Adaptive Cluster Sampling

Week 9

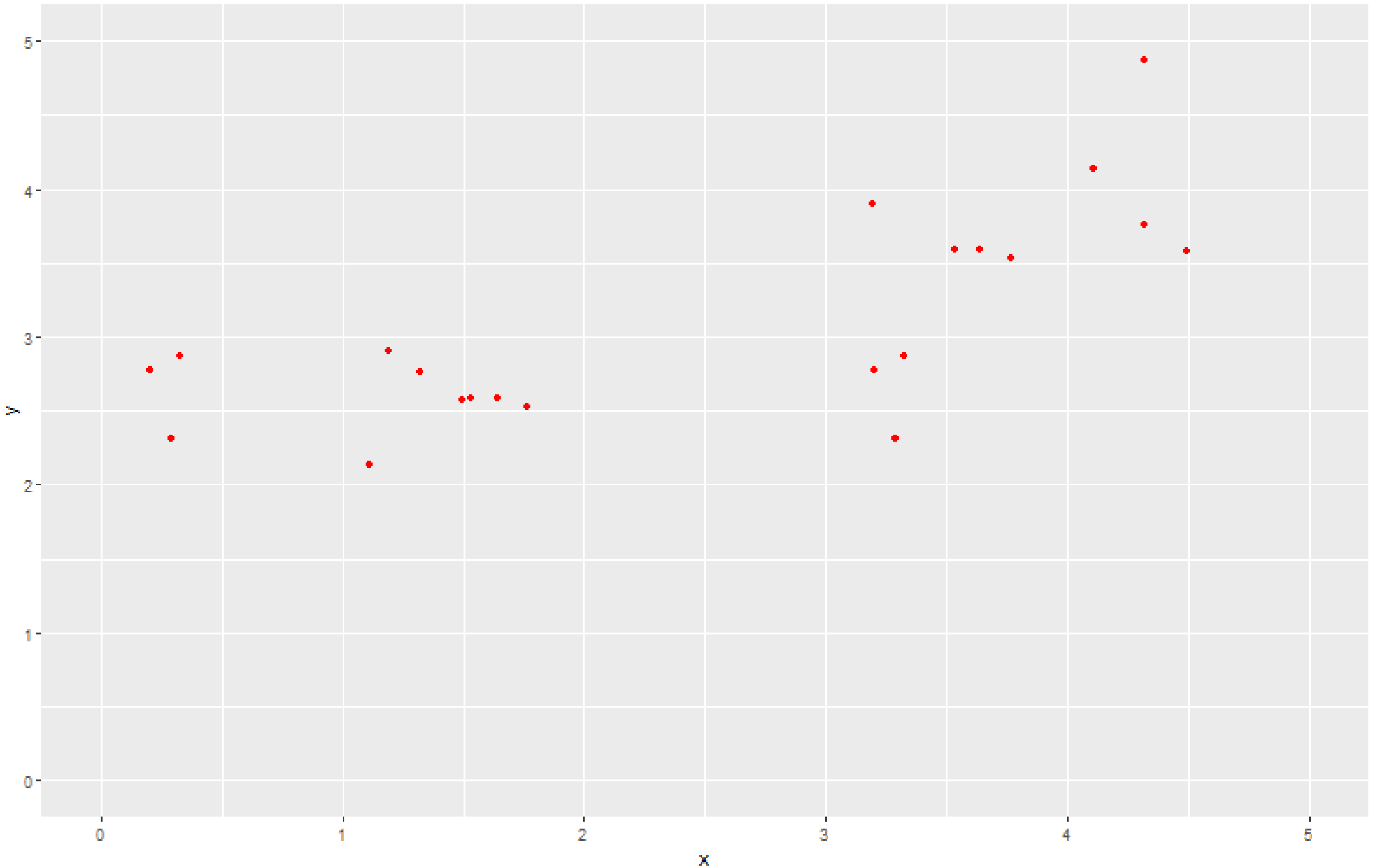
Stat 260, St. Clair

Adaptive designs

- *Adaptive* design: sampling units are determined "on the fly" based on observed characteristics of previously sampled units
- Adaptive *cluster* sampling (ACS): goal is to sample rare, clustered populations
- *cluster* denotes a spatial, social or genetic "closeness"
    - rare animal/plant species that are spatially clustered
    - rare disease/trait that are spatially *or* socially or genetically clustered

Example

A simplified "strawberry field" example: how many plants (dots) in the field?



Adaptive designs

- Idea of ACS:

  (1) get an initial SRS of units

  (2) if sampled unit i meet a condition C and add is neighboring units to the sample

  (3) repeat (2) until no more units can be added

Example

ACS design:

- Divide the region into grid plots to create a sampling frame
- Sampling unit: grid plot (N=25)

