California Environmental Contaminant Biomonitoring Program (CECBP)

Randy Curtin, Ph.D.
National Center for Health Statistics
Centers for Disease Control and Prevention

Sample Design
Concepts, Issues and Options
10-24-2008

Presentation Overview

Introduction to Sample Design
Designing a State-wide Study
Designing a Community Study
Questions and Discussion

Is a probability sample really needed?

Non Probability or Convenience Samples:

- Maybe cheaper, easier to conduct
- Internal Validity

- Results may not be generalized
  - Specific populations excluded
  - Results subject to selection bias
- Potential bias cannot be ascertained

Probability samples

Can control some aspects of selection

Randomness and sample size protect against non-controlled aspects of the design

Results can be generalized (unbiased)

Non-response bias can be ascertained

Coverage of entire target population

Valid statistical inference and comparisons

Scientific scrutiny of results

Sample Design is part of the overall Total Study Design

Choice of Population to Survey

Choice of Survey Objectives and Mode

Sample Design to meet objectives/budget

Measurement Design

Design of Control System

Design of Evaluation System

Basic Steps in Any Sample Design

- Study objectives (content, how to measure)
- Target population(s)
- Analytic Sample size
  - Statistics, Analysis
  - Required Precision, Power
- Design options
  - Data Collection mode
  - Cost and Variance components
- Practical/operational issues
- Budget Restrictions

Characteristics of a Sample Design

Controlled Selection/Multi-stage

- Area Frame: Stratification/Clusters
- Density stratification of sample units
- Screening - “over-sampling”

