

8 Cluster Sampling

8.1 Introduction

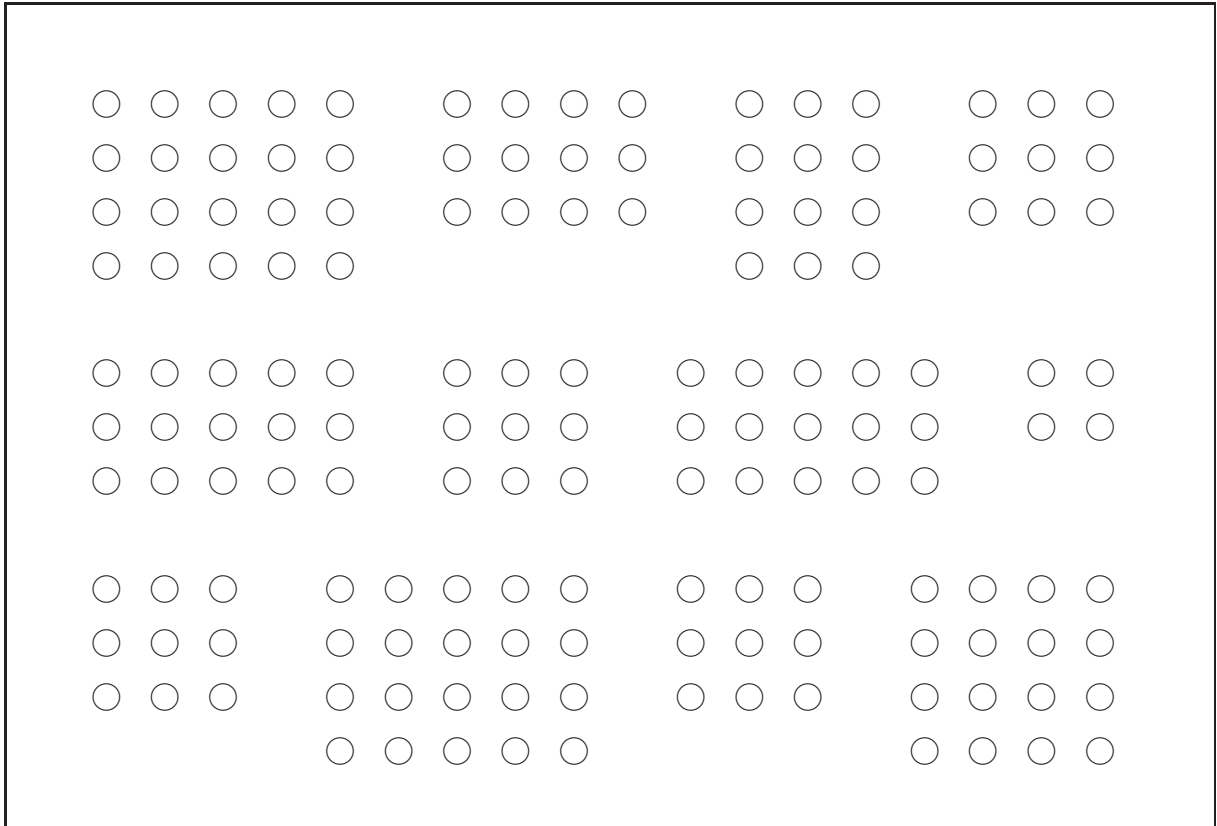
What is a cluster sample?

When should a cluster sample be used?

