Survey Data Analysis

Recap of many ideas already explored

ehsan.karim@ubc.ca

Sept 24, 2019

SPPH 504/007

Pearson's chi-squared test is used for

A test of goodness of fit establishes whether an observed frequency distribution differs from a theoretical distribution.

A test of homogeneity compares the distribution of counts for two or more groups using the same categorical variable.

A test of independence assesses whether observations consisting of measures on two variables, expressed in a contingency table, are independent of each other.

What is Hosmer–Lemeshow test?

explains the proportion of variance in the dependent variable that is explained by the predictors

a test that is used to evaluate the statistical significance of each coefficient in the model

a statistical test for goodness of fit for logistic regression models

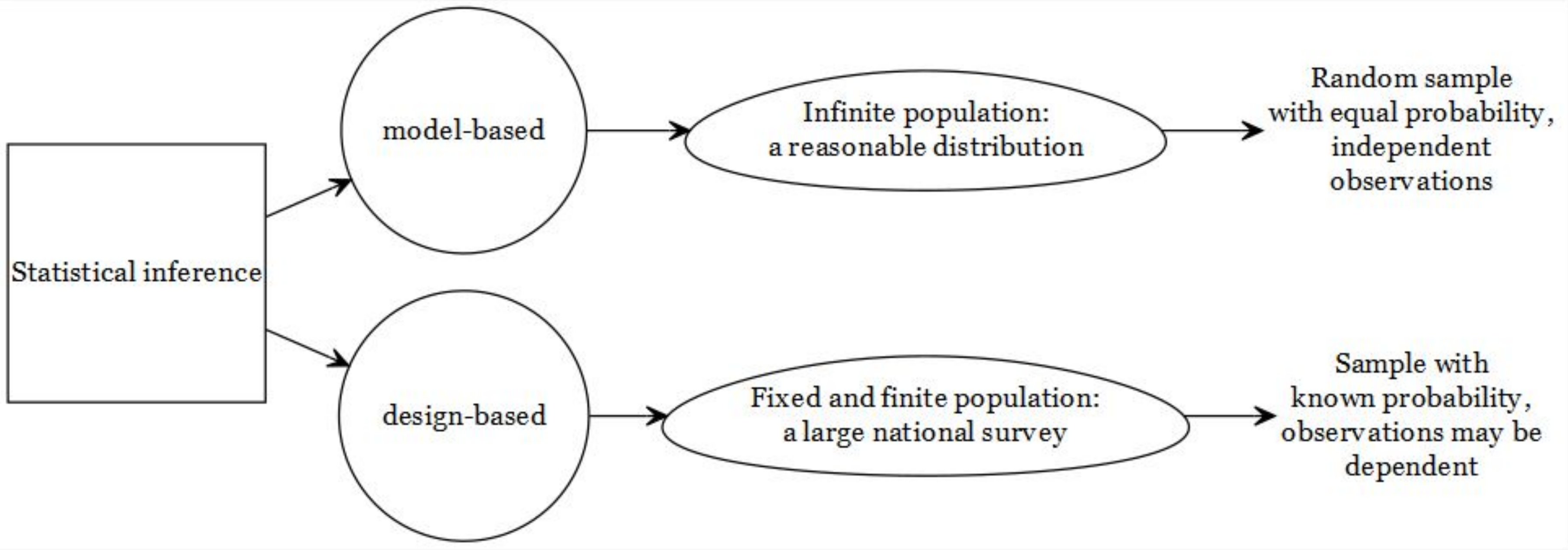
assesses the relative importance of individual predictors in the model

Reference

- Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). **Applied survey data analysis**. [edition 1] Chapman and Hall/CRC. [eBook available at UBC]
    - All of the blue page numbers are from here
- 2nd edition available (2017), but eBook not available now at UBC.

Review

From first class



Design-based vs. Model based inference

- The difference between
  - Design-based (finite population/data could be unknown but fixed population/does not support generalization to other population) and
  - Model-based (infinite population/a random process that generates data) inference

is the population to which the results can be **generalized**.

- Analysis of complex survey is: **design-based** (usually).