Example

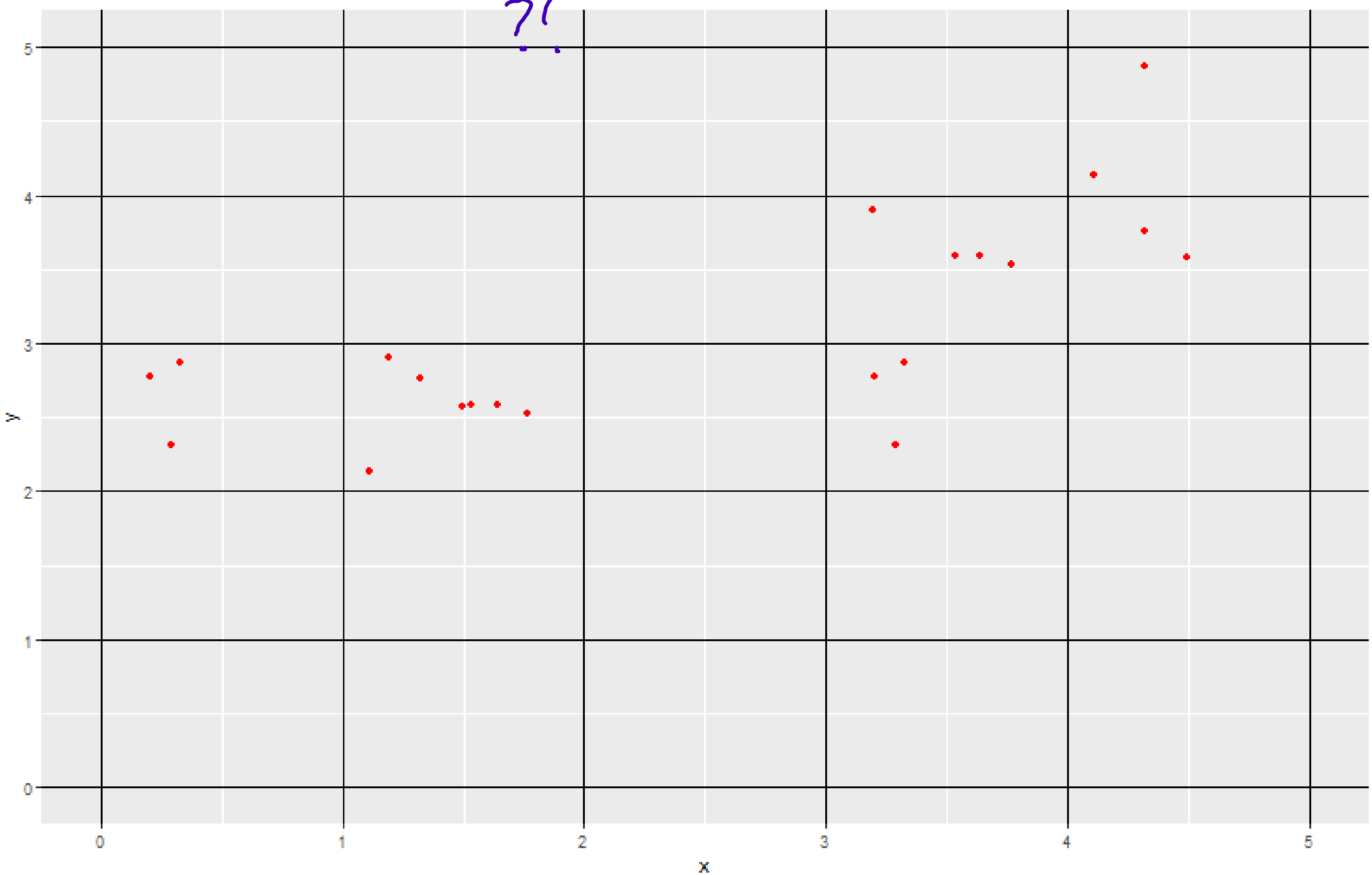
$N = 25$

$y_i = \#$ plants in ~~unit~~ cell i

$$t = \sum_{i=1}^{25} y_i \quad (= 21)$$

??

Sampling frame grid:



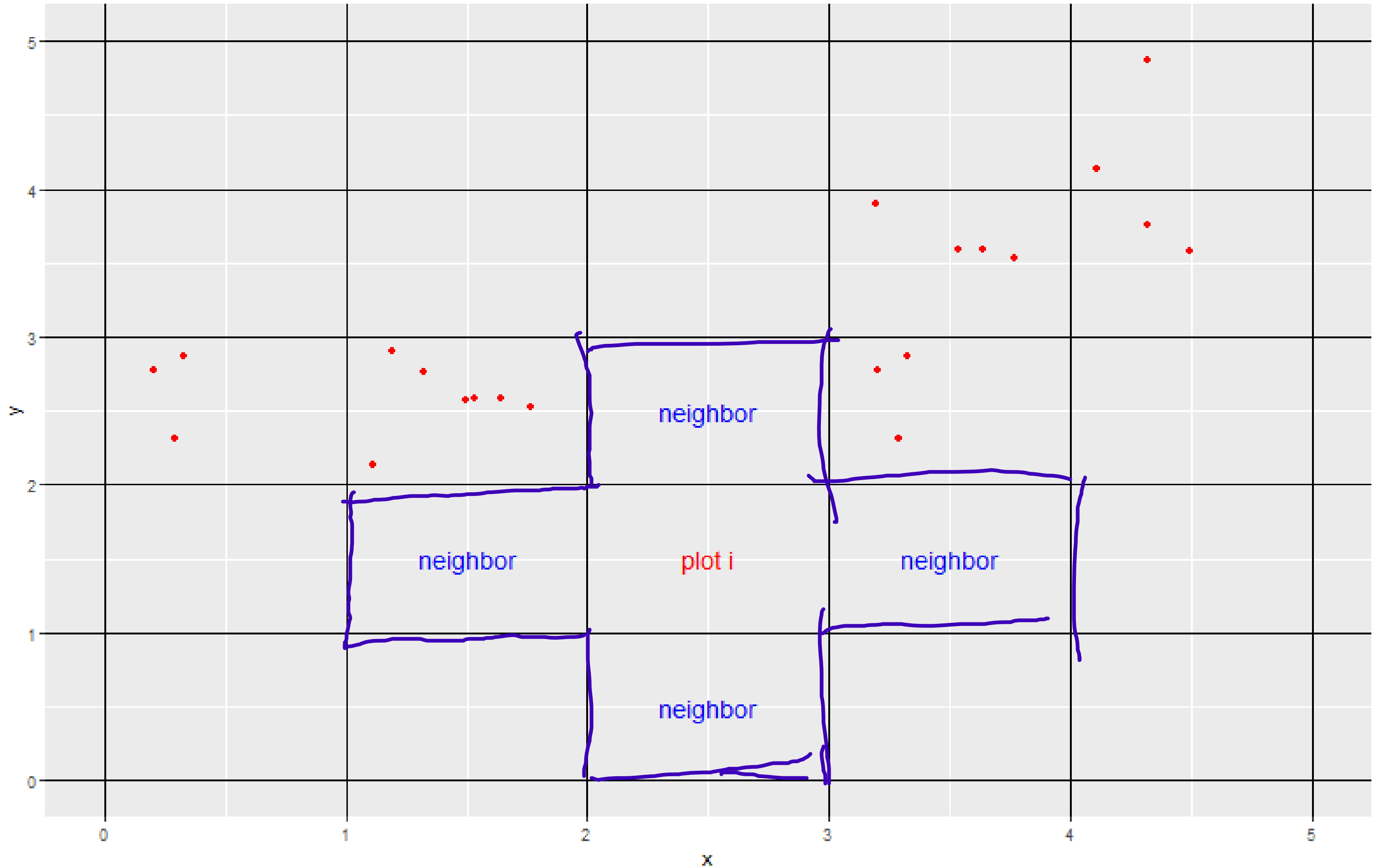
Example

ACS design:

- define a neighborhood:
  - plot i's neighbors are cells to the north/south/east/west

Example

Neighborhood of plot i



Example

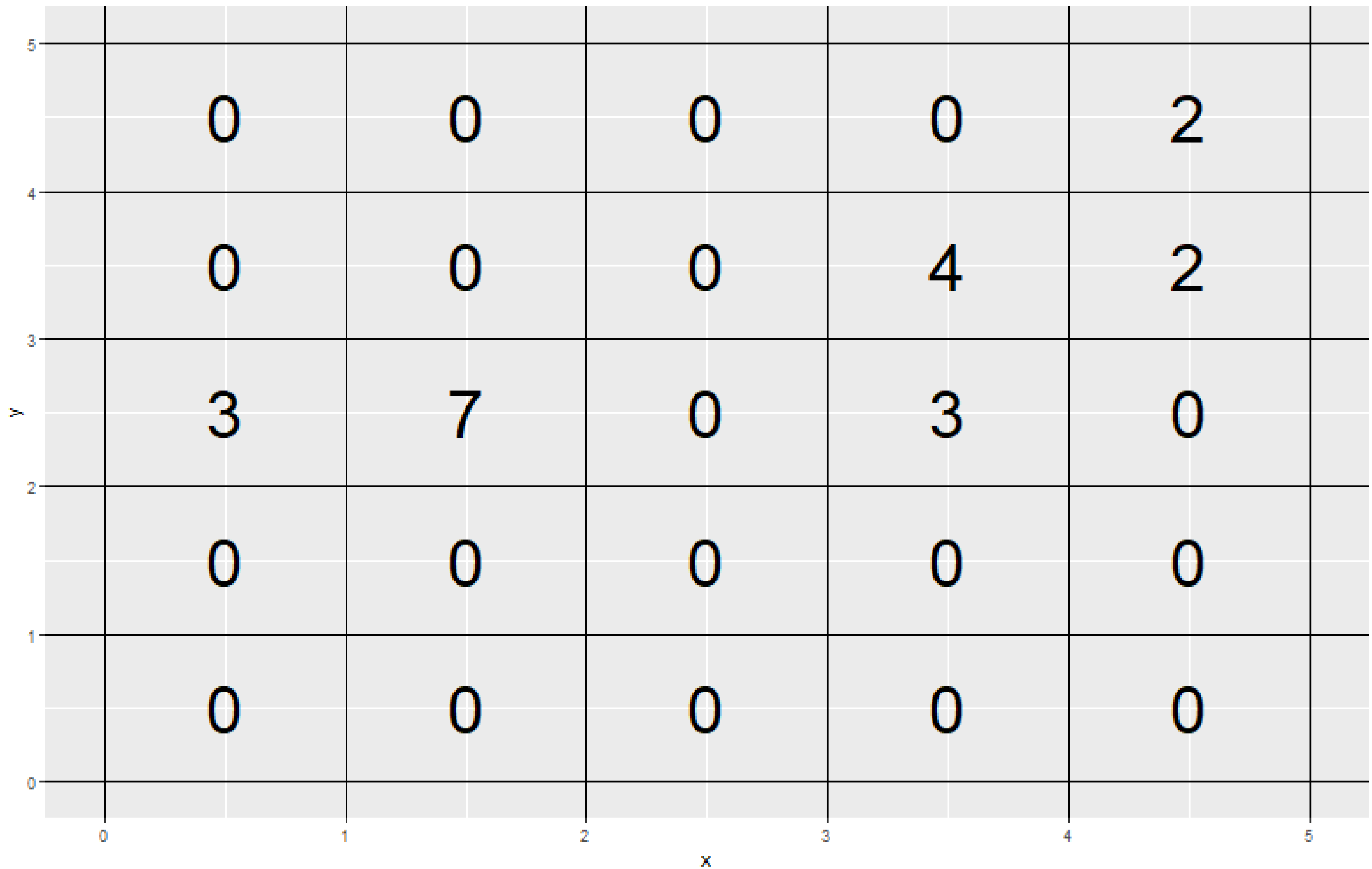
ACS design:

- Determine condition C: we want to find the plots with plants!
  - $y_i =$ number of plants in plot i
  - $C: y_i > 0$, add neighbors if plot i contains plants
- All that matters in the population is units, neighborhoods and y_i values

$C = \{y_i > 0\}$

Example

Let's just look at y_is



Example

ACS design:

(1) Initial SRS of $n_1 = 3$ plots and count plants y_i

(2) **adaptively add units:** If $y_i > 0$, add plot i's neighbors

(3) Repeat (2) until no more neighbors are adaptively added

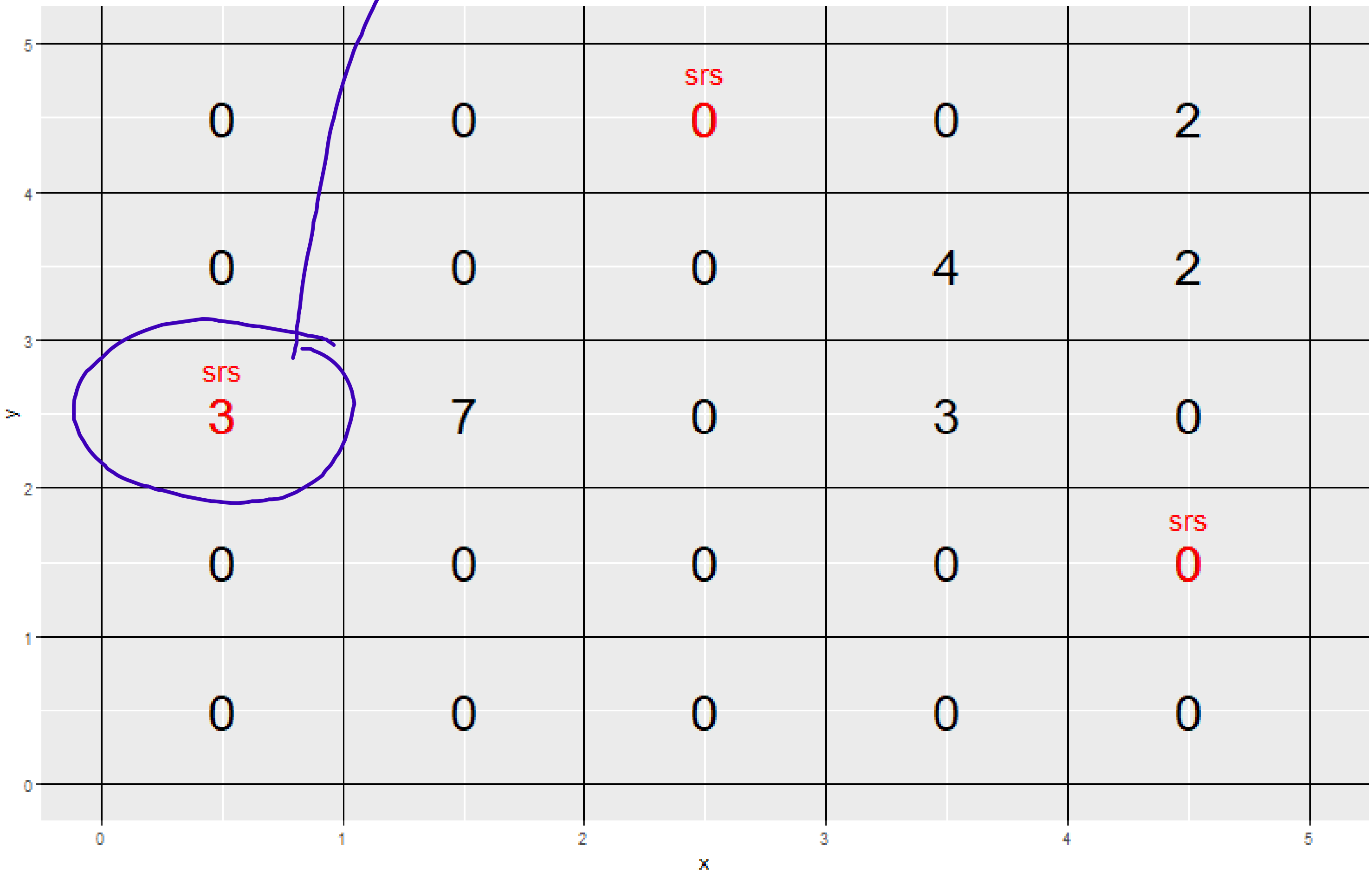
- all neighbors have $y_i = 0$

final sample size is unknown.

Example

$\{y_i > 0\}$ is met

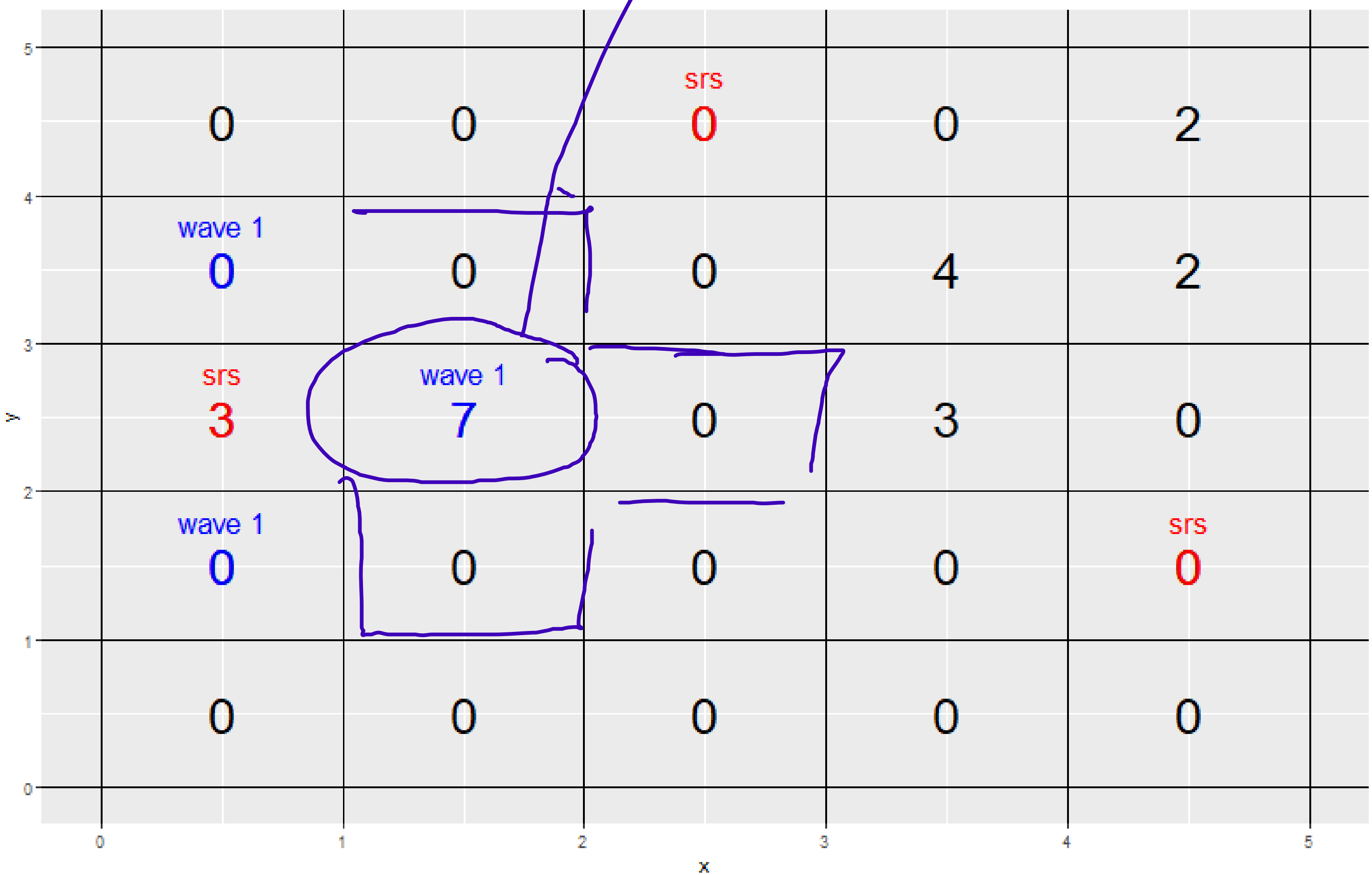
(1) Initial SRS of size 3 is highlighted



