



The Hows and Whys of SAS/STAT Survey Procedures




SAS User Association of Victoria, May 2008
Presented by Tim Trussell, SAS Canada

**THE
POWER
TO KNOW®**

Objectives

- Explain the reasons why the survey suite is needed.
- Examine the correction of the survey procedures in a NHANES case study.

Recall From Fall (Survey Select)



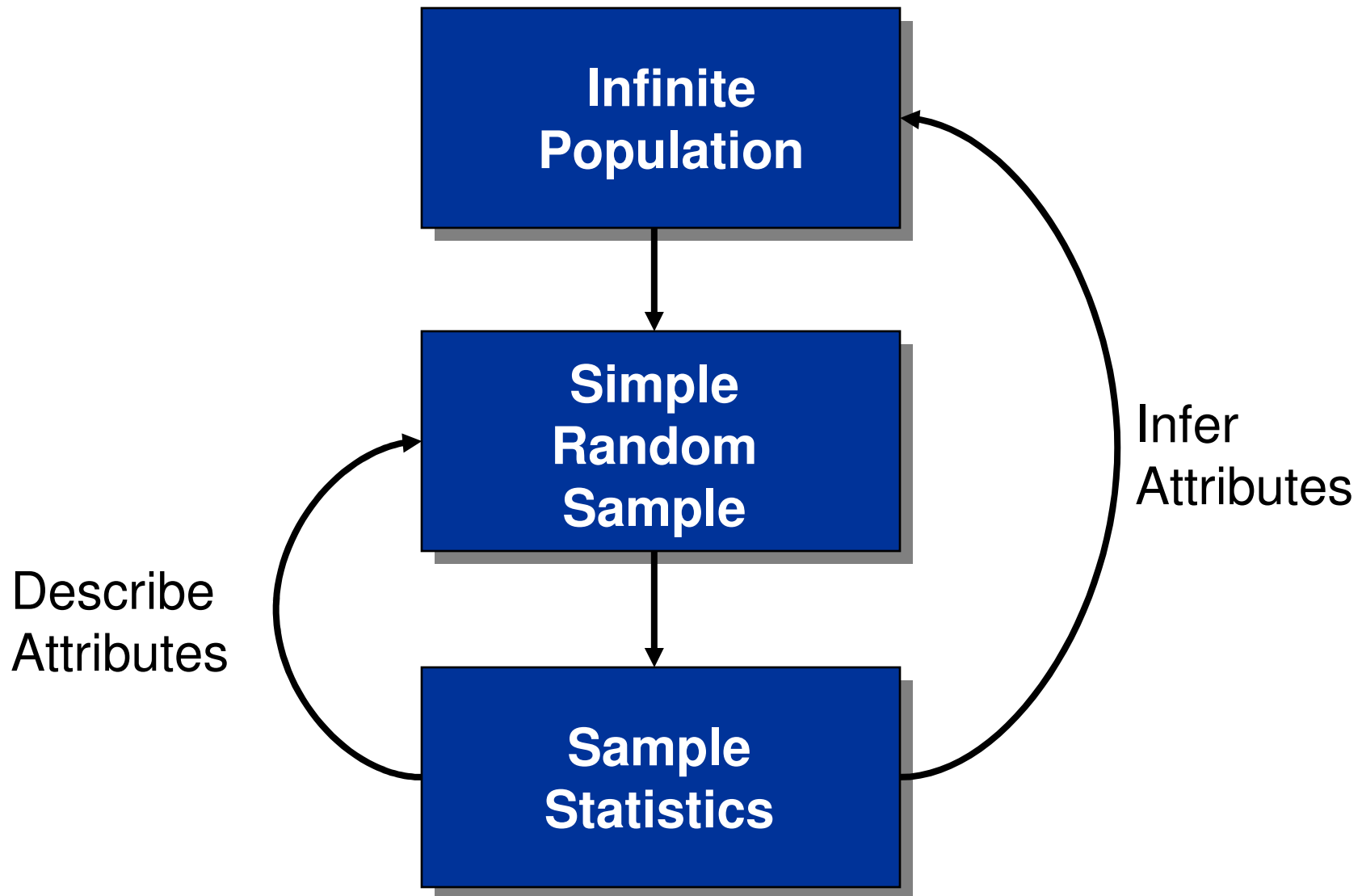
**Sample selection using
Proc SURVEYSELECT**