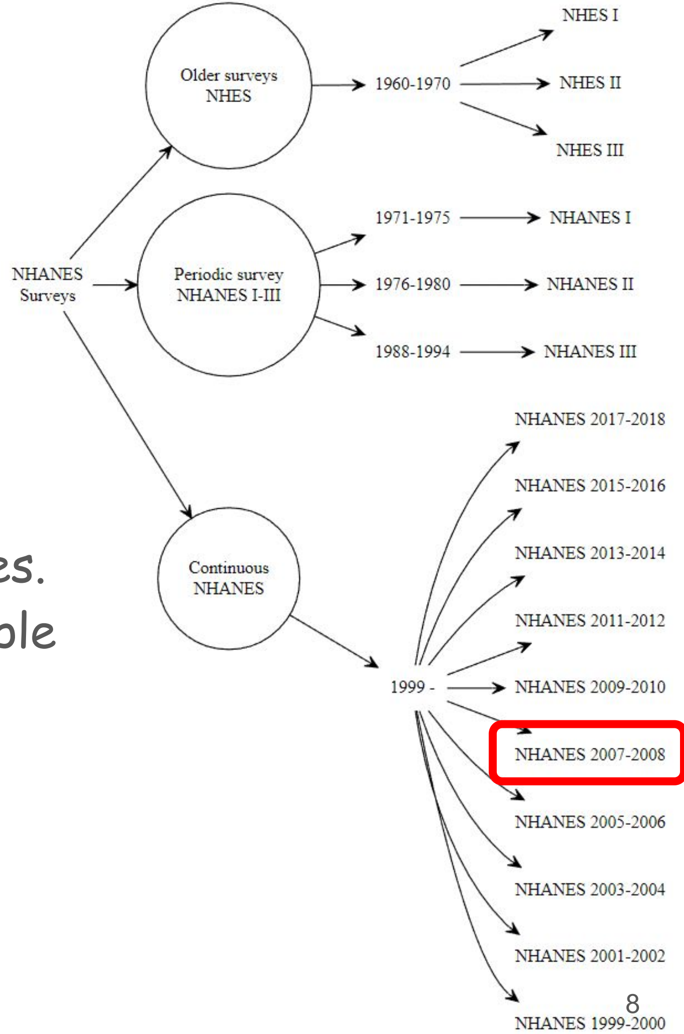
General survey data sources

- CCHS
- NHANES
- BRFSS
- KNHANES
- HRS
- NCS-R
- ESS

Survey Example review

NHANES:

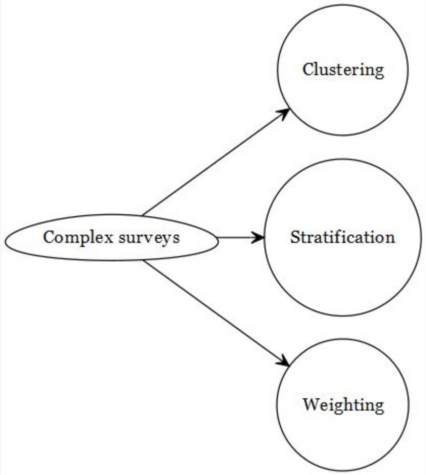
- survey of the adult, noninstitutionalized population of the United States
- Let's say, we are interested in cycle 2007-08
- contains data from 10,149 individuals of all ages.
- an in-home medical history interview with sample respondents
- (there is also a medical examination part)



Sampling Procedure review

NHANES:

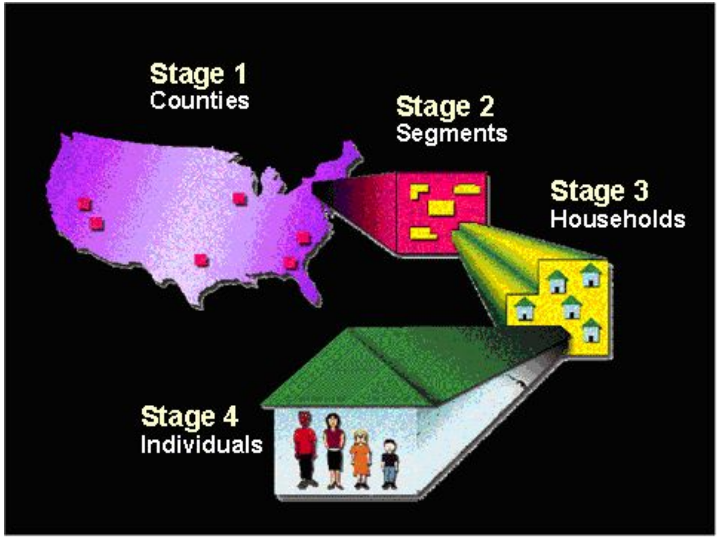
- NOT obtained via simple random sample.
- multistage sample designs
    - to increase the convenience of the data collection process (convenience - use clustering)
    - To ensure that we can estimate from each groups of interest with reasonable precision (gender groups / income levels, etc. - use stratification)
    - Generally know as complex survey design



Sampling Procedure review

NHANES:

- NOT obtained via simple random sample.
- multistage sample designs

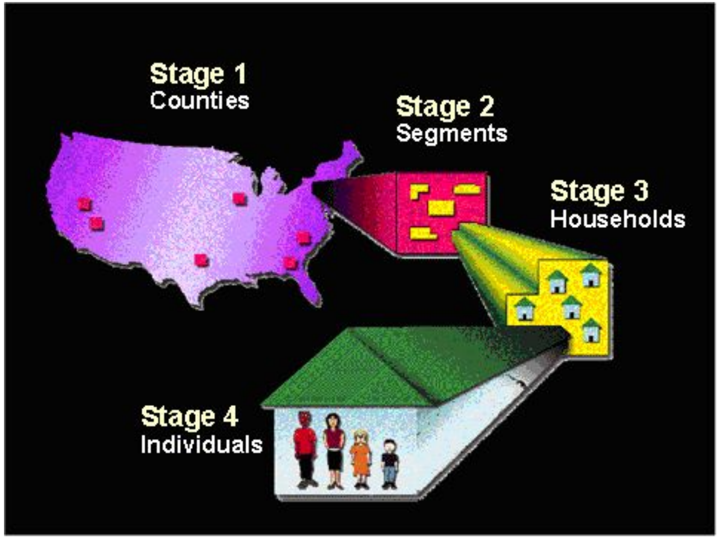


Stage 1: PSU/clusters = geographically contiguous counties. 50 states - divided into ~3100 counties. Each PSU is assigned to a strata (e.g., urban/rural or PSU size etc.). The counties are randomly/PPS selected using a 2-per-stratum design. Complex sample variance estimation requires PSU + strata (masking involved).
Stage 2: each selected *county is broken into segments* (with at least ~50-100 housing units). Segments are randomly/PPS selected.
Stage 3: each selected *segment is divided into households*. Households are randomly selected.
Stage 4: Within each sampled household, an individual is randomly selected.

Sampling Procedure review

NHANES:

- NOT obtained via simple random sample.
- multistage sample designs



Stage 1:

Strata (e.g., urban/rural) → **PSU** (County)

Stage 2: **Segment** (~50-100 housing units)

Stage 3: **Household**

Stage 4: **Individual**

Sampling Procedure review

NHANES:

- It is a probabilistic sample (we know probability of getting selected for all individuals)
- This sample is unlikely to be representative of the entire population, as
  - some under/oversampling occurs (unlike SRS),
  - samples may be dependent (due to proximity of some samples)
- For example, household with the following characteristics may be oversampled in NHANES:
  - African Americans
  - Mexican Americans
  - Low income White Americans
  - Persons age 60+ years

Sampling Procedure review

- probabilistic sample
    - under/oversampling occurs (unlike SRS),
    - samples may be dependent (due to proximity of some samples)

Note that, when complex sampling designs are used to collect data, that invalidates our usual models as the observations are not independent anymore! **Beta coefs, p-values, CIs, SEs all are useless in inferring about the population.**

interview / sample weight review

NHANES:

- A sample weight is assigned to each sample person.
- Weight = the number of people in the target population represented by that sample person in NHANES.
  - A respondent's interview weight = 50 means that person represents 50 people in the target population (US).
- Weights reflect
  - the unequal probability of selection,
  - nonresponse adjustment, and
  - adjustment to independent population controls.

