

WEIGHTING AND VARIANCE ESTIMATION FROM COMPLEX SURVEYS



Mehdi Nassirpour, Ph.D.
Illinois Department of Transportation

Presented at the Transportation Research Board conference Washington DC in
January 2011.

OBJECTIVES



- **Brief Background**
- **Weighting the Sample**
- **Variance Estimation**
- **Examples (Safety Belt Survey)**

Stages in the Selection of a Sample

Define the target Population

Select a sampling frame

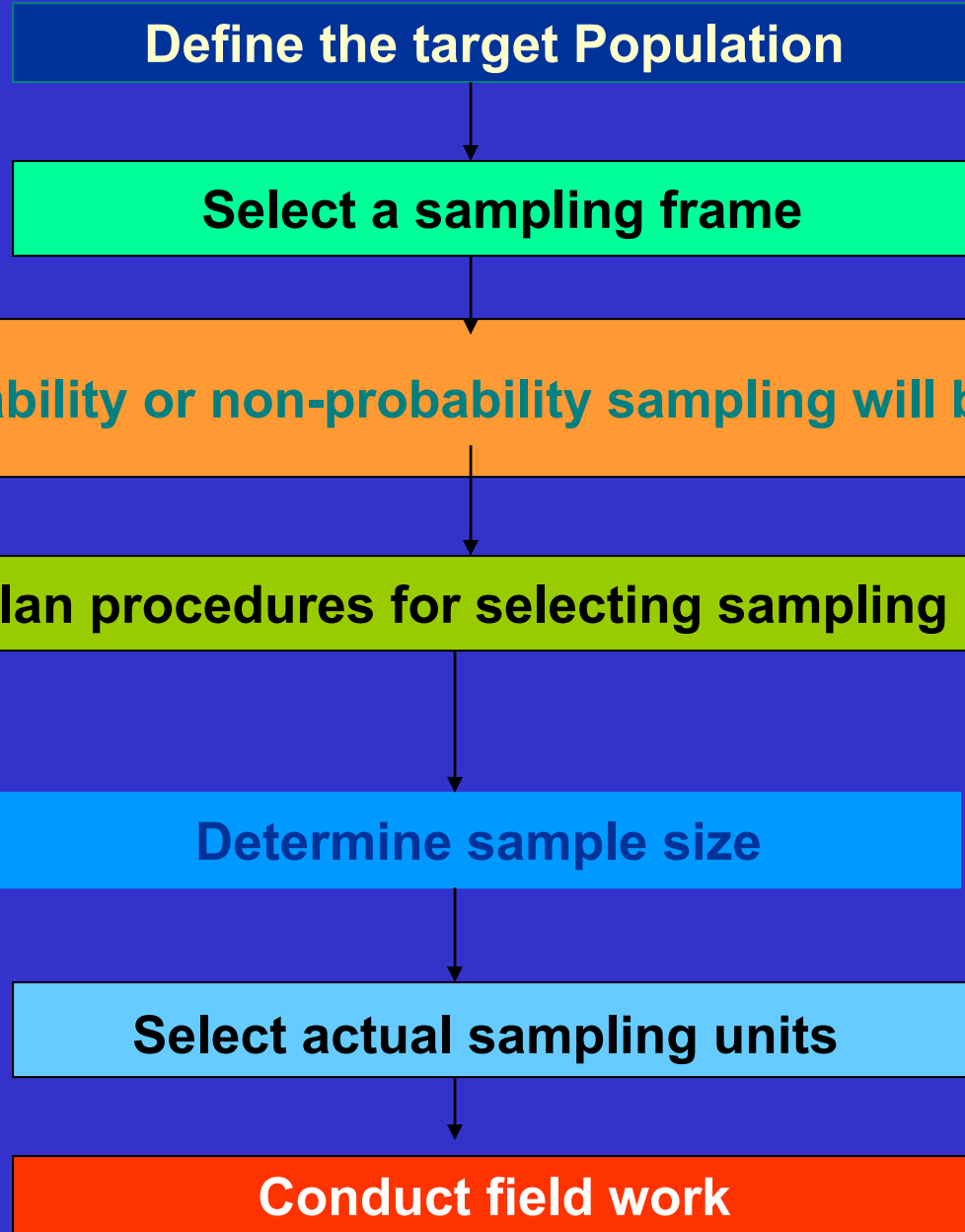
Determine if probability or non-probability sampling will be chosen

Plan procedures for selecting sampling units

Determine sample size

Select actual sampling units

Conduct field work




HOW GOOD MUST THE SAMPLE BE?