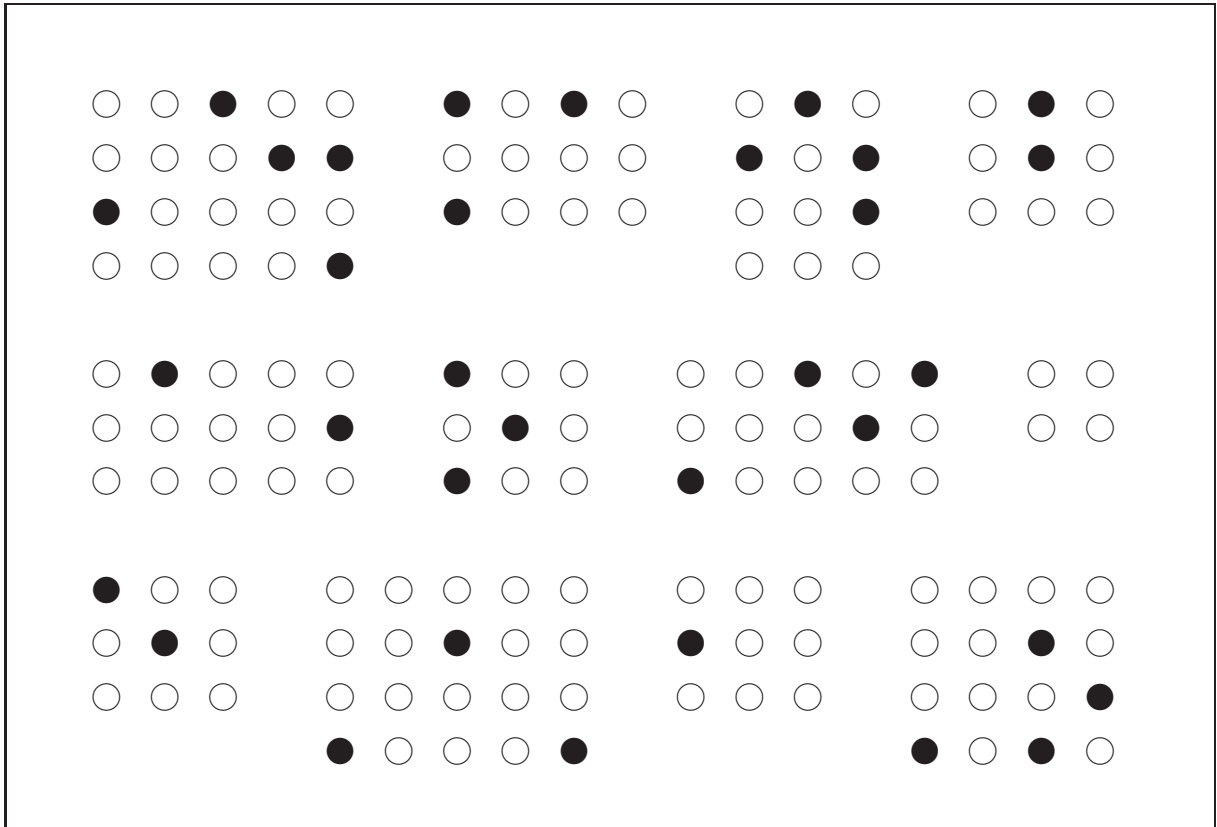
Recall the example from chapter 3: Suppose you want to survey Lutheran church members in Minneapolis to inquire about church donations, but you do not have a list of all church members in the city, so you cannot take a simple random sample of church members.

Why should cluster sampling be performed?