Survey features review

NHANES (2007-08):

- Interview weight
  - Another weight is for MEC.
- Strata
- PSU/cluster

cluster ids/PSUs

nested within strata

WTINT2YR - Full Sample 2 Year Interview Weight

Variable Name: WTINT2YR

SAS Label: Full Sample 2 Year Interview Weight

English Text: Interviewed Sample Persons.

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
2359.373828 to 186295.50665	Range of Values	10149	10149	
.	Missing	0	10149	

SDMVSTRA - Masked Variance Pseudo-Stratum

Variable Name: SDMVSTRA

SAS Label: Masked Variance Pseudo-Stratum

English Text: Masked Variance Unit Pseudo-Stratum variable for variance estimation

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
59 to 74	Range of Values	10149	10149	
.	Missing	0	10149	

SDMVPSU - Masked Variance Pseudo-PSU

Variable Name: SDMVPSU

SAS Label: Masked Variance Pseudo-PSU

English Text: Masked Variance Unit Pseudo-PSU variable for variance estimation

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1 to 2	Range of Values	10149	10149	
.	Missing	0	10149	

Is it possible to tell from this table how many participants are 90 years old?

RIDAGEYR - Age in years at screening

Variable Name: RIDAGEYR

SAS Label: Age in years at screening

English Text: Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count
0 to 79	Range of Values	9595
80	80 years of age and over	376
.	Missing	0

Yes

No

Estimates of interest

- We are generally interested in population
- We can, however, make statements about the sample

	Population	Sample
ATE	PATE	SATE

Estimates of interest

- Exposure = rheumatoid arthritis (RA),
- Outcome = myocardial infarction / heart attack (MI)

In a regression: MI ~ RA + covariates, we get OR = 1.54

- No survey features weights or cluster/strata were used in the fitting.

Interpretation: Those who had RA exhibited increased odds of prevalent MI compared to non-RA individuals after controlling for baseline covariates.

OR 1.54 applies to

US population in 2007-08

Surveyed people who were interviewed in NHANES 2007-08

Design effects review

design effect (DE) =

$$D^2(\hat{\theta}) = \frac{SE(\hat{\theta})^2_{complex}}{SE(\hat{\theta})^2_{srs}} = \frac{Var(\hat{\theta})_{complex}}{Var(\hat{\theta})_{srs}}$$

