Using Nationally Representative Data from Complex Surveys in the Classroom

Michael R. Jiroutek, DrPH
Campbell University
jiroutekm@campbell.edu

Matthew Hayat, PhD
Georgia State University
mhayat@gsu.edu

MyoungJin Kim, PhD
Illinois State University
mkim2@ilstu.edu

Todd A. Schwartz, DrPH
University of North Carolina at Chapel Hill
Todd_Schwartz@unc.edu

Background

- The National Center for Health Statistics (NCHS), in conjunction with the Centers for Disease Control and Prevention, conducts national surveys across a variety of health topics.
- Sampling is done utilizing complex probability samples
- Special attention must be paid to analyze the complex survey data correctly
- Data are commonly available in SAS, SPSS and Stata dataset formats, along with label and format code and analysis templates
- **Consistent with GAISE, analyses allow incorporating real-world data into the classroom**

Datasets

- These freely available survey data are warehoused at the cdc.gov/nchs website. The following are current:
  - National Ambulatory Medical Care Survey (NAMCS)
  - National Hospital Ambulatory Medical Care Survey Emergency and Outpatient Departments (NHAMCS-ED and OPD)
  - National Survey of Family Growth (NSFG)
  - National Health Interview Survey (NHIS)
  - National Immunization Child and Teen Surveys (NIS-Child and NIS-Teen)
  - National Health and Examination Nutrition Survey (NHANES)
  - National Vital Statistics System (NVSS)

Analysis Methodology

Utilization of survey weighting, cluster and stratification variables and domain construction are required to generate accurate national estimates

Sample weighting:
- Reflects probability sampling
- Can account for nonresponse & calibration to target population

Clustering:
- Reflects 2-stage (or higher) sampling structure
- Clusters may be randomly selected in the first stage, followed by more refined sampling (e.g., households or individuals) in subsequent stage(s)

Stratification:
- Reflects partitioning of the sampling frame into mutually exclusive and exhaustive subgroups, strata
- Should be accommodated in analysis

Inference:
- Adjusted tests needed to account for complex sampling (e.g., Rao-Scott, etc.)
- Focus can be on estimation or testing

Examples: Unweighted and Weighted Analyses

Study Description
- Retrospective, cross-sectional, observational study investigating gender disparities in patient education provided during patient visits with a diagnosis of coronary heart disease.
- Utilizes National Ambulatory Medical Care Survey (NAMCS) data collected between 2005 and 2014, inclusive.
- Patient education defined as one or more of diet/nutrition, exercise, tobacco use/exposure, or weight reduction education received at patient visit.

		Unweighted Counts and Percents (N = 17,332)[1]			Weighted Counts and Percents (N = 40,642,262)[2]		
		Health Education			Health Education		
		Received N (%)	Not Received N (%)	95% CI for Percent Receiving Health Education[3]	Received N (%)[2]	Not Received N (%)[2]	95% Wald CI for Percent Receiving Health Education[2]
Gender	Female	1,455 (21.2)	5,399 (78.8)	(20.3, 22.2)	3,687,093 (22.3)	12,874,271 (77.7)	(20.4, 24.1)
	Male*	2,335 (22.3)	8,143 (77.7)	(21.5, 23.1)	6,015,516 (25.0)	18,065,383 (75.0)	(23.2, 26.8)

Multivariable Logistic Regression Models

	Unweighted		Weighted	
Predictor Variable	Adjusted OR (95% Wald CI)	p-value	Adjusted OR (95% Wald CI)	p-value
Gender (Female vs. Male*)	0.93 (0.85 – 1.02)	0.1178	0.85 (0.75 – 0.97)	0.0160
Age group (≥75 vs. 18-44*)	0.93 (0.70 – 1.22)	0.5805	0.91 (0.61 – 1.36)	0.6508
(65-74 vs. 18-44*)	1.13 (0.86 – 1.48)	0.3885	1.16 (0.77 – 1.75)	0.4809
(45-64 vs. 18-44*)	1.21 (0.62 – 1.71)	0.1632	1.38 (0.96 – 1.97)	0.0806
Tobacco use (Current vs. Non-current*)	2.29 (2.05 – 2.56)	<0.0001	2.05 (1.69 – 2.49)	<0.0001
Primary care provider seen (Yes vs. No*)	0.66 (0.60 – 0.72)	<0.0001	0.63 (0.52 – 0.75)	<0.0001
Diabetes (Yes vs. No*)	1.09 (1.00 – 1.20)	0.0611	1.16 (0.99 – 1.35)	0.0617
Hypertension (Yes vs. No*)	1.13 (1.02 – 1.24)	0.0197	1.28 (1.11 – 1.48)	0.0007
Obesity (Yes vs. No*)	2.82 (2.51 – 3.17)	<0.0001	2.60 (2.14 – 3.16)	<0.0001
Insurance type ('Other' vs. Private*)	0.75 (0.61 – 0.92)	0.0050	0.69 (0.49 – 0.96)	0.0289
(Medicaid/SCHIP vs. Private*)	0.84 (0.69 – 1.03)	0.0886	0.78 (0.58 – 1.07)	0.1204
(Medicare vs. Private*)	0.84 (0.75 – 0.95)	0.0039	1.00 (0.82 – 1.21)	0.9609
[Etc.: Not all variables in model shown]				

*Denotes reference category
CI: Confidence Interval; OR: Odds Ratio: SCHIP: State Children's Health Insurance Program
1. Raw, unweighted survey sample size; 2. Accounting for sampling weights and clusters; 3. Clopper-Pearson exact confidence intervals

Noteworthy Results
- Proportion of males receiving health education was substantially higher for males in the weighted (versus unweighted) calculation
- Point estimate and confidence interval suggest that females were substantially less likely to receive health education in the weighted (versus unweighted) model
- Statistical significance in the comparison for Medicare vs. Private insurance was attenuated to the null in weighted (versus unweighted) model

Example SAS Code

```
proc surveymeans data=final_1602f nomcar;
cluster cpsum;
weight patwt;
strata cstratm;
var age;
domain inflag;
```

```
proc surveyfreq data=final_1602f nomcar;
cluster cpsum;
weight patwt;
strata cstratm;
table inflag*(agenew age01 age02
      age03 age12 age13 age23)*ptedu
      / row chisq cl or;
```

```
proc surveylogistic data=final_1602f;
weight patwt;
cluster cpsum;
strata cstratm;
class inflag gender agenew paytypenew
      usetobacnew primcarenew diabnew
      hyplipid htn obesity / param=ref;
model ptedu (event=first) = gender agenew
      paytypenew usetobacnew primcarenew
      diabnew hyplipid htn obesity / rsquare;
domain inflag;
```

Conclusions

- Incorporating survey weights in the analysis is needed to produce correct standard error estimates
- National studies provide opportunities for teaching of advanced statistical concepts and learning importance of sampling
- Working with nationally representative data lends to fun and interesting active learning class activities and exercises
- Data are publicly available and easy to access
- Getting started is enabled by freely available template code for the mainstream statistical software packages
- **Implementation of GAISE recommendations using real-world data can foster student enthusiasm and interest and allow for demonstration of advanced statistical concepts and methods**

References

1. https://www.cdc.gov/nchs/index.htm
2. Hilleary RS, Jabusch SM, Zheng B, Jiroutek MR, Carter CA. Gender Disparities in Patient Education Provided During Patient Visits with a Diagnosis of Coronary Heart Disease. Women's Health. Accepted.