Sylvain Tremblay
SAS Canada – Education Group

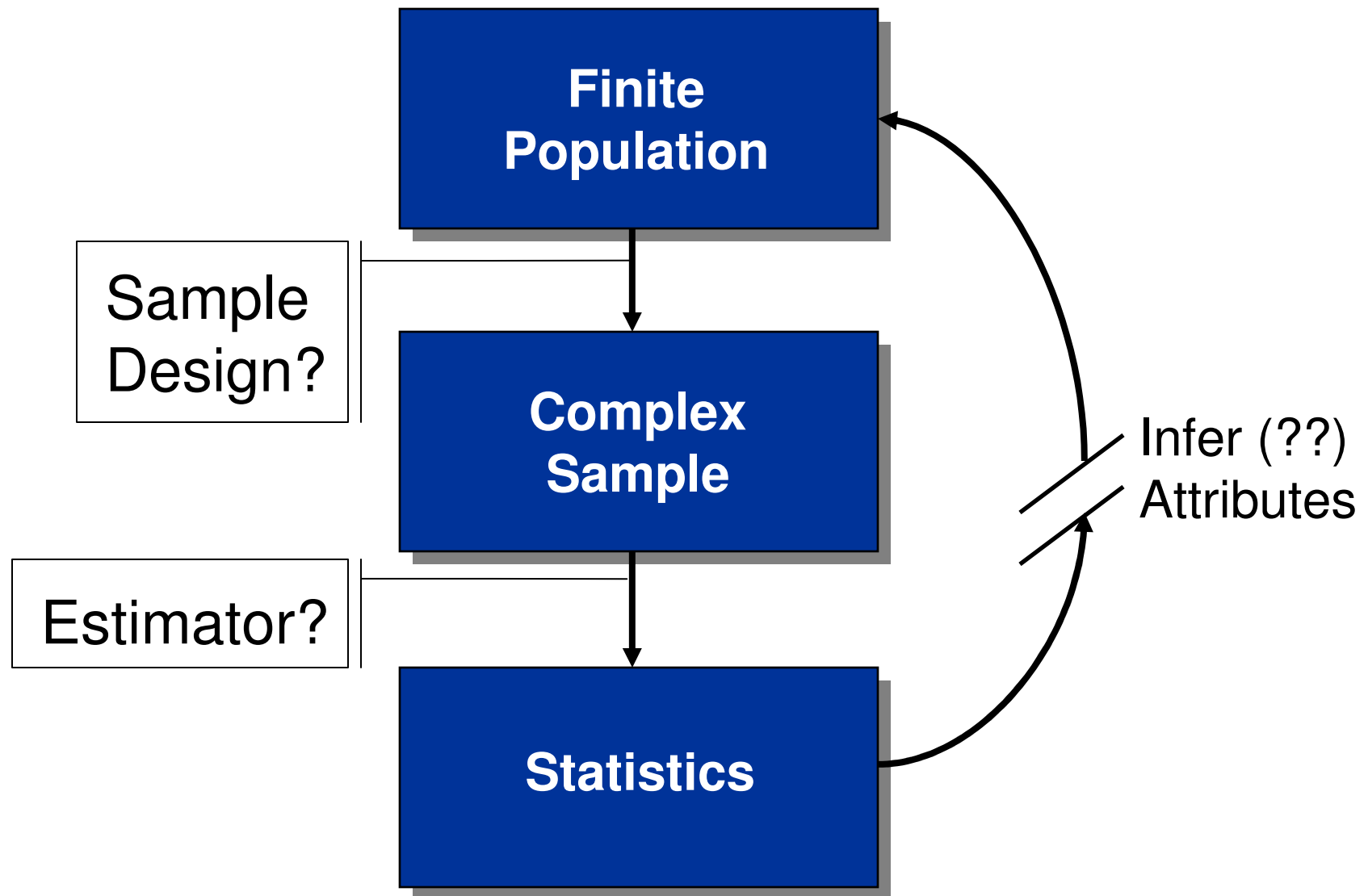
**THE
POWER
TO KNOW.**

Copyright © 2009, SAS Institute Inc. All rights reserved.