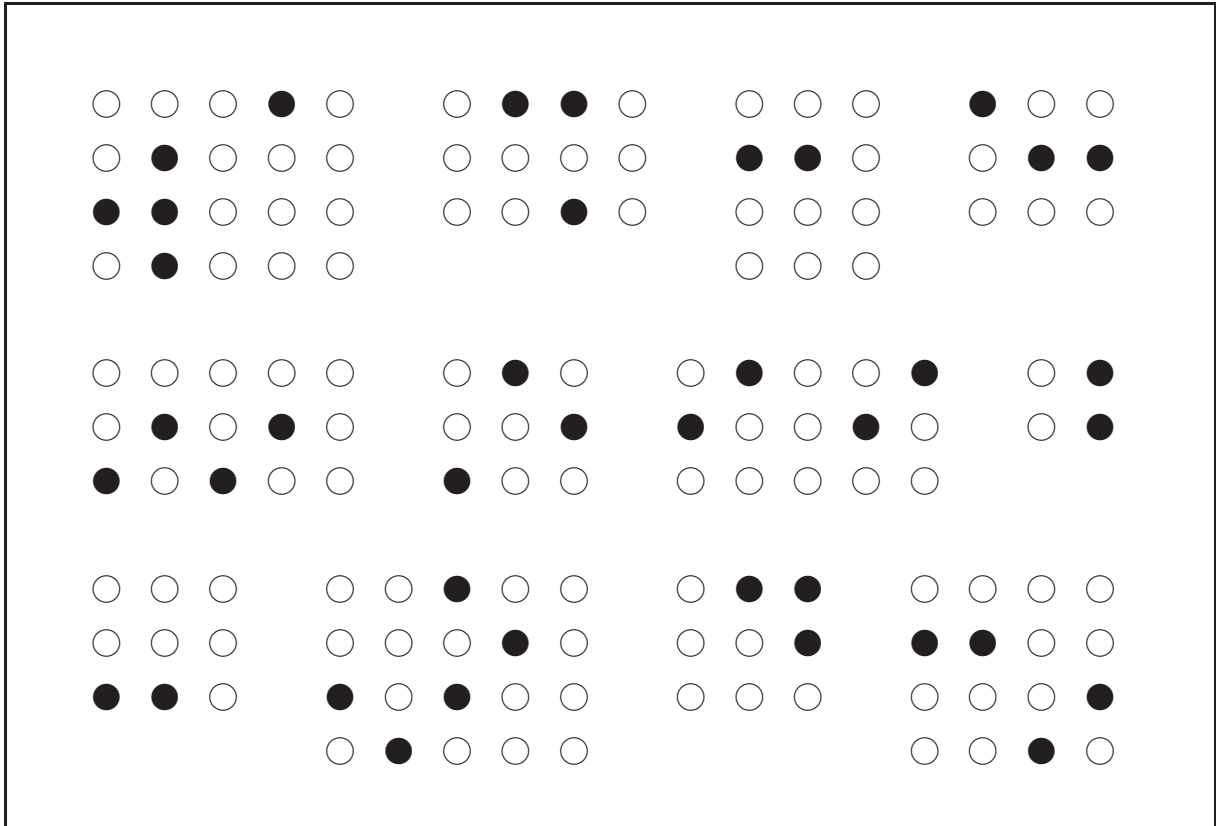
Consider the following target population.



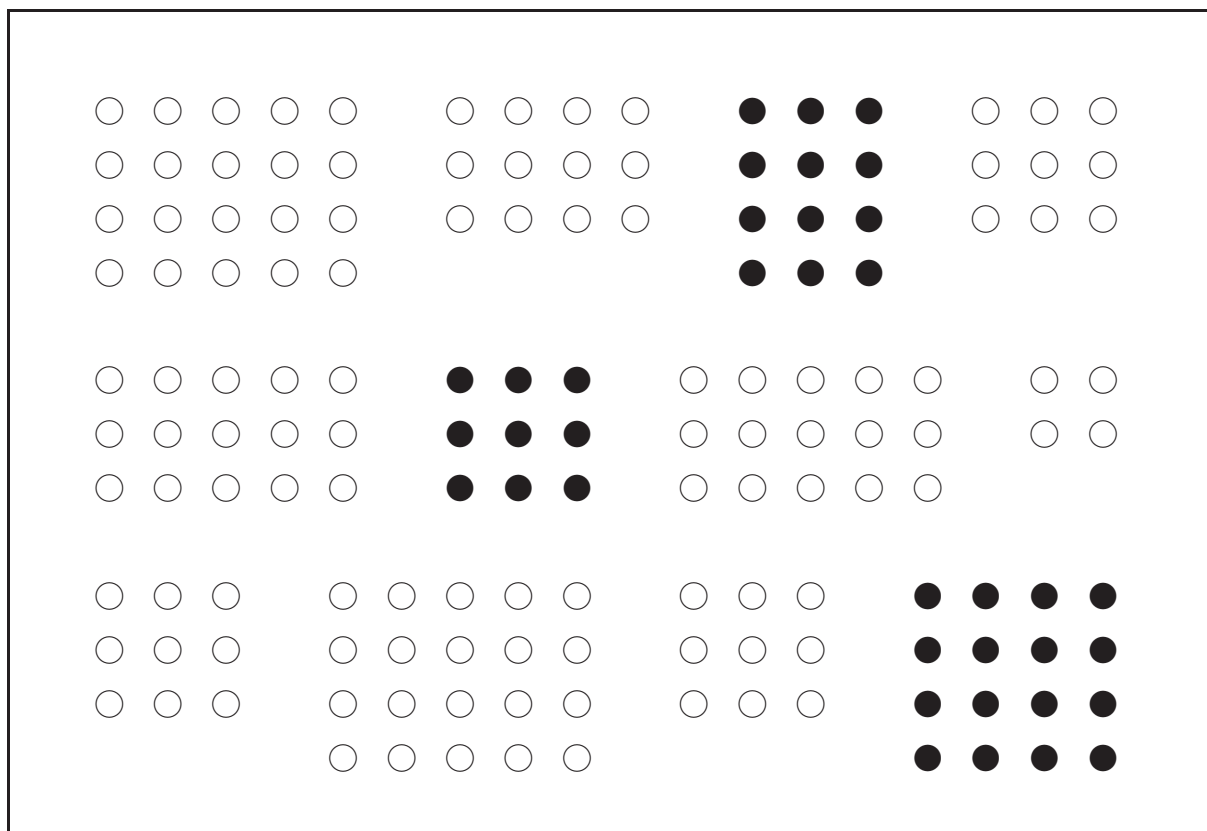
What type of sampling is represented here?



