

# Sampling and Inference

---

**Michel Boudreaux**

Graduate Research Assistant, SHADAC

SHADAC Data User Workshop

Minneapolis, MN

October 22, 2010

# Acknowledgements

---

- Large portions of this presentation were taken from lectures notes developed by Davern and Johnson

# Overview

---

- Intuition for design and analysis of complex surveys
- Knowing when to trust an estimate
- Data processing
- Data Documentation

# What is Statistics?

---

- **Summarize Information**
  - Means, Medians, Modes
- **Make inferences about a population from sample data**
  - What decision should I make with this information?

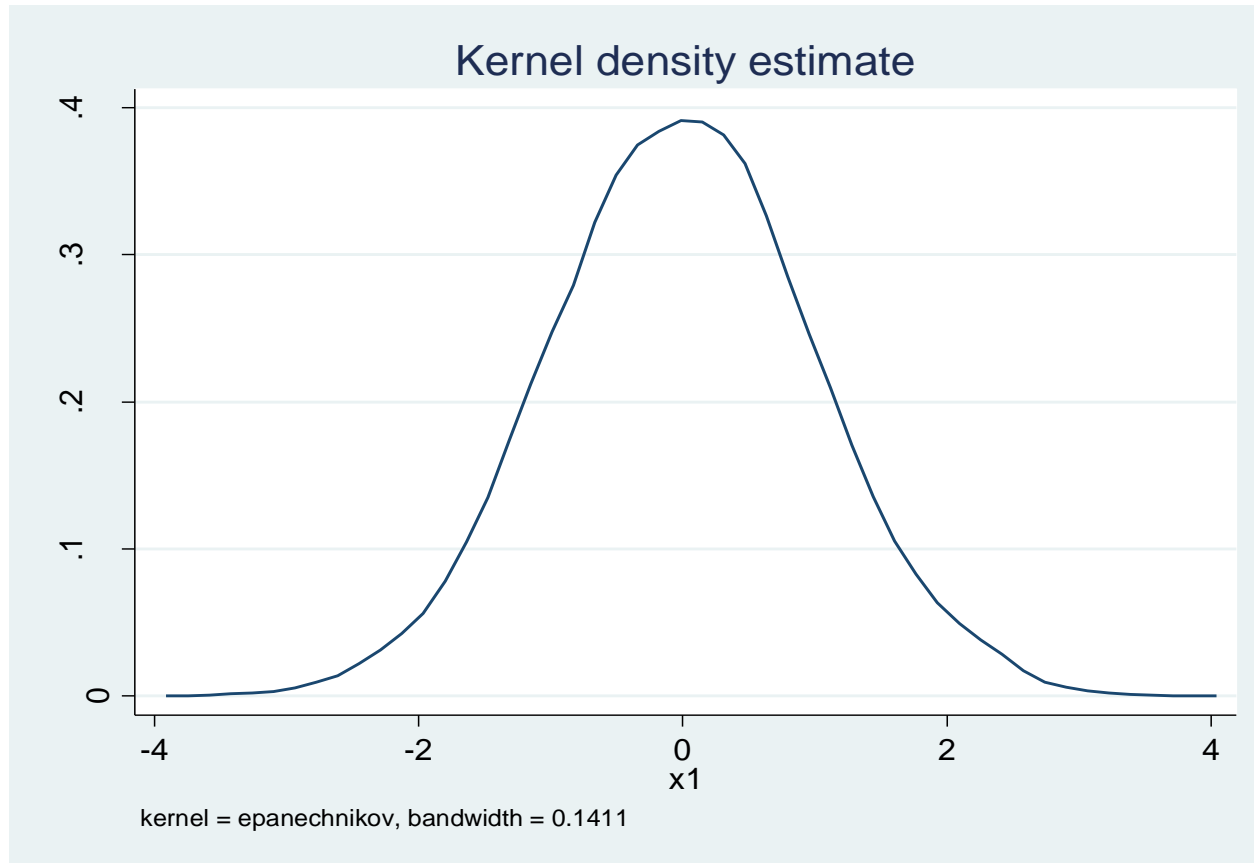
# General Sampling Terms

---

- **Probability Sample**
  - Every element in the population has a non-zero chance of being in the sample
- **Sampling Error**
  - The difference between a sample statistic and the true population parameter.
- **Standard Error**
  - A measure of sampling error

# Hypothetical Sample Distribution

---



# Sampling Terms

---

- **Sample Frame**
  - A list of elements in the population
  - Example: The Master Address File
- **Coverage**
  - The agreement of the sample frame and the target population
- **Coverage Error**
  - Disagreement of sample frame and target population
  - Either missing elements or including the wrong elements

# POPULATION

## SAMPLE FRAME

### SAMPLE

### DATA

Non-Respondents

Respondents

**Coverage Error:** When the sample frame misses part of the population or includes elements outside of the target population.

