodemographic variables used in post-stratification of weight.

Analysis of the CPES

- This section will focus on analysis of complex survey data, correct analysis of the CPES and provide a review of current software tools.

Analysis of Complex Sample Data

- Data originates from sample designs that include features such as non-response adjustments, clustering, stratification, and differing probabilities of selection
 - **CPES data is complex and variance estimation adjustments are required for proper analysis**
- Complex samples do not assume independence of observations, clustering and homogeneity are present
- Assuming a Simple Random Sample (SRS) generally results in underestimation of variance estimates due to effective loss of sample size due to clustering within strata

Design Variables

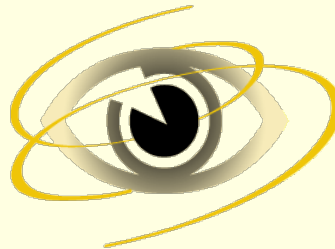
- Strata and Cluster variables are specified to represent the complex sample of the data, each data set has the appropriate design variables in the demographic dataset
- From SAS documentation:
 - The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample. The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata

CPES Design Variables

- Along with the CPES weights:
 - 2 key variables represent the complex sample design
 - SESTRAT
 - SECLUST
 - Understanding the distributions of the 2 variables in the CPES is very important and a first step is to examine distributions of the two variables

CPES Strata and Cluster Variables

- An analysis of the SESTRAT * SECLUSTER shows there are a total of 180 stratum and values of 1 or 2 per cluster
- Every case in the data set falls into 1 and only 1 of the SESTRAT*SECLUSTER cells, no missing data on these variables
- This demonstration shows the full distribution of the SESTRAT / SECU variables



Software for Analysis of Complex Sample Data

- Analysts should take the complex nature of the design into account by using one of the software packages that includes complex sample analysis routines
 - SAS survey procedures
 - Sudaan
 - Stata's svy procedures
 - Complex Samples module of SPSS
 - Other options are IVEware, R, Wesvar, and Mplus

Common Statistical Techniques for Complex Sample Variance Estimation

- The Taylor Series Linearization Approach is used by default in all SAS Survey Procs, Stata svy commands, SPSS CS routines, and in Sudaan (see SAS, Stata, SPSS, or Sudaan documentation for details on this technique)
- Other methods include Repeated Replication Techniques such as Jackknife Repeated Replication and Balanced Repeated Replication, these will not be demonstrated during this training session but are included in the references and can also be performed in SAS (v9.2 +), Sudaan, and STATA plus additional software not covered today such as WesVar, R, IVEware
- The next few slides show a grid of the major software with variance estimation methods, options included and analysis techniques available

Software for Survey Data Analysis

| | Stata | SAS | Sudaan | SPSS |
|---------------------------------|-------|------|--------|----------|
| Sample Design/Weights | | | | |
| Range of sample designs handled | Wide | Wide | Wide | Moderate |
| Use of replicate weights | Yes | Yes | Yes | No |
| Variance estimation | | | | |
| Taylor Series | Yes | Yes | Yes | Yes |
| JRR | Yes | Yes | Yes | No |
| BRR | Yes | Yes | Yes | No |
| Single Unit per Strata | Yes | No** | Yes | No |
| Subpopulations | | | | |
| Subpopulation analysis | Yes | Yes | Yes | Yes |

** Means that cannot be done without some programming effort.

Descriptive Techniques

| | Stata | SAS | Sudaan | SPSS |
|-----------------------|--------------|------------|---------------|-------------|
| | | | | |
| Means | Yes* | Yes* | Yes* | Yes* |
| Totals | Yes* | Yes* | Yes* | Yes* |
| Ratios | Yes* | Yes* | Yes* | Yes* |
| Percentiles | No | Yes* | Yes* | No |
| Contingency Tables | Yes* | Yes* | Yes* | Yes* |
| | Stata | SAS | Sudaan | SPSS |

* Means appropriate hypothesis test option available with complex sample adjustment.

Regression Techniques

| | Stata | SAS | Sudaan | SPSS |
|--------------------------|-------|------|--------|------|
| Regression | | | | |
| Linear | Yes* | Yes* | Yes* | Yes* |
| Binary Logistic | Yes* | Yes* | Yes* | Yes* |
| Orginal Logistic | Yes* | Yes* | Yes* | Yes* |
| Multinomial Logistic | Yes* | Yes* | Yes* | Yes* |
| Poisson Regression | Yes* | No | Yes* | No |
| Probit | Yes* | Yes* | No | Yes* |
| Survival Analysis | | | | |
| Cox PH Model | Yes* | No | Yes* | Yes* |

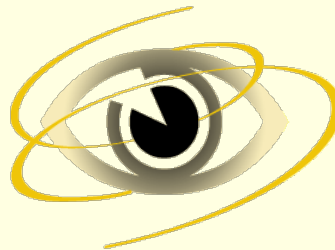
*Means appropriate hypothesis test option available with complex sample adjustment.

SAS v9.2

- SAS has full range of data management, analysis, graphing, reporting, complex design correction procedures and more
- Complex design procs available in v9.2
 - Proc surveymeans-means, univariates, ratios, totals, quantiles
 - Proc surveyfreq-frequency tables, 1 way and nway
 - Proc surveyreg-linear dependent variables,ANOVA
 - Proc surveylogistic-binary, ordinal, nominal logistic regression
 - Proc surveyselect-sampling procedure

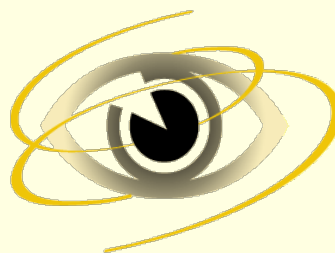
SPSS v16

- SPSS is another very good option for data management and basic analysis that also includes survey data analysis commands in the “Complex Samples” module
- Widely used around world, compatible with many users
- Extensive online tutorials, very easy to use with either the point and click approach or by using commands
- See SPSS documentation for command syntax and list of complex sample commands
 - Online demonstration of commands



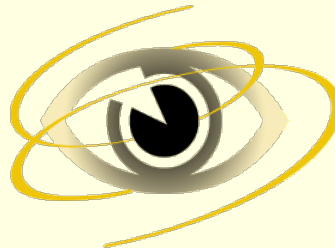
STATA 10

- Stata offers full range of data management and analysis tools including many survey procedures for complex design corrections, offers varied options for complex design corrections
- Most extensive and up-to-date software for survey data analysis
- Not as widely used as SAS or SPSS at this point
- See Stata documentation for complete list of svy procedures
 - Online demonstration