Example

meets the condition $\{y_i > 0\}$

(2) add first round of neighbors:

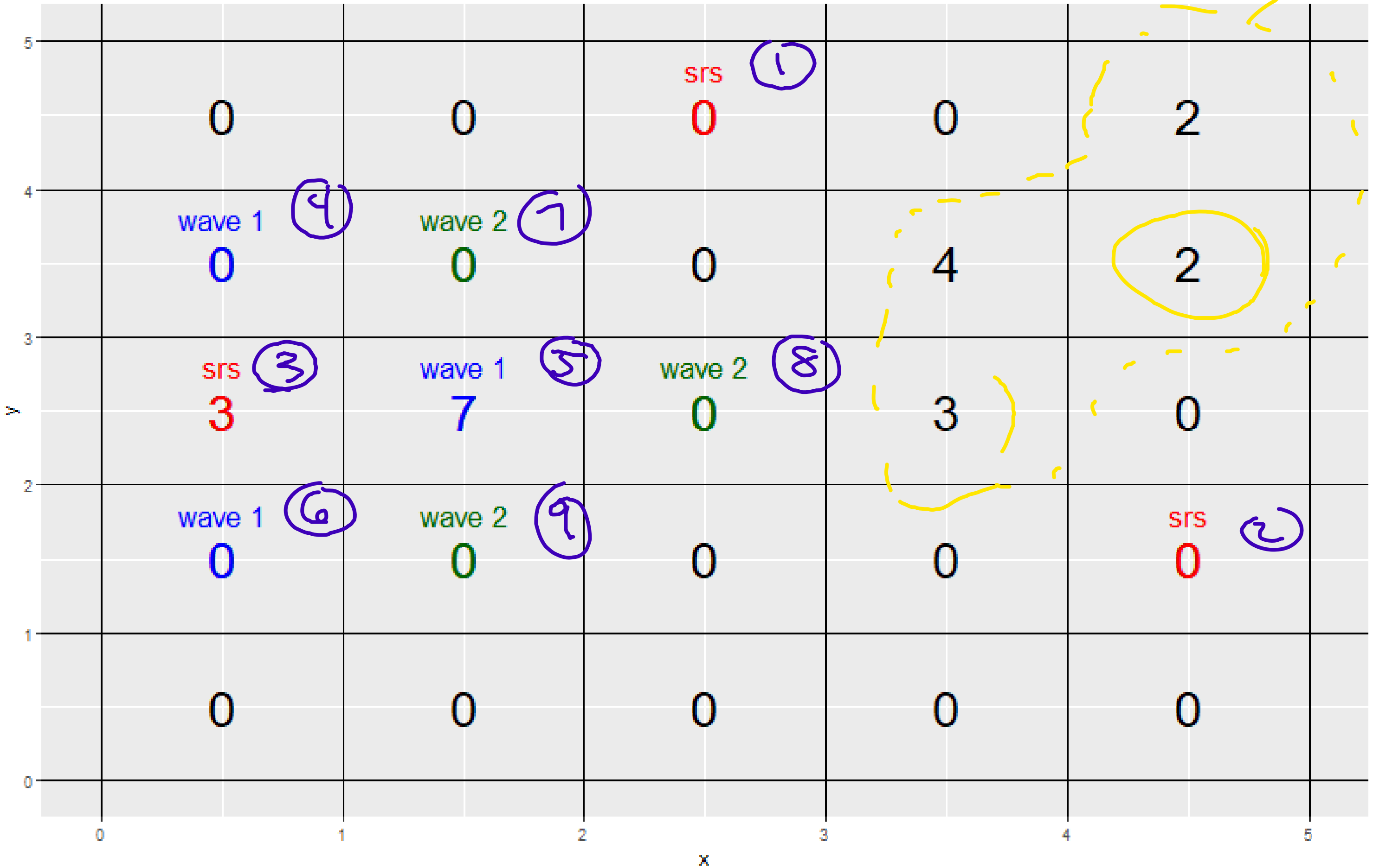


Example

all wave 2 units have $y_i = 0$
$\Rightarrow$ no new units added.

(2) add second (and final!) round of neighbors:

unseen info



Estimation

data : {3, 7, 0, 0, 0, 0, 0, 0, 0}

Should we use the SRS estimate? Will it be biased?