Inferences from Simple Random Samples



Inferences from Complex Survey Samples



Features of Sample Designs

- I. Equal Probabilities
 - a) Equal probabilities at all stages
 - b) Equal final probabilities after multiple stages

- II. Element Sampling:
Single stage, elements are sampling units

- I. Unequal Probabilities
 - a) Caused by irregularities in frame or procedures
 - b) Caused by disproportionate allocation

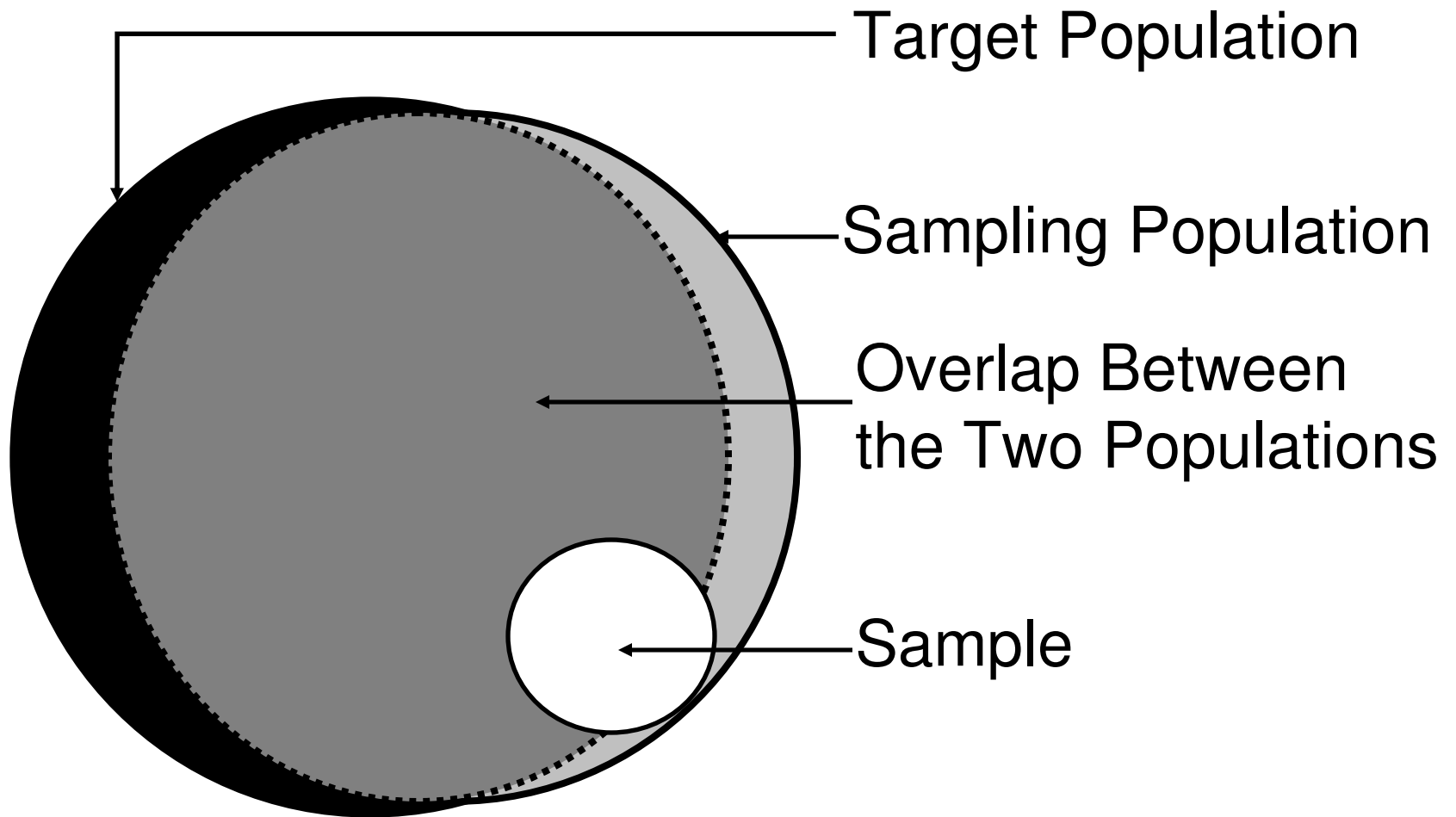
- II. Cluster Sampling:
Sample units are groups of elements
 - a) One-stage
 - b) Multi-stage

continued...