Sudaan v9.0

- Sudaan is another excellent tool for survey or complex sample data analysis
- It offers many analytic techniques with a wide range of sample designs are accepted
- Sudaan includes extensive hypothesis testing options appropriate for each analytic technique
- Can perform a full range of variance estimation methods
- See Sudaan documentation for list of procedures
 - Online demonstration of Sudaan procedures



Analysis Topics

- This section focuses on special topics that arise during analysis
 - Hardware
 - Data transfer
 - Data preparation
 - Multiple imputations
 - Subpopulations

Considerations for Analysis of the CPES

- What software and hardware tools are needed to carry out analysis?
 - Local machine
 - Remote server
 - SDA approach
- What part of the CPES instruments do the variables of interest come from?
 - This will have an impact on weights used if part 2 NCSR variables
- What weights should be used, full sample or subset?
- Has something very similar already been published by someone else? Is this a realistic topic?

Data Transfer Software

- DBMS Copy- allows easy and accurate movement of datafiles between all major packages such as SAS, SPSS, Excel, STATA
- StatTransfer-another good tool for moving data between software packages
- SAS built-in options- procs import and export, engine architecture allows reading of some types of external data sources, WIZARD enables point and click for data transfer
- SPSS- produces SAS files as output in SPSS16
- Stata-some options to read in files from other programs
- Sudaan-also includes options to read in data sets from other programs

Common Analysis Techniques

- Data analysis usually consists of 80-90% preparation for analysis and actual analysis phase is about 10-20% of task
- Of the 10-20% of analysis performed, about 80% of analysis would be covered by the following types of techniques:
- Descriptive analysis
 - Means/Univariates
 - Frequency tables
 - Graphs
 - Survival Curves
- Inferential analysis
 - Ordinary Least Squares
 - Logistic Regression with varying types of dependent variables (binary, ordinal, multinomial)
 - Survival analysis, mixed models/hierarchical models, Latent Class Analysis, and other more advanced methods represent a small portion of analyses

Standardized Approach to Analysis

- Standardized approach to analysis work
 - If working with others, all team members use same software product, eliminates inefficiencies and confusion
- Use coding rather than point and click, save programs for others to use, allows sharing of knowledge and programs
- Replication of results can be achieved with organized setup and coding
- Datasets stored in shared space and maintained by data manager
 - Shared computing resources, allows general sharing of programs and data, streamlined and less expensive licensing

Data Preparation

- Preparing data for analysis: missing data issues and common techniques for imputation: means, medians within subgroups, regression based imputation
- Recoding vars from 1 to 5 to 0/1, reversing scales, adding scales, arrays to process data iteratively, cleaning data by looking at outlier, wild codes, all missing right or wrong?
- Variable construction, example of complex variable such as time varying education or suicide ideation and onset derived from multiple questions

Missing or Inconsistent Data

- Check for structural missing versus missing due to refusal or don't know by examining skip patterns and section flow
- Examine key variables needed for analysis to identify problems to fix or impute
- Examples are missing age of onset or age of recency along with disorder diagnosis
- Logical inconsistencies such as age of onset of disorder is later than age of first treatment for same disorder
- Check distributions for all variables needed for analysis to check for outliers or other problems prior to analysis phase

Missing Data

■ Simple imputation

- Use of overall statistic for a common crossing of demographic variables such as age*gender*education group that person falls within: assign mean or median for crosstab group for imputation of personal income
- Imputation based on other variables that might give clues that will help assign a realistic imputed value, for example, one approach is to use age of first treatment or age of last episode to do best guess of age of onset if that is missing
- Etc. many other simple approaches
- Check carefully for checkpoint problems or skip pattern inconsistency that might indicate data collection problem
- Check actual interview or respondent comments from text file to see if any further information available

■ Multiple Imputation

- Use regression based imputation tool such as IVEware or SAS Proc MI, MICE, SOLAS etc. to impute values

Multiply Imputed Data Sets

- Most software packages can analyze multiply imputed data sets produced by the multiple imputation process
- Analysis of multiple data sets along with complex sample corrections can be done at once thus including variability introduced from imputation and variance corrections for complex sample surveys

| | Stata | SAS | Sudaan | SPSS |
|--|-------|-----|--------|------|
| Ability to Analyze Multiply Imputed Data | Yes | Yes | Yes | No |

Subpopulation Analysis

- Commonly done by most analysts interested in analysis within subpopulations such as gender/age groups etc.
- Issues arise in how the subpopulations are handled by each software
- Make sure to use the “domain” (SAS) or “subpopn” (Sudaan) or “subpop” (Stata) options when analyzing data from the CPES
- Why? This will ensure that you are not deleting strata/secu combinations incorrectly and will utilize the full degrees of freedom for significance tests
- Subsetting the data set rather than use of the subpopulation options often results in empty or singleton strata/secu cells and a loss of the original complex sample design representation

Subpopulation Analysis in SAS

- From SAS Documentation v9.2:
 - The DOMAIN statement of SAS requests analysis for subpopulations, or domains, in addition to analysis for the entire study population. The DOMAIN statement names the variables that identify domains, which are called domain variables.
 - It is common practice to compute statistics for domains. The formation of these domains may be unrelated to the sample design. Therefore, the sample sizes for the domains are random variables. In order to incorporate this variability into the variance estimation, you should use a DOMAIN statement.
 - Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently (from SAS v9.2 documentation)

Sudaan/SPSS/Stata Subpopulation Analysis

- Sudaan software offers domain type analysis for every procedure via the “subpopn” statement
- Use of a “subpopn” statement will allow Sudaan to subset the data for analysis but include all design information in the analysis, similar to the domain analysis in SAS
- Results from Sudaan with a subpopn statement will use the full degrees of freedom for the entire sample even though the actual analysis may not use all records, why? Because those people theoretically could exist in each cell of the strata*secu grid but don’t happen to be in this particular sample
- Stata 10 also offers a subpopulation analysis for each of the “svy” procedures (subpop), see the Stata documentation for help
- SPSS includes an optional subpopulation feature for every command

