

# Statistical Aspects on the Best Practices for Analyzing TUS-CPS Data

Benmei Liu

Division of Cancer Control and Population Science

September 17, 2013

# Outline

- Use of standard errors in analyzing survey data
- Ways to compute standard errors for TUS-CPS estimates
  - Generalized variance functions (SE parameters)
  - Replication methods (replicate weights)
- Replicate weights on merged data sets

# Importance of Standard Errors

- Indicate the accuracy of survey estimates
- Construct confidence interval
- Conduct statistical tests

# Confidence Intervals

- Formula:  $\hat{y} \pm Z \times SE(\hat{y})$
- Example: The 95% CI for prevalence of males 18+ who currently smoke in 2010-2011 ( $\hat{y} = 18.0\%$ ,  $SE = 0.19\%$ ):

$$18.0\% \pm 1.96 \times 0.19\% = (17.7\%, 18.4\%)$$

# Hypothesis Testing

- Formula (two group t-test)
- $\frac{|\hat{X}-\hat{Y}|}{SE(\hat{X}-\hat{Y})} > Z \implies$  Statistical significance
- Example (current smoking prevalence 2010/2011)

	P	SE(P)		
Male	18.0%	0.19%		
female	14.2%	0.15%	t-stat	p-value
diff	3.8%	0.20%	18.9	<0.0001

# Estimating Standard Errors TUS-CPS

- Generalized variance functions(GVF) (SE parameters)
- Replication methods (replicate weights)

# Standard Errors Using 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)
- For more details, see the link:  
<http://www.census.gov/prod/techdoc/cps/cpsaug10.pdf>

# Estimating Standard Errors Using Replication

- Select subsamples from whole sample
- Form estimates from full sample and replicates
- Measure variation between full sample and replicate estimates



# Replication Methods

- Jackknife
- Balanced repeated replication (BRR)
  - Variant: Fay's method  
Note: TUS-CPS uses Fay's method
- Bootstrap

# Balanced Repeated Replication (BRR) Based on Replication Weights

- Replicate weights not on TUS-CPS public use file (2010-11 available from Census Bureau:  
[http://thedataweb.rm.census.gov/ftp/cps\\_ftp.html#cpsrepwgt](http://thedataweb.rm.census.gov/ftp/cps_ftp.html#cpsrepwgt),  
earlier files upon request from NCI)
- Requires special software (SUDAAN, WesVar, etc.)
- Provides a more accurate standard error than GVF

# Replication SE formula

$$SE(\hat{Y}) = \sqrt{c \sum_{r=1}^R (\hat{Y}_{(r)} - \hat{Y}_{(0)})^2}$$

where:

$R$  = total number of replicates

$c$  = a constant that depends on replication method

Note:  $c = 4/R$  for TUS-CPS

# Replication SE formula for TUS-CPS

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

$R = 48$  (for 1980-based designs)

$R = 80$  (for 1990-based designs)

$R = 160$  (for 2000-based designs)

# Replication SE Example

$$\hat{Y}_{(0)}=10, \hat{Y}_{(1)}=8, \hat{Y}_{(2)}=11, \hat{Y}_{(3)}=12$$

$$\begin{aligned} SE(\hat{Y}) &= \sqrt{\frac{4}{3} [(8 - 10)^2 + (11 - 10)^2 + (12 - 10)^2]} \\ &= \sqrt{\frac{4}{3} (4 + 1 + 4)} \\ &= 3.46 \end{aligned}$$

# Implementing Replication

- Create weights for the full-sample
- Form replicates (or subsamples) of the full-sample and create replicate weights
- Attach weights to survey data set
- Compute estimates and standard errors using special software

# Replicate Weights for Combining Multiple Years of Data

Adjust replicate weights to account for merging data

- Within Sample design
- Across Sample designs
- 1980 based – 48 replicates
  - 1990 based – 80 replicates
  - 2000 based – 160 replicates

# Adjust Replicates for Combined Data

- Within same sample design
  - No special adjustment for replicate weights
  - Still use Fay factor of 4
- Across Sample design
  - Stack replicates (Number of replicates= $R1+R2$ )  
Ex.  $48+80=128$
  - Adjust replicate weights to account for stacking
  - Original replicate weights adjusted to reflect new R
  - New replicate weights set equal to full-sample weight
  - Change Fay factor from 4 to 16



# Talk Recap

- Use of standard errors in analyzing data
- Ways to compute standard errors for TUS-CPS estimates
  - Generalized variance functions (SE parameters)
  - Replication methods (replicate weights)
- Replicate weights on combined data sets