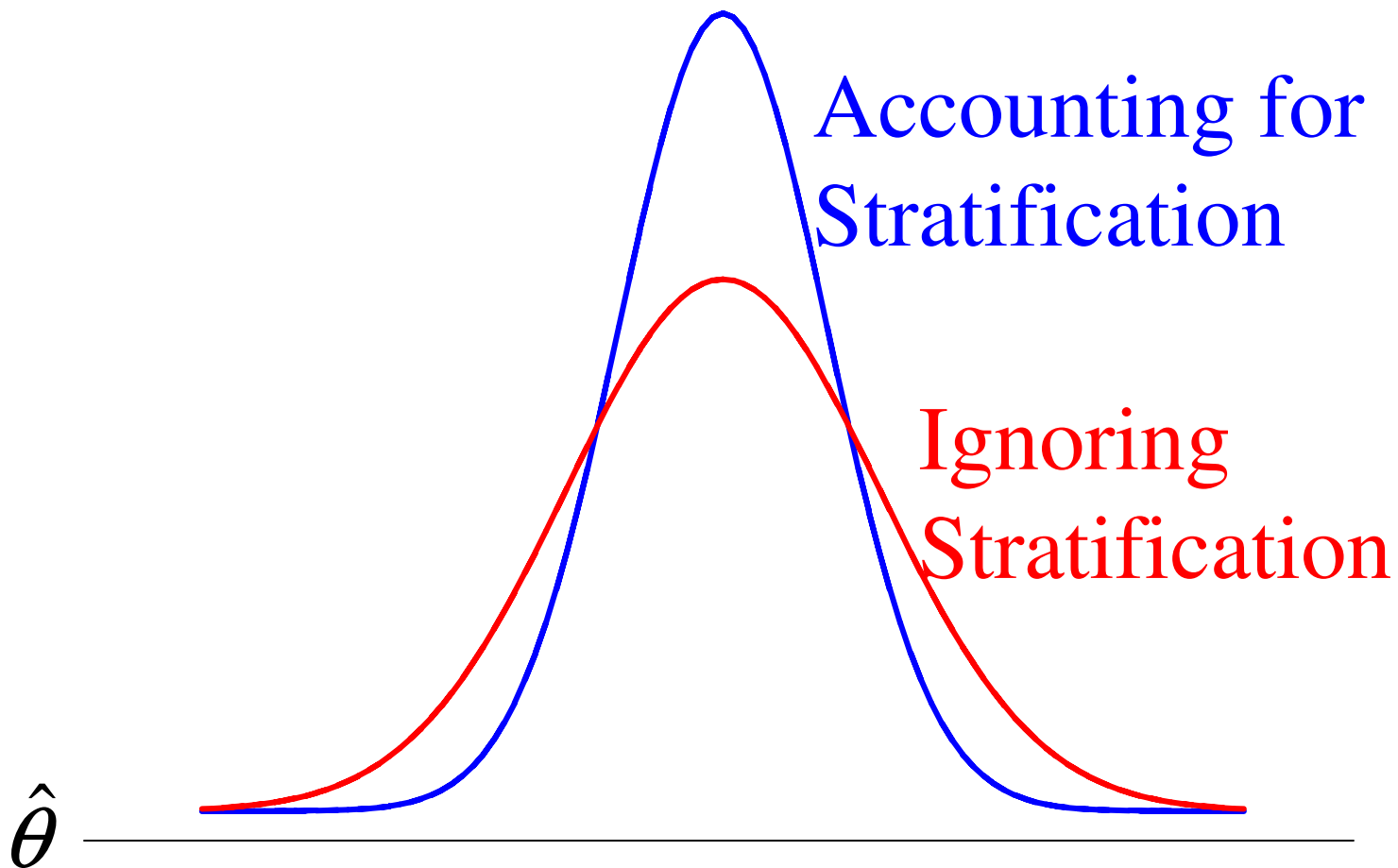
Features of Sample Designs

<p>III. Unstratified selection: sample units selected from population</p>	<p>III. Stratified selection: separate selection from partitions of population</p>
<p>IV. Random selection of sampling units</p>	<p>IV. Systematic selection of sampling units</p>
<p>V. One-phase sampling: final sample selected directly from population</p>	<p>V. Multi-phase sampling: final sample selected from an earlier sample</p>