Summary of Analysis of Complex Survey Data

- CPES data is complex due to each survey including stratification, clustering and sampling techniques which are simple random samples
- Special techniques should be used to incorporate these features of the designs
- SAS, Stata, SPSS, and Sudaan each offer very good complex sample survey data analysis features to account for the complex sample design
- CPES users should make full use of the weights, strata, and cluster/SECU variables when analyzing the data set

Survival Analysis with CPES

- This section focuses on using survival techniques for analysis of the CPES data set

Key Elements for Survival Analysis

- Timing of events is studied in event history analysis, analysis of timing of when and if events occur
- Predictors can be either time-varying (marital status, education) or time-invariant (race)
- Censoring is another important concept for event history analysis, censoring occurs when follow-up of individual ends and we can no longer determine whether or not event of interest occurs, such cases that do not yet have the outcome of interest are called censored
- Right censoring occurs naturally in our surveys, censored at time of interview

Discrete-Time and Continuous Time

- Continuous time is measured as a positive, continuous variable
- Discrete-Time is appropriate for situations in which events can only occur at regular point in time: presidential elections every four years, yearly medical examination, yearly interview
- Discrete-time can also handle situations where an event can occur at any point in time yet data is available or collected at a certain discrete point in time, our surveys typically deal with data of this type that asks only if a marital status change occurred during this year or did you have onset of a disorder in this year
- The presentation demonstrates use of a
 - descriptive technique for survival analysis - Lifetable survival/failure curves using SAS PROC LIFETEST
 - Model based approach for survival analysis – discrete time logistic regression using SAS PROC SURVEYLOGISTIC

Data Structure

- For survival analysis you need a few key variables
 - Time at which event of interest occurred and time at which last observed if event did not occur, example is age of onset of a given disorder or age at interview if no disorder
 - Status of event occurring or yes/no type variable, for example dummy variable for disorder 0=no, 1=yes
 - With time and status or age at onset or age at interview and yes/no for disorder we now have the key variables to use in survival curve analysis

Descriptive Survival Curve

- A common approach is to use an age of onset for a given disorder or group of disorders and graph the cumulative percentages
- Key variables
 - yes/no indicator of having the disorder of interest
 - age at onset of the disorder
 - age censored / CPES might use age of interview
- Data is structured with 1 record per person and SAS PROC LIFETEST is used to examine the survival/failure distribution
- See SAS documentation on PROC LIFETEST for details about the statistical approaches used
- Note that PROC LIFETEST does not include a complex sample correction, still has value as descriptive approach
- Sudaan and SPSS do offer a survey data Kaplan-Meier option for corrected significance testing for descriptive survival analysis

Sample data for survival curve

- Sample records with key variables detailed:

| Sampleid | dsm_mde | mde_ond | age |
|----------|---------|---------|-----|
| 1 | 0 | 0 | 50 |
| 2 | 1 | 20 | 44 |
| etc. | | | |

- Person number 1 would be followed from years 1 to 50 and would never have a “yes” on the outcome of dsm_mde
 - Person number 2 would be followed from years 1 to 20 (age of onset of dsm_mde) and has a “yes” on dsm_mde as well as an age of onset for mde
-
- Note that every person will be included in this type of analysis since we have year 1 and age at interview for every respondent

Survival Curve: Major Depressive Disorder

- Outcome variable is v07877 (major depressive disorder)
- Age of onset of disorder is v08771 (mde_ond)
- Age at interview is used as the censoring variable if no disorder present
- Prior to producing survival curve, examine data and means to obtain “picture” of data

Overall Prevalence for Major Depressive Disorder and Age of Onset Distribution

| Variable | Mean | N | NMiss | Std Dev |
|----------|------------|-------|-------|------------|
| dsm_mde | 0.1873074 | 16423 | 0 | 43.9585612 |
| V08771 | 25.9979886 | 3208 | 13215 | 1534.05 |

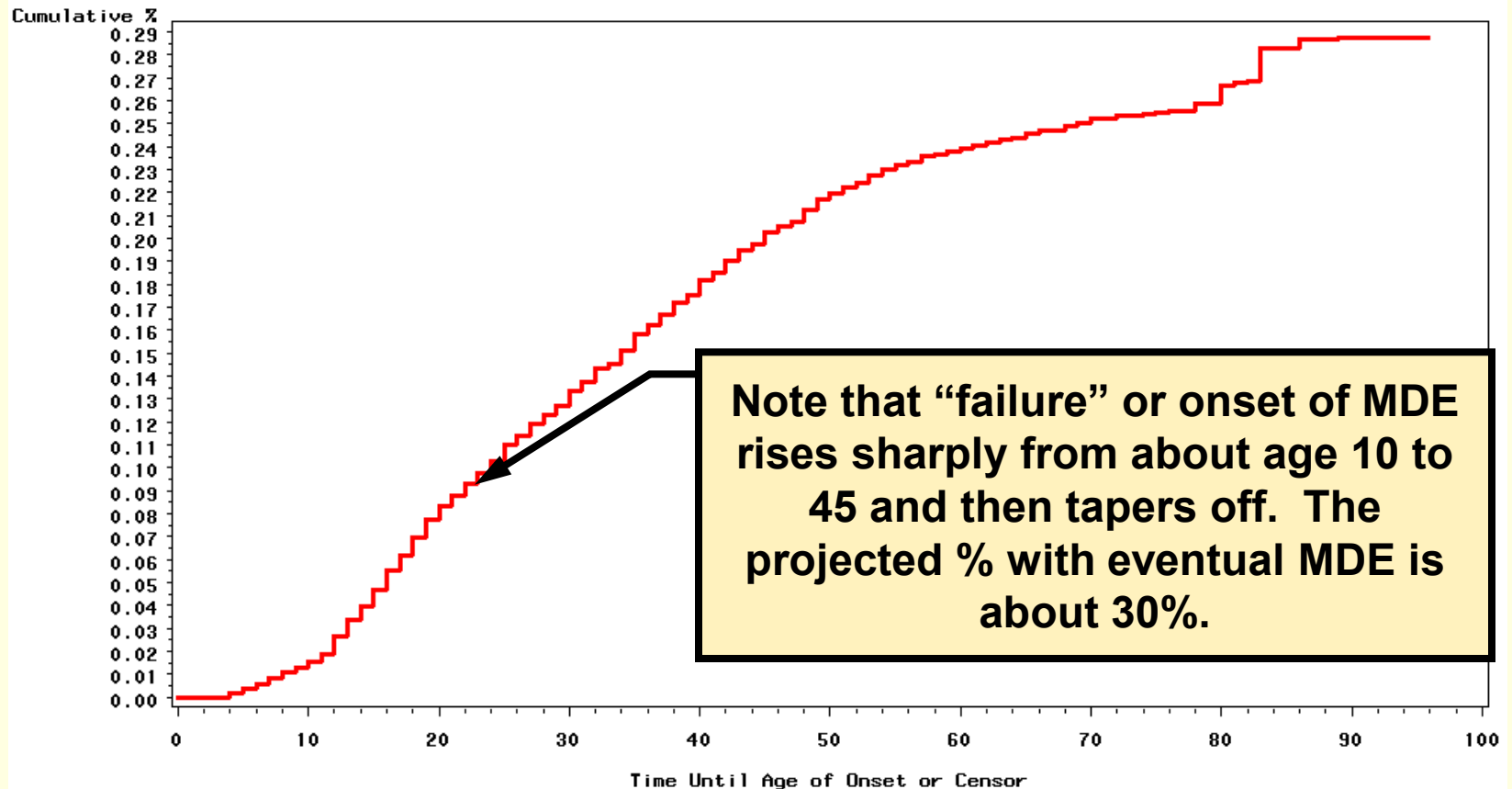
```
* prevalence of mde and age of onset ;  
proc means mean n nmiss stddev ;  
var dsm_mde v08771 ;  
weight cpeswtlg ;  
run ;
```


