JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey Weighting & Adding Questions

Module 1
Lecture 2

Objectives

- Understand the concept of survey weights and how survey weights affect survey estimates
- Describe commonly used methods for assigning survey weights
- Explain statistical methods for calculating survey weights to enhance representativeness to the target population
- Describe how survey questions are tested/validated

- This is NOT needed for Homework #1

Surveys in Population Health Research

- Understand burden of disease and risk factors in a population
  - How that differs among specific subpopulations
- Trends over time
- Assist communities/nations in targeting prevention, screening and treatment efforts
- Inform screening guidelines and other health service practices
- In order to do these things, we need our survey data to supply inferences about our entire target population

What we are NOT covering

- Survey design
- Sampling strategies
- Quality control/Quality assurance
- Interviewer training
- Methods for gathering data, *etc.*

So what are we doing?

- Assume you have a survey that has been done
  - Most of our publicly available and properly done survey data will come to you WITH weights
- You need to know a little bit about weighting and how it was done in order to apply that information to your analysis
- How might you ask to add questions to a survey?
- This is not work you will need to replicate for this class
  - This provides some background knowledge that will be useful
- Next lecture we will discuss accessing publicly available surveys and analyzing weighted survey data
  - That is what you will be doing for Assignment 1

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey Weighting

Survey Design – why do we need weights?

- Simple Scenario: Target an appropriate number within a full population and survey them
- Most samples are not “miniatures” of your population
  - Non-response bias: Rates not the same in each sub-population
  - Coverage bias: Your frame doesn’t include certain sub-populations
  - Information bias: Some subgroups are so small that estimates would be imprecise
- Selection bias even within appropriate probability sampling
  - People with more than 1 phone?
  - Who chooses to participate within a household?
- Strategies to overcome various biases include things like over-sampling in certain populations, stratifying the frame, etc.

What is a Survey Weight

- A value assigned to each individual in the dataset
- Goal: Make the statistics you compute from your data more representative of your target population
  - So each weight value indicates how much each case will count towards your overall estimates
- Values: Always non-zero, and always positive
  - 1 means that case contributes information as 1 case in the dataset
  - 2 means that case contributes information as 2 cases in the dataset
  - 0.5 means that case contributes information as a half of a case in the dataset
  - Weight of 0 would exclude that person from the dataset
- Three types: 1) Design, and 2) Non-Response, and 3) Post-Stratification (calibration)

What does weighting do in analysis?

- Important note: Weights have a larger effect on your descriptive statistics (prevalence) than on your regression coefficients (association)

- Tend to affect your standard errors more than your actual estimate/coefficient
- We use software to specify weights and design characteristics
- Some surveys offer numerous replicate weights – so you do your analysis numerous times, once with each weight, and then pool results

Design Weight

- Used in multi-stage sampling schemes where sample selection probability may differ
  - If you are selecting 50 households from each district, and district size differs
- Used to account for over- or under-sampling specific sub-populations
  - What is Over-Sampling:
    - Certain groups are so small that using normal methods would yield so few people in that group that your estimates would be unreliable
    - Over-sampling means that people in that group have a higher chance of being selected than others
  - Then you down-weight those individuals so that overall they represent the same proportion of your survey population as they do in the full target population
  - But your precision is improved due to underlying larger numbers
- Calculated as the inverse of the sample selection probability

“Weight” surveys within a Cohort Study

- If you nest a survey within an ongoing cohort, can you “weight” the data to be more representative of a broader population?
  - Is a Design Weight appropriate here?
- The survey literature says no
  - This is a non-probability sample (just like a social media survey, volunteers, etc.)
- Several methods have been proposed both for generating weights or for directly generating population-level prevalence estimates using this data

Non-Response Weights

- “Correct” for the fact that some subgroups responded to your survey more than others

- Look at non-response rates by various categorizations
  - Geographic Are people living in rural areas more or less likely to respond. People living in particularly difficult to reach, say, high mountainous areas or something like that.
  - Demographic

- Need to up-weight those subgroups that had higher non-response rates

- Calculated as the inverse of the response rate
  - Response rate of 80%, weight is 100/80.

Post-Stratification Weights: Calibrate to known population

- Compensate for the fact that once you incorporate design and non-response weights, your sample may not fully reflect your target population.
- Characteristics include: compare or make match with your target population
  - Age
  - Education
  - Race/Ethnicity & Language
  - Sex (females more likely to respond)
  - Gender (gender minorities may be less likely to respond, and we may not know their actual identity due to not asking properly)
- More complicated to calculate – more to come…

Calculating Weights

- Can only use one weight per case
- There are ways that design, non-response and post-stratification weights are combined into one
- Simple approach:

$$Wt = DWt * PSWt * NRWt$$

Wt = Final weight

DWt = Design weight

PSWt = Post-stratification weight

NRWt = Non-response weight

Specific Statistical Models

Specific Techniques

- Using survey data to answer causal questions
- Might need to use unweighted data
  - Could throw in the weight as a "covariate"
  - Include as covariates characteristics that might affect underlying non-response
- Certain multi-level models and structural equation models do not allow weights