- There is no uniform standard of quality that must be reached by every sample.
- The quality of the sample depends entirely on the stage of the research and how the information will be used.

INAPPROPRIATE SAMPLE DESIGN



- Whether or not a sample design is appropriate depends on how it is used and the resources available. It may be fair to say that the sample generalizations made from the sample go too far.

WHAT IS THE APPROPRIATE SAMPLE DESIGN?



- Degree of Accuracy
- Resources
- Time
- Advanced Knowledge of the Population
- National versus Local
- Need for Statistical Analysis

SMALL-SCALE SAMPLE WITH LIMITED RESOURCES



- **Generalizability**
- **Sample size**
 - Too small for a meaningful analysis
 - Adequate for some but not all major analyses
 - Adequate for the purpose of study
- **Sample Execution**
 - Poor response rate
 - Careless field work
- **Use of resources**

PROBABILITY SAMPLING



- Simple Random Sample
- Systematic Random Sample
- Stratified Random Sample
- Cluster Sample
- Multi-stage Random Sample

WEIGHTING THE SAMPLE



- Reason for weighting is to correct problems associated with sample bias (sampling and non-sampling).
- Known Sampling biases, such as household selected by random digit dialing will have more than one phone number.

WEIGHTING PROCESS



- Assign a weight that is equal to the inverse of its probability of selection. In this case, where all sample elements have had the same chance of selection, given the same weight: 1. (This is called self-weighting sample)

WEIGHTING EXAMPLE

Unweighted Sample				Expected Sample (Based on Population)		
	Non- white	White	Total	Non- white	White	Total
Female	12.3%	56.7%	69%	7.2%	57.7%	65%
Male	9.8%	21.2%	31%	3.8%	31.2%	35%
Total	22%	78%	100%	11%	89%	100%

Nonwhite Female weight = $7.2/12.3=0.59$

Nonwhite Male weight = $3.8/9.8=0.39$

White Female weight = $57.7/56.7=1.02$

White Male weight = $31.2/21.2=1.47$

WEIGHTING FORMULA FOR STRATIFIED SAMPLE



$$\bar{X} = \sum_{i=1}^k W_i \bar{X}_i$$

DATA FOR COMPUTING PARAMETER ESTIMATES FROM STRATIFIED SAMPLES

	County			Total
	1	2	3	
Size of County (M_i)	10,000	15,000	25,000	50,000 (=M)
Weight (W_i)	.2	.3	.5	1.00
Size of sample (N_i)	50	50	50	150
Sample Mean (\bar{X}_i)	3,100	4,300	3,800	
Sample Standard Deviation (S_i)	500	400	300	

ESTIMATED STANDARD ERRORS



County 1:

$$\frac{S_1}{\sqrt{N-1}} = \frac{500}{\sqrt{49}} = 71.4$$

County 2:

$$\frac{S_2}{\sqrt{N-1}} = \frac{400}{\sqrt{49}} = 57.1$$

County 3:

$$\frac{S_3}{\sqrt{N-1}} = \frac{300}{\sqrt{49}} = 42.9$$

ESTIMATED MEAN AND VARIANCE



$$\bar{X} = .20(3,100) + .30(4,300) + .50(3,800) = 3,810$$

$$\hat{\sigma}_{\bar{X}} = (.20)^2 (71.4)^2 + (.30)^2 (57.1)^2 + (.50)^2 (42.9)^2 = 957.5$$

CLUSTER SAMPLING



- Divide population into a large number of groups, called clusters and then sample among clusters. Finally select all individuals within those clusters.
- The main reason for cluster sampling is to sample economically while retaining the characteristics of a probability sample.