Survival and Failure Calculations

- The Lifetable approach assumes that any censored cases will be censored in the midpoint of the time interval
 - This approach is appropriate for situations with a large number of observations and many unique event times
 - The probability of surviving to $t(i)$ or beyond is calculated as a series of cumulative probabilities
 - For example for t_4 :
 - a=survival to t_2 or beyond
 - b=survival to t_3 or beyond
 - c=survival to t_4 or beyond
- So, $prC = pr(A, B, C)$ or $pr(C) = (1-q_3)(1-q_2)(1-q_1)$
- Y axis is then cumulative % rather than regular % of a fixed denominator

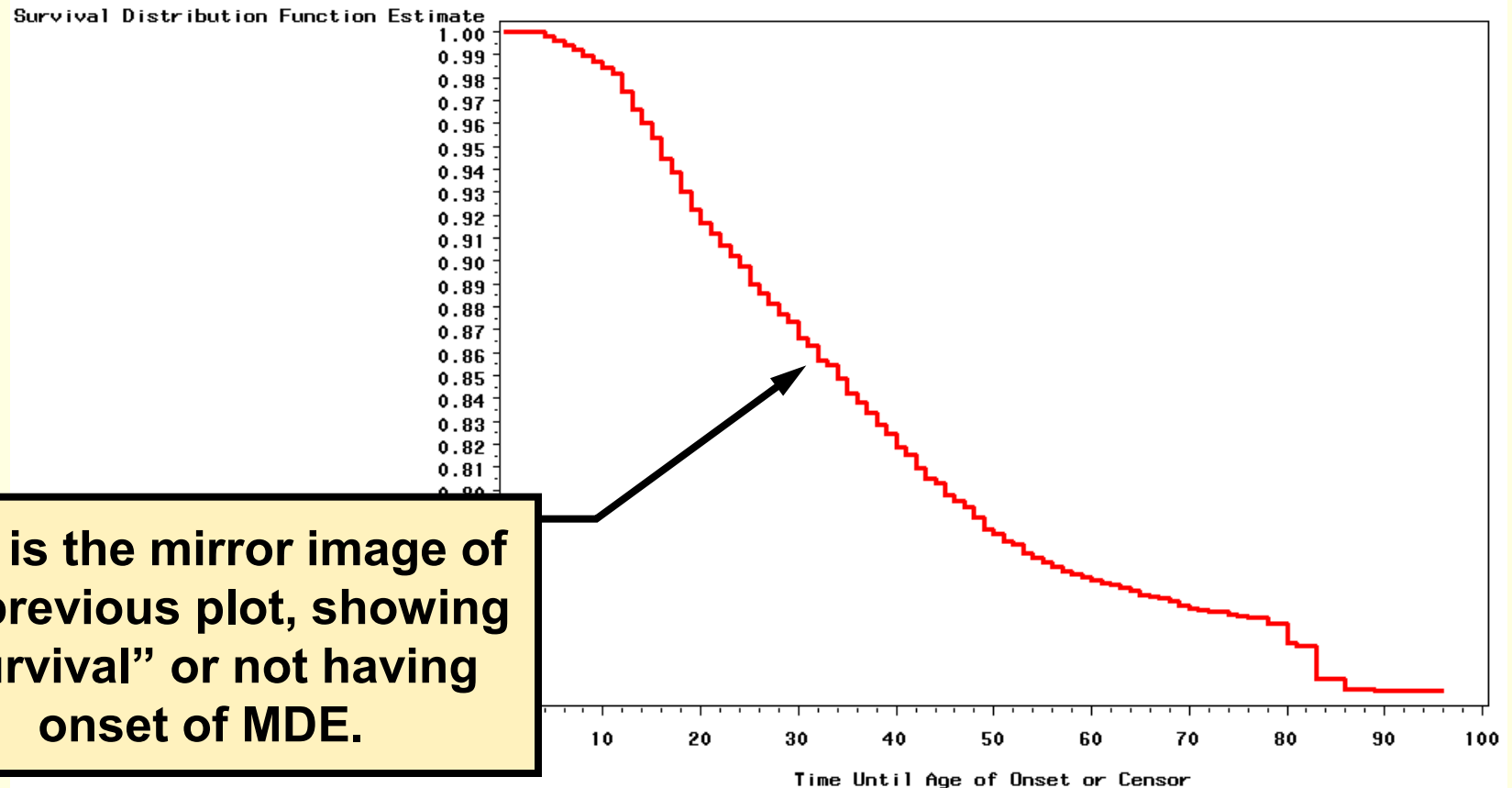
Failure Curve for MDE

Failure Curve for Age of Onset of Major Depressive Episode



Survival Curve for MDE

Survival Curve for Age of Onset of Major Depressive Episode



SAS Code for Data Preparation

```
*recode dsm_mde* ;  
if v07877 ne 1 then dsm_mde=0 ; else dsm_mde=1 ;  
age=v07306 ;  
*create age at onset or age at censor ;  
*note that each person in file receives an age: either onset or age at  
interview ;  
if dsm_mde=1 then ageevent=v08771 ; else ageevent=age ;  
  
*multiply weight by 100 for proc lifetest freq statement ;  
finalp1w100=cpeswtlg*100 ;
```