Sample Size/allocation

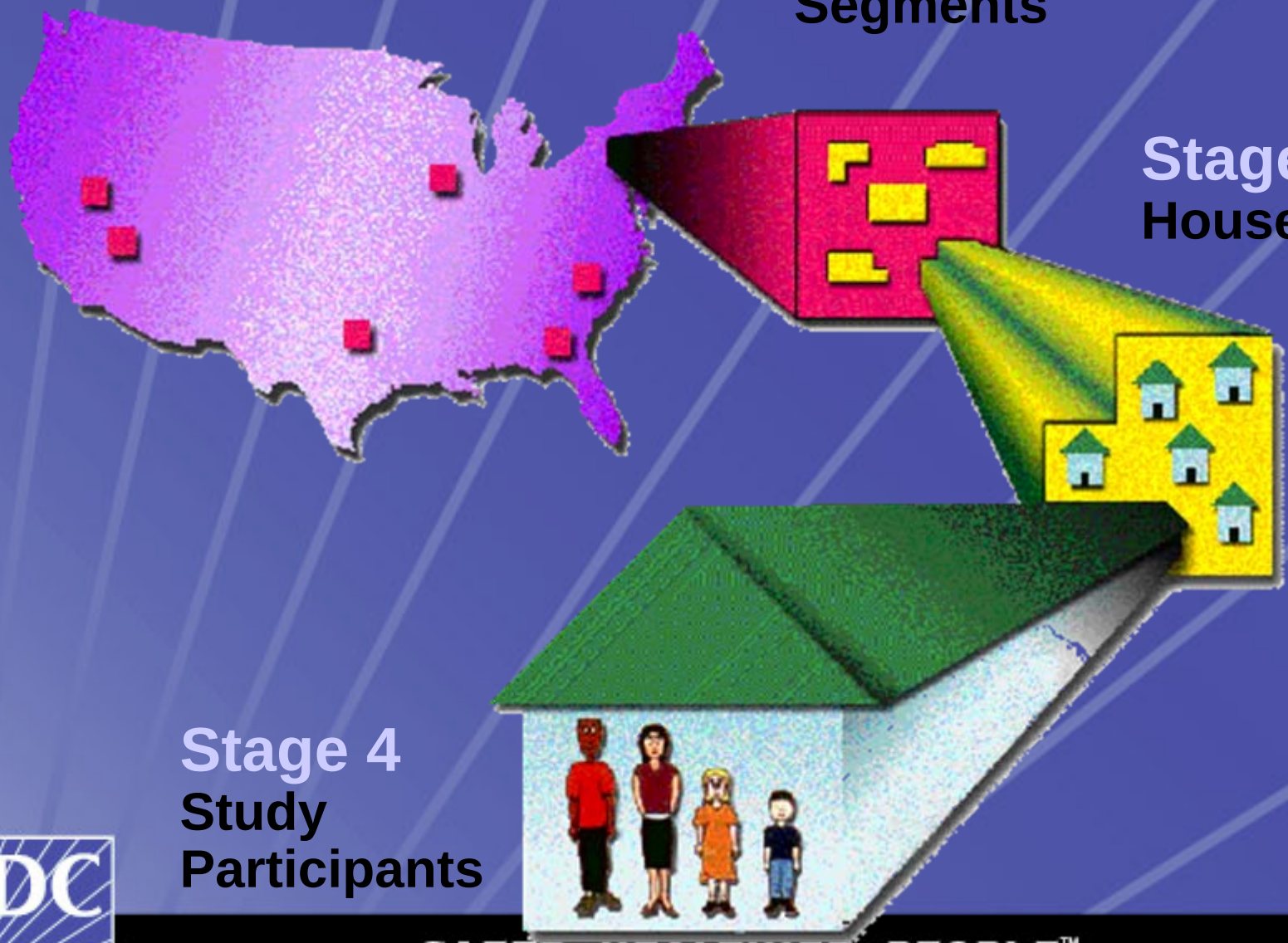
sample units, size of units, persons

All selections (stages) at random

Stage 1
Counties

Stage 2
Segments

Stage 3
Households



Stage 4
Study Participants

The Sample Design Trade-off: Cost versus Statistical Considerations

Precision – to get better precision (decrease variance) typically costs more

- Larger sample size/More areas
- Larger or Smaller cluster sizes
- Yes or No - Differential sampling

Reliability – to measure and improve reliability typically costs more

Fixed budgets tend to fix sample size/content

Two Terms of Immediate Interest

Relative Standard Error (RSE)

Standard error of mean / mean

Coefficient of Variation (CV)

Variance of mean / mean^2

Design Effect (DEFF or VIF)

$DEFF = Var_{complex}/Var_{srs}$

Weights and Clustering

$DEFF = (1 + CV^2_{wts}) * (1 + (m-1)p)$

Effective sample Size

$n_e = n/DEFF$

$100 = 150/1.5$

Design Effects (means) for Total Population, NHANES 99-00

Glucose, serum	2.24	(RSE=0.37)
Creatinine. serum	2.77	(RSE = 0.67)
Total Cholesterol	3.42	(RSE = 0.46)
C-reactive protein	4.49	(RSE = 3.58)
Creatinine, urine	5.45	(RSE = 1.69)
Measles Antibody	8.01	(RSE = 2.69)
Blood Lead	9.50	(RSE = 2.28)
Total Mercury	10.59	(RSE = 8.05)
Calcium	25.63	(RSE = 0.29)
Chloride	34.10	(RSE = 0.22)

Design effects by age

	Var 1	Var 2
All Ages	2.6	9.5
0 – 4	1.3	1.7
5 – 11	1.1	1.5
12 – 19	1.4	1.7
20 - 39	1.2	1.9
40 – 59	1.1	2.1
60 +	1.5	2.6

How to do “Sample” size

Statistic	Proportion = 10 %
Reliability	RSE < 30 % CI = (4,16)
Analytic Sample	100 (SRS or n_e)
DEFF (1.5)	150 (respondents or n_R)
RR (.75)	200 (Selected n_s)
Domains (K)	200*K Total sample or n_t

Multiple Objective Studies

One size does not fit all

Variability in required sample sizes

Design efficient overall, not specific

- Different stratification (geographic) variables
- Efficient for one implies inefficient for other

Single survey versus continuous survey

- More than one year required for objective
- More than one year for demographic detail

Study Design for Multiple Objective Surveys

State all objectives

Translate into statistical measures/sample size

Determine most demanding for sample size

Determine efficient design for key variables

Estimate cost/budget

Implement or revise (decrease sample/cost)

- Change statistical/subdomain requirements
- Drop components/objectives
- Change time line (number of years needed)

Questions?

Getting Started on the CECBP Sample Design

Goals and objectives for CECBP

Things to consider

Characteristics of California

Operations and Maximum Capacity

A sample of a sample design for a State study

Goals and Objectives for CECBP

The CECBP is authorized under State law.