# Simple Random Samples

---

- Standard statistical methods are based on simple random samples
  - Given a complete list of the population you randomly choose a desired number of elements
- Each observation is ‘independent’
- Each new observation adds another unit of information

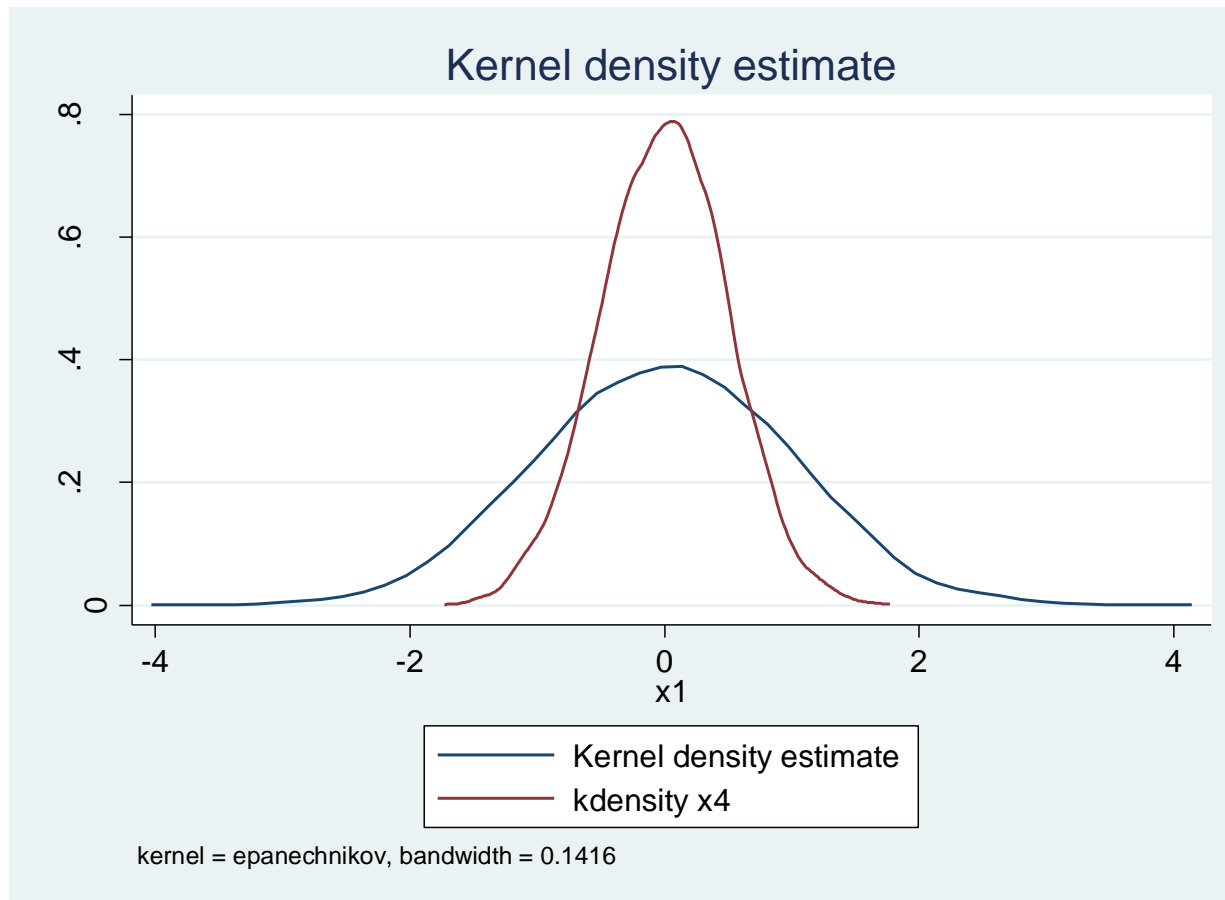
# Complex Sample Design

---

- Population Surveys are based on complex sample designs
- Standard methods don't work
  - Observations are not independent
  - **If you use SRS methods you will get the wrong answer and make the wrong decision.**
- Complex samples have larger standard errors
  - They contain less information per observation