TYPES OF CLUSTER SAMPLING



- Single -Stage Cluster sampling--Divide population into several hundred census tracts and then select 40 tracts as a sample. Then select every individuals within selected census tracts.
- Multistage Cluster Sampling--Take a random sample of census tracts within a city. Then within each selected census tract we take a simple random sample of blocks (smaller clusters). Finally we might select every third house and interview every second adult within each of these households

CLUSTER SAMPLING

Probability Proportionate to Size (PPS)



- Arrange clusters in a desired order (not necessarily by size)
- Obtain the size data
- Sum up the size measures over clusters
- Determine sampling interval
- Select a random start

DIFFERENCE BETWEEN CLUSTER SAMPLE AND STRATIFIED SAMPLE



- Although both types of sample involve divide population into groups, they involve in a opposite sampling operations.
- In a stratified sample, we sample individuals within every stratum. The sampling errors involve variability within strata. Strata are supposed to be homogeneous as possible and as different as possible from each other.
- In (single-stage) cluster sampling, we have no source of sampling error within the clusters because every case is being used. The variability is between the clusters.

DIFFERENCE BETWEEN CLUSTER SAMPLE AND SIMPLE RANDOM SAMPLE



- Cluster sample is less efficient than the simple random samples of the same size. But it may cost considerably less.
- The efficiency can be measured in terms of the size of standard error of estimate, a small standard error indicates high efficiency.

VARIANCE ESTIMATION FROM COMPLEX SAMPLES



- Sampling variance means differences in sample statistics when repeated samples are selected from the same population.
- Since actual replication is uncommon (due to time and budget), replication is done through sub-sampling or by the use of pseudo-replication procedures after the data are collected. This needed to be done along with the balanced replication procedures.

Sampling Variance Estimation (Example)



Death Rate per 1000 Population for Counties in four Selected Sub-Samples			
Sub-sample	Death Rate per 1000 Population	Sub-sample	Death Rate per 1000 Population
1	14.7	1	12.8
2	10.3	2	10.2
3	10.1	3	9.2
4	9.1	4	8.6
5	8.2	5	7.8
Total	10.2	Total	9.7
1	16.1	1	16.9
2	12.3	2	11.3
3	11.3	3	11
4	10.2	4	10.2
5	9.7	5	9.7
Total	11.9	Total	11.8

$$S^2 = \frac{\sum (xi - \bar{x})^2}{k(k-1)}$$

$$S^2 = \frac{0.5^2 + 1.3^2 + 0.9^2 + 0.8^2}{4(3)}$$

$$S^2 = 0.28$$

$$\bar{x} = 11$$

BALANCED REPLICATION PROCEDURE



- This procedure groups sample PSUs into pairs and selects half-samples by choosing one PSU from each pair.
- Since individual replications are subject to high variability in their estimates, repeated replications by computer are required for reliability. The high number of replications will increase the precision of the variance estimates.