It will collect information on environmental exposures and health outcomes for the population of California. The study population will be a representative sample with respect to age, race/ethnicity and income. Collection of information will include interviews, examinations and biological specimens.

Things to Consider

Target Population

Subdomains of interest

Statistical precision required

Budget limitations on sample size

Potential Stratification Variables

Alternative Frames

Field work – mapping/listing

Race/ethnicity for California (Equal Probability Sample of 2,000)

Race/ethnic	Cal %	US %	sample
White	77.0	80.2	1,540
- Hisp/Latino	35.2	14.4	704
- White, Non-Hisp	43.5	66.9	870
Black	6.7	12.8	133
Asian	12.2	4.3	244
Am Ind/Alaska Native (AIAN)	1.2	1.0	24
Hawaiian	0.4	0.2	8
Multi-racial	2.4	1.5	48

Percent Distribution: US Population compared to NHANES sample

Group	Population	Sample
Black	12.1	22.9
Mexican American	7.8	31.5
Less than 20 years	29.5	47.3
20-39 years	30.4	17.5
40-59 years	24.8	14.7
60 years and over	15.4	20.4

Pros/Cons when over-sampling

"Drives" the design towards density clusters – geographic, urban "bias"

Increased DEFF for "total"

Increased cost (for fixed total sample size)

Better precision for subdomains

Testing for disparities

Example of Disproportionate sample

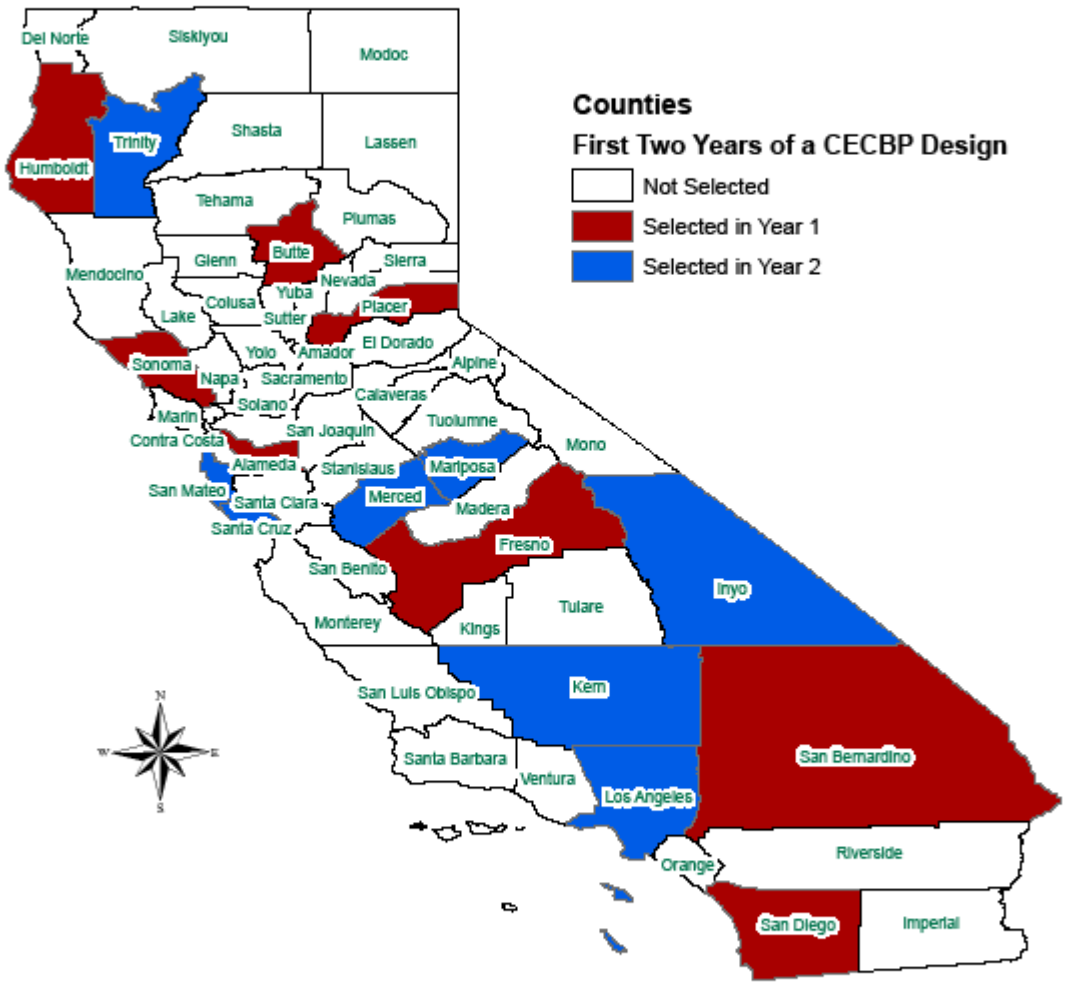
Race/ethnic	Cal %	equal	unequal	weight
Hisp	35.2	704	500	25,466
White, NonHisp	43.5	870	600	26,380
Black	6.7	133	400	6,085
Asian	12.2	244	400	11,060
Am Ind	1.2	24	100	4,260