Page 24

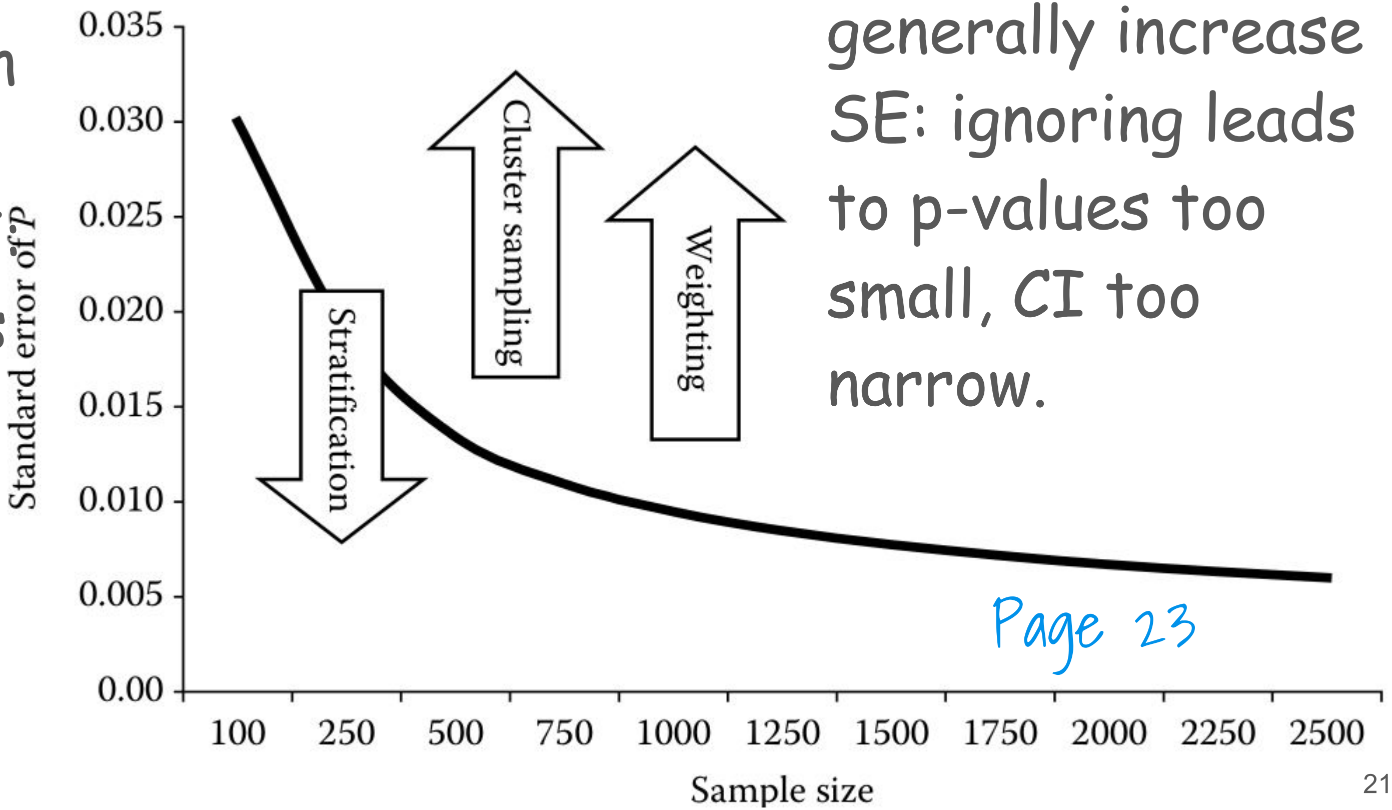
the ratio of

- a variance of an estimate (e.g., correlation/beta) in a complex sample to
- the variance of the same estimate in a SRS

DE = 2 implies that the variance from complex survey is twice as large as we would expect with SRS. That also implies that if we used complex survey instead of SRS, we would have to use twice the sample size.

Design effects review

Stratification generally decreases SE: ignoring leads to p-values too high, CI too wide.



Clustering and weighting generally increase SE: ignoring leads to p-values too small, CI too narrow.

Page 23

Variance estimation

- Standard approaches to calculate SE assume SRS
  - consequently p-values and CIs get distorted
- Doesn't take into account of
  - Stratification,
    - requires SE to be computed separately within each stratum, and then combined
  - Clustering,
    - strata nested within clusters in NHANES
      - 2 counties per stratum
  - Weighting
    - Unequal probability
    - non-response

Variance estimation

- Taylor series approximations [strata and cluster info provided]
  - Sampling error stratum and cluster info are required for this method for variance estimation
- Replication Methods [replicate weights provided]
  - JRR
  - BRR
  - Fay's BRR
  - Bootstrap (Rao-Wu) [Not the same procedure that we have seen earlier]

Comparison of the TSL, JRR, BRR, and Bootstrap Variance Estimation Methods for the Estimation of Descriptive Population Parameters

			Standard Error by Method						
Statistic	*n*	Estimated Mean	SRS	W-BRR	W-JRR	S-TSL	S-BRR	S-JRR	R-Boot
Years of school	9,759	12.31	0.031	0.078	0.077	0.077	0.077	0.077	0.076
Body weight (lbs)	9,759	172.35	0.374	0.447	0.432	0.435	0.434	0.435	0.404
Words recalled	9,759	7.57	0.028	0.064	0.064	0.066	0.066	0.066	0.067

Source: 1992 HRS.