Longitudinal Work

- Example: Four wave panel, done in 2000, 2005, 2010, and 2015
- Typical Strategy:
  - If you are using all 4 waves in your fixed or random effects model, then use the baseline weights
  - If you are analyzing each person from a specific wave forward, use the weight for that specific wave (each person's baseline)

Post-Stratification Weighting

- Goal: Have the core sociodemographic profile of your surveyed population be the same as the target population you are aiming to represent.

target population

- Finding good estimates of population characteristics can be difficult
- Sources of population data:
  - Census (or American Community Survey, for the US)
  - Other large population surveys
  - Health department or other government population profile
  - Reports from an organization/agency, school/university

Example: A Single Characteristic

Sex	Population Proportions	Sample Proportions	Weight
Female	0.5	0.6	0.5/0.6=0.833
Male	0.5	0.4	0.5/0.4=1.25
Total	1	1	

- If your population has a 50/50 sex distribution, and your sample had 60% females, you can weight your sample so that the females count less.
- Now what about doing this for more than one characteristic?

Adjusting for Multiple Characteristics

- Create a single table with combined characteristics
- You need to have information available at this level of granularity AND your numbers in each cell start to get quite small

Subgroup	Population Proportion	Sample Proportion	Weight
Male, 18-40, <HS			
Male, 18-40, ≥HS			
Female, 18-40, <HS			
Female, 18-40, ≥HS			
Male, 41-65, <HS			
Male, 41-65, ≥HS			
Female, 41-65, <HS			
Female, 41-65, ≥HS			

Adjusting for Multiple Characteristics: Manual

- Create separate tables for each characteristic (*i.e.*, sex, education, urbanicity)
- How to combine?
  - Compute a weight for each characteristic then multiply all weights together – not recommended – less accurate with more variables
  - Instead, sequential weighting
    - Calculate weights for sex
    - Then generate frequency distribution for education, using the data weighted by sex
    - Create weights by sex and education by multiplying them
    - Generate table for your next variable…
  - Note: When sequential weighting, by the end your characteristics of your earlier variables become less like your total population
    - More of an issue when characteristics are correlated (*i.e.*, income and education)

Automated Post-Stratification Adjustment of Multiple Variables: Logistic Regression

- Extract an individual dataset for the total population (*i.e.*, prior registry) with just the variables you want to weight on
- Add in a variable called “Survey” and set it to 0.
- Extract a subset of your survey data with the same variables, and set “Survey” to 1.

total population subdata (survey = 0) + survey data (survey = 1)

- Combine your two datasets.
- Conduct a logistic regression model with “Survey” as your dependent variable, and your characteristics as independent variables.

Automated Post-Stratification Adjustment of Multiple Variables: Logistic Regression

survey = 1

- Save the predicted probability (probability of "being in your survey") for each individual.
- Reminder: What is the predicted probability?
  - What is the probability of being in the survey for a 20 year old person?

    $\text{Log-Odds}_{survey} = \text{Constant} + (20 * \beta_{age})$
  - Exponentiate to get Odds

    Probability = Odds / (1+Odds)
- Use this to calculate weight

  Wt = 1/Probability
- See code in coding supplemental slides

Automated Post-Stratification Adjustment of Multiple Variables: Raking

- An automated iterative process of adjusting on multiple variables, available in most programs
- Example: BRFSS switched to raking (and cell phone sampling) in 2011
- The algorithm basically repeatedly estimates weight across each set of the variable in turn until the weights converge/stop changing
  - Convergence takes longer if there are more categories of responses for each variable
  - Also difficult if you have very few or no responses from a particular subpopulation
  - You can limit the time it takes by setting a max number of iterations
- Ideally you rake on those variables most strongly associated with non-response or non-coverage
- You need to know breakdowns for each variable, but not cross-tabs for each
- See code in coding supplement

Weight Trimming

- Sometimes you will have large variations in survey weights (outliers on either end)
- Occurs due to many reasons that can occur from design to data collection to post-stratification weighting
- These will affect your point estimates (for means, *etc.*) for your target population
- Trimming or truncating large weights can reduce this variability
  - Numerous methods exist for doing this
  - Usually you redistribute the “excess weight” among the non-trimmed units
- Many surveys will trim weights

Coding Supplement

- You have been provided with a coding supplement that shows you commands used for raking in Stata and R
    - You don’t need to be able to replicate these for this class

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Question Addition and Validation

Motivation

- Who might want to modify a survey questionnaire
  - Extensive user who wants to understand a topic better or obtain information on additional aspects/nuances
  - Expert/investigator in a field that is currently not covered in the survey
  - New/emerging issue or public health problem that needs more information

- The goal is not to teach HOW to develop and validate a question, but how to explain/justify the work that has been done for a question you propose to add

Associated Costs

- To the researcher:
  - 2023 BRFSS in Texas reports:
    - Cost to add a question to the 2023 BRFSS: $5,500 per question
    - Cost to add a geographic area or oversampling: $105 per survey
    - Additional costs may be added for analyses/report writing
  - 2023 BRFSS in Delaware:
    - $4,000 per question

- To the survey administrator
  - Example: NHANES added 20 minutes of occupational questions in 1988
    - Estimated cost: $1 – 1.4 million

Process to Propose a New Question/Exam Component/Lab Test

- Submit a Letter of Intent
  - 2 pages
  - Include technical requirements, issues of safety and privacy of participants, and the public health significance.