What type of sampling is represented here?



What type of sampling is represented here?



Example: Use cluster sampling via door-to-door sampling to estimate the population size of a city.

Suppose the city has 100 neighborhoods. Sample, say, 20 of the neighborhoods at random. Go door-to-door to inquire about household size. Suppose the sample average neighborhood size is 230 people. How should we estimate the population size of this city?

□

8.3 Estimation of a Population Mean and Total

Notation:

- ⊙ N = number of *clusters* in the population
- ⊙ n = number of clusters selected in a simple random sample
- ⊙ m_i = number of elements in cluster i , for $i = 1, \dots, N$
- ⊙ $\bar{m} = n^{-1} \sum_{i=1}^n m_i$ = average cluster size for the sample
- ⊙ $M = \sum_{i=1}^N m_i$ = number of elements in the population

- ⊙ Y_i = total of all observations in the i th cluster

Estimator of the population mean μ :

$$(8.1) \quad \hat{\mu} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n m_i}$$

What type of estimator is $\hat{\mu}$?

Estimated variance of $\hat{\mu}$:

$$(8.2) \quad \hat{V}(\hat{\mu}) = \left(\frac{N - n}{Nn(M/N)^2} \right) s_r^2,$$

where

$$(8.3) \quad s_r^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu} m_i)^2}{n - 1}$$

Suppose M , the population size, is unknown. How can M be estimated?

Exercise 8.2, p. 269: A manufacturer of band saws wants to estimate the average repair cost per month for the saws he has sold to certain industries. He cannot obtain a repair cost for each saw, but he can obtain the total amount spent for saw repairs and the number of saws owned by each industry. How should the clusters be selected?

The manufacturer selects a *simple random sample* of $n = 20$ from the $N = 96$ industries he services. The data on total cost of repairs per industry and number of saws per industry are as given in the table.

(a) Plot the data to determine if cluster sampling is appropriate.

```
> saws = scan2( "EXER8.2.DAT", T )  
> m = saws[ , 2] # vector for number of saws  
> y = saws[ , 3] # vector for cost
```

(b) Estimate the *average* repair cost per saw for the past month.

(c) Estimate the variance of $\hat{\mu}$ from part (a).

```
> N = 96
```

```
> n = length(y)
```

sr2 is the sample variance of $(y - \text{mu.hat} * m)$.

(d) Compute a bound on the error of estimation.

(e) Compute a 95% confidence interval on μ , the *average* repair cost per saw for the past month.

We are 95% confident that the population mean repair cost per saw for the past month is between \$17.95 and \$21.51. \square

Since the population total is $\tau = M\mu$ (with M known), then how should we estimate τ and its variance?

Estimator of the population total τ when M is known:

$$(8.4) \quad \hat{\tau} = M \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n m_i}$$

Estimated variance of $\hat{\tau}$ when M is known:

$$(8.5) \quad \hat{V}(\hat{\tau}) = N^2 \left(\frac{N-n}{Nn} \right) s_r^2$$

Suppose M is unknown when estimating τ in (8.4).