Maximum Sample Capacity

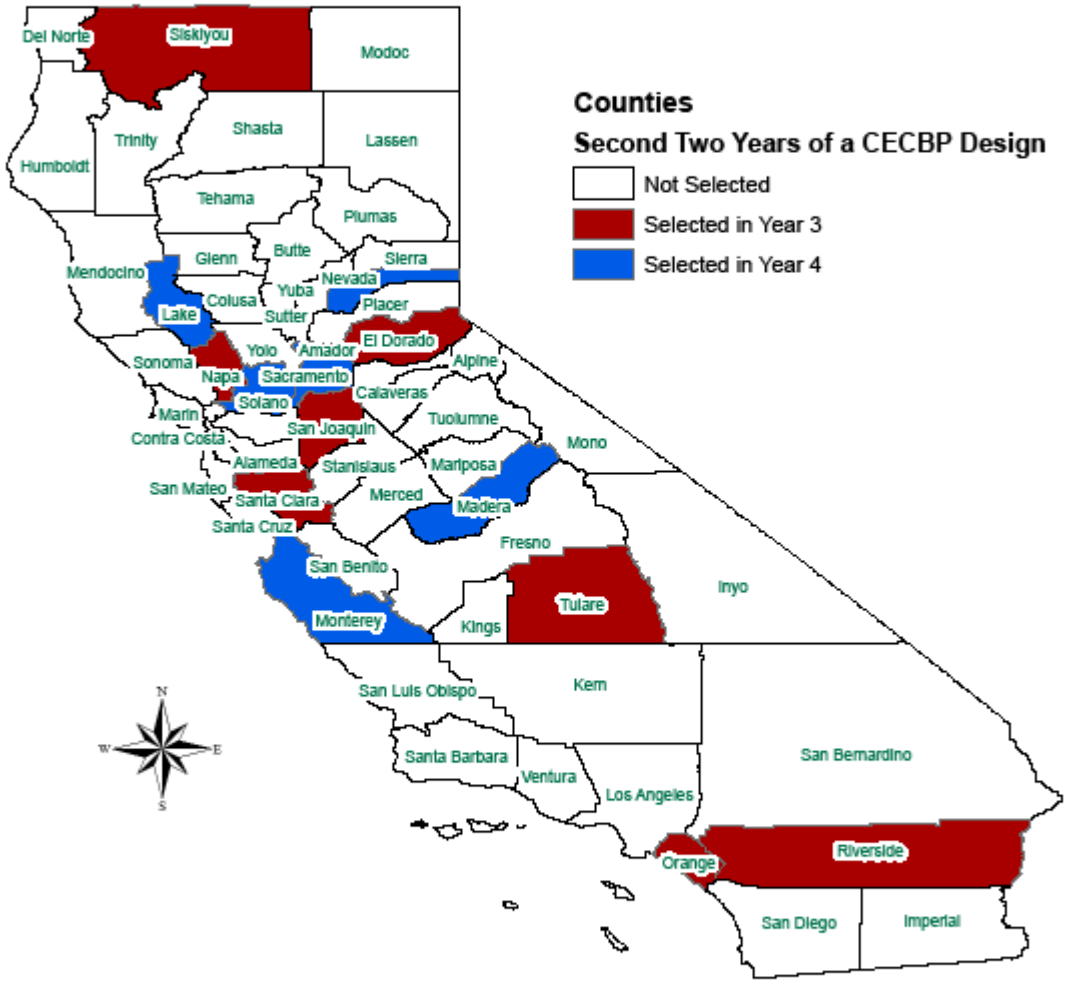
50 working weeks per year
2 weeks down time between Areas/Communities/PSU
5 working days per week
12 persons maximum per day
Equal allocation over PSUs

