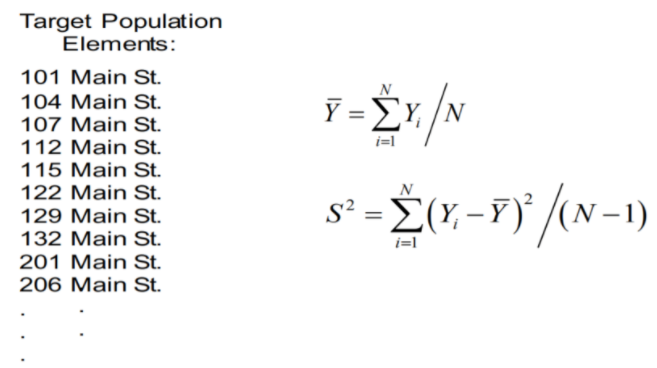
**W3 Saving money using cluster sampling**

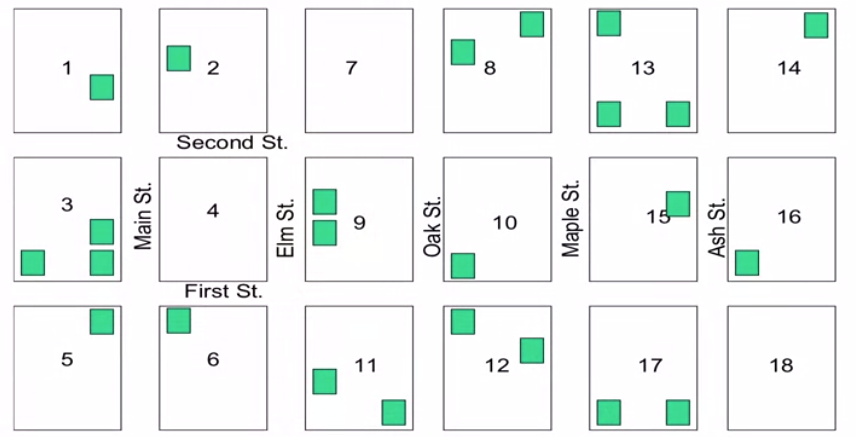
**3.1 Simple Complex Sampling – choosing entire clusters**

A population



Simple Random Sampling

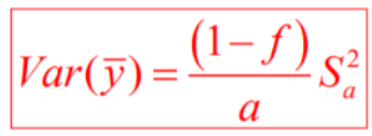
• Doesn’t have to include a sample from every block in the neighborhood



• If I don’t have a list of addresses… then we build our “own list” 🡺 list addresses by hand 🡺 but that list creation activity incurs “costs”

• We can still get “list of blocks” that is used in Census (but doesn’t list individual addresses often due to confidentiality/privacy issues) 🡺 we have the CLUSTER but don’t have the addresses

Cluster Sampling



• Populations often distributed geographically like this

- cannot afford to create an element frame

- cannot afford to visit n units drawn randomly from the entire area

• Cluster selections are used to reduce listing costs

- select clusters and list elements only for selected clusters

• Clusters are used to reduce travel costs

• Clusters are often already listed

- makes them “naturally occurring units”

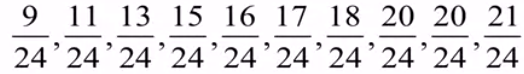
- seldom equal size

• Suppose we select an SRS of a = 10 classrooms from A=1000, and examine the immunization history of all b = 24 children in selected classrooms

• Here N = A x B = 1000 x 24 and n = a x b = 240

• We refer to the A classrooms as **primary sampling units** or **PSUs**

• For each of the (a = 10) selected PSU’s, we record the number of children immunized:

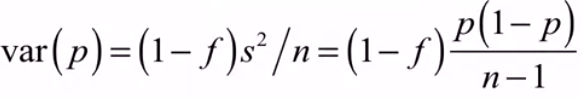


• Adding the numerators, there are 160 immunized children

• The overall proportion immunized is p = 160/240 = 0.67

• Recall from SRS (without replacement selection n elements), the sample proportion was 

