Complex Surveys

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

3/19/25

Complex surveys

- Complex sample design features: Strata, clusters and weights
- The weighted mean $\hat{\bar{y}}_w = \frac{u}{w} = r$ is no different from what we've seen before; important to recognize that this is a form of a ratio mean
- The sum of the weights across all strata and clusters yields an estimate of the population size!
- If there is a positive correlation between the weights and the variable of interest (informative weights), the weighted estimate will be higher than the unweighted estimate (same idea for a negative correlation; the weighted estimate will be lower)
- No correlation: no difference in estimates!

Implementation: Selecting a stratified two-stage sample

```
# number of psus selected n =2, size=rep(n=2,3 strata) (srswor)
# number of students selected m_i =3 size=rep(m_i= 3,6 classes) (srswor)
numberselect<-list(table(classeslong2$strat),rep(2,3),rep(3,6))
# select a stratified two-stage cluster sample
set.seed(75745)
tempid<-mstage(classeslong2,stage=list("stratified","cluster","cluster"),
               varnames=list("strat","class","studentid"),
               size=numberselect, method=list("","srswor","srswor"))
# get data
sample3<-getdata(classeslong2,tempid)[[3]]  #3rd stage
sample3$finalweight<-1/sample3$Prob
# check sum of weights, should be close to number of students in population
# (but not exactly equal, since psus not selected with prob proportional to M_i)
sum(sample3$finalweight)
```

```
[1] 624
```

```
sample3[1:3,] # print the sample
```

```
     class class_size strat studentid ID_unit Prob_ 3 _stage       Prob
5.31     5         76     1        32      32     0.03947368 0.03947368
5.42     5         76     1        43      43     0.03947368 0.03947368
5.61     5         76     1        62      62     0.03947368 0.03947368
     finalweight
5.31    25.33333
5.42    25.33333
5.61    25.33333
```

Multi-stage selection

```
sample1<-getdata(classeslong2,tempid)[[1]]  #1st stage
sample2<-getdata(classeslong2,tempid)[[2]]  #2nd stage
names(sample1)
```

```
[1] "class"          "class_size"     "studentid"      "strat"
[5] "ID_unit"        "Prob_ 1 _stage" "Stratum"
```

```
table(sample1$`Prob_ 1 _stage`)
```

```
  1
647
```

```
table(sample2$strat,sample2$`Prob_ 2 _stage`) # Selection probs for psus in strata
```

```
    0.285714285714286 0.333333333333333   1
  1                 0                 0 176
  2                 0                98   0
  3                44                 0   0
```

```
table(sample3$class,sample3$`Prob_ 3 _stage`) # Selection probs for ssus in psus
```

```
    0.03 0.0394736842105263 0.0555555555555556 0.0681818181818182 0.125 0.15
  5    0                  3                  0                   0     0    0
  7    0                  0                  0                   0     0    3
  8    0                  0                  0                   3     0    0
  9    0                  0                  3                   0     0    0
  12   0                  0                  0                   0     3    0
  14   3                  0                  0                   0     0    0
```

Inference with complex sample designs

- Variance estimation proceeds as we've seen before for ratio means given stratified cluster samples; we now take the variances and covariances of weighted cluster totals

$$var(\hat{\bar{y}}_w) \approx \frac{1}{w^2}[var(u) + r^2 var(w) - 2 * r * cov(u, w)]$$

- Note that the W_h values from our discussion of estimating sampling variance for stratified samples has been absorbed into the element weights used for estimation
- Degrees of freedom = # clusters - # strata
- Syntax in the R *survey* package: *svydesign(id = ~psu, strata=~strata, weights=~finalweights, data)*

Increase in variance

- The use of weights in estimation can increase the sampling variance of weighted estimates (part of the design effect)
- The "1+L" method (assuming 0 correlation of weights and survey variables) provides a potential increase in variance
- L = loss of precision due to weighting $= CV^2(\text{weights})$
- If $1 + L = 1.41$, there is a potential increase in sampling variance of about 41% due to the use of weights in estimation
- One can treat $1 + L$ like a DEFF and compute an effective sample size, but this does not account for cluster sampling or stratification