Define

$$(8.6) \quad \bar{Y}_t = n^{-1} \sum_{i=1}^n Y_i$$

Estimator of the population total τ when M is unknown:

$$(8.7) \quad N\bar{Y}_t = \frac{N}{n} \sum_{i=1}^n Y_i$$

Estimated variance of $N\bar{Y}_t$ when M is unknown:

$$(8.8) \quad \hat{V}(N\bar{Y}_t) = N^2 \hat{V}(\bar{Y}_t) = N^2 \left(\frac{N-n}{Nn} \right) s_t^2,$$

where

$$(8.9) \quad s_t^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_t)^2}{n-1}.$$

Note that s_t^2 is a sample variance based on simple random sampling.

Exercise 8.3, p. 293: Use the data in exercise 8.2.

(a) Estimate the *total* amount spent by the 96 industries on band saw repairs.

(b) Estimate the variance of $\hat{\tau}$.

(c) Compute a bound on the error of estimation.

(d) Compute a 95% confidence interval on τ , the *total* repair cost for the past month.

We are 95% confident that the total repair cost for the past month is between \$9,136.93 and \$15,487.07. \square

Exercise 8.4, p. 323 After checking his sales records, the manufacturer of exercise 8.2 finds that he sold a total of 710 band saws to these industries.

(a) Using this additional information, estimate the *total* amount spent on saw repairs by these industries.

> M = 710

(b) Estimate the variance of your estimator in part (a).

(c) Compute a bound on the error of estimation.

(d) Compute a 95% confidence interval on τ , the *total* repair cost for the past month.

We are 95% confident that the total repair cost for the past month is between \$12,898.06 and \$15,119.63. \square

8.5 Selecting the Sample Size for Estimating Population Means and Totals

Here, the sample size is n , the number of *clusters*.

Recall from section 8.3, we have ONE estimator for μ but TWO estimators for τ (corresponding to M known and M unknown).

When M is *unknown*, how should we estimate M ?

What is the estimator of μ , the population mean?

Approximate sample size required to estimate μ with a bound B on the error of estimation:

$$(8.12) \quad n = \frac{N\sigma_r^2}{ND + \sigma_r^2},$$

where σ_r^2 is estimated by s_r^2 (from a preliminary sample) and $D = B^2M^2/(4N^2)$.

$$(8.3) \quad s_r^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu} m_i)^2}{n-1}$$

⊙ For estimating τ , the population total, first consider the case where M is **KNOWN**.

What is the estimator of τ , the population total?

Approximate sample size required to estimate τ with a bound B on the error of estimation (when M is known):

$$(8.13) \quad n = \frac{N\sigma_r^2}{ND + \sigma_r^2},$$

where σ_r^2 is estimated by s_r^2 (from a preliminary sample) and $D = B^2/(4N^2)$.

⊙ Next consider the case where M is **UNKNOWN**.

Approximate sample size required to estimate τ with a bound B on the error of estimation (when M is unknown):

$$(8.15) \quad n = \frac{N\sigma_t^2}{ND + \sigma_t^2},$$

where σ_t^2 is estimated by s_t^2 (from a preliminary sample), and where $D = B^2/(4N^2)$.

$$(8.9) \quad s_t^2 = \sum_{i=1}^n \frac{(Y_i - \bar{Y}_t)^2}{n-1}$$

with $\bar{Y}_t = n^{-1} \sum_{i=1}^n Y_i$

Example (revisit exercise 8.2): Consider the original data in exercise 8.2 to be a preliminary sample, regarding the number of band saws sold and the total repair cost for each industry.

(a) What sample size is needed to estimate μ , the mean repair cost of a saw, within error of estimation bounded by \$1.1 (where M is UNKNOWN)?

$$n > N = 96$$

$$D = B^2 M^2 / (4N^2)$$

How do we determine D since M is unknown?

- (b) Assume, as in exercise 8.4, that the population size (number of saws) is KNOWN to be 710. What sample size is needed to estimate μ , the mean repair cost of a saw, within error of estimation bounded by \$1.1?

$$D = B^2 M^2 / (4N^2)$$

- (c) What sample size is needed to estimate τ , the total repair cost of all saws, within error of estimation bounded by \$781, where M is KNOWN to be 710?