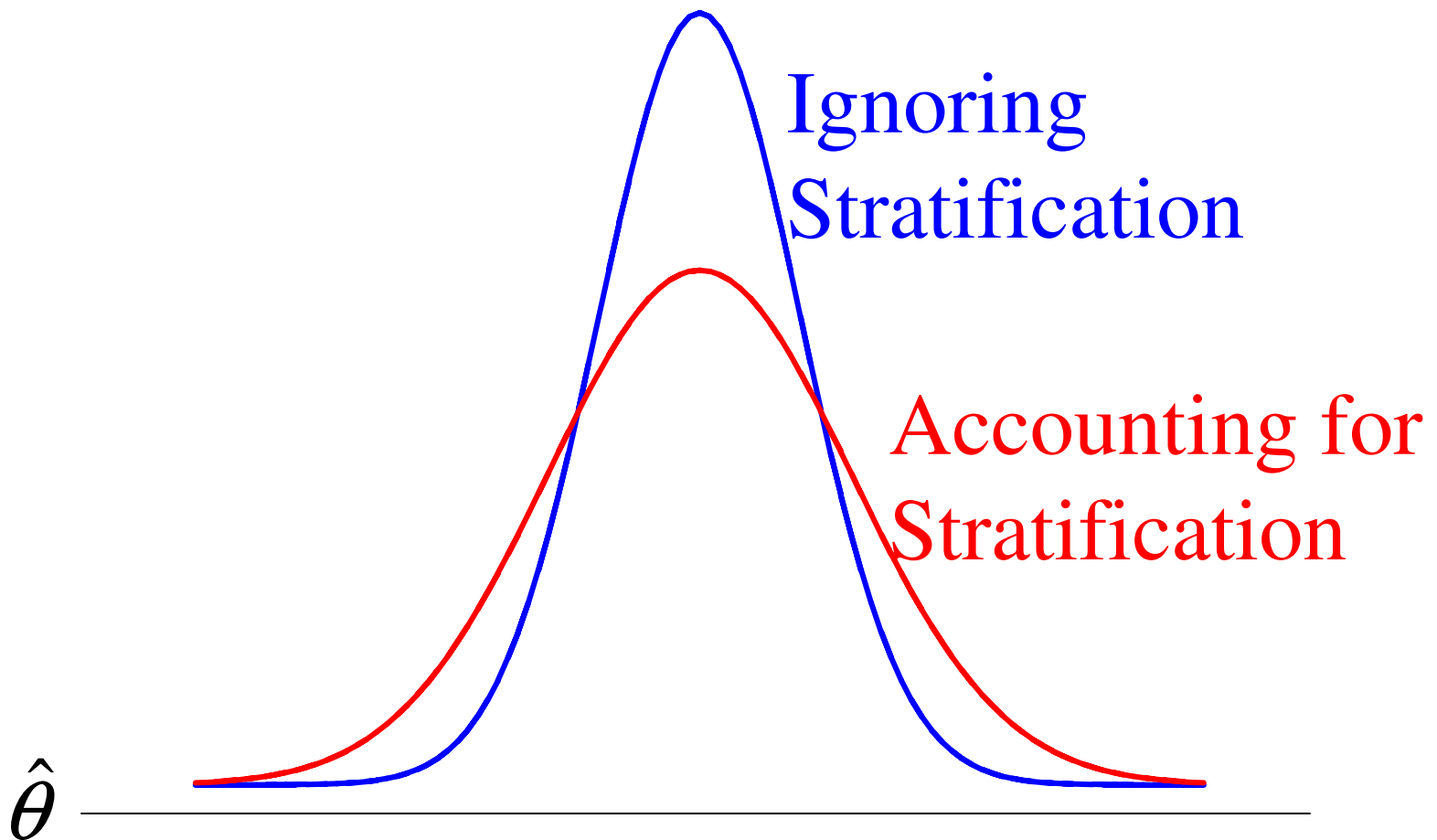
Target versus Sampling Populations

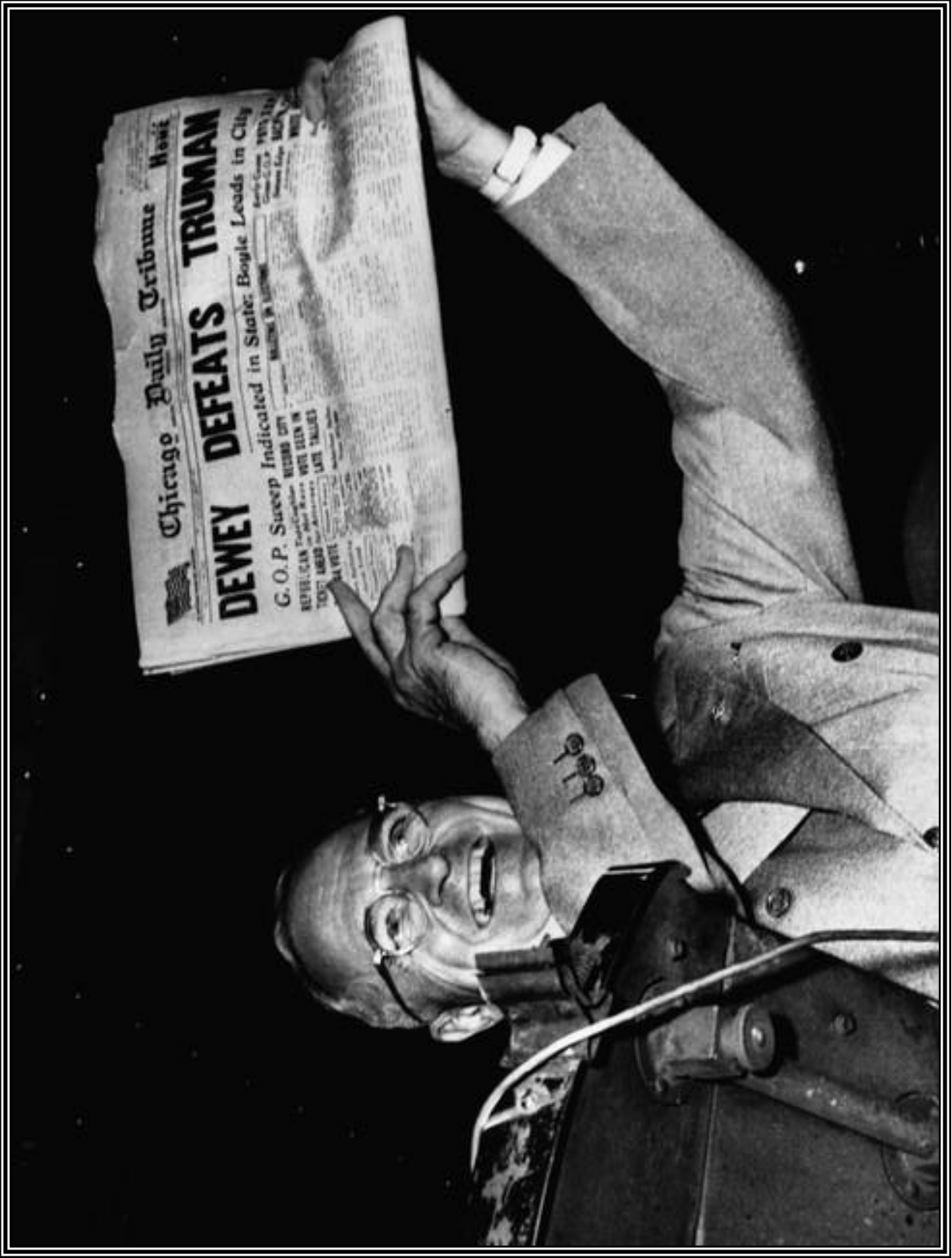


(Possible) Effect of Stratification for the Primary Survey Item



(Possible) Effect of Stratification for Secondary Survey Item





The SURVEYFREQ Procedure

Selected syntax for PROC SURVEYFREQ:

```
PROC SURVEYFREQ DATA=SAS-data-set <options>;  
  TABLES requests  
           / CHISQ CHISQ1 LRCHISQ LRCHISQ1  
             DEFF;  
  CLUSTER variables;  
  STRATA variables </LIST>;  
  WEIGHT variable;  
RUN;
```