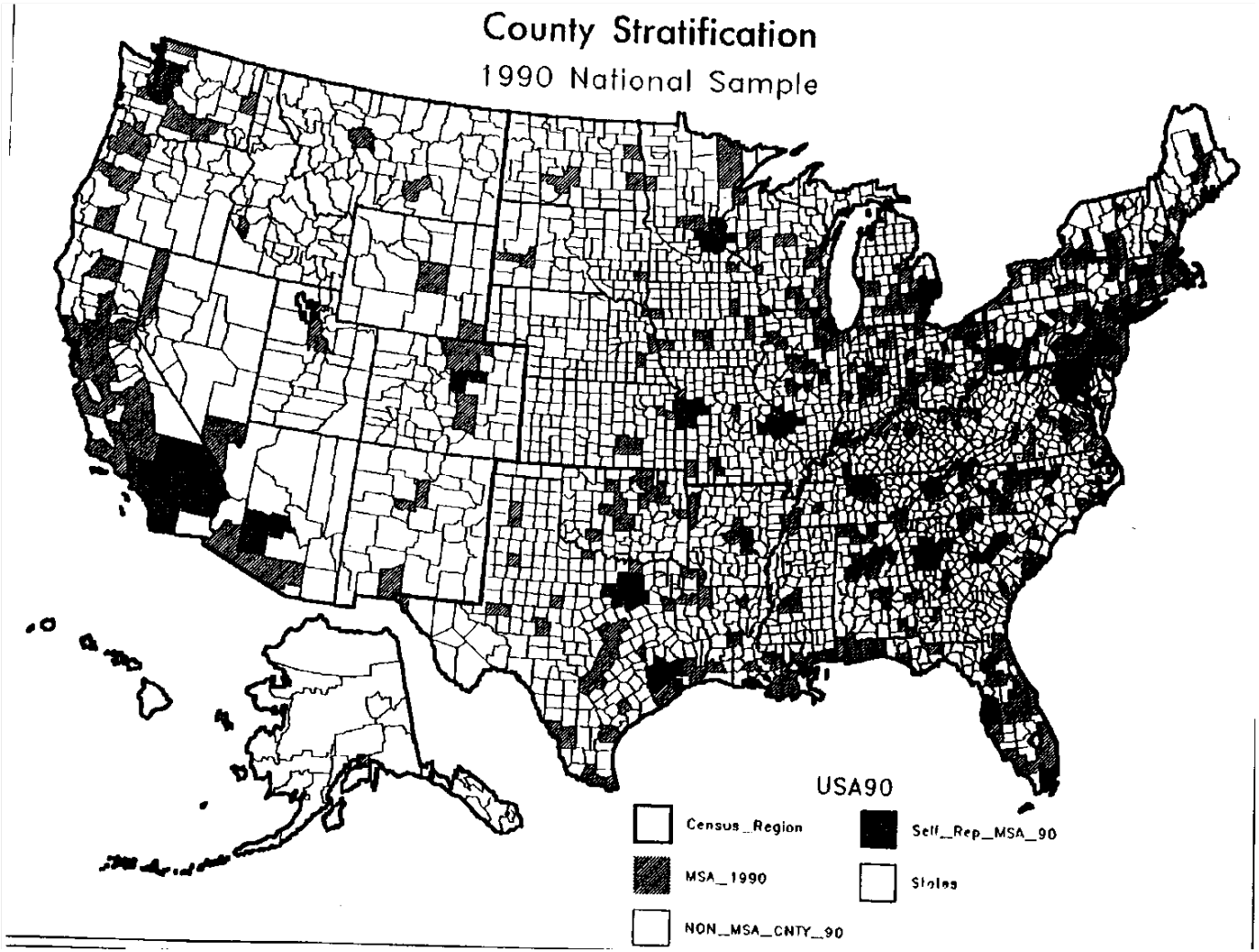
$$n_{ess} = \frac{n}{deff} = \frac{n}{1+L} = \frac{(\sum w_i)^2}{\sum w_i^2} = \frac{10}{1.41} = 7.09$$

(v.s. nominal $n = 10$)