# What A larger Standard Error Looks Like

---



# So Why Use Complex Samples?

---

- Cheaper data collection costs
  - Less interviewer travel
- Increase confidence in rare subgroups by oversampling
- Correct for non-response by using weights.
- Some information can only be obtained by a complex sample
  - Family dynamics

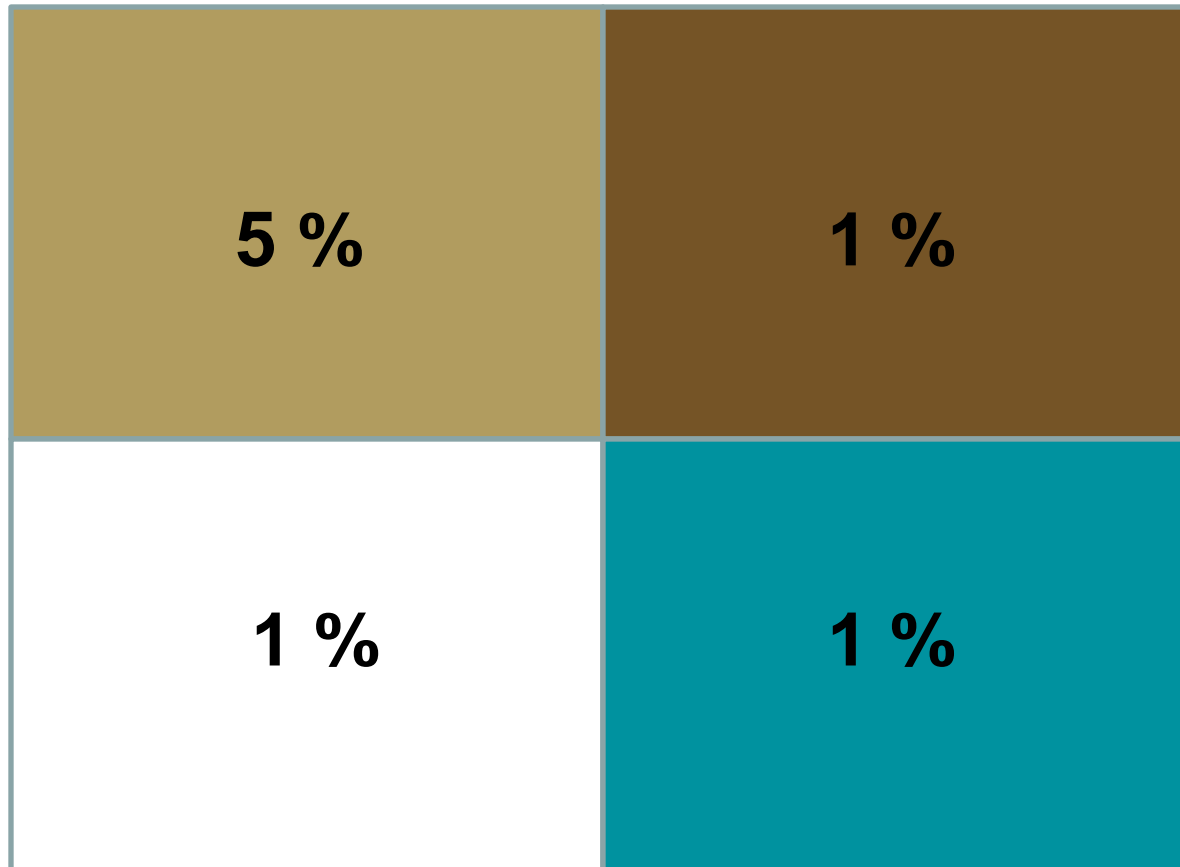
# Complex Sample Concepts

---

- **Stratification**
  - Divide population into homogenous groups and select more units from a certain group
- **Clustering**
  - Natural grouping of sampled elements
    - Households, city blocks, clinics, etc.
  - Some clusters chose at the exclusion of others
  - Observations inside a cluster are correlated
    - Examples?

# Stratification

---



# Cluster

---


# Complex Sample Concepts

---

- **Multi-stage**
  - Often complex samples are drawn in several stages
  - Randomly select a cluster (call it a PSU)
  - Randomly select an individual in each cluster (call it the USU)
  - Generally, you don't need to know how many stages there are.