SAS Code for Survival Curve

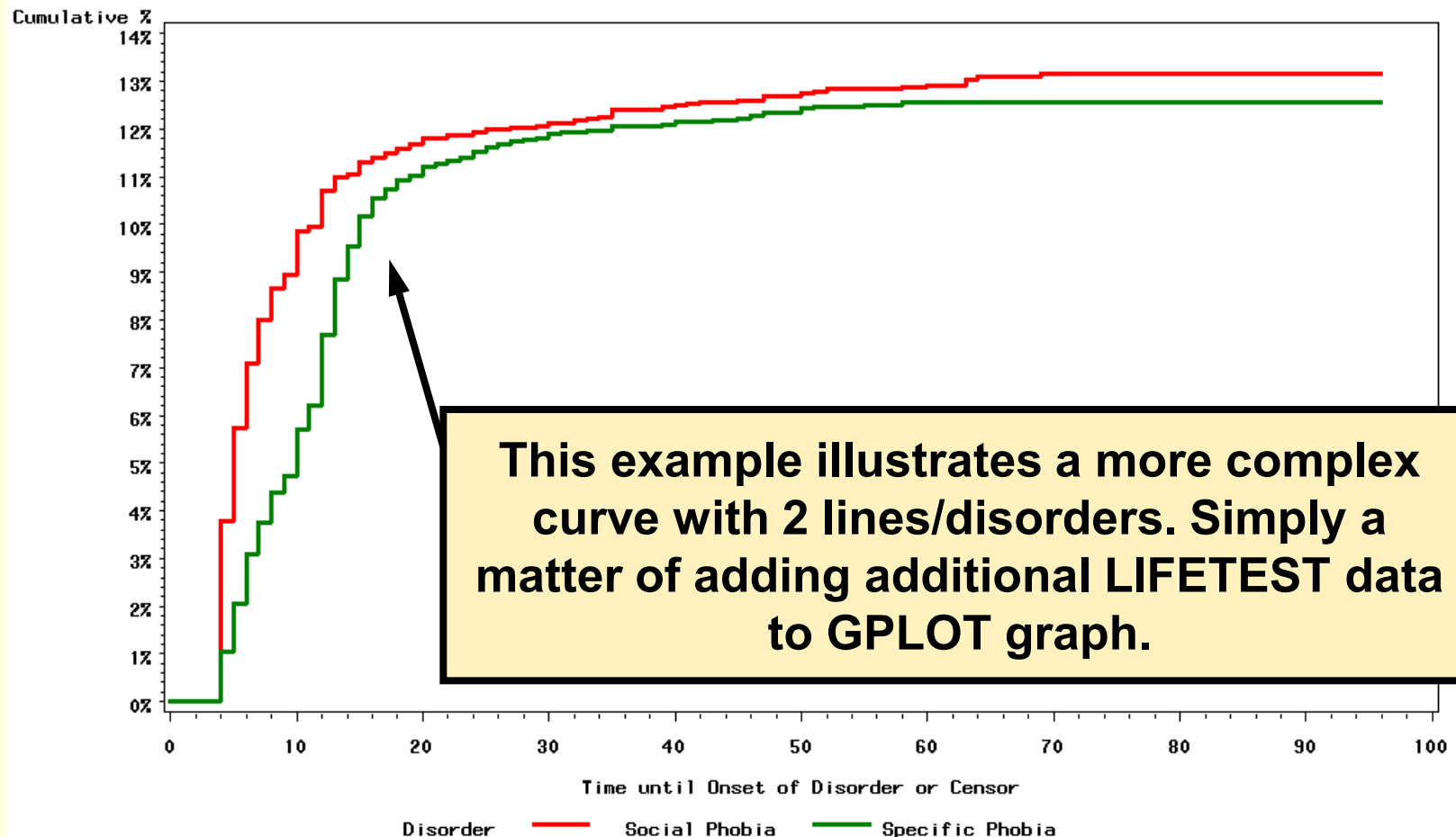
```
proc lifetest data=one
    method=lt
    intervals=(1 to 96 by 1)
    plots=(s)
    outs=out (keep=ageevent survival) ;
time ageevent * dsm_mde (0) ;
freq finalp1w100 ;
run ;

data survival ;
    set out ;
failure=1-survival ;
label failure="Cumulative %" ageevent="Time Until Age of Onset or Censor" ;
proc print ;
run ;

symbol c=red i=steprj w=3 ;
proc gplot ;
plot failure*ageevent ;
Plot survival*ageevent;
format fail percent10. ;
run ;
```