- Full Proposal
  - Invited to submit after review of LOI
  - Additional components:
    - History of cognitive and validation testing
    - History of prior use
    - Analytical plan
    - Pertinent to Healthy People objectives or a priority public health issue
  - Will still go through technical review, cognitive testing and field testing

Accepted Steps in Question Development

- Cognitive Testing

- Pilot testing
  - Reliability
    - Internal Consistency
    - Test-Retest reliability
    - Inter-rater reliability
  - Validity
    - Face validity
    - Content validity
    - Construct validity

Cognitive Testing

- Goal: Assess if respondents understand the question correctly and can provide accurate answers

- Does the question capture the scientific intent of the query and at the same time make sense to respondents?

- Uses qualitative study methods:
  - In-depth semi-structured interviews
  - Small purposive sample
  - Includes understanding what the person is thinking/feeling when responding

- Also useful if translating a validated question to a new language/culture/context

Cognitive Testing: Question-Response Process

	Cognitive Stage	Definition	Identified Errors/Problems
Stage 1	Comprehension	Respondent interprets question	Unknown/ambiguous terms Long/complex questions
Stage 2	Retrieval	Respondent recalls relevant information	Recall difficulty – length, complexity
Stage 3	Judgement	Respondent evaluates/estimates response	Estimation difficulty Sensitive/judgmental question
Stage 4	Response	Respondent provides information in the requested format	Incomplete response options Awkward format

https://www.cdc.gov/nchs/data/washington_group/meeting5/WG5_Appendix4.pdf

Example: Cognitive Testing of Physical Activity Questions

- Finger, JD et al., “How well do physical activity questions perform? A European cognitive testing study.” Archives of Public Health 2015; 73: 57-65.

- Study of the physical activity questions in the NHIS 62 people across 4 countries.

- Findings:
  - Overall the questions performed well
  - Problems understanding concepts of light, moderate and vigorous exercise
  - Problems recalling instances of “normal” activity (walking, sitting)
  - Problems calculating total duration of more than one activity

- Many publications talk about its use during the development of the questionnaire

Reliability

- Assesses the consistency of survey results
  - Internal consistency
  - Test-retest reliability
  - Inter-rater reliability

- Internal Consistency
  - Extent to which questions are consistent in measuring the same construct
    - Example: Patient Health Questionnaire (PHQ-9) has 9 questions measuring depressive symptoms
  - Split-half reliability: Divide the questions into two halves and administer them. See how well the answers are correlated.
  - Cronbach's alpha: Mean of all possible split-halves.
    - Alpha >0.7 is generally considered "adequate".

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5463570/

Reliability

- Test-retest reliability
  - Extent to which respondent’s answers remain consistent across multiple administrations
  - Usually tested using the Pearson’s correlation coefficient (Pearson’s r)
  - How much time in between?
    - Too short and they “learn” or “memorize” the questions
    - Too long and there might be actual changes that occur in between

- Inter-rater reliability
  - Extent to which multiple raters consistently evaluate the same questionnaire
  - Usually tested using the Kappa statistic
    - There are other more complicated measures of intra-class correlation

Validity

- Goal: Does the question measure what it is intended to measure?

- Face validity
  - How does the item “appear” to the respondent
  - Helps with PR and support for your survey

- Content Validity:
  - Do the questions measure the intended underlying construct?
  - Measured by experts in the field, more theoretically based, tested using things like the Content Validity Ratio or Content Validity Index
  - Example: Kim *et al*. (J Alt Comp Med, 2008) tested the content validity of expressions to describe sensations of the needle at different stages of acupuncture.

Validity

- Criterion Validity:
  - How does your question compare with a “gold standard” question or test.
  - Example: How does the PHQ-9 questionnaire compare with the “gold standard” of psychiatrist diagnosis?

- Construct Validity:
  - How well does the question associate with other variables or questions measuring the same construct?
  - Do the answers to your questions match well with questions on a related construct, and appropriately differ from questions measuring a separate construct?
    - PHQ-9 matches well with the Center for Epidemiologic Studies Depression Scale (CESD), but not with Social Responsiveness Scale (SRS) used to measure social impairment in autism-related disorders.
  - Less common outside of the social sciences

Source to Find Questions and Testing Information

- Q Bank (https://wwwn.cdc.gov/QBANK/Home.aspx)
- Established in 2002 by NCHS/CDC
- NOT “just” a database of “good” questions
- Provides:
  - Access to question testing/evaluation reports
  - Example questions and response options from a variety of surveys
  - Reports of issues/problems faced by investigators in specific situations, subpopulations, etc.

Next Class

- Explain the main population-based surveys done routinely in the US and how to access the publicly available data

- Understand basic survey sampling and design characteristics in terms of how they affect survey analysis

- Describe how to insert survey design and weighting information into Stata/R in preparation for analysis