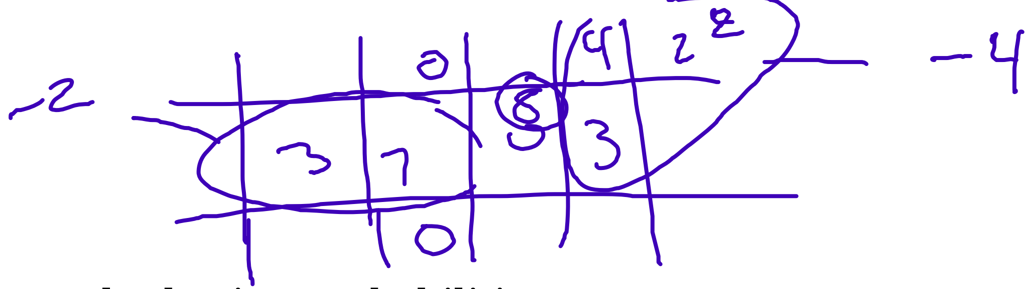
$$N\bar{y} = 25\frac{3+7+0+\cdots+0}{9} \approx 27.78$$

ACS design $\longrightarrow$ designed to add lots of units with plants ($y_i > 0$) $\longrightarrow$ raw data & treat it like an SRS

$\Longrightarrow$ tend to overestimate mean/total

* need to account for (unequal) inclusion probs. with HT-estimator.

Estimation



- Units have unequal selection probabilities
  - need to use a Horvitz-Thompson estimate of total!
  - units with $y_i > 0$ have higher inclusion probabilities
- Inclusion probability for unit i looks like

$$\pi_i = P(\underline{\text{unit } i \text{ is in the SRS}} \text{ or } \underline{\text{adaptively added}})$$

$\pi_8 = 1 - P(\text{unit 8 not in SRS \& not adapt. added})$

$$= 1 - \frac{\text{\# of SRS that don't contain } \underline{\text{unit 8}}^{-1}\text{, and don't contain any unit from the "clusters" that neighbor unit 8}}{\binom{25}{3}}$$

$$= 1 - \frac{\binom{25-1-2-4}{3}}{\binom{25}{3}} = \boxed{.645}$$

unknown!

Not possible to compute with our ACS data

Estimation

Problem: unless we see the *entire population*, we can't compute all π_i for observed units

- can't tell if unit i borders a cluster unless we've seen all units around it

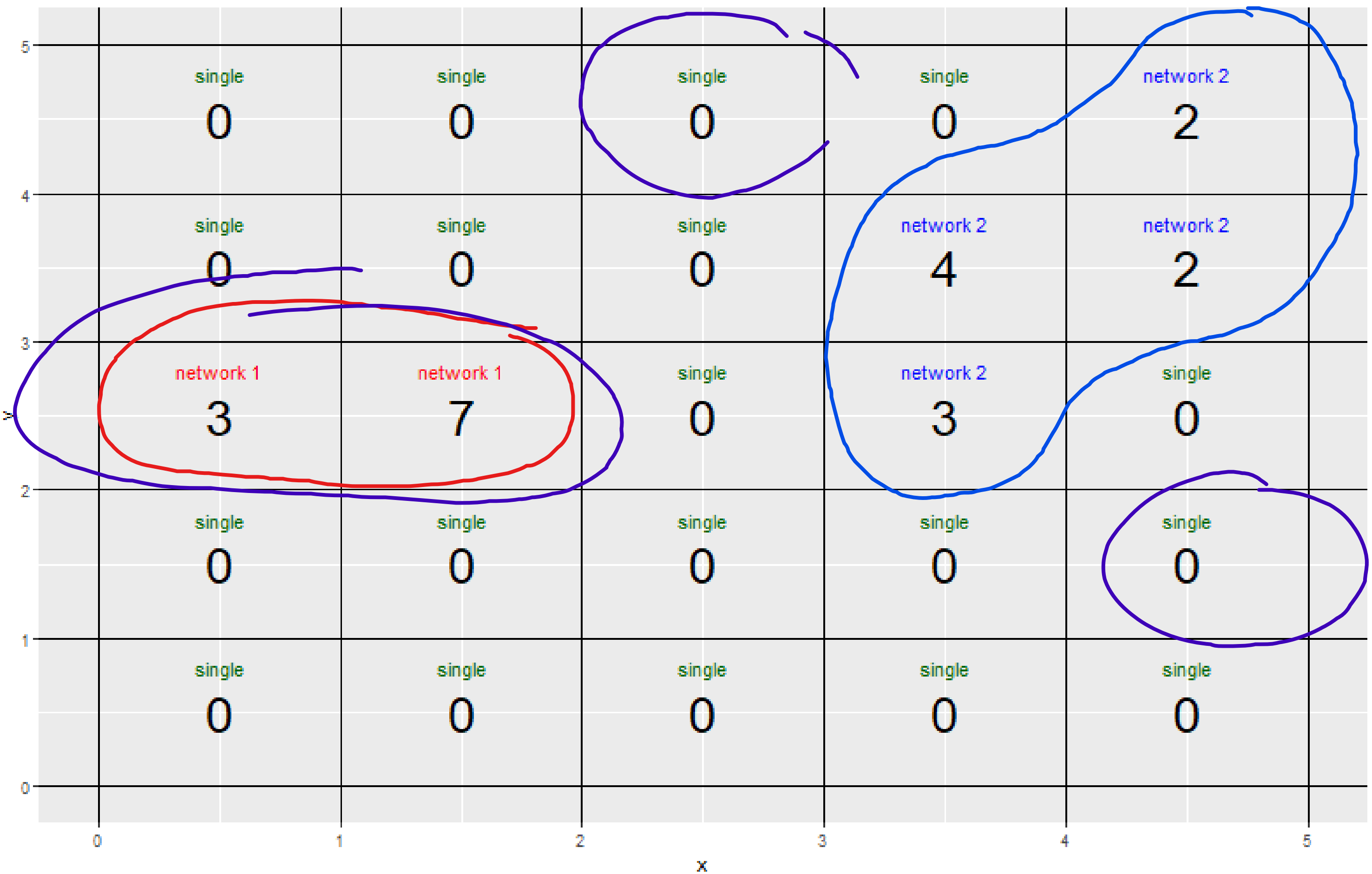
Networks instead of plots

Solution: define observations in terms of *networks*

- **Network:** a cluster of units generated by selection of any of the units within the cluster
- networks either
    - contain units that all satisfy condition C
    - are a single unit where C is not satisfied

Example

Two networks satisfy $y_i > 0$, 19 other networks are single plot with $y_i = 0$.