Example: Multi-stage area sampling

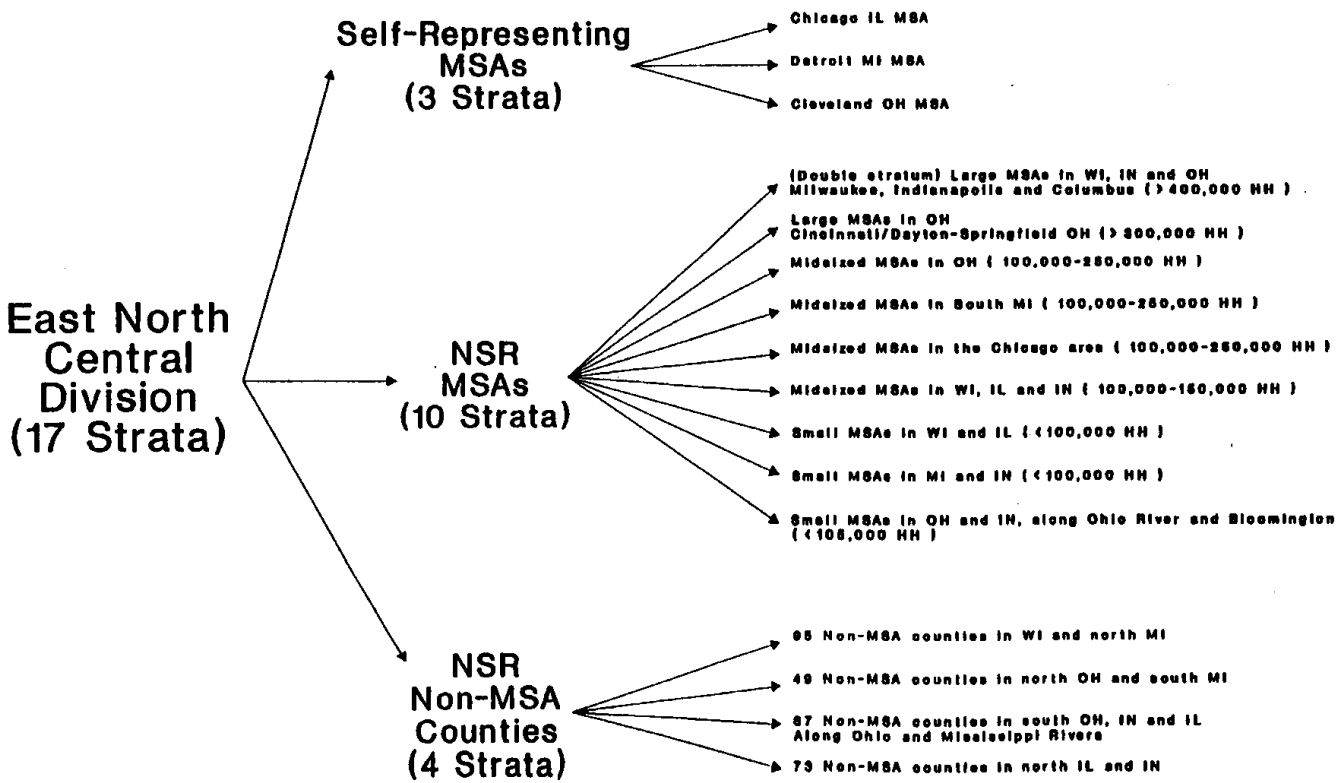
1. Select counties (PSUs) within strata
2. Select Census blocks (SSUs) within selected counties
3. Select housing units from each block
4. Select one individual from each housing unit

