# Tobacco Use Supplement to the Current Population Survey

User's Workshop
June 9, 2009

Tips and Tricks Analyzing TUS-CPS Data

Lloyd Hicks
Westat

# Talk Outline

1. Uses of Standard Errors in Analyzing Data
2. Methods to Compute Standard Errors for TUS-CPS estimates
   - – Generalized variance functions (SE parameters)
   - – BRR replication – Fay's method (replicate weights)
3. Special Topics for Analysts
   - – Change in Race/Ethnicity Questions
   - – 2002-03 Overlap Sample
   - – Replicate Weights when Data Sets Merged

# Uses of Standard Errors

- Constructing confidence intervals
  - reflects the accuracy of survey estimates
- Hypothesis testing
  - compare estimates between subgroups (within same year)
  - compare estimates across time

# Uses of SEs: Confidence Intervals

Formula: $\hat{X} \pm z \times SE(\hat{X})$

- X = estimate
- SE(X) = standard error
- z = confidence interval coefficient (e.g. 1.645 for 90% CI)

Example: 90% CI for males 18+ smokers (20%)

$20\% \pm 1.645 \times 0.15\% = 20\% \pm 0.25\%$

$= [19.75\%, 20.25\%]$

# Uses of SEs: Hypothesis Testing

Formula:

$$\frac{(\hat{X} - \hat{Y})}{SE(\hat{X} - \hat{Y})} > z \Longrightarrow \text{ statistical significance}$$

- X is the estimate for the 1st group
- Y is the estimate for the 2nd group
- SE(X – Y) is the standard error of difference
- z = critical value threshold

# Hypothesis Testing: Example 1

| | P | SE (P) | |
|---|---|---|---|
| group 1 | 21% | 0.15% | |
| group 2 | 20% | 0.15% | t-stat |
| diff | 1% | 0.212% | 4.71 |

Note: difference is statistically significant

(since 4.71 is greater than z where z =1.645 at 90% confidence level)

# Hypothesis Testing – Example 2

| | P | SE (P) | |
|---|---|---|---|
| group 1 | 25% | 2.50% | |
| group 2 | 20% | 2.00% | t-stat |
| diff | 5% | 3.202% | 1.56 |

Note: difference is not statistically significant

(since 1.56 is less than z  where z =1.645 at 90% confidence level)

# Methods of Estimating Standard Errors for TUS-CPS

1. Generalized variance functions (GVF)

   – Fast, easy but only approximate
   – More practical for large number of survey items
   – Requires a and b parameters from source and accuracy statements
   – Standard errors formulas for means, totals, percentages and their differences
   – Standard errors for complex estimates not possible (e.g. regression)

# GVF Example

Standard error for a percentage

$$S_{x,p} = \sqrt{\frac{b}{x} p(100 - p)}$$

- p is the estimate of the percentage
- x is the estimate of the base of the percentage
- b is the b parameter obtained from S&A statement

# GVF Example

P = percentage of male smokers 18+ = 20.7%

X = 101,244,000

b parameter = 1,575 (from S&A table)

$$S_{x,p} = \sqrt{\frac{1{,}575}{101{,}244{,}000} \times 20.7 \times (100 - 20.7)} = 0.16\%$$

Note: Data from 2003 TUS-CPS

# Methods of Estimating Standard Errors for TUS-CPS

2. Balanced repeated replication (BRR) based on replication weights

   – Replicate weights not on TUS-CPS public use file (available from NCI on request)
   – Requires special software (Sudaan, WesVar, etc.)
   – Provides a more accurate standard error than GVF
   – Standard errors for medians and other quantiles can be problematic

# SE Formula for CPS-TUS Using BRR (Fay's Method)

$$SE(\hat{X}) = \sqrt{\frac{4}{R}\sum_{r=1}^{R}\left(\hat{X}_{(r)} - \hat{X}_{(0)}\right)^2}$$

X(r) = replicate estimate

X(0) = full sample estimate

R = number of replicates

48 for 1992 – 1993 files (1980 decennial based samples)

80 for 1995 – 2003 files (1990 decennial based samples)

160 for 2006 – 2007 files (2000 decennial based samples)

4 = Fay Adjustment Factor (required in Sudaan)

# Special Topics for Analysts

1. Changes in Race/Ethnicity Data
2. 2002/2003 Overlap Sample
3. Merging Data Sets

# Special Topics 1: Changes to CPS Race/ethnicity data starting in 2003

- Respondents can now select more than one race when answering the survey.
- Asian or Pacific Islander (API) category split:

  1. Asian
  2. Native Hawaiian or Other Pacific Islander (NHOPI)

- The ethnicity question asked directly whether the respondent was Hispanic
- Ordering of race and ethnicity reversed

# Implication of Race/ethnicity Change On TUS-CPS data

1. No effect on estimates and trends for entire nation
2. Potential impact on estimates and trends by race/ethnicity

# Issues when Analyzing TUS-CPS Data By Race/ethnicity

1. Can't use race data for post-2003 data in same manner as pre-2003
   - Use single race = "only" category
   - Use "any mention" category
   - Neither group same as pre-2003 group
2. Analyzing Trends for single race groups spanning pre-2003 and post-2003
   - NCI developed "race bridge" approach to construct single-race estimates for post-2003 data

# TUS-CPS Race bridging approach

- NCI developed model to predict pre-2003 race/ethnicities given post-2003 value (using May 2002 CPS data supplied by Census)
- Paper summarizing the approach on website (http:/riskfactor.cancer.gov/studies/tus-cps/race bridging 071307.pdf).
- Paper summarizing application of approach on TUS-CPS data on website (http://www.fcsm.gov/07papers/Davis.VII-C.pdf)

# Special Topic 2: 2002/2003 Overlap Sample (for Limited Longitudinal Analysis)

- Persons in overlap sample (respondents in both)
  - TUS-CPS in Feb. 2002
  - TUSCS-CPS in Feb. 2003
  - Approximately 22,000 in overlap sample
- Responses from both studies can be analyzed as a longitudinal study
- New weights were developed for overlap sample

# Development of Overlap Sample Weights

- New weights for the overlap sample developed from 2003 weights
- New weights were derived to reflect 2003 population for gender, race/ethnicity, age, and geography
- Overlap sample weight

  $w^* = r * w$

  Overlap weights = (adjustment factor) * (2003 weights)
- Full sample and replicate weights using same approach

# Overlap Sample Weights: Derivation of Adjustment factor

- Choose adjustment factor so that sums of overlap sample weights match sums of 2003 sample weights in groups defined by
  - Census region (4)
  - Gender (2)
  - Race/ethnicity (4)
  - Age categories (19)
- Details in http://riskfactor.cancer.gov/studies/tus-cps/TUS-CPS_overlap.pdf

# Special Topic 3: Replicate Weights for Merged Data

## Within Same Sample Design (Correlated)

- Blend replicates (no new replicate weights needed)
- Still Use Fay Factor of 4

## Across Sample Design (Uncorrelated)

- Stack replicates (add replicate weights together)
  - Ex. 80 + 160 = 240
- Adjust replicate weights to account for stacking
- Change Fay Factor from 4 to 16

# Talk Recap

1. Uses of Standard Errors in Analyzing Data
2. Methods to Compute Standard Errors for TUS-CPS estimates
   - – Generalized variance functions (SE parameters)
   - – BRR replication – Fay's method (replicate weights)
3. Special Topics for Analysts
   - – Change in Race/Ethnicity Questions
   - – 2002-03 Overlap Sample
   - – Merged Data Sets