• The estimated sampling variance is:

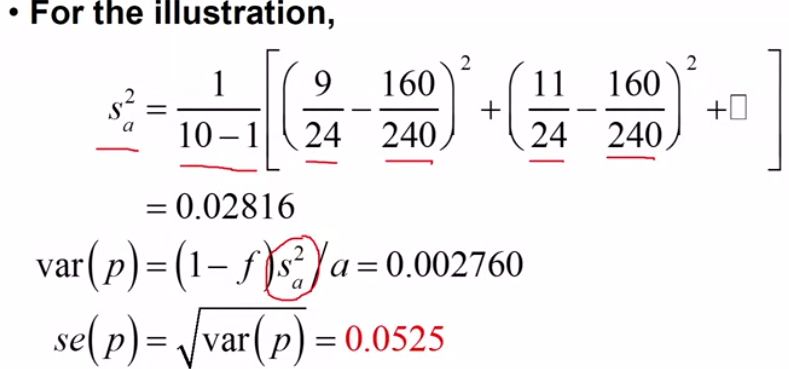
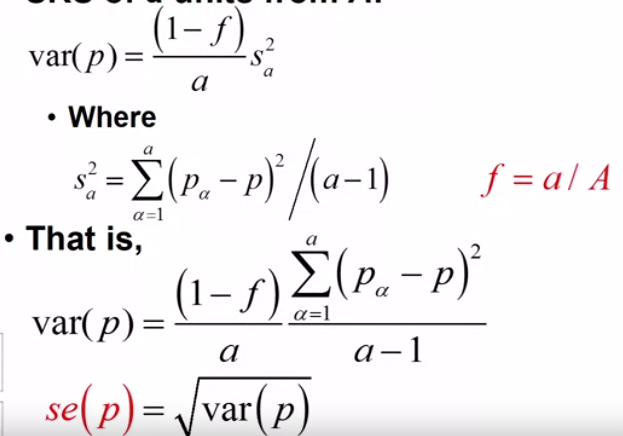


• But for an SRS of “a” equal sized clusters from “A”, we have sampled clusters not elements

• Randomization occurs at the cluster level

• We have a P\_a for each selected PSU

• In cluster sampling, treat the sample as an SRS of “a” units from “A”:



• Of course, the standard error is then used in a confidence interval

• But there is an important adjustment to what we’ve done up till now

• Recall that we briefly introduced he idea of using the t-distribution rather than the normal in confidence intervals

• That is much more important here with cluster samples than for simple random samples of elements

• That’s because the confidence interval is built on a standard error that depends on the number of random events in the sample

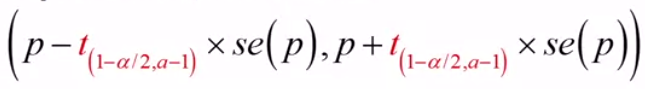
• The number of random events in simple cluster sampling is “a”, not “n”

• Hence we need to be worried about not “n” degrees of freedom, but “a” degrees of freedom

• And “a” is much smaller than “n”

• As a result, we will use “t-statistic” instead of the “z”

• In particular, we will use



• It’s the same as p, and the same standard error, but the multiplier for the standard error is from the t-distribution

• It’s because we have only a random events in the sample, not n – a much smaller number

• We need to use a larger multiplier for the confidence interval when the number of random events is smaller

**3.2 Design Effects**

• A question is how did the cluster sample compare to a simple random sample?

• Need to establish grounds for comparison

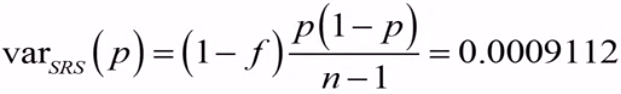
- compare precision since both designs are unbiased, and yield the same mean on average

- on what basis should the precision be compared?

- usually equal sample size

- And a comparison of sampling variances

• If the sample had instead been an SRS of n = 240 children from all schools, then p = 160/240



• Compared to cluster sampling, the estimated variance of p is considerably smaller for SRS