SAFETY BELT SURVEY DESIGN IN ILLINOIS

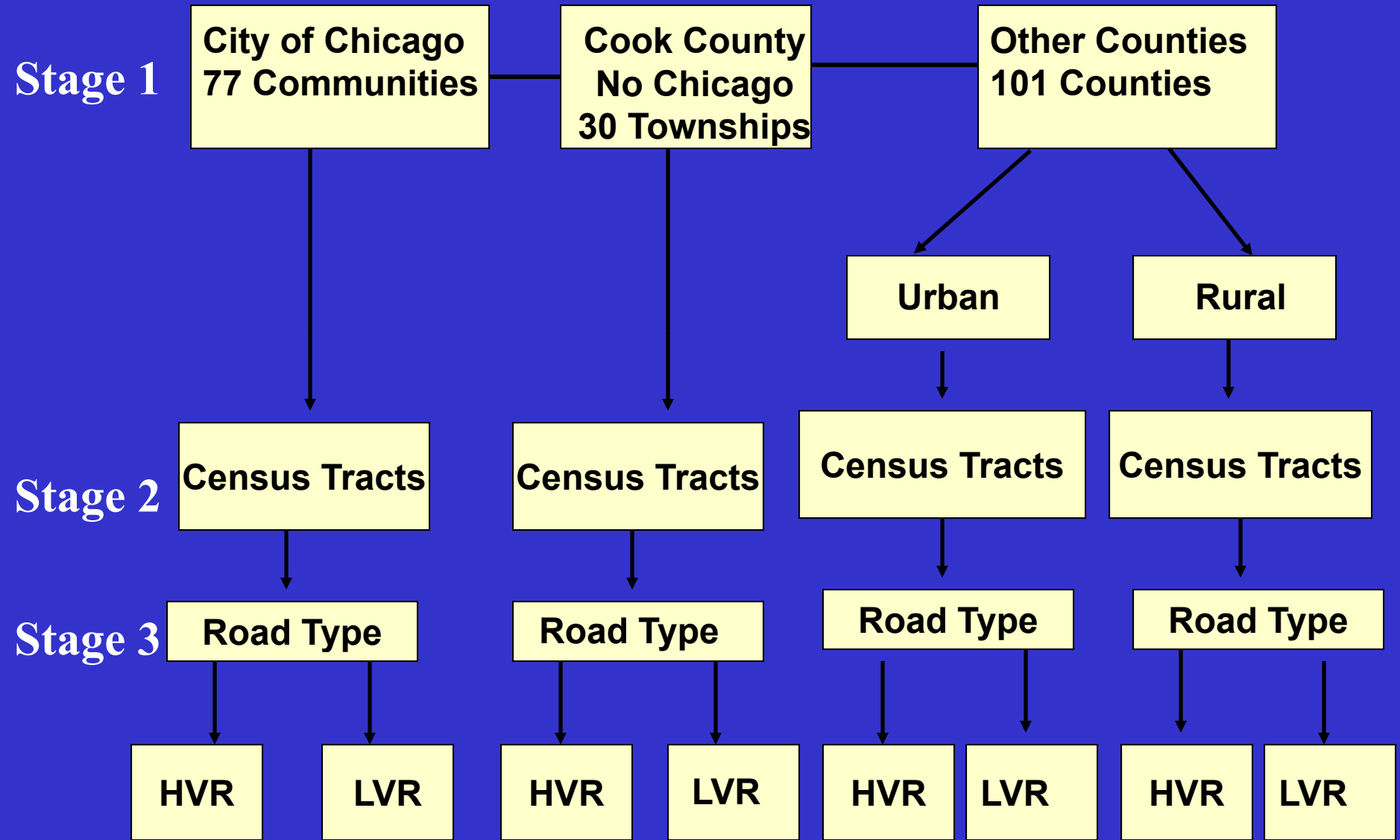
EXAMPLE



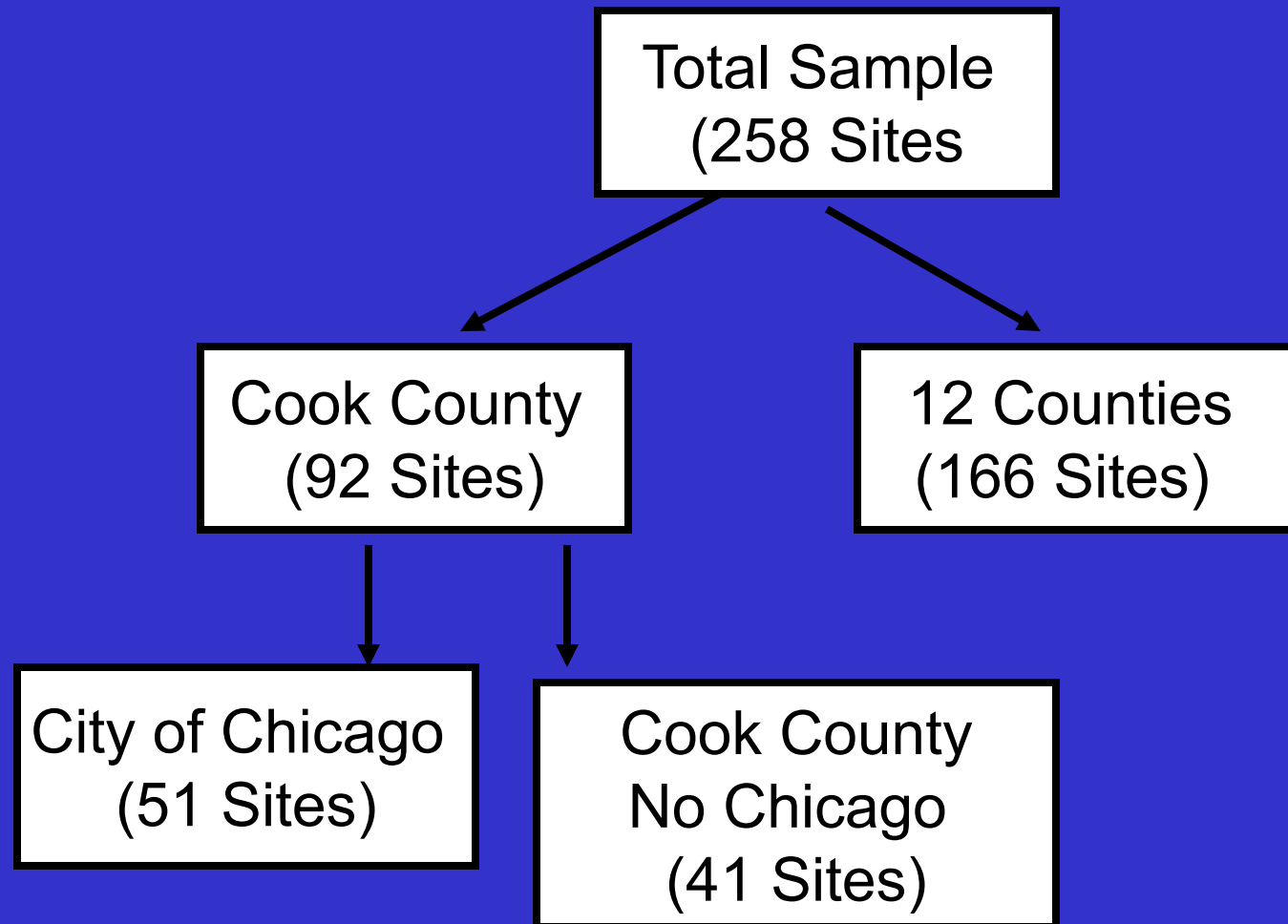
OBJECTIVE OF THE SURVEY

- To design a multi-stage cluster random sample in order to estimate the statewide safety belt usage rate in Illinois.

SAMPLE DESIGN SUMMARY IN ILLINOIS



ALLOCATION OF SAMPLE SITES



STAGE 1 - PSU SELECTION

- Sampling Units - 20 PSU
 - 5 Community Areas in the City of Chicago
 - 3 Townships in Cook County
 - 11 Counties (DuPage County was selected twice)
 - 4 out of 11 counties were rural counties.
 - 47 counties were eliminated due to a small number of VMTs and population.
 - 54 counties contains about 90 percent of VMT and population
- Selection - Probability Proportional to Vehicle Miles of Travel and Population.
- Selection is systematic (1/Ns)
 - After Random Start