Networks

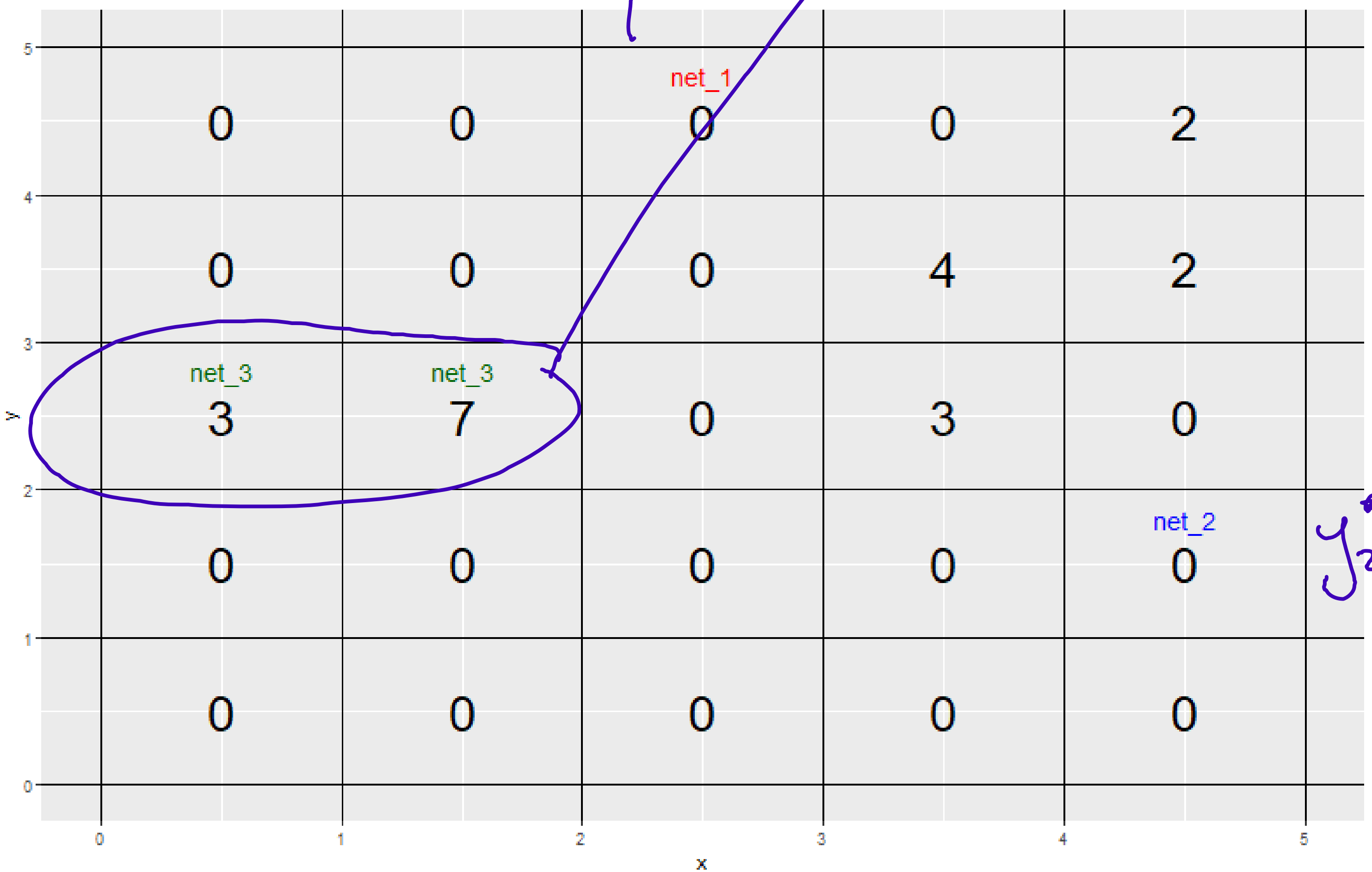
- Population has K distinct networks
  - Ex: $K = 19 + 2 = 21$
- Sample has κ distinct networks
  - Ex: $\kappa = 3$

Example

$y_1^* = 0$

$y_3^* = \sum_{i \in net3} y_i = 3 + 7 = 10$

Three networks were sampled



$y_2^* = 0$

Networks

- Let y_k^* be the total response of all units in network k

$$y_k^* = \sum_{i \in net_k} y_i$$

- We still have the same population total:

$$t = \sum_{i=1}^{N} y_i = \sum_{k=1}^{K} y_k^*$$

- Use HT estimator to weight the observed network totals y_k^*

Network inclusion probabililities

- α_k is the inclusion probability for network k:
    - network k is included in the ACS if **at least one** of its units is in the **initial SRS**

- Compute α_k using the complement rule:

$$\alpha_k = 1 - P(\text{no units in } k \text{ are in the initial SRS})$$

$\boxed{\text{Net\_3}}$

2 units

$\alpha_3 = 1 - P(\text{no units in net 3 are in initial SRS})$

network size

$$= 1 - \frac{\binom{25-2}{3}}{\binom{25}{3}} = .23$$

Network inclusion probabililities

- There are $\binom{N}{n_1}$ possible SRS of size n_1
- $x_k =$ number of units in network k
- There are $\binom{N-x_k}{n_1}$ possible SRS of size n_1 that **don't contain any units** in network K

$$\alpha_k = 1 - P(\text{no units in } k \text{ are in the initial SRS}) = 1 - \frac{\binom{N-x_k}{n_1}}{\binom{N}{n_1}}$$