Variance estimation

- That means, all of the Hypothesis testing we were doing under SRS are now invalid
- MLE can't be defined if the samples are not even random
    - Pseudo-likelihood
- Goodness of fit tests also get impacted

Other useful statistic

Design-based tests, properly utilises survey features

Page 84

Comparisons of Hypothesis Testing Procedures for Simple Random Sample Data versus Complex Sample Survey Data

Simple Random Sample (iid) Data	Complex Sample Survey Data
Student t-tests of hypotheses $(H_0 \mid H_A)$ for means, proportions, single model parameters; for example, $\bar{Y} = \bar{Y}_0$; $\beta_j = 0$	Design-adjusted Student t-test • Correct standard error in denominator • Design-adjusted degrees of freedom
Student t-tests of simple hypotheses concerning differences of linear combinations of means; for example, $(\bar{Y}_1 - \bar{Y}_2) = 0$; $\Sigma_j a_j P_j = 0$	Design-adjusted Student t-test • Correct standard error in denominator reflecting separate estimates of variance and covariance of component estimates • Design-adjusted degrees of freedom
X^2 test of independence (association) in bivariate and multiway tables: • Pearson X^2, Likelihood Ratio G^2	Design-adjusted X^2 and F-tests • Rao–Scott first- and second-order corrections adjust for design effects in $\hat{\Sigma}(p)$ • X^2 transformed to F-test statistic
Full and partial F-tests for hypotheses for linear regression model goodness of fit and full versus reduced model, for example, $H_0: \beta = \{\beta_1, \beta_2, \ldots, \beta_p\} = 0$; $H_0: \beta_{(q)} = \{\beta_{p-q}, \ldots, \beta_p\} = 0$.	Design-adjusted Wald X^2 or F-test • Correct $\hat{\Sigma}(\hat{\beta})$ under complex design • Adjusted degrees of freedom
F-tests based on expected mean squares for ANOVA-type linear models	Linear regression parameterization of the ANOVA model. Design-adjusted Wald X^2 or F-tests as in linear regression above
Likelihood Ratio X^2 tests for maximum likelihood estimates of parameters in generalized linear models, for example, $H_{0,MLE}: \beta = \{\beta_1, \beta_2, \ldots, \beta_p\} = 0$; $H_{0,MLE}: \beta_{(q)} = \{\beta_{p-q}, \ldots, \beta_p\} = 0$.	Design-adjusted Wald X^2 or F-test • Correct $\hat{\Sigma}(\hat{\beta})$ under complex design • Adjusted degrees of freedom Design-adjusted likelihood ratio test (Lumley and Scott, 2014)

Other useful statistic

- **Rao-Scott Chi-Square Test**: a design-adjusted version of the Pearson chi-square test. Page 166
    - 2 versions are available (F and chi-square)
- **Pseudo-R2**: Nagelkerke and Cox-Snell pseudo-R2 statistics for logistic regression: survey featured statistics are available. Page 243
- **Archer-Lemeshow goodness of fit test** for survey data is the counterpart of Hosmer-Lemeshow goodness of fit test. Page 244

Generalizability

- An extreme example
- H0 concludes differently

Page 59

Regression Models of Log-Transformed Household Income for the 2006 HRS Black Subpopulation ($n = 2,465$)

Independent Variable	Regression Parameter Estimate (Standard Error, p-value)	
	Unweighted	**Weighted (Design-Based SE)**
Age (Continuous)	0.0026 (0.0058, $p = 0.66$)	0.0056 (0.0093, $p = 0.54$)
Gender		
Female	–0.4629 (0.1246, $p < 0.001$)	–0.3034 (0.2199, $p = 0.17$)
Male	Reference category	Reference category
Education		
Grade 0–11	–1.5585 (0.1991, $p < 0.0001$)	–1.9016 (0.2610, $p < 0.0001$)
Grade 12	–1.0304 (0.2011, $p < 0.0001$)	–1.5177 (0.2871, $p < 0.0001$)
Grade 13–15	–0.5145 (0.2152, $p < 0.0001$)	–0.7114 (0.1330, $p < 0.0001$)
Grade 16+	Reference category	Reference category
Region		
Northeast	0.0804 (0.2743, $p = 0.77$)	0.1462 (0.1680, $p = 0.38$)
Midwest	–0.3331 (0.2635, $p = 0.21$)	–0.2423 (0.2614, $p = 0.36$)
South	–0.2525 (0.2476, $p = 0.31$)	–0.3519 (0.2405, $p = 0.15$)
West	Reference category	Reference category
Urbanicity		
Urban	–0.0697 (0.1690, $p = 0.68$)	–0.0553 (0.3751, $p = 0.88$)
Suburban	0.05878 (0.1965, $p = 0.76$)	–0.0262 (0.4764, $p = 0.58$)
Rural	Reference category	Reference category

