Complex survey data: NHANES

Ehsan Karim

ehsan.karim@ubc.ca

When poll is active, respond at **PollEv.com/ehsank878**

Text **EHSANK878** to **22333** once to join

I am a

1st year SPPH PhD student

1st year SPPH PhD student, but a transfer from MSc

2nd year SPPH PhD student

SPPH MSc student

Non-SPPH student

None of the above, I got this zoom link from a friend

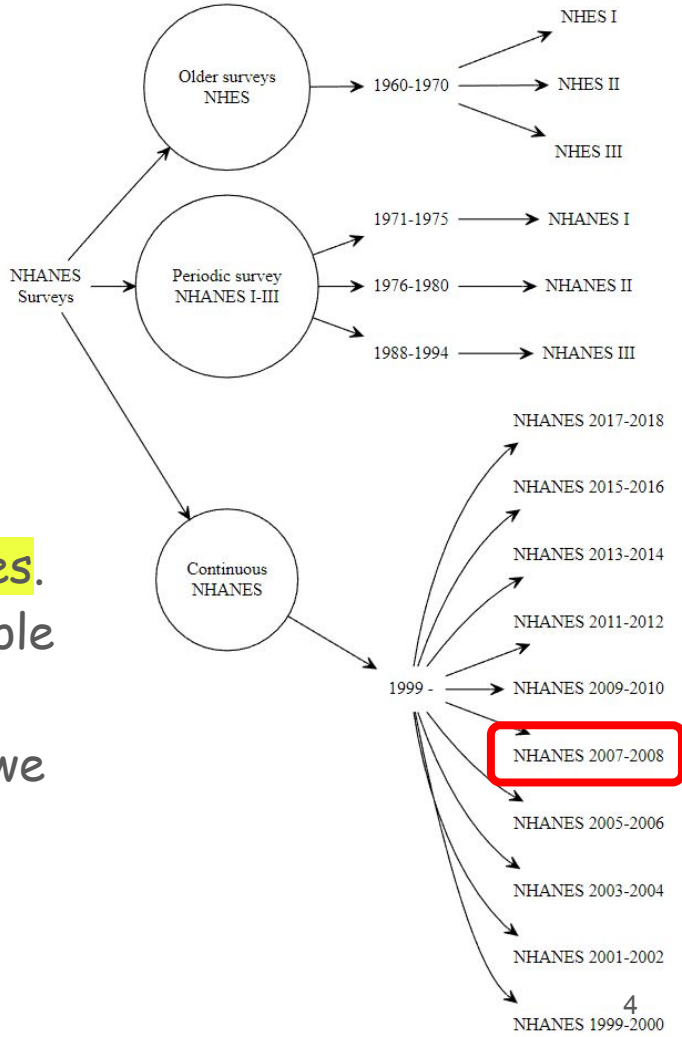
Some video and materials are posted on Canvas prior to this class. How much of it did you cover (read, watch) already?

0% (Wait, what is posted?)
1-20% (browsed a bit)
21-40% (some)
41-60% (near the half)
61-80% (most)
Almost everything (81-100%)

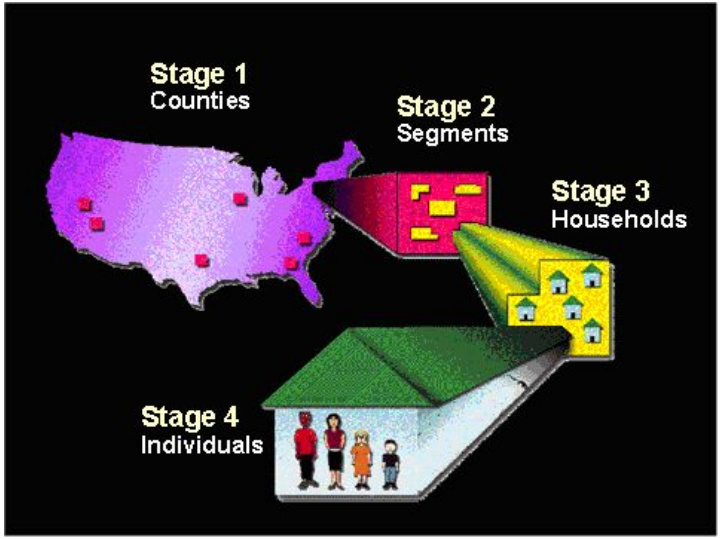
Dataset: Source

NHANES:

- survey of the adult, noninstitutionalized population of the United States
- We will use cycle of 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, but we will not use that in this example)



Dataset: Sampling Procedure



NHANES:

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

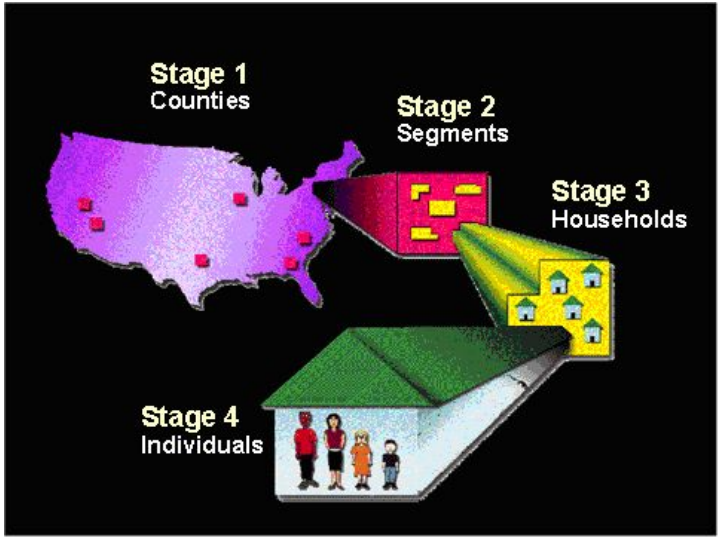
Stage 1:

- 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.).
  - Often strata is created to ensure capturing sub-population of interest
- The counties are randomly selected via PPS using a 2-per-stratum design.
  - PPS sampling = sampling units with larger populations are more likely to be selected than those with smaller populations.

Dataset: Sampling Procedure

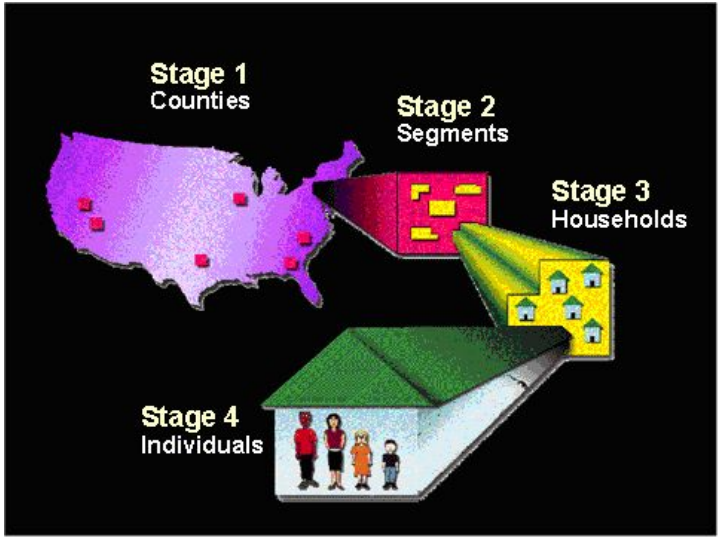
NHANES:

- 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.

Dataset: Sampling Procedure



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**

Dataset: Sampling Procedure

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

Dataset: interview / sample weight 

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.

Sampling design of NHANES is

Simple random sampling

Stratified sampling

Cluster sampling

Multistage sampling

What is NHANES?

Dataset: Survey features

NHANES (2007-08):

- Interview weight
- 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

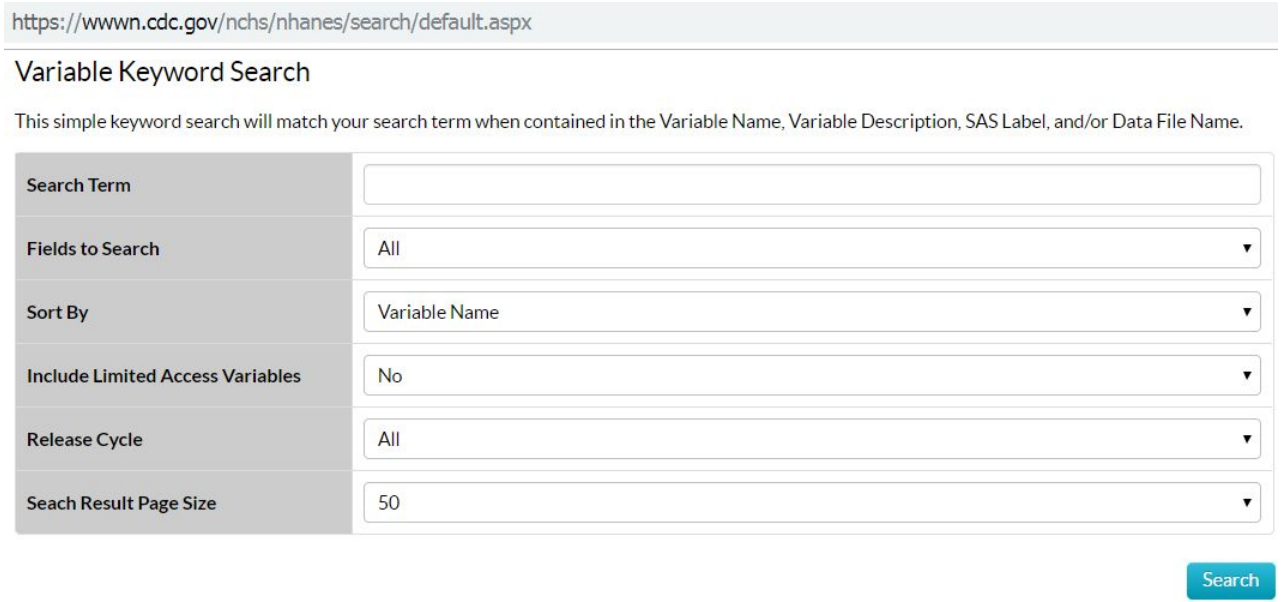
Target: Both males and females 0 YEARS - 150 YEARS

