# `Survey` package: unequal probabilities of selection (WOR)

## Week 9

Stat 260, St. Clair

# Unequal probability sampling

What you need:

- $\pi_i$ inclusion probabilities for selected units
- $\pi_{ij}$ joint inclusion probabilities for selected units
    - optional, if not included you get *with* replacement overestimation of variance

$\widehat{Var}_{WR}(\hat{t}_{HT})$

# Design object: Unequal probability sampling

- `id` PSU id (often just `id = ~1`)
- `fpc` inclusion probabilities $\pi_i$ for your sampled units
- `pps` a `ppsmat` object that contains the inclusion probability matrix
  - something like `pps=ppsmat(mat)` where `mat` is a matrix with $\pi_{ij}$ on the off-diagonals and $\pi_i$ on the diagonals.
  - if left out, you get the with[redacted]replacement SE (Lohr equation 6.24)
- `variance` (optional) specifies variance formula
  - default is Horvitz-Thompson SE estimate (Lohr equation 6.22)
  - `YG` is Sen-Yates-Grundy (SYG) formula (Lohr equation 6.23)
- `weights` are not specified!

# Lohr Example 6.10: Ag census data

**Design**

- Sample 15 counties in 1992 using initial selection probabilities that are proportional to the 1987 farming acreage (`acres87`) in each county (WOR)

```
> agpps<- read.csv("http://math.carleton.edu/kstclair/data/agpps.csv'
> dplyr::glimpse(agpps)
Rows: 15
Columns: 24
$ county   <chr> "ST MARTI", "HALIFAX", "DOUGLAS", "WASHINGT", "COCHR
$ state    <chr> "LA", "NC", "MO", "IL", "TX", "IA", "MN", "WY", "TX'
$ farms92  <int> 274, 346, 1187, 831, 227, 812, 947, 207, 1576, 429,
$ largef92 <int> 18, 70, 33, 57, 122, 73, 74, 61, 116, 177, 48, 81, 3
$ acres92  <int> 70936, 204443, 300970, 297003, 370572, 353683, 39502
$ acres87  <int> 73265, 208333, 295392, 315971, 330711, 344010, 36811
$ sizemeas <int> 73265, 208333, 295392, 315971, 330711, 344010, 36811
$ pii      <dbl> 0.001137612, 0.003234862, 0.004586659, 0.004906196,
$ weight   <dbl> 879.03419, 309.13220, 218.02364, 203.82389, 194.7393
$ jtprob1  <dbl> 0.0000e+00, 3.4500e-06, 4.9000e-06, 5.2400e-06, 5.48
$ jtprob2  <dbl> 3.45e-06, 0.00e+00, 1.39e-05, 1.49e-05, 1.56e-05, 1.
$ jtprob3  <dbl> 0.000004900, 0.000013900, 0.000000000, 0.000021100,
$ jtprob4  <dbl> 0.000005240, 0.000014900, 0.000021100, 0.000000000,
$ jtprob5  <dbl> 0.000005480, 0.000015600, 0.000022100, 0.000023700,
```