Survival Curves for Two Disorders

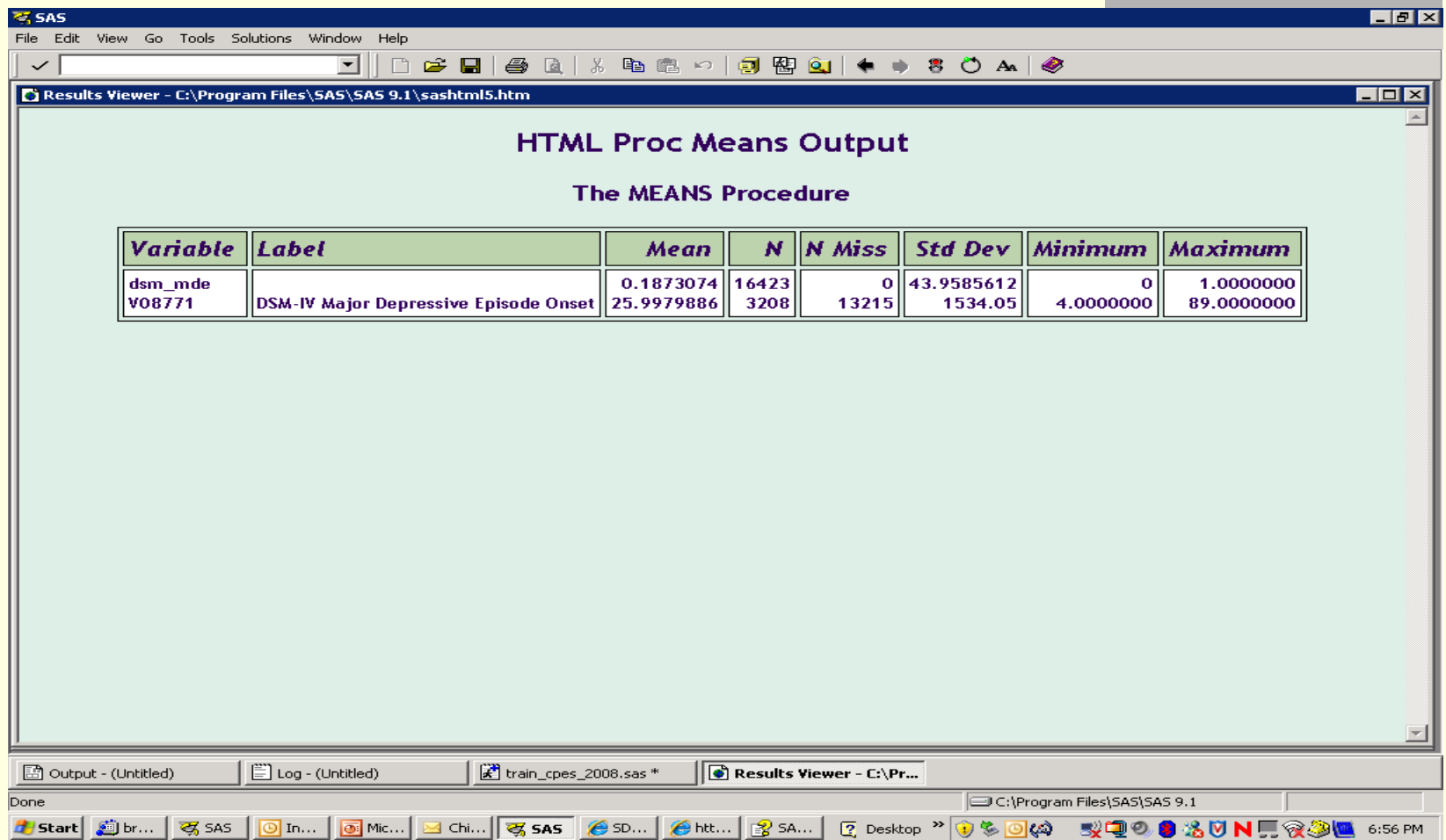
Survival Curves for Social and Specific Phobias



Using SAS ODS to Transfer Analysis Output to External Software

- The Output Delivery System of SAS allows various output delivery destinations such as various files types (HTML, PDF, RTF) as well as output datasets for each part of the procedure output
- Use of ODS can make moving analysis output into software of choice automated and error-free
- SAS graphing and reporting tools are fully capable of all types of reports but many journals request other formats such as Word or PDF files
- ODS offers a number of methods for moving tabular and graphical output into files of a type that Excel can read, HTML, tagsets.msoffice2k, GIF/BMP, and other files types

Example of HTML Output from Output Delivery System



HTML Proc Means Output

The MEANS Procedure

| Variable | Label | Mean | N | N Miss | Std Dev | Minimum | Maximum |
|----------|---------------------------------------|------------|-------|--------|------------|-----------|------------|
| dsm_mde | | 0.1873074 | 16423 | 0 | 43.9585612 | 0 | 1.0000000 |
| V08771 | DSM-IV Major Depressive Episode Onset | 25.9979886 | 3208 | 13215 | 1534.05 | 4.0000000 | 89.0000000 |

Preparation for Discrete-Time Logistic Regression

- Thus far, we have used a person level data set consisting of 1 record per person but for logistic regression using a discrete time approach, use a multiple record per person approach that reflects the years a person is at risk for outcome of interest
- Create survival dataset using an “output” statement in SAS turns person-level file into person-year file or equivalent multiple record per individual type file
- Create time-varying covariates and dependent variables as well as person-years or “ints”
- Check printouts of data to know exactly what is happening with coding, make no assumptions about coding without examination of data

Person-Year Dataset

- Create a person-year or person-day or some type of person-unit of analysis dataset from a person-level dataset, organizes records for correct analysis of timing of event of interest
- This is easily done using a “do” loop with an output statement in SAS
- For example, we use person-year as our unit of analysis and expand our dataset to a 1 record per person to multiple records per individual, number of records depends on the variables in the do loop

Creating a Multiple Record File from a Single Record File

```
data personyear ;  
set two ;  
do int = 1 to age ;  
output ;  
end ;
```

This code creates an output data set for every person with records ranging from 1 to age at interview (v07306). Use of the output statement forces an explicit output of each record between 1 and age at interview thus creating multiple records per person.

Distribution of Person Years

Partial Output of “int” variable

Distribution of Ints for NCS-R Person Year File

| int | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|-------------------------|-----------------------|
| 1 | 9282 | 2.24 | 9282 | 2.24 |
| 2 | 9282 | 2.24 | 18564 | 4.47 |
| 3 | 9282 | 2.24 | 27846 | 6.71 |
| 4 | 9282 | 2.24 | 37128 | 8.94 |
| 5 | 9282 | 2.24 | 46410 | 11.18 |
| 6 | 9282 | 2.24 | 55692 | 13.41 |
| 7 | 9282 | 2.24 | 64974 | 15.65 |
| 8 | 9282 | 2.24 | 74256 | 17.88 |
| 9 | 9282 | 2.24 | 83538 | 20.12 |
| 10 | 9282 | 2.24 | 92820 | 22.35 |

And so on.....

Outcome of Interest

- Preparation includes creation of time-varying variables as needed for the analysis
- We want a time varying outcome which indicates the year of life of onset of MDE
 - Why? Given the structure of the data set with multiple records per person, the interest changes from an outcome of MDE at any point in time to MDE in the year that onset occurred
 - To capture this information use an indicator of MDE onset in the year it occurred

Outcome of Interest

- Create outcome for model: onset of Major Depressive Episode
- The outcome variable is set to yes or 1 only in the year of onset of substance use with all other person years set to 0

```
data personyear ;  
    set personyear ;
```

```
*time varying mde* ;  
mde_ond=v08771 ;  
region=v08992 ;
```

```
if 4<=mde_ond<=89 and int=mde_ond then mdeonset=1 ;  
    else mdeonset=0 ;  
if dsm_mde=1 then ageevent=mde_ond ;  
    else ageevent=age ;
```

Sample Person Year Records

- For example, this person has 38 records so age at interview was 38 but onset of MDE at age 4, indicated by the 1 in the “mdonset” variable followed by all zeros for the rest of the data array
- Note that once a person has the event of interest they are no longer at risk
 - These records will not be used in predicting time to onset of MDE in the logistic model framework

| Obs | CASEID | int | MDE_OND | mdeonset |
|-------|--------|-----|---------|----------|
| 38023 | 848 | 1 | 4 | 0 |
| 38024 | 848 | 2 | 4 | 0 |
| 38025 | 848 | 3 | 4 | 0 |
| 38026 | 848 | 4 | 4 | 1 |
| 38027 | 848 | 5 | 4 | 0 |
| 38028 | 848 | 6 | 4 | 0 |

More records follow...