# More Complex Sample Concepts

---

- **Weights**

- Usually want to know how many units in the population have a characteristic, not how many units in your sample
- Usually the characteristics of the sample do not match characteristics of the population
  - Non-response
  - Oversampling
- If you don't use weights you won't get a population estimate, you will get a sample characteristic – and that misses the point.

# Analyzing Survey Data

---

- **Taylor Series**
  - Uses explicit information about the sample design
  - This information not available in Census data, so we proxy sampling information using available geography.
- **Replicate Weights**
  - Make several estimates using different weights and use the spread to estimate the standard error
  - The weights contain all the sample design information
- **Variance Parameters**
  - Data producer provides generic variance inflators that analysts use to adjust SRS estimates.

# What Can Go Wrong?

**Table 1. State health insurance coverage rates and standard error computation comparisons by year: 2001**

	2001 health insurance coverage estimate (%)	Survey design-based standard error on internal census file (%)	Ratio of method to survey design-based on internal census file			
			Robust	Survey design-based on the public use file	Generalized variance estimation	Simple random sample (SRS)
United States	85.4	.18	.52	.77	.39	.42
Alabama	86.9	1.20	.53	.72	.43	.48
Alaska	84.3	1.11	.62	.87	.48	.57
Arizona	82.1	1.52	.47	.79	.41	.46

# Pros and Cons

---

- Replicate Weights
  - Pro: Gold standard
  - Pro: Can drop observations in your sample
  - Con: Can take very long time to run
  - Con: Only one package automates (Stata 11)
- Taylor Series
  - Pro: Near Gold
  - Pro: Faster processing
  - Con: Can only drop observations in special circumstances

# Rule of Thumb: Only Drop Cases By State When Using Taylor Series

Health Insurance Rates, Standard Error Methods and Bias, CPS 2009								
	Full Population Subpop by state or drop				Children < 100% Poverty Subpop by POV or Drop			
	%	S.E. (SUBPOP)	S.E. (DROP)	Ratio	%	S.E. (SUBPOP)	S.E. (DROP)	Ratio
Alabama	88.1	0.96	0.96	1	91.5	3.17	2.91	0.92
Alaska	80.2	1.3	1.3	1	74.5	7.96	8.04	1.01
Arizona	80.5	1.26	1.26	1	77.8	5.2	5.21	1
Arkansas	82.2	1.21	1.21	1	87.3	3.66	3.68	1.01
California	81.4	0.44	0.44	1	84.3	1.65	UNDEF	
Colorado	84.1	0.83	0.83	1	71.4	4.77	4.62	0.97

# When to Trust the Estimate

---

- You can always make an estimate, but you should know when not to trust it
- These are called suppression rules
- **2 rules of thumb**
  - The estimate is based on 50 cases or fewer
  - The ratio of the standard error to the mean is more than 0.5
- This tells you when the sampling error is too high, non-sampling error is much harder

# Dealing with Imprecision

---

- In ACS, pooled data files are released for 3 and 5 years of data collection.
  - $n_1 = 4.5$  mil;  $n_3 = 13.6$  mil;  $n_5 = 22.5$  mil
  - Period Estimates: Experience on an average day in period
  - American FactFinder will provide data to Tract level in 5 year release
  - PUMS file will be limited to PUMA
  - Health Insurance Estimates are on alternate release schedule.

# More Dealing With Imprecision

---

- In CPS, Census recommends averaging estimates from 2-3 Years.
- Should let your own suppression rules guide when to use averages
  - Generally: subgroups within states, and state-specific differences across time
- The point estimate is just the straight average
- The standard error is more complex, because the CPS uses a rotating panel sample design
  - ~30% of the sample is the same in one year to the next.

# Standard Errors for Averaged CPS

---

- S.E. for a 2 year average is:

$$S.E._{A,B} = \frac{SQRT(S.E._A^2 + S.E._B^2 + r * S.E._A * S.E._B)}{2}$$

- S.E. for the difference of two non-overlapping 2-year averages:

$$S.E._{A-B} = SQRT(S.E._{A,B}^2 + S.E._{C,D}^2 - \frac{1}{2} * r * S.E._B * S.E._C)$$

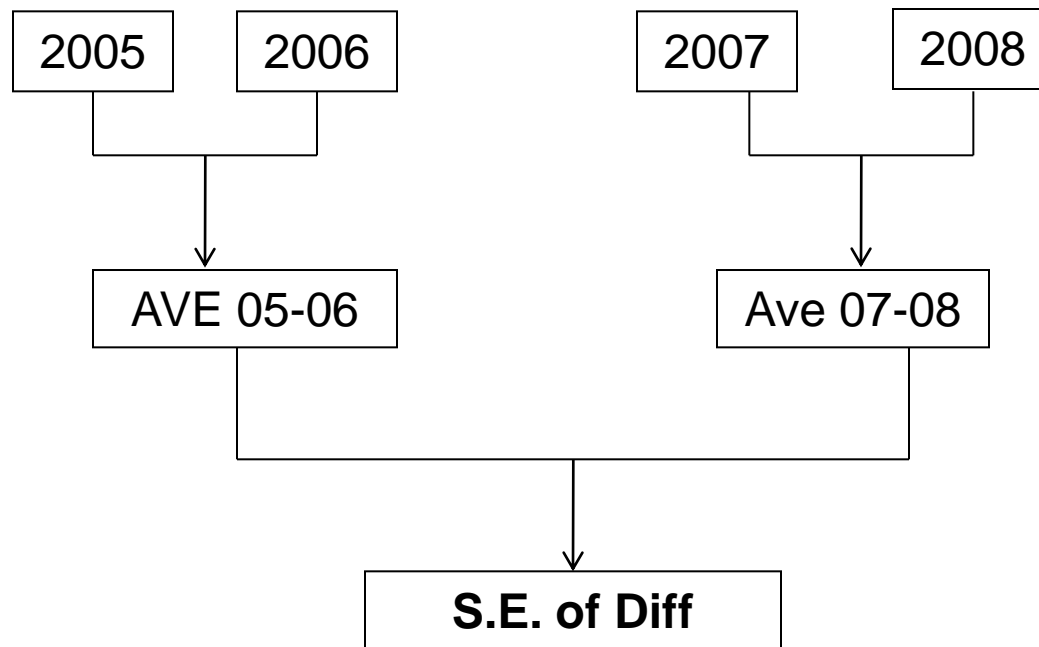
- Where  $r$  = Census provided correlation coefficient

## Intuition

- Average data for overlapping samples: penalize ourselves (+  $r$ ).
  - Now we have 2 clusters households *and* people by year
- Difference of overlapping samples: reward ourselves (-  $r$ ).
  - When the same person is different in two periods we *know the difference is real* and not sampling error.

# What it Looks Like

---



# Be Excited About ACS!

**Table 4 (Supplement). Comparison of ACS and CPS Sample Size and Uninsurance Standard Error, Non-Elderly U.S. Civilian Non-institutional Population (Obs in thousands)**

	ACS		1-Year CPS-ASEC			2-Year CPS-ASEC			3-Year CPS-ASEC		
	Obs	S.E.	Obs	S.E.	Ratio of ACS to CPS S.E.	Obs	S.E.	Ratio of ACS to CPS S.E.	Obs	S.E.	Ratio of ACS to CPS S.E.
United States	2,489	0.06	187	0.15	0.37	372	0.12	0.48	558	0.11	0.54
Alabama	38	0.35	2	1.34	0.26	4	0.92	0.38	6	0.83	0.43
Alaska	6	1.02	2	1.86	0.55	5	1.22	0.84	8	0.96	1.06
Arizona	51	0.32	2	1.48	0.22	5	1.19	0.27	8	1.02	0.31

# Sampling Review

---

- Demographic surveys, like the CPS and ACS and based on complex samples
- You must take this into account when analyzing data or you will reach the wrong conclusions.
- Be cautious when drawing inferences from estimates supported by small cell sizes or when the standard error is half the size of the mean.