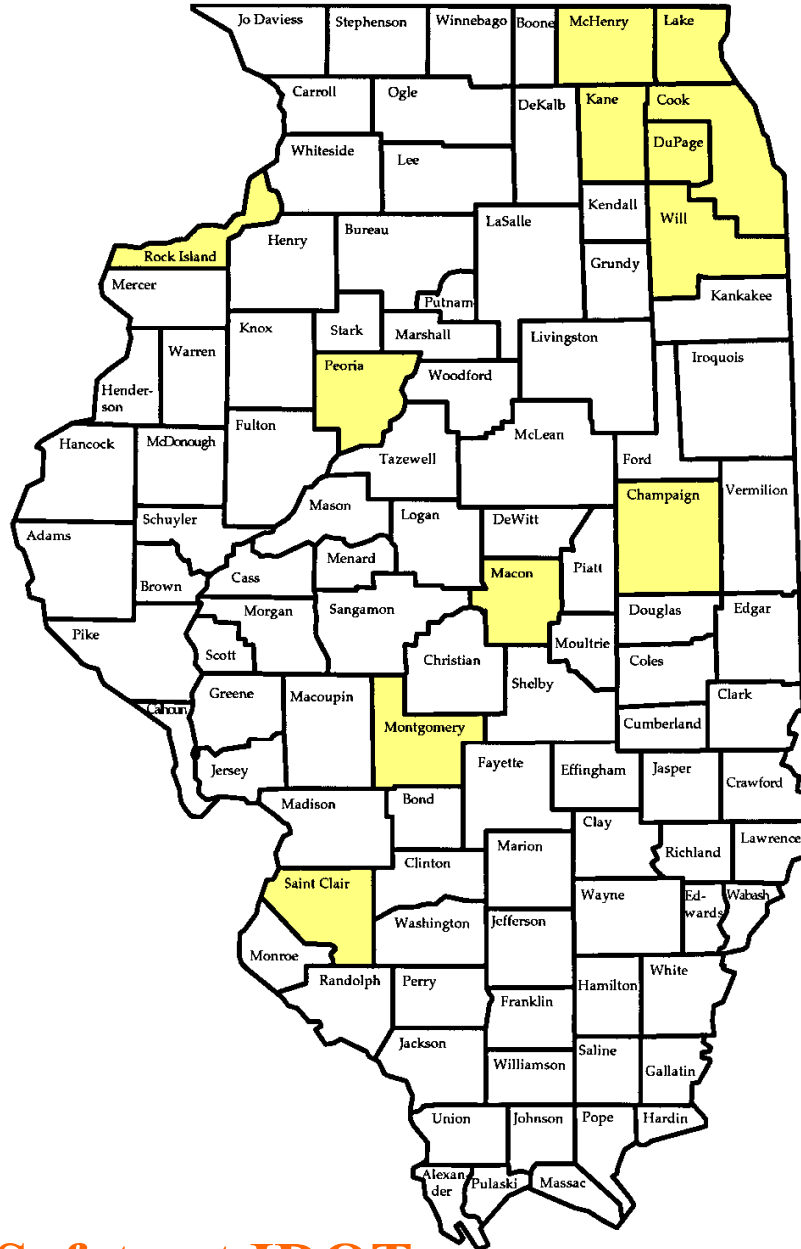
SAFETY BELT LOCATIONS IN IL

ROCK ISLAND CO. LOCATIONS (7)

PEORIA COUNTY LOCATIONS (9)

MONTGOMERY COUNTY LOCATIONS (2)

ST. CLAIR COUNTY LOCATIONS (14)



MCHENRY COUNTY LOCATIONS (6)

LAKE COUNTY LOCATIONS (31)

KANE COUNTY LOCATIONS (14)

WILL COUNTY LOCATIONS (18)

CHICAGO LOCATIONS (49)

COOK COUNTY LOCATIONS (40)

DUPAGE COUNTY LOCATIONS (42)

CHAMPAIGN COUNTY LOCATIONS (9)

STAGE 2 - CENSUS TRACT SELECTION

- Census Tracts
 - A total of 96 census tracts (10-12 percent of total tracts) within those selected counties were selected randomly.
- Selection - Probability Proportional to Population.
- Selection is systematic (1/Ns).
 - After Random Start

STAGE 3 - ROAD TYPE SELECTION

- Road Type
 - All roads (Interstates, freeways, US/IL highways, and residential streets) within selected tracts were identified using tract maps.
 - All Interstates and IL/US highways within these tracts were selected with certainty and other roads were selected randomly.