Example: 8 PSUs per year, 2 years for a sample
34 weeks for data collection
6 /4 weeks per PSU
2,000 max Sample persons per year

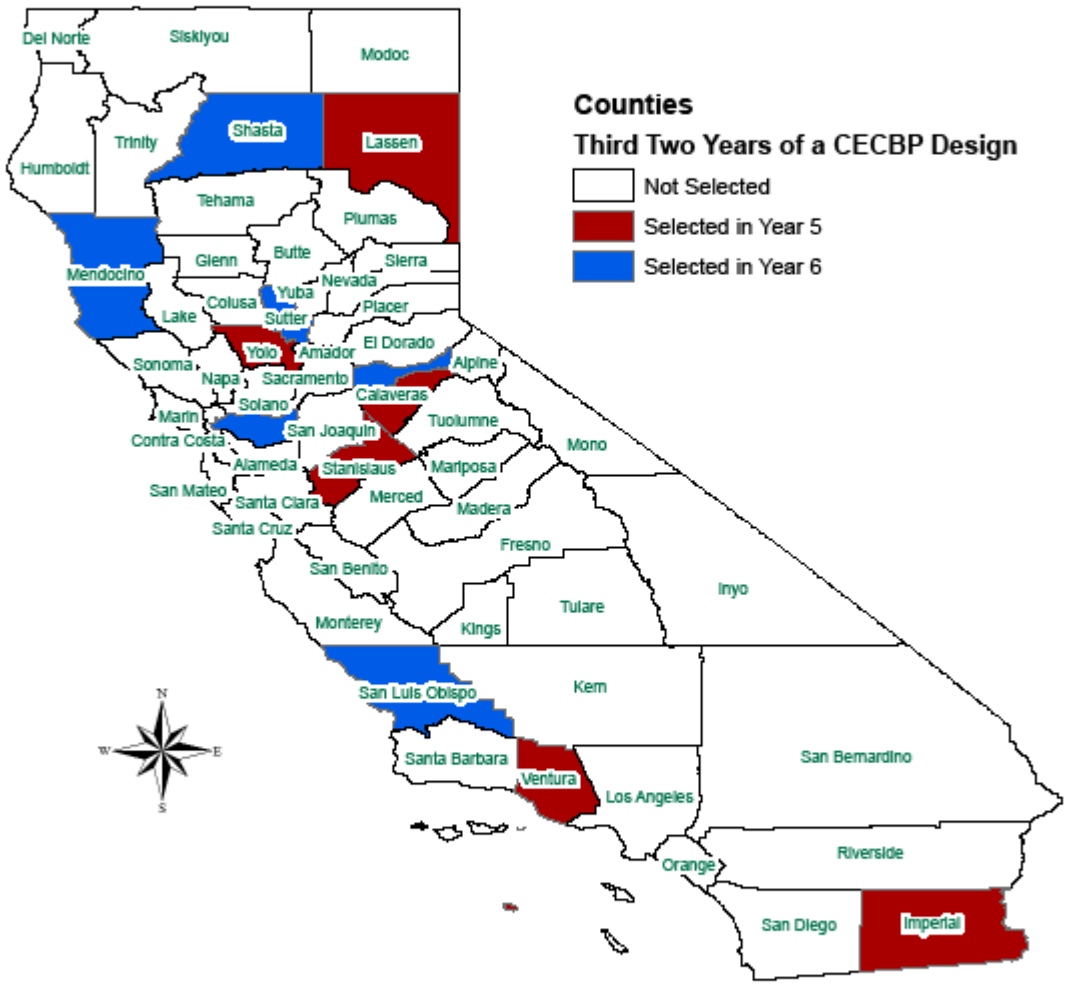
California


Counties
First Two Years of a CECBP Design
Not Selected
Selected in Year 1
Selected in Year 2
Del Norte
Siskiyou
Modoc
Trinity
Shasta
Lassen
Humboldt
Tehama
Plumas
Butte
Glenn
Sierra
Mendocino
Nevada
Yuba
Colusa
Placer
Lake
Sutter
El Dorado
Yolo
Amador
Sonoma
Alpine
Napa
Sacramento
Calaveras
Solano
Marin
Tuolumne
Contra Costa
San Joaquin
Mono
Alameda
Stanislaus
Mariposa
San Mateo
Merced
Santa Clara
Madera
Santa Cruz
Fresno
San Benito
Inyo
Tulare
Monterey
Kings
Kern
San Luis Obispo
San Bernardino
Santa Barbara
Ventura
Los Angeles
Riverside
Orange
San Diego
Imperial

California


Counties
Second Two Years of a CECBP Design
Not Selected
Selected in Year 3
Selected in Year 4
Del Norte
Siskiyou
Modoc
Trinity
Shasta
Lassen
Humboldt
Tehama
Plumas
Mendocino
Glenn
Butte
Sierra
Nevada
Colusa
Yuba
Placer
Lake
Sutter
El Dorado
Yolo
Amador
Sonoma
Napa
Sacramento
Alpine
Solano
Calaveras
Marin
Tuolumne
Contra Costa
San Joaquin
Mono
Alameda
Stanislaus
Mariposa
San Mateo
Santa Clara
Merced
Madera
Santa Cruz
Fresno
San Benito
Tulare
Inyo
Monterey
Kings
San Luis Obispo
Kern
Santa Barbara
Ventura
San Bernardino
Los Angeles
Riverside
Orange
San Diego
Imperial
N
W
E
S

California


Counties
Third Two Years of a CECBP Design