The SURVEYMEANS Procedure

Selected syntax for the SURVEYMEANS procedure:

```
PROC SURVEYMEANS DATA=SAS-data-set  
                                <options>;  
    CLASS variables;  
    STRATA variables ;  
    VAR variables ;  
    WEIGHT variable ;  
RUN;
```

The SURVEYREG Procedure

Selected statements of the SURVEYREG procedure:

```
PROC SURVEYREG <options>;  
  CLASS variables;  
  STRATA variables ;  
  MODEL dependent=<effects> </options> ;  
  WEIGHT variable ;  
  ESTIMATE 'label' effect values  
              <effect values ...>  
RUN;
```

The SURVEYLOGISTIC Procedure

Selected syntax for PROC SURVEYLOGISTIC:

```
PROC SURVEYLOGISTIC DATA=SAS-data-set <options>;  
  CLASS variable <(options)>;  
  CLUSTER variables;  
  STRATA variables </ LIST>;  
  WEIGHT variable;  
  MODEL variable <(options)> = effects </options>;  
  CONTRAST 'label' effect values, ...,  
    effect values </options>;  
RUN;
```

Population Parameters

A *population parameter* is a function of (usually unobserved) values of an attribute for all target population units:

- Total $\Sigma_P y_k$
- Mean $\Sigma_P y_k / N$
- Ratio $\Sigma_P y_k / \Sigma_P z_k$
- Regression Coefficient $(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$

where summation is over P, the target population.

Sample Statistics

A *statistic* is a function of observed survey values for all sample units:

- Total $\sum_S w_k y_k$
- Mean $\sum_S w_k y_k / \sum_S w_k$
- Ratio $\sum_S w_k y_k / \sum_S w_k z_k$
- Regression Coefficient $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{y}$

where summation is over all sampled units, S , and w_k is the sampling weight for unit k .

Sampling Variability

The variability of an estimator is induced by sample-to-sample variability.

$$t = \sum_P y_k \quad \hat{t}_\pi = \sum_S \frac{y_k}{\pi_k} = \sum_P I_k \frac{y_k}{\pi_k}$$

$$I_k = \begin{cases} 1 & \text{if unit } k \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

$$E[I_k] = \pi_k = \text{prob}(\text{unit } k \text{ is selected})$$

$$V[I_k] = \pi_k (1 - \pi_k)$$

The National Health and Nutrition Examination Survey



Department of Health and Human Services
Centers for Disease Control and Prevention



<http://www.cdc.gov/nhanes/>

The National Health and Nutrition Examination Survey

NHANES

- has been conducted on a periodic basis since 1971
- was most recently started in 1999
- annually completes about 7,000 individual interviews
- annually completes about 5,000 medical exams in Mobile Examination Centers (MEC)
- targets the civilian, non-institutionalized U.S. population.

The National Health and Nutrition Examination Survey

The objectives of NHANES are as follows:

- estimate the number and percent of persons in the U.S. population and designated subgroups with selected diseases and risk factors
- monitor trends in the prevalence, awareness, treatment, and control of selected diseases
- monitor trends in risk behaviors and environmental exposures
- analyze risk factors for selected diseases

continued...

The National Health and Nutrition Examination Survey

NHANES objectives (continued):

- study the relationship between diet, nutrition and health
- explore emerging public health issues and new technologies
- establish a national probability sample of genetic material for future genetic research
- establish and maintain a national probability sample of baseline information on health and nutritional status

The NHANES Sample Design

The NHANES uses a stratified multistage design that includes the selection of the following:

- PSUs, which are counties or small groups of contiguous counties
- SSUs within PSUs, which comprise one or more blocks of households
- households within the SSUs
- one or more participants within the households