PROCEDURE USED TO ASSIGN OBSERVATION SITES TO OBSERVATION TIME PERIOD

- Selected roads randomly assigned to time of day and day of week.
- To minimize travel time and distance required to conduct the observations, the sampled sites were grouped into 29 geographic clusters, with each cluster containing about between 6 and 12 road sites.
- Each cluster is randomly (without replacement) assigned to a day of the week. Each day was divided into eight one-hour time periods between 7:00am to 6:30pm when the light was adequate for observation.

WEIGHTING PROCEDURES

$$W_{ctr} = 1/P_{ctr}$$

$$P_{ctr} = P_c \cdot P_{t/c} \cdot P_{r/ct}$$

c identifies a sample of county

t identifies a sample of tract

r identifies a sample of road site

P_c probability of selecting county **c**

$P_{t/c}$ probability of selecting tract **t** within selected county **c**

$P_{r/ct}$ probability of selecting road **r**, conditional to the tract and county

FORMULA FOR ESTIMATING SAFETY BELT USAGE RATE

$$R = \frac{\sum_{ctr} W_{ctr} \cdot X_{ctr}}{\sum_{ctr} W_{ctr} \cdot Y_{ctr}}$$

R = Safety belt usage rate

W_{ctr} = weight associated with road r in tract t in county c ;

X_{ctr} = number of safety-belted front seat occupants of passenger cars and pick-up trucks observed in county c , tract t , and road site r ; and

Y_{ctr} = number of front seat occupants observed in county c , tract t , and road site r .

FORMULA FOR ESTIMATING SAFETY BELT USAGE RATE

F_{ctr} = adjustment for lanes not observed in road site r , tract t in county c ;

L_{ctr} = total number of lanes in road site r , tract t in county c ; and

O_{ctr} = total number of observed lanes in road site r within tract t in county c

$$R = \frac{\sum_{CTR} W'_{ctr} . X_{ctr}}{\sum_{CTR} W'_{ctr} . Y_{ctr}}$$

$$W'_{ctr} = W_{ctr} . F_{ctr}$$

$$F_{ctr} = \frac{L_{ctr}}{O_{ctr}}$$

VARIANCE ESTIMATION

Estimated safety belt usage rate is subject to sampling error since a small number of front seat occupants were observed. These sampling errors are expressed as coefficients of variation (relative error), that are the ratios of the relative standard errors of estimates to the estimates themselves.

VARIANCE ESTIMATION FORMULA

$(Y_{ijkl}, X_{ijkl}, Y_{ijk2}, X_{ijk2}, \dots, Y_{ijkp}, X_{ijkp})$

where $i = 1, 2, \dots, l$; $j = 1, 2, \dots, n_j$; $k = 1, 2, \dots, m_{ij}$; i is the stratum identification, j is the cluster identification, k is the element-within-cluster identification.

The above vector is used in the analysis and estimator is

$$R_s = \frac{Y_s}{X_s} - \frac{\sum_{i=1}^L \sum_{j=1}^{n_j} \sum_{k=1}^{m_{ij}} W_{ij} \cdot Y_{ijks}}{\sum_{i=1}^L \sum_{j=1}^{n_j} \sum_{k=1}^{m_{ij}} W_{ij} \cdot X_{ijks}}, s = 1, 2, 3, \dots, p.$$

SAFETY BELT USAGE RATES IN ILLINOIS, JULY 2010

Statewide (# of Sites)	Safety Belt Estimate (Ratio)	Standard Error	Lower Limit %	Upper Limit %
258	92.6	0.4	91.8	93.4

VARIANCE ESTIMATION (PROC SURVEY)

The SURVEYMEANS Procedure in SAS was used to estimate population means and population totals from sample survey data.

PROC SURVEYMEANS also provides domain analysis (subgroup or subpopulation analysis).

- The procedure uses the Taylor series expansion method to provide estimates of design-based variances for the quantities of interest.