Use equation (8.13).

$$D = B^2 / (4N^2)$$

Compare your answers in parts (b) and (c).

(d) What sample size is needed to estimate τ , the total repair cost of all saws, within error of estimation bounded by \$781, where M is UNKNOWN?

Use equation (8.15).

$$D = B^2/(4N^2)$$

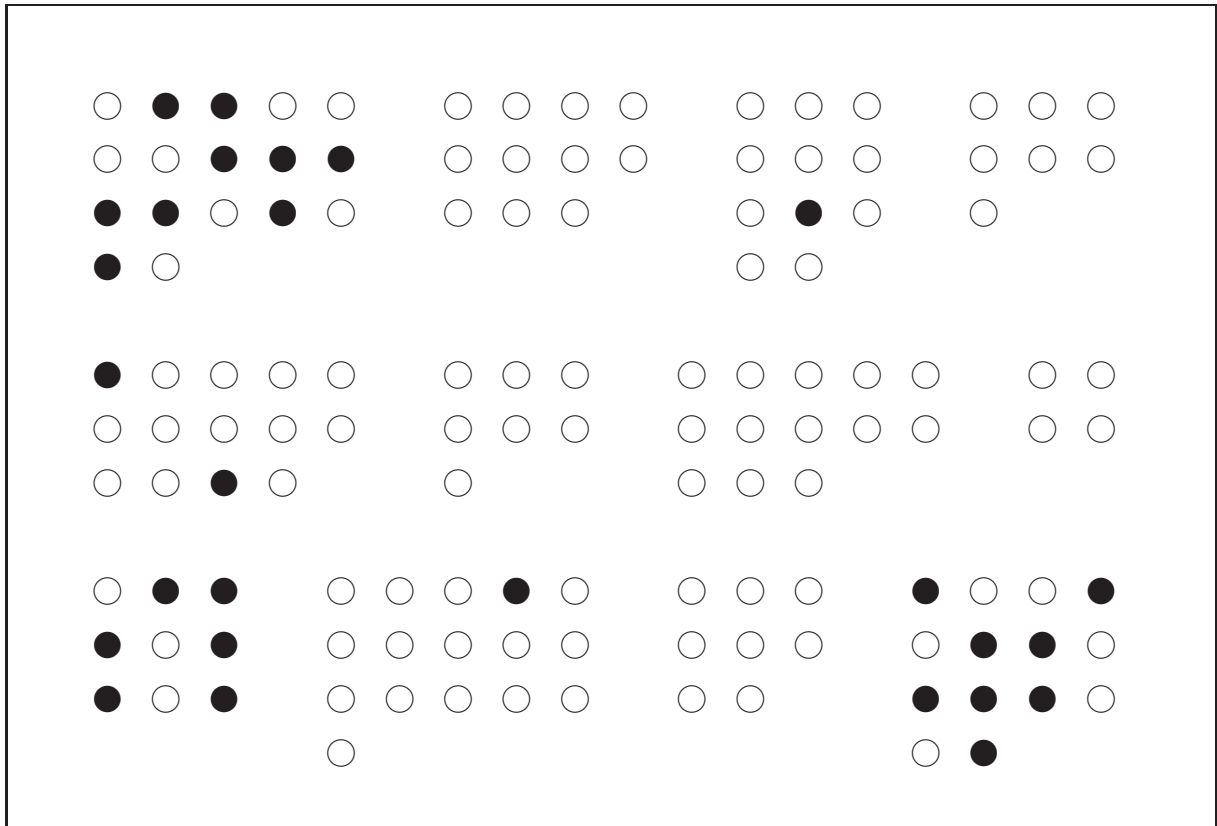
Compare your answers to parts (c) and (d).

□

8.6 Estimation of a Population Proportion

Example: Consider the following *data set*, not *target population*. The groupings represent *clusters*. Each open circle represents a zero, and each closed circle represents a one.

The data represent mice from a particular strain. Each cluster represents a pregnant mouse, a *dam*, exposed to some toxin. Each *open* circle represents a viable implant (i.e., a would-be *living* fetus). Each *closed* circle represents a non-viable implant (i.e., a would-be *dead* fetus). We are interested in the population proportion of non-viable implants for this strain.