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

Dataset: variables

NHANES (2007-08):

See R logbook of how the analytic data was created from the workshop website



Illustrative example: Research question

Research Question: Whether or not adult patients with rheumatoid arthritis (RA) are at increased risk for heart attack (or myocardial infarction) in US.

Outcome (Y): heart attack (MI)

Exposure (A): rheumatoid arthritis (RA)

Comparison group: People without RA.

Exclusion criteria: Patients with

Osteoarthritis or other arthritis, young subjects (age < 20).

Dataset: variables

NHANES (2007-08):

MI,

RA,

age, BMI, diabetes, smoking, sex, race, education, marital status, income, origin, physical activity, access to medical services, hypertension/high blood pressure and diet

MCQ160E - Ever told you had heart attack

Variable Name: MCQ160E

SAS Label: Ever told you had heart attack

English Text: Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had a heart attack (also called myocardial infarction)?

Target: Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	282	282	
2	No	5642	5924	MCQ160F
7	Refused	0	5924	MCQ160F
9	Don't know	11	5935	MCQ160F
.	Missing	3731	9666	

MCQ190 Which type of arthritis

Variable Name: MCQ190

SAS Label: Which type of arthritis

English Text: Which type of arthritis was it

Target: Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Rheumatoid arthritis	346	346	
2	Osteoarthritis	531	877	
3	Other	219	1096	
7	Refused	1	1097	
9	Don't know	658	1755	
.	Missing	7911	9666	

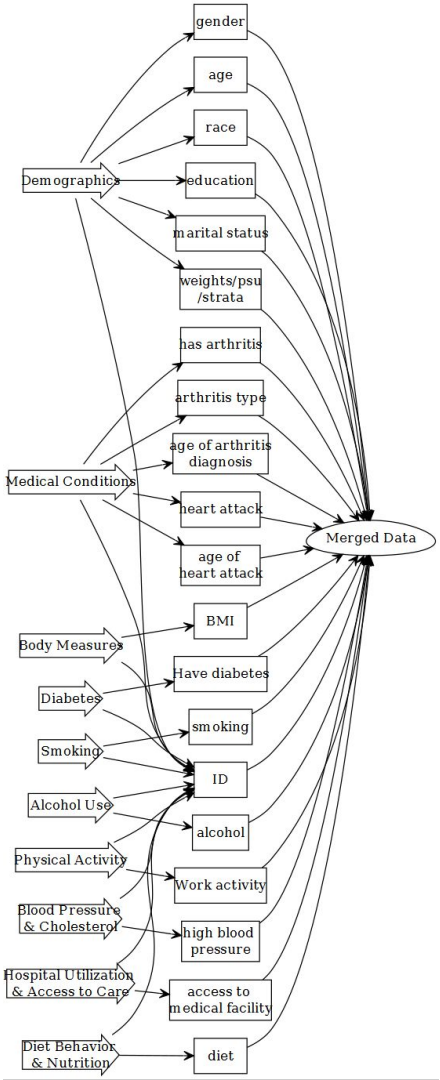
Dataset: variables

All of these variables are coming from different components that are connected with unique IDs

Codebook and Frequencies

SEQN - Respondent sequence number

Variable Name:	SEQN
SAS Label:	Respondent sequence number
English Text:	Respondent sequence number.
Target:	Both males and females 0 YEARS - 150 YEARS



Textbook List

- Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). Applied survey data analysis. Chapman and Hall/CRC.
- 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/workshop