HINTS: Discussion of Statistical Weights and Their Use

William Davis, PhD

Richard P. Moser, PhD

February 9, 2004

National Cancer Institute

HINTS Survey Carried out by Westat

- List of telephone exchanges purchased
- Exchanges and numbers sampled using random digit dialing (RDD)
  - Screens out unwanted exchanges (e.g., business exchanges)
  - Exchanges with high minority representation were oversampled (HINTS stratification)
- For more information see L. Rizzo's document on our website
  - "NCI HINTS Sample Design and Weighting Plan"

NATIONAL CANCER INSTITUTE

HINTS Statistical Weight

- Statistical weight:
  - number of people in the population that the sampled person represents
- HINTS Statistical weights derived from
  - selection probabilities,
  - response rates,
  - post-stratification adjustment.

NATIONAL CANCER INSTITUTE

HINTS: Race Ethnicity

Race Eth	N	%		Wgt N	Wgt %	Diff %
Hispanic	764	12.0%		23,340,239	11.1%	**0.9%**
White	4276	67.1%		143,031,482	68.3%	-1.1%
Afr Amer	716	11.2%		20,905,523	10.0%	**1.3%**
Others	312	4.9%		12,028,337	5.7%	-0.8%
Missing	301	4.7%		10,148,812	4.8%	-0.1%
Total	**6369**	**100.0%**		**209,454,391**	**100.0%**	

Reflects the planned oversampling of minority exchanges.

NATIONAL CANCER INSTITUTE

HINTS: Average Weights

Race Ethnicity	N	Avg Wgt
African American	716	29,198
Hispanic	764	30,550
Total	**6369**	**32,887**
White	4276	33,450
Others & Missing	613	36,178

Reflects the planned oversampling of minority exchanges.
Categories are ordered by their average statistical weight.

NATIONAL CANCER INSTITUTE

HINTS: Age Groups

Age Grp	N	%		Wgt. N	Wgt %	Diff %
18-34	1656	26.0%		65,185,716	31.1%	**-5.1%**
35-39	656	10.3%		21,238,207	10.1%	0.2%
40-44	684	10.7%		22,367,629	10.7%	0.1%
45+	3373	53.0%		100,662,838	48.1%	**4.9%**
Total	**6369**	**100.0%**		**209,454,391**	**100.0%**	

Older folks participated at a higher rate.

NATIONAL CANCER INSTITUTE

HINTS: Gender

	N	%		Wgt N	Wgt %	Diff %
Male	2521	39.6%		100,707,025	48.1%	**-8.5%**
Female	3848	60.4%		108,747,366	51.9%	**8.5%**
Total	**6369**	**100.0%**		**209,454,391**	**100.0%**	

Females participated at a higher rate.

NATIONAL CANCER INSTITUTE

HINTS: Weighted vs. unweighted analyses

- If we want a combined estimate for the whole population – not just specific groups then

- Unweighted HINTS analyses would have
  - Too many African Americans and Hispanics
  - Too many 45+ and too few 18-34 year olds
  - Too many females and too few males

NATIONAL CANCER INSTITUTE

HINTS website directions

Use the following code in SUDAAN procedures:

proc *procedurename* data=*datasetname* design=jackknife;

weight fwgt;

jackwgts fwgt1-fwgt50/adjjack=.98;

- fwgt: final statistical weight
- fwgt1-fwgt50: set of 50 replicate weights
- Descript, Crosstab and Regress are valid SUDAAN procedures

Replicate weights

- What are replicate weights?
  - HINTS 50 replicate weights were obtained by deleting 1/50th of the subjects in the full sample (and reweighting)
- Why do we need replicate weights?
  - Used to estimate the variance of estimates obtained from the full sample -- for example a mean or a regression coefficient
- For more information see the SUDAAN manual or
  - Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. John Wiley, p. 29.

NATIONAL CANCER INSTITUTE

Examples of HINTS Weights

Sub	fwgt	fwgt1	fwgt2
1	14,367	14,693	14,837
2	109,694	111,069	111,021
3	14,767	0	14,859
4	18,467	19,301	0

Full sample and 2 replicate weights for 4 sampled people
First two subjects are in both replicates while other two are not.

Jackknife estimate of variance

Full Sample estimate	$\hat{\theta}$
Replicate estimate	$\hat{\theta}_i$
Jackknife estimate of variance	$Var(\hat{\theta}) = \frac{49}{50}\sum_{i=1}^{50}\left(\hat{\theta}_i - \hat{\theta}\right)^2$

NATIONAL CANCER INSTITUTE

SAS vs. SUDAAN

- SAS
  - Many procedures not designed for survey data
  - Procedures that are valid for survey data do not allow use of replicate weights
- SUDAAN
  - All procedures explicitly for survey data
  - Though point estimates (e.g., means) will be the same when comparing weighted SAS vs. SUDAAN, standard errors are usually larger in SUDAAN

SAS and SUDAAN Procedures

	SAS	SUDAAN
Mean	MEANS	DESCRIPT
Crosstab	FREQ	CROSSTAB
Multiple regression	REG or GLM	REGRESS
Logistic regression	LOGISTIC	RLOGIST

Analysis of HINTS Data Using SUDAAN

- SUDAAN is specifically designed to analyze complex survey designs that utilize samples that may be multi-staged, stratified, unequally weighted and/or clustered
- One of several statistical programs able to analyze complex survey designs (e.g., Westvar, STATA)

NATIONAL CANCER INSTITUTE

Example: Testing Whether Smokers, Who Answer Regarding Self or the Average Smoker, Differ in Their Perceived Risk of Developing Lung Cancer

- **Outcome (tu15/tu16):**
  - How likely do you think it is that you (or the average cigarette smoker) will develop lung cancer in the future?
    - 1= Very low
    - 2= Somewhat low
    - 3= Moderate
    - 4= Somewhat high
    - 5= Very high

- **Predictor:**
  - **Current smokers** randomly assigned to report on either:
    - 1= Self
    - 2= Average smoker

Comparing Means-SAS vs. SUDAAN*

Group	SAS Proc Reg Unweighted	SAS Proc Reg Weighted	SUDAAN Proc Regress
Average	3.75	3.77	3.77
Self	3.19	3.16	3.16

Notice: Point estimates (means) are equal when comparing weighted SAS and SUDAAN

*This is an example from an ongoing analysis—please do not cite

Regression Results-SAS vs. SUDAAN

Value	SAS Proc Reg Unweighted	SAS Proc Reg Weighted	SUDAAN Proc Regress
Beta	.55	.61	.61
Standard Error	.0563	.0562	.0782
t	9.8	10.8	7.8
p	<.0001	<.0001	<.0001

Notice: Larger standard error in SUDAAN as compared to SAS

SUDAAN Linear Regression Syntax

```
proc regress data=tobacco design=jackknife;
weight fwgt;
jackwgts fwgt1-fwgt50/adjjack=.98;
subpopn smokegrp=3/name="Current Smokers";
subgroup abstype;
levels 2;
model abs=abstype;
setenv decwidth=9 colwidth=20;
test adjwaldf;
lsmeans abstype;
run;
```

NATIONAL CANCER INSTITUTE