Do these data appear to be a simple random sample?

Does it make sense to perform cluster sampling for this biological scenario?

What is a reasonable estimate of p , the population proportion of implants which are non-viable?

Let n = number of pregnant mice in the sample.

Let m_i = number of implants in the i th litter.

Let Y_i = number of non-viable implants in the i th litter.

$$\hat{p} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n m_i}$$

□

Recall (again!): A proportion is a special case of a mean, using zeros and ones as the population values.

In section 8.3, when estimating means, Y_i was the total of all observations from the i th cluster.

If 0 represents a *failure* and 1 represents a *success*, what does Y_i represent for a particular cluster?

The above formula for \hat{p} is equivalent to the formula for $\hat{\mu}$ in (8.1), p. 269.

Example (revisit the mice!):

(a) Estimate p , the population proportion of implants which are non-viable.

> m = c(17, 11, 11, 7, 14, 7, 13, 4, 9, 16, 8, 14)

> y = c(9, 0, 1, 0, 2, 0, 0, 0, 6, 1, 0, 8)

> # Apply equation (8.1), p. 269.

(b) Estimate the variance of your estimator of p (although $n < 20$).

> # Apply equations (8.2) and (8.3), p. 269.

$$(8.2) \quad \hat{V}(\hat{\mu}) = \left(\frac{N-n}{nN(M/N)^2} \right) s_r^2$$

Take M and N to be huge such that the population mean litter size M/N is fixed.

What is $\frac{N-n}{N}$, the finite population correction?

How should we estimate M/N ?

(c) Compute a bound on the error of estimation.

(d) Compute a 95% confidence interval on p .

We are 95% confident that the population proportion of implants which are non-viable is between 4.6% and 36.6%.

(e) What would have been the estimated variance of \hat{p} , if the observations (implants) had been **erroneously** treated as a simple random sample?

□

8.7 Selecting the Sample Size for Estimating Population Proportions

Again, p is a special case of μ , so use equation (8.12), p. 279, involving cluster sampling.

$$(8.12) \quad n = \frac{N\sigma_r^2}{ND + \sigma_r^2},$$

where σ_r^2 is estimated by s_r^2 and $D = B^2M^2/(4N^2)$.

$$(8.3) \quad s_r^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu} m_i)^2}{n-1}$$

How should we estimate M/N ?

Example (revisit the mice!): What sample size, n (the number of pregnant mice), is needed to estimate p , the population proportion of implants which are non-viable, within error of estimation bounded by 3% (using the previous sample as a preliminary sample for initial estimates)?

> # Apply equation (8.3), p. 269.

Taking N to be huge, equation (8.12) becomes approximately what formula?