Discrete Time Logistic Regression

- The model will use the following predictors
 - Sex (sexf)
 - Age at interview in 4 categories
 - Region in 4 categories
 - Note that each of the predictors do not vary with time although you can use time-varying predictors as well in this type of model (time varying education or marital status)
- A survival model should include a control variable representing year of life or actual records in the data set
- This was developed in the output “do loop” as the “int” variable (continuous in this example but could be categorical)

Defining Years at Risk

- Ints can be used as either single year dummy variables or as collapsed variables
- Other issues are years of risk and what particular years of risk should be included in the models
- In the case of modeling the outcome of MDE
 - use the years from age 1 of life to year of event occurring (age of onset of MDE) or year censored (age of interview)
 - note that once the outcome occurs that person is no longer at “risk” and person years after that point in time are no longer included in the analysis

Discrete-Time Logistic Regression

- Here is a simple example of a discrete time logistic regression using PROC SURVEYLOGISTIC
- “where” statement selects person years from age 1 to event (onset of MDE)/age at censor (age at interview)
 - Excludes person years after the event of interest has occurred or uses all person years until censor (age at interview)
- Also note that the “ints” or years of life are included as a categorical predictor, need these to represent time units in model

Data Preparation and Variable Creation

```
data personyear ;  
    set personyear ;  
*time varying mde* ;  
mde_ond=v08771 ; region=v08992 ;  
if 4<=mde_ond<=89 and int=mde_ond then mdeonset=1 ; else mdeonset=0 ;  
if v09036=2 then sexf=1 ; else sexf=0 ;  
if dsm_mde=1 then ageevent=mde_ond ; else ageevent=age ;  
agecat=. ;  
if 18<=age<=29 then agecat=1 ;  
else if 30<=age<=45 then agecat=2 ;  
else if 46<=age<=59 then agecat=3 ;  
else agecat=4 ;  
if 1<=int<=12 then intcat=1 ;  
else if 13<=int<=19 then intcat=2 ;  
else if 20<=int<=29 then intcat=3 ;  
else if 30<=int<=39 then intcat=4 ;  
else intcat=5 ;
```

SURVEYLOGISTIC Code

```
proc surveylogistic ;  
title "Discrete Time Logistic Regression Example" ;  
strata   sestrat ;  
cluster  seclustr ;  
weight   cpeswtlg ;  
class    intcat agecat region / param=reference ;  
model    mdeonset (event='1') = intcat agecat sexf region ;  
where     int <= ageevent ;  
format    agecat agecf. intcat intf. region regf. ;  
run ;
```

Note use of class statement for categorical predictors and use of where statement to restrict person years up to and including year of onset of event.

Partial SURVEYLOGISTIC Results

The SURVEYLOGISTIC Procedure
Analysis of Maximum Likelihood Estimates

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-----------|-------|----|----------|----------------|-----------------|------------|
| Intercept | | 1 | -6.4277 | 0.1026 | 3925.9115 | <.0001 |
| intcat | 1-12 | 1 | -1.4130 | 0.1075 | 172.8571 | <.0001 |
| intcat | 13-19 | 1 | -0.1437 | 0.1137 | 1.5982 | 0.2062 |
| intcat | 20-29 | 1 | -0.3422 | 0.0957 | 12.7865 | 0.0003 |
| intcat | 30-39 | 1 | -0.0781 | 0.0871 | 0.8045 | 0.3698 |
| agecat | 18-29 | 1 | 2.1023 | 0.1035 | 412.6790 | <.0001 |
| agecat | 30-45 | 1 | 1.6363 | 0.1031 | 252.1036 | <.0001 |
| agecat | 46-59 | 1 | 1.1157 | 0.0994 | 126.0877 | <.0001 |
| sexf | | 1 | 0.4627 | 0.0566 | 66.7564 | <.0001 |
| region | MW | 1 | -0.0148 | 0.0807 | 0.0339 | 0.8540 |
| region | NE | 1 | 0.0767 | 0.1204 | 0.4060 | 0.5240 |
| region | S | 1 | -0.1418 | 0.0719 | 3.8864 | 0.0487 |

Results, continued

Odds Ratio Estimates

Point Estimates/OR's indicate younger age groups are significantly more like to have an onset of MDE as compared to those 60+. Women are 1.6 times more likely to have onset of MDE than men, and none of the regions are significant.

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---------------------|----------------|----------------------------|--------|
| intcat 1-12 vs 40+ | 0.243 | 0.197 | 0.300 |
| intcat 13-19 vs 40+ | 0.866 | 0.693 | 1.082 |
| intcat 20-29 vs 40+ | 0.710 | 0.589 | 0.857 |
| intcat 30-39 vs 40+ | 0.925 | 0.780 | 1.097 |
| agecat 18-29 vs 60+ | 8.185 | 6.682 | 10.025 |
| agecat 30-45 vs 60+ | 5.136 | 4.197 | 6.286 |
| agecat 46-59 vs 60+ | 3.052 | 2.512 | 3.708 |
| sexf | 1.588 | 1.421 | 1.775 |
| region MW vs W | 0.985 | 0.841 | 1.154 |
| region NE vs W | 1.080 | 0.853 | 1.367 |
| region S vs W | 0.868 | 0.754 | 0.999 |

Summary of Logistic Regression

- The simple example presented illustrates how to set up a dataset for a person-year format, create time-varying variables as both outcome and predictors, and specify the model correctly
- These concepts can obviously be extended to include interactions, linear contrasts and subpopulation analyses
- Time does not permit extensive examples in this area but many of the CPES publications includes analysis techniques of this type
- Discussion/questions

References

References for data preparation and descriptive analysis

Cody, R.P. and Smith, J.K., “Applied Statistics and the SAS Programming Language”, Cary, NC: SAS Institute Inc.