Software availability

Available Analytic Techniques in the Software Packages

Analytic Technique	Stata	SAS	SUDAAN	SPSS	IVEware	WesVar	Mplus	R
Descriptive								
Means	Yes	Yes	Yes	Yes	Yes	Yes	NA[a]	Yes
Totals	Yes	Yes	Yes	Yes	Yes	Yes	NA[a]	Yes
Ratios	Yes	Yes	Yes	Yes	Yes	Yes	NA[a]	Yes
Percentiles	No	Yes	Yes	No	No	Yes	NA[a]	Yes
Contingency tables	Yes	Yes	Yes	Yes	Yes	Yes	NA[a]	Yes
Regression								
Linear	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Binary logistic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ordinal logistic	Yes	Yes	Yes	Yes	Yes[b]	No	Yes	Yes
Multinomial logistic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Poisson regression	Yes	No	Yes	No	Yes	No	Yes	Yes
Probit	Yes	Yes	No	Yes	Yes[b]	No	Yes	Yes
Cloglog	Yes	Yes	No	Yes	Yes[b]	No	No	Yes
Survival Analysis								
Cox proportional hazards model	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Kaplan–Meier estimation	No	No	Yes	Yes	No	No	Yes	Yes
Missing Data								
Multiple imputation of missing data	Yes	Yes	Yes	Yes[c]	Yes	Yes[d]	Yes	Yes
Analysis of multiply imputed data sets	Yes	Yes	Yes	Yes[c]	Yes	Yes[d]	Yes	Yes

Page 404

How to make inference about target population?

- Interview weight should be used
  - to make statistical inference at the population level.
- PSU/cluster + strata information need to be used
  - to get correct SE.
  - Otherwise variance is incorrectly calculated under SRS assumption.

analytic.data

Filter

	arthritis.type	heart.attack	interview.weight	PSU	strata	bmi	diabetes	smoke	age	race
2	Non-arthritis	No	8727.798	1	70	(25,80]	No	No	(50,70]	Other
3	Non-arthritis	No	24342.505	2	68	(0,25]	No	No	(0,50]	Black
4	Non-arthritis	No	9811.075	2	65	(25,80]	No	Yes	(50,70]	Other
5	Rheumatoid arthritis	No	8058.685	2	66	(25,80]	Yes	No	(50,70]	Black
6	Non-arthritis	No	19055.790	2	71	(25,80]	No	No	(0,50]	Other
7	Non-arthritis	No	89655.454	1	59	(0,25]	No	No	(0,50]	Other
8	Non-arthritis	No	25274.173	1	66	(25,80]	No	No	(0,50]	Other
9	Non-arthritis	No	7782.354	1	62	(25,80]	No	No	70+	White
11	Non-arthritis	No	14685.353	1	65	(25,80]	No	Yes	(0,50]	Other

Estimating treatment effect in analytic sample

- Interview weight should NOT be used to make inference about the study sample.
- PSU/cluster + strata information need to be used to get correct SE. Otherwise variance is incorrectly calculated under SRS assumption (i.e., sample dependency ignored).

Generalizability of regression estimates

- OR = <u>1.66 </u>(95% CI 0.71, 3.89) applies to <u>US population</u>.
  - used weights and cluster+strata option
- OR = <u>1.54</u> (95% CI 0.82, 2.89) applies to <u>survey sample</u>.
  - used cluster+strata option, no weights
- OR = <u>1.54</u> (95% CI 0.95, 2.51) has a misleading/somewhat smaller CI / SE estimate.
  - no survey features used; assuming SRS

Inappropriate survey data analysis

KNHANES is a

- complex,
- stratified,
- multistage,
- probability-cluster survey

of a representative sample of the non-institutionalized civilian population in Korea.