Example: Multi-stage area sampling

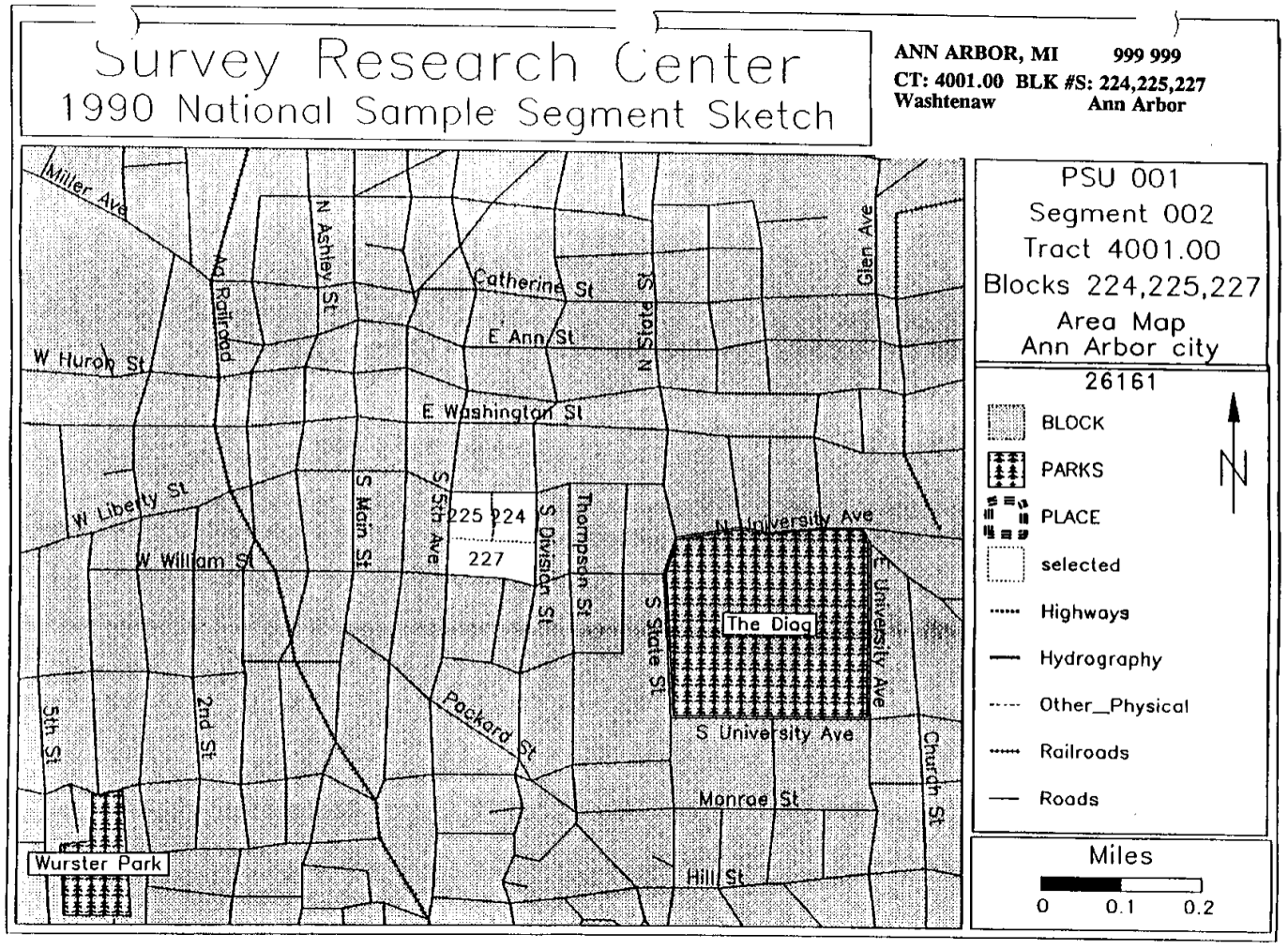


Example: Multi-stage area sampling

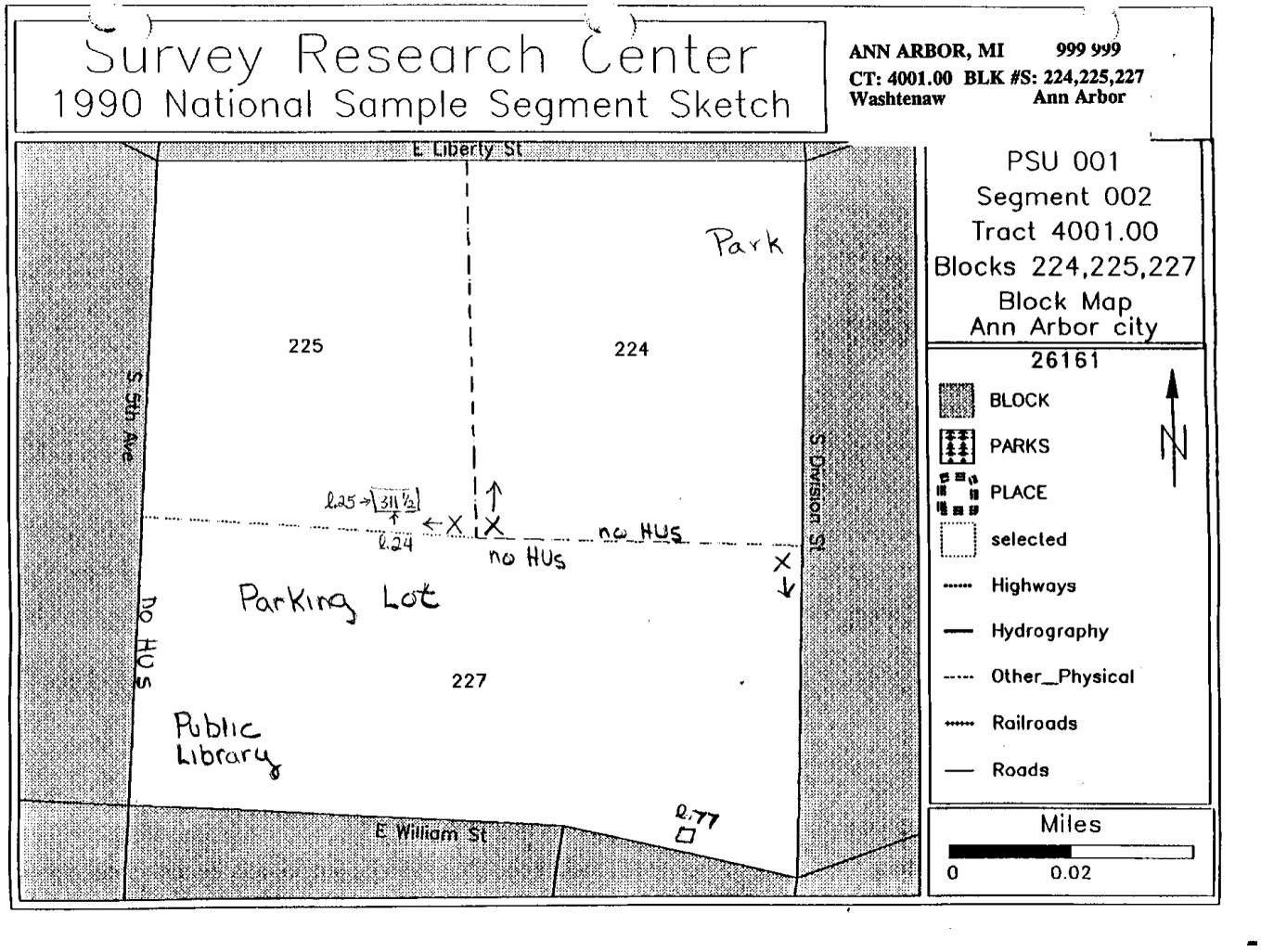
Figure 5: Primary Stage Stratification Criteria
For the 1990 SRC National Sample Design
East North Central Division of Midwest Region