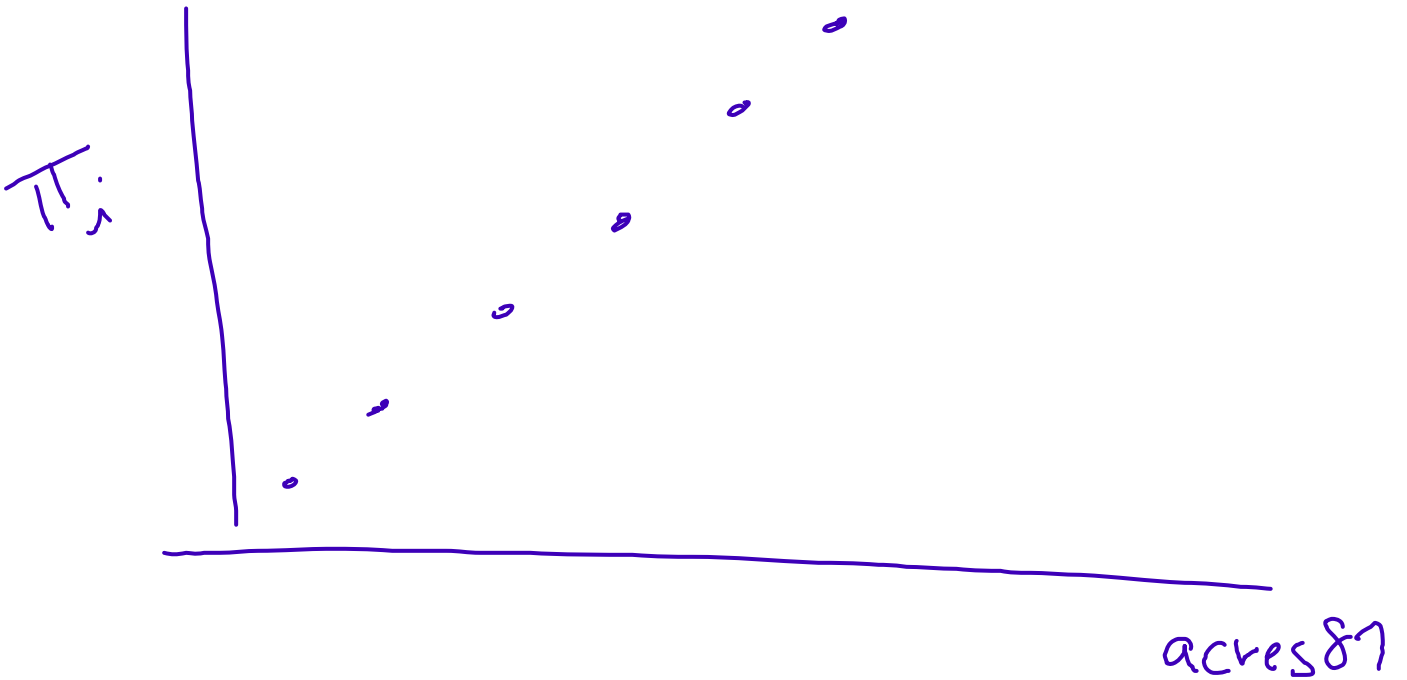
# Lohr Example 6.10: Ag census data

- `id`: counties so could be `~1` or `~county`

# Lohr Example 6.10: Ag census data

- `fpc`: `pii` gives the county inclusion probabilities

```
> ggplot(agpps, aes(x = acres87, y = pii)) +
+   geom_point()
```



# Lohr Example 6.10: Ag census data

- `pps`: joint inclusion probs are given by the columns `jtprob1` - `jtprob15`
- For example, $\pi_{12,15} = \pi_{15,12} = 0.000377359$

```
> agpps$jtprob12[15]   # pi_12,15
[1] 0.000377359
> agpps$jtprob15[12]   # pi_15,12
[1] 0.000377359
```

# Lohr Example 6.10: Ag census data

- pps: we need a matrix
- Columns 10 - 24 are the joint probs:

```
> incl_mat<- as.matrix(agpps[,10:24])
```

- Then we need to fill in unit inclusion probs on the diagonal:

```
> diag(incl_mat) <- agpps$pii
> incl_mat
          jtprob1     jtprob2     jtprob3     jtprob4     jtprob5
 [1,] 0.001137612 0.000003450 0.000004900 0.000005240 0.000005480 0.0
 [2,] 0.000003450 0.003234862 0.000013900 0.000014900 0.000015600 0.0
 [3,] 0.000004900 0.000013900 0.004586659 0.000021100 0.000022100 0.0
 [4,] 0.000005240 0.000014900 0.000021100 0.004906196 0.000023700 0.0
 [5,] 0.000005480 0.000015600 0.000022100 0.000023700 0.005135069 0.0
 [6,] 0.000005700 0.000016200 0.000023000 0.000024600 0.000025700 0.0
 [7,] 0.000006100 0.000017400 0.000024600 0.000026300 0.000027600 0.0
 [8,] 0.000006480 0.000018400 0.000026100 0.000028000 0.000029300 0.0
 [9,] 0.000008530 0.000024300 0.000034400 0.000036800 0.000038500 0.0
[10,] 0.000009660 0.000027500 0.000038900 0.000041700 0.000043600 0.0
[11,] 0.000011146 0.000031700 0.000045000 0.000048100 0.000050300 0.0
```

# Lohr Example 6.10: Ag census data

Estimating total farm acres in 1992: $\hat{t}_{HT} = 992,665,088$ with a SE of $SE_{HT}(\hat{t}_{HT}) = 73,550,378$

```
> ag_pps<- svydesign(id = ~1,
+                    fpc = ~pii,
+                    pps = ppsmat(incl_mat),
+                    data = agpps) ## HT
> svytotal(~acres92, ag_pps, deff=T)   # H-T SE
             total        SE   DEff
acres92 992665088  73550378 0.0063
```

↓ indicates this PPS design is much more precise than a SRS of $n=15$.

# Lohr Example 6.10: Ag census data

Changing how the SE is estimated to the Yates-Grundy: $\hat{t}_{HT} = 992,665,088$ with a SE of $SE_{YG}(\hat{t}_{HT}) = 11,015,154$

```
> ag_pps<- svydesign(id = ~1,
+                    fpc = ~pii,
+                    pps = ppsmat(incl_mat),
+                    variance = "YG",
+                    data = agpps)
> svytotal(~acres92, ag_pps, deff=T)  # H-T SE
            total        SE  DEff
acres92 992665088  11015154 1e-04
```

# Lohr Example 6.10: Ag census data

Changing how the SE is estimated to the With Replacement version:
$\hat{t}_{HT} = 992,665,088$ with a SE of $SE_{WR}(\hat{t}_{HT}) = 11,508,289$

```
> ag_pps<- svydesign(id = ~1,