Delwiche, L. and Slaughter, S., “The Little SAS Book, a Primer”, Cary, NC: SAS Institute Inc.
The SAS Language Guide, Reference, SAS Institute Publishing

References for Analysis of Categorical data:

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Stokes, M.E., Davis, C.S., and Koch, G.G. (2000), *Categorical Data Analysis Using the SAS System, Second Edition*, Cary, NC: SAS Institute Inc.

Reference for Linear Regression:

Freund, R.J. and Littell, R.C. (1986), *SAS System for Regression, 1986 Edition*, Cary, NC: SAS Institute Inc

References

References for Survival Analysis

Allison, P.D., Survival Analysis using the SAS System, A Practical Guide, Cary, NC: SAS Institute Inc

Allison, P.D. 1982 "Discrete-time methods for the analysis of event histories." Pp. 61-98 in Samuel Leinhardt (ed.), *Sociological Methodology 1982*. San Francisco: Jossey-Bass

Bradley Efron, Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve Journal, J Amer Stat Assn, Volume 83, pages 414-425 1988

References for analysis of complex survey data

Heeringa, S., and Liu, J. (1997), Complex sample design effects and inference for mental health survey data, *International Journal of Methods in Psychiatric Research*, 7, Whurr Publishers Ltd. – Pages 221 – 230.

Rust, K., (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics*, 1 (4), Statistics Sweden Publishing Service. Pages 381 –397

Landis, J., Lepkowski, J., Eklund, S., and Stehouwer, S., (1984) A Statistical Methodology for Analyzing Data from a Complex Survey: The first National Health and Nutrition Examination Survey, *Vital Health Statistics*, 21 (10), MD: U.S. National Center for Health Statistics.

Little, R.J.A., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R.C., (1997), Assessment of Weighting Methodology for the National Comorbidity Survey, *American Journal of Epidemiology*, 146 (5), –Pages 439 – 449.



NIMH Collaborative Psychiatric Epidemiology Surveys

*Examples of Complex Design-
Corrected Analyses Using the CPES*

Myriam Torres
NSAL Analyst



Outline

- Demonstrations using 3 software packages
- Specifying weights and complex design
- Crosstabulations
- Linear Regressions
- Means and testing differences in means
- Logistic Regressions
- Means adjusted for Covariates



Specifying weights and complex design – SPSS (see Page 1)

- SPSS version 14, 15 or 16 should be used.
- Need to create a Complex Sample Plan
- **Ex:** Using all cases, and all variables in analyses are in NCS-R's Short form (Part 1).
 - ❖ Analyze → Complex Samples → Prepare for Analysis
 - ❖ Create a plan file: Browse to the location where you want to save your plan, and give it a name. Next.
 - ❖ Strata = *sestrat*, Clusters = *seclustr*, Sample Weight = *CPESWTSH*.
 - ❖ Finish.
- **Recommendation:** Set up as many CS Plans you need using the weight name to name the CSPlan to avoid confusion.



Specifying weights and complex design - STATA

- Stata version 9.2 or higher should be used.
- **Ex:** When using all cases and all variables in analyses are in NCS-R's Short form.
- Run this command at the start of your Stata session:

```
svyset seclustr [pweight = CPESWTSH], strata(sestrat) || _n
```
- **Recommendation:**
 - ❖ Have as many svyset commands in your do-file as needed depending on the appropriate weights you will use.
 - ❖ For each analysis run, be sure to check if you need to change the weight!!!



Specifying weights and complex design - SAS

- SAS 9.1.3 or higher is recommended. Chi-sqr degrees of freedom are calculated correctly in these versions.
- Include these lines in each of your SAS PROC survey commands:
 - ❖ Weight *weightname*; (replace *weightname* with the appropriate weight for each analyses)
 - ❖ Strata *sestrat*;
 - ❖ Cluster *seclustr*;



Survey Commands: Cross-tabulations (see Pages 2-9)

- Complex design commands very straightforward.
- Standard errors (se's) and Confidence Intervals (CI's) are corrected for sample design. Rates are weighted.
- Pearson Adj F (SPSS), Design-based F (Stata), and F from Rao-Scott Chi-sqr (*chisq* option) (SAS) provide comparable design corrected tests
- SPSS and Stata calculate degrees of freedom in same manner (not whole #'s). SAS uses whole #'s.



Survey Commands: Linear Regressions (see Pages 10-13)

- Complex design commands very straightforward.
- SPSS: Use CSGLM to run linear regressions.
- Stata: Use SVY: REGRESS
- SAS: Use PROC SURVEYREG
- Each software package has different methods for assigning the contrast levels for each categorical predictor.



Survey Commands: Means, T-tests & ANOVAS (see Pages 14-21)

- Identical means, se's, and CI's across all three softwares.
- For tests of differences in means:
 - SPSS: Use CSGLM command, changing the contrast as needed. The overall Wald F provided is equivalent to an ANOVA F-statistic.
 - Stata: Use *test* commands right after the *svy: mean* with *over(subpopulation var)* command. In this manner, the tests will be using design-corrected s.e.'s.
 - SAS: Use PROC SURVEYREG where the only predictor is the categorical subpopulation variable, changing the contrast as needed.



Survey Commands: Logistic Regressions (see Pages 22-29)

- Complex design commands very straightforward.
- Be aware that diagnostics are coded 1,5. It affects the interpretation of results.
- SPSS: Be sure to correctly assign the reference category for the dependent variable. (see handout).
- Stata: Dependent variable must be 0-1. Use *char* commands to assign reference category for the categorical predictors.
- SAS: To assign the “Event” value (D.V.), either specify the value of the event category in quotes, use the default (EVENT=FIRST), or EVENT=LAST.



Survey Commands: Other Regression commands

- SPSS: Version 15 and 16 have CS Ordinal Regression command.
- Stata has various commands besides svy: logit: ologit, mlogit, probit, oprobit.
- SAS PROC SURVEYLOGISTIC allows you to run logit and probit.



Survey Commands: Means adjusted for Covariates (see Pages 30-37)

- SPSS: Use CSGLM and follow similar steps as tests of differences in means.
- Stata: Use *adjust* command right after the *svy: mean* with *over(subpopulation var)* command followed by the *test* command.
- SAS: Use PROC GLM to get adjusted means followed by a PROC SURVEYREG.
- Both Stata and SAS require several additional steps to obtain desired results