Example

- Three sampled networks:
- $net_3 = \{3, 4\}, y_3^* = 10, x_1 = 2, \alpha_3 = 1 - \frac{\binom{25-2}{3}}{\binom{25}{3}} = 0.23$
- $net_1 = \{1\}$

$y_1^* = 0 \quad x_1 = 1 \quad \alpha_1 = 1 - \frac{\binom{25-1}{3}}{\binom{25}{3}} = .12$

- $net_2 = \{2\}$ $\quad x_2 = 1$

$\alpha_2 = .12$

Example

- Estimated total: $\hat{t}_{HT} = ?$

$$\hat{t}_{HT} = \sum_{k=1}^{3} \frac{y_k^*}{\alpha_k} = \frac{0}{.12} + \frac{0}{.12} + \frac{10}{.23} = 43.5$$

networks

Network inclusion probabililities

- Joint inclusion probability that both networks j and k are in the ACS
  - Use the rule: $P(j \text{ or } k) = P(j) + P(k) - P(j \text{ and } k)$ → α_{jk}
- So the probability of j **and** k is

$$
\begin{aligned}
P(j \text{ and } k) = \alpha_{jk} &= \alpha_j + \alpha_k - P(j \overset{\text{or}}{\cancel{\text{and}}}\ k \text{ in ACS}) \\
&= \alpha_j + \alpha_k - \left(1 - P(\text{neither } j, k \text{ in ACS})\right) \\
&= \alpha_j + \alpha_k - \left(1 - \underbrace{\frac{\binom{N - x_j - x_k}{n_1}}{\binom{N}{n_1}}}_{\text{neither } j,k}\right)
\end{aligned}
$$

→ complement

Example

- Joint prob for networks 1 and 2:

$x_1 = 1 \quad x_2 = 1$

$$\alpha_{12} = \overset{\alpha_1}{0.12} + \overset{\alpha_2}{0.12} - \left(1 - \frac{\binom{25-1-1}{3}}{\binom{25}{3}}\right) = 0.01$$

- Joint prob for networks 1 and 3, and also 2 and 3:

$x_1 = x_2 = 1$

$x_3 = 2$

$$\alpha_{13} = \alpha_{23} = \underset{\alpha_1,\ \alpha_2}{0.12} + \underset{\alpha_3}{0.23} - \left(1 - \frac{\binom{25-1-2}{3}}{\binom{25}{3}}\right) = 0.01957$$

options $\underline{\underline{HT}}$ or SYG est. of variance

Example

y^*

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} t_i^2 + 2 \sum_{i} \sum_{\substack{k \\ i<k}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k}$$

all pairs of networks => 3 pairs

- SE for $\hat{t}_{HT}$: only need to sum over non-zero network responses (and all joint products are 0!)

$\rightarrow \frac{1-\alpha_3}{\alpha_3^2}(10)^2$

$$SE_{HT}(\hat{t}_{HT}) = \sqrt{\frac{1-0.23}{0.23^2} 10^2 + 2\underline{(0)}} = 38.15$$

$$\Rightarrow \hat{V}_{HT} = \sum_{k=1}^{K} \frac{1-\alpha_k}{\alpha_k^2} (y_k^*)^2 + 2 \sum_{j} \sum_{k} \frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_{jk}} \frac{y_j^*}{\alpha_j} \frac{y_k^*}{\alpha_k}$$

pairs networks

$y_1^* = y_2^* = 0$

1, 2
1, 3
2, 3

Example

- To estimate in R, enter network level data: y_k^* and x_k

```
> acs_data <- data.frame(
+   y_net = c(0,0,10),
+   x_net = c(1,1,2) )
```

y_k^*

x_k

- Then get single network inclusion probabilities:

$1 - \binom{N-x_k}{3} \Big/ \binom{N}{3}$

```
> n1 <- 3
> N <- 25
> acs_data$pi_single <- 1- choose(N - acs_data$x_net,n1)/choose(N,n1)
> acs_data
  y_net x_net pi_single
1     0     1      0.12
2     0     1      0.12
3    10     2      0.23
```

α_k

Example

- Joint inclusion probabilities take more work
  - `jnt_fun` computes α_{jk} for all $k = 1, \ldots, \kappa$

```
> jnt_fun <- function(xj,x=acs_data$x_net,N=25,n1=3)
+ { 1- choose(N - xj,n1)/choose(N,n1) -
+     choose(N - x,n1)/choose(N,n1) +
+     choose(N - xj-x,n1)/choose(N,n1)}
> jnt_fun(xj = 1)
[1] 0.01000000 0.01000000 0.01956522
> jnt_fun(xj = 2)
[1] 0.01956522 0.01956522 0.03826087
```

$\Rightarrow \pi_{jk}$ for each x_k in data

$x_k = 1 \Rightarrow (x_1 = 1, x_2 = 1, x_3 = 2$

$\pi_{12} = .1 \quad \pi_{12} = .1 \quad \pi_{13} = .0196$

Example

- Fill the rows of the inclusion matrix:

```
> jnt_mat <- matrix(
+   c(jnt_fun(acs_data$x_net[1]),
+   jnt_fun(acs_data$x_net[2]),
+   jnt_fun(acs_data$x_net[3])),
+   byrow=TRUE, nrow=3)
> diag(jnt_mat) <- acs_data$pi_single αk # fix diagonals
> jnt_mat
           [,1]       [,2]       [,3]
[1,] 0.12000000 0.01000000 0.01956522   row 1
[2,] 0.01000000 0.12000000 0.01956522   row 2
[3,] 0.01956522 0.01956522 0.23000000   row 3
```

Example

- Then use "pps" design:

$\propto_k$

```
> library(survey)
> acs_design <- svydesign(id = ~1, fpc= ~pi_single,
+                         pps=ppsmat(jnt_mat), data=acs_data)
> svytotal(~y_net, acs_design)
       total     SE
y_net 43.478 38.152
```

- Again, get $\hat{t}_{HT} = 43.48$ and SE of 38.15.
- Note: Unless n is very large and clusters not "too clustered", you can't trust conventional confidence intervals for ASC data!