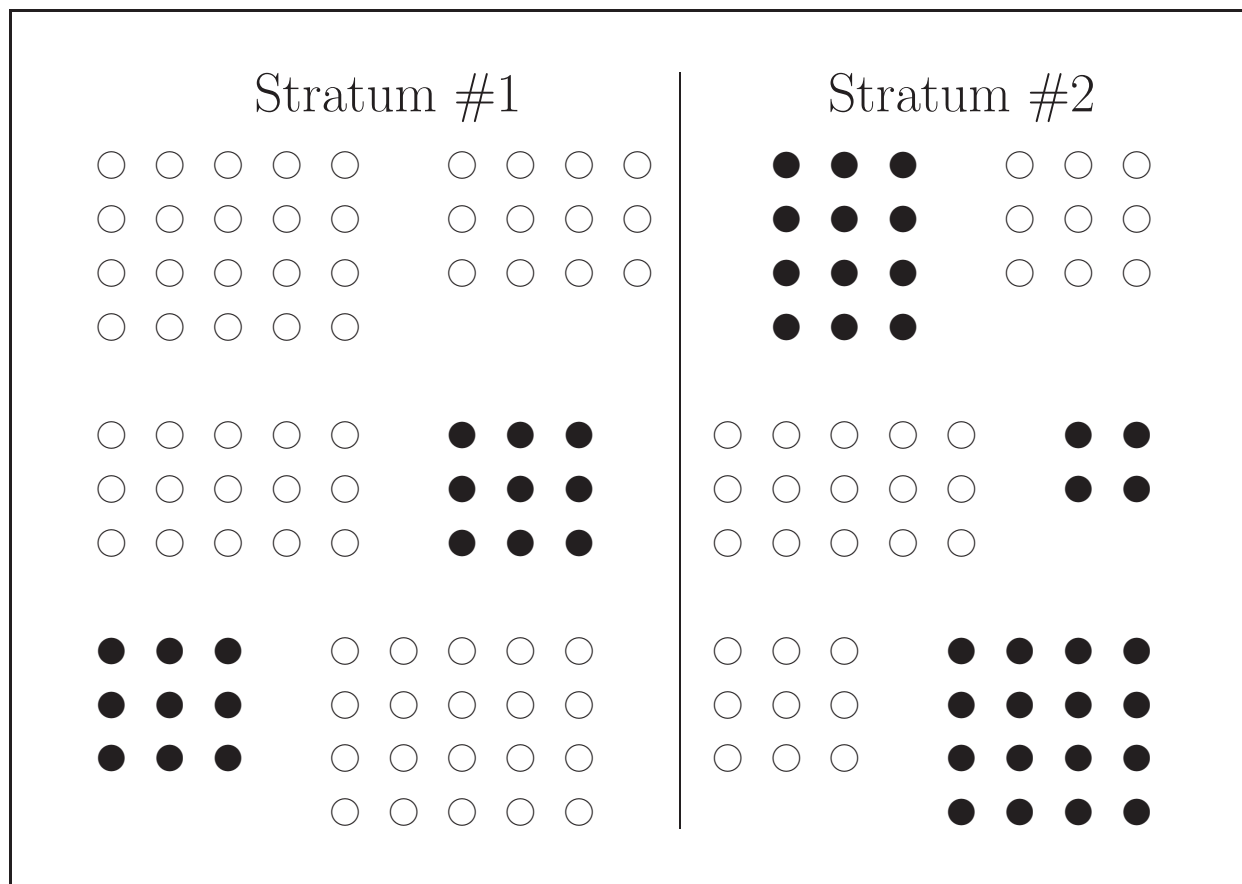
□

8.8 Cluster Sampling Combined With Stratification

Scenario:

- ⊙ Each stratum is divided into clusters.
- ⊙ Within each stratum, some clusters are sampled according to a simple random sample.

- ⊙ Within each cluster sampled, many or all of the elements are sampled.



To estimate a mean, use the stratified estimator (i.e., an estimator weighted by N_i/N).

Then, one may estimate a ratio, using the ratio of the stratified estimators.

Recall section 6.5, "Ratio Estimation in Stratified Random Sampling."

Exercise 8.19, p. 297: A certain firm specializing in the manufacture and sale of leisure clothing has 80 retail stores in Florida and 140 in California. With each state as a stratum, the firm wishes to estimate average sick-leave time per employee for the past year. Each outlet can be viewed as a cluster of employees, and total sick leave time for each store can be determined from records. Simple random samples

of 8 stores from Florida and 10 stores from California are taken, where m_i is the number of employees and Y_i is the total days sick leave for the i th store. Estimate the average amount of sick leave per employee.

```
> y.mat = scan2( "EXER8_19.DAT", T, T )
> y.fl = y.mat[ 1:8, 2 ]
> m.fl = y.mat[ 1:8, 1 ]
> y.ca = y.mat[ 9:18, 2 ]
> m.ca = y.mat[ 9:18, 1 ]
> N = c( 80, 140 )
> y.mean = c( mean(y.fl), mean(y.ca) )

> m.mean = c( mean(m.fl), mean(m.ca) )

> y.bar.st = sum( N * y.mean ) / sum(N) # Estimate mean days sick leave per store

> m.bar.st = sum( N * m.mean ) / sum(N) # Estimate mean number of employees per
store

> mu.hat = y.bar.st / m.bar.st # Combined stratified estimator
```

The average amount of sick leave per employee is estimated to be 2.69 days.

It can be shown that the estimated variance is 0.05595481 squared days, so the bound on the error of estimation is $2\sqrt{0.05595481} = 0.05595481$ days. \square