Not Selected
Selected in Year 5
Selected in Year 6
Del Norte
Siskiyou
Modoc
Humboldt
Trinity
Shasta
Lassen
Tehama
Plumas
Mendocino
Glenn
Butte
Sierra
Nevada
Colusa
Yuba
Placer
Lake
Sutter
Yolo
Amador
El Dorado
Sonoma
Napa
Sacramento
Alpine
Calaveras
Solano
Marin
Tuolumne
San Joaquin
Contra Costa
Mono
Alameda
Stanislaus
Mariposa
San Mateo
Santa Clara
Merced
Madera
Santa Cruz
Fresno
San Benito
Inyo
Tulare
Monterey
Kings
Kern
San Luis Obispo
Santa Barbara
Ventura
San Bernardino
Los Angeles
Riverside
Orange
San Diego
Imperial
N
W
E
S

California


Del Norte
Siskiyou
Modoc
Trinity
Shasta
Lassen
Humboldt
Tehama
Plumas
Glenn
Butte
Sierra
Mendocino
Nevada
Yuba
Colusa
Placer
Lake
Sutter
El Dorado
Yolo
Amador
Alpine
Sonoma
Napa
Sacramento
Calaveras
Solano
Marin
Tuolumne
San Joaquin
Contra Costa
Mono
Alameda
Stanislaus
Mariposa
San Mateo
Merced
Santa Clara
Madera
Santa Cruz
Fresno
San Benito
Inyo
Tulare
Monterey
Kings
Kern
San Luis Obispo
San Bernardino
Santa Barbara
Ventura
Los Angeles
Riverside
Orange
San Diego
Imperial
N
W
E
S
Counties
Counties in a 6-Year CECBP Design
Not Selected
Selected

Cumulating sample over time

Cumulating sample yields a larger total sample size

Some events may be “rare” and require larger sample size

May desire very detailed race/ethnic/age subdomains

Prevalence for subdomain versus statistical power to detect “differences”

Questions?