Inappropriate survey data analysis

A study

Original Article

J Prev Med Public Health 2013;46:96-104 • http://dx.doi.org/10.3961/jpmph.2013.46.2.96

pISSN 1975-8375 eISSN 2233-4521

Journal of Preventive Medicine & Public Health

Inappropriate Survey Design Analysis of the Korean National Health and Nutrition Examination Survey May Produce Biased Results

Yangho Kim[1], Sunmin Park[2], Nam-Soo Kim[3], Byung-Kook Lee[4]

[1]*Department of Occupational and Environmental Medicine, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan;* [2]*Department of Food and Nutrition, Hoseo University, Asan;* [3]*Institue of Environmental & Occupational Medicine, Soonchunhyang University, Asan;* [4]*Korea Industrial Health Association, Seoul, Korea*

Table 1. Number (%) of research articles published in peer-reviewed journals cited in PubMed using Korean National Health and Nutrition Examination Survey data during 2007 to 2012[1]

Published year	No. of research articles (%)		
	Ordinary statistical analysis	Design-based analysis	Total
2007	12 (92.3)	1 (7.7)	13
2008	15 (88.2)	2 (11.8)	17
2009	24 (88.8)	3 (11.2)	27
2010	40 (87.0)	6 (13.0)	46
2011	58 (76.3)	18 (23.7)	76
2012[1]	57 (73.1)	21 (26.9)	78
Total	206 (80.2)	51 (19.8)	257

[1]Until the end of June.

Survey weights are

useless at all times

useful and irreplaceable

often problematic and should be handled with care

useful only if we are interested in population

Word of caution about "weights"! ⚠

- Most people agree that weights should be used to get population based descriptive estimates (e.g., prevalence, means).
- Not everybody agrees that 'weights' should be used beyond descriptive statistics (e.g., in regressions). Some arguments
  - Reduced precision / Inflated SE / loss of efficiency.
  - Weights can't be handled in many software packages.
  - Correct specification of model may still produce valid results
- Why not check results from both approaches and investigate?
- Popular survey data analysis textbooks seem to use weights.

Word of caution about "stata/cluster"! ⚠

- Consequences of omitting survey features (weights, cluster, strata)
    - Biased estimates
    - Underestimated SE
    - Smaller CI
    - Overstated significance levels
- There may be some arguments for omitting weights, but none for cluster/strata
- Unfortunately, in **CCHS public access data** cluster/strata info are not provided (only weights provided)
    - Can estimate OR (for sample or population), but CIs are wrong for both.

Word of caution about "Subpopulation analysis" ⚠

- Analysis with subpopulation (data restricted to only male group) will lead to bias, if you simply delete part of the population (the female population)
  - Point estimates will be fine
  - It is the SE estimation that will not be able to take proper consideration of
    - # of strata
    - # of cluster
    - complete info about these are essential for SE calculations
- Solution?
  - Prepare the design object first (based on all data, strata, cluster, weight)
  - Subset within the design (not just the data) for the subpopulation (e.g., male only)

Additional Textbook List

- Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). Applied survey data analysis. Chapman and Hall/CRC.
- Kim, Y., Park, S., Kim, N. S., & Lee, B. K. (2013). Inappropriate survey design analysis of the Korean National Health and Nutrition Examination Survey may produce biased results. Journal of preventive medicine and public health, 46(2), 96.
- Lewis, T. H. (2016). Complex survey data analysis with SAS. Chapman and Hall/CRC.
- Lumley, T. (2011). Complex surveys: a guide to analysis using R (Vol. 565). John Wiley & Sons.
- Lumley T. (2016). Survey: Analysis of Complex Survey Samples . R package version 3.31. https://cran.r-project.org/web/packages/survey/index.html.

Thanks!

ehsan.karim@ubc.ca

www.ehsankarim.com