# Compiling Datasets

---

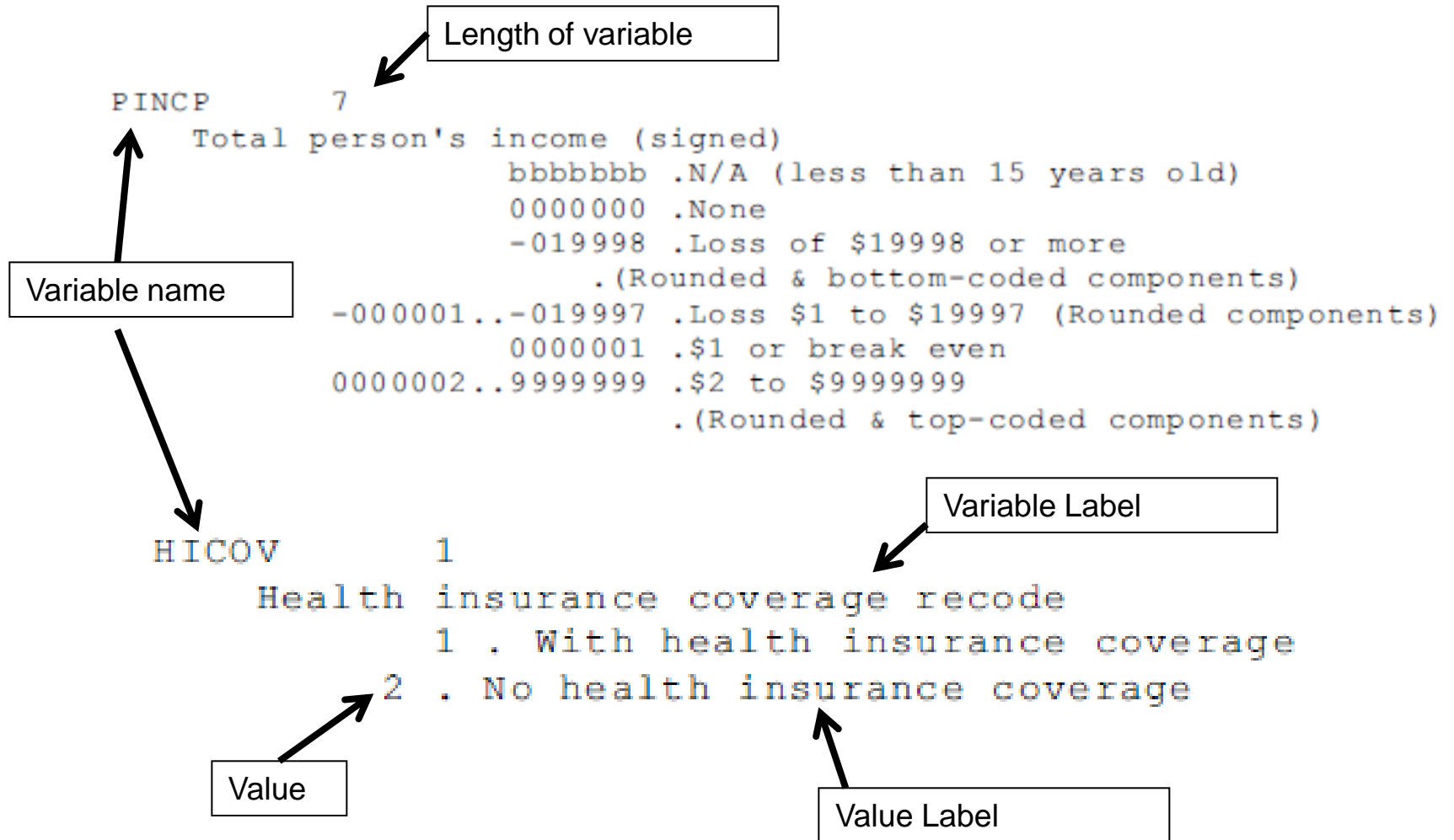
- Before Census releases a product it cleans the data (Thank you Census!!)
  - Imputation
    - Fill in missing data
  - Editing
    - Derive variables that people usually don't know and make responses as internally consistent as possible.
  - Disclosure Avoidance
    - Protect confidentiality
- Data users should be aware of these processes
  - be skeptical, but not cynical

# Learning About Your Data

---

- Every dataset comes with documentation
  - Accuracy Statement
  - Data Dictionary
  - Sample Code (merging, formatting)
- Documentation is hidden, confusing, incomplete, and frustrating
- IPUMS is an example of good documentation

# Reading ACS Data Dictionary



# Reading CPS Data Dictionary

---

name	length	position	range
D SS-YN	1	290	(0:2)
Item 56b - Did ... receive s.s.?			
U P-STAT =	1 or 2		
V	0	.Not in universe	
V	1	.Yes	
V	2	.No	

Who received the question?

# Where to Find Help

---

- Google
- Colleagues
- SHADAC
- Census

# Contact information

---

- Michel Boudreaux, State Health Access Data Assistance Center (SHADAC)
  - boudr019@umn.edu
  - 612-626-1640