Design Issues for Communities

- How to define a community
- How to select community
- In a State design, communities are selected at random and represent other similar communities
- If Community is a “PSU” then a community design is simply the within PSU design in the State Sample

For a Community or a Within PSU Selection

Define Smaller areas or Segments (SSU)

Number of Segments per Community or PSU

- Segment size - Number of Households/Persons

Stratification of Segments

Number of segments selected

Number HH per segment

Number Persons per Household

An example of sample size by stage for NHANES 1999-2000

Number of PSU	26
Number of Stands	27
Number of Segments	681
Number HH Screened	22,839
Number HH with identified SP	6,005
Number selected SP's	12,160
Number interviewed (82%)	9,965
Number completing MEC (76%)	9,282

Example of segment size

Segment is a group of Census Blocks

100 households per segment

Select 25 households to be screened

Average between 6 and 7 Households with at least one sample person

Average about 14 people per segment

Impact of Fixed to Variable costs

Suppose fixed costs of $4M and variable costs of $1,000 per person

Then $5M to collect n = 1,000
And $6M to collect n = 2,000

Good for planning: An increase in 16% budget increases sample size by 100%

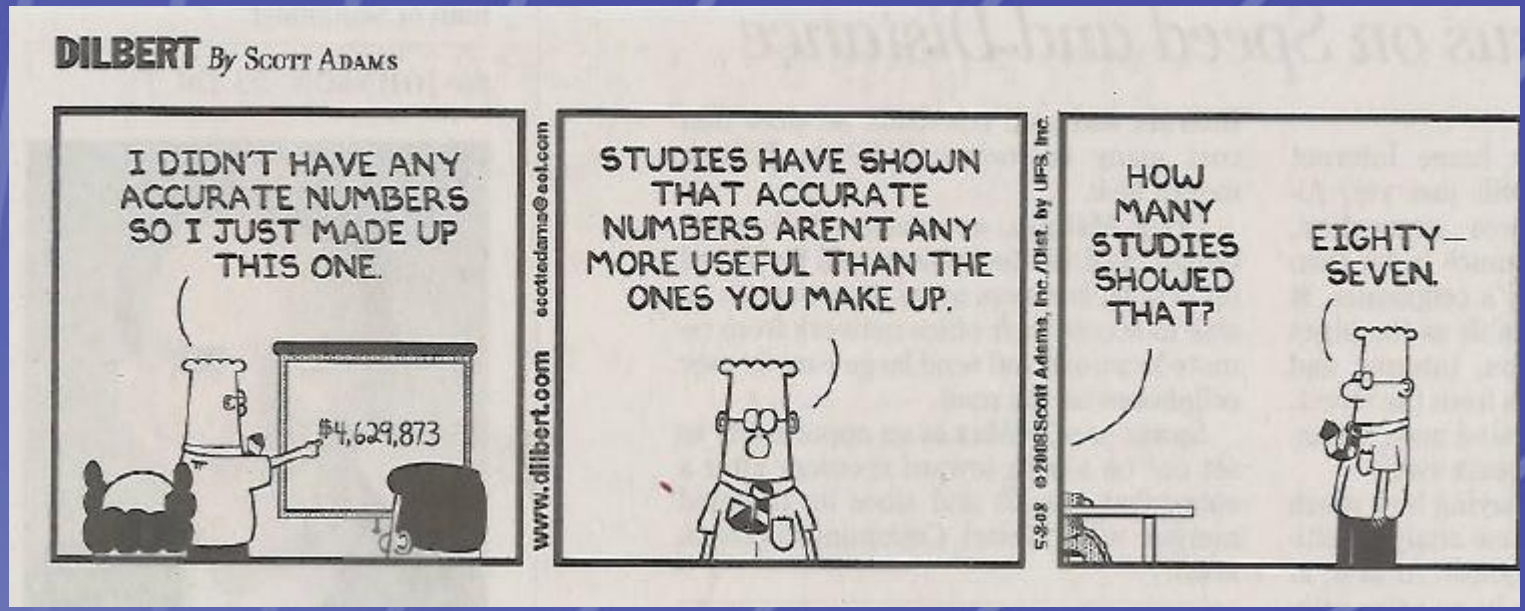
Bad for budget cuts: To cut budget by 16% have to cut sample by 50%

Final Design Parameters required

Size/magnitude of prevalence estimate (5%, 10%, 15%)
Desired Level of Statistical Reliability (RSE = 20, 30 %)
Subdomains/Communities of Interest
- Demographic (Age, Race/Ethnic, Sex)
- Geographic (urban/rural, specific counties)

Comparisons needed (tests of statistical significance)

Variance inflation factors/ Design Effects by stage
Geographic coverage/stratification
Operational/practical considerations


DILBERT By Scott Adams
I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

End of Presentation

- Questions?