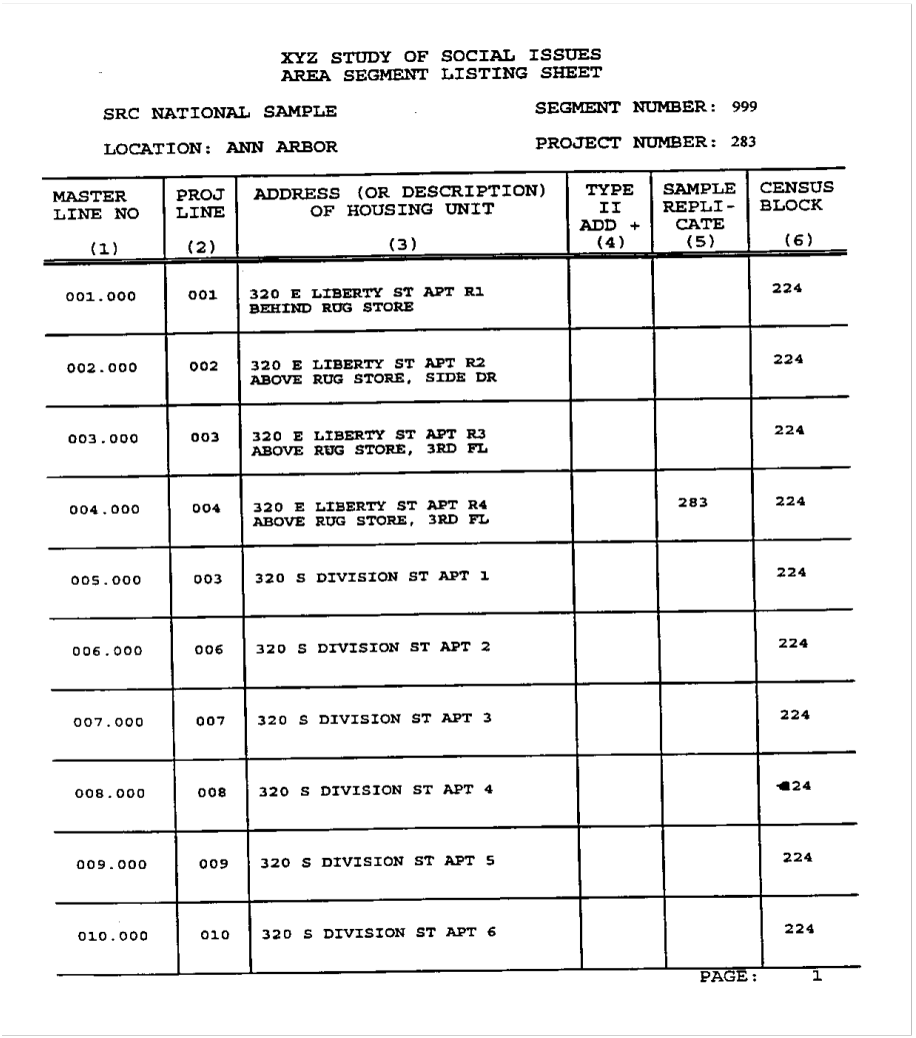
Example: Multi-stage area sampling



Example: Multi-stage area sampling



Example: Multi-stage area sampling



XYZ STUDY OF SOCIAL ISSUES
AREA SEGMENT LISTING SHEET

SRC NATIONAL SAMPLE — SEGMENT NUMBER: 999

LOCATION: ANN ARBOR — PROJECT NUMBER: 283

MASTER LINE NO (1)	PROJ LINE (2)	ADDRESS (OR DESCRIPTION) OF HOUSING UNIT (3)	TYPE II ADD + (4)	SAMPLE REPLI-CATE (5)	CENSUS BLOCK (6)
001.000	001	320 E LIBERTY ST APT R1 BEHIND RUG STORE			224
002.000	002	320 E LIBERTY ST APT R2 ABOVE RUG STORE, SIDE DR			224
003.000	003	320 E LIBERTY ST APT R3 ABOVE RUG STORE, 3RD FL			224
004.000	004	320 E LIBERTY ST APT R4 ABOVE RUG STORE, 3RD FL		283	224
005.000	003	320 S DIVISION ST APT 1			224
006.000	006	320 S DIVISION ST APT 2			224
007.000	007	320 S DIVISION ST APT 3			224
008.000	008	320 S DIVISION ST APT 4			224
009.000	009	320 S DIVISION ST APT 5			224
010.000	010	320 S DIVISION ST APT 6			224

PAGE: 1

Example: Multi-stage area sampling

Example: Multi-stage area sampling

Example: Multi-stage area sampling

Examples: National complex sample surveys

- ACS (Team Cochran)
- CPS (Team Groves)
- FoodAPS (Team Heeringa)
- GSS (Team Hess)
- HRS (Team Lepkowski)
- NHANES (Team Little)
- NHIS (Team Valliant)

Review: Design

- Theoretically ideal: SRS (random, representative, objective)
- Reduce cost: cluster sampling (roh, equal/unequal sizes, one-/two-stage, subsample size)
- Borrow auxiliary information: stratified sampling (allocation)
- Easy to implement: systematic sampling (fractional interval)
- Practice: stratified cluster sampling, PPS/PPeS
- Key evaluation criteria: design effect, variance, budget
- New design projection: portability of roh, mean, and element variance

Review: Practical implementation

- Frame problems: duplicates, blanks, coverage errors
  - Proper use of available (good) materials
  - No available frames: non-probability surveys
- Always try to minimize variance per unit cost

Review: Mean estimate

- Goal: unbiased estimate of the population mean
- Desire to use the sample mean if epsem
- If not epsem, account for the unequal probabilities of selection/response, use weighted mean, especially when the weights are correlated with the survey measure
- When the sample size is random, use the ratio mean; however, remember that the ratio mean is a (slightly) biased estimate

Review: Variance estimation

- SRS: fpc, $p(1-p)/(n-1)$ for binary outcome
- Cluster sampling: ultimate cluster approximation, variance-a function of cluster totals
- Stratified sampling: sum of within-stratum variances
- Systematic sampling: unmeasurable, use variance estimation models to approximate – SRS, stratified, paired and successive differences
- Stratified unequal-sized cluster sampling: approximation of the sampling variance of ratio means ($cv(x)$ for adequacy diagnosis), depend on the methods used to select clusters within each stratum and rely on cluster totals - multiple, paired and successive difference models

Review: Degrees of freedom

- General rule: DF for CI construction / test statistics $= n–H$
  - Number of primary stage clusters, minus number of primary stage strata
- This rule relies on assumptions about the sampling distribution, but works fairly well in general
- Adapting the rule to different designs
  - Cluster sample, no stratification: $DF = n–1$ (only one stratum)
  - Stratified element sample, no cluster sampling: $DF = M–H$ (each individual is their own primary stage cluster!)
  - Simple random sample: $DF = M–1$ (one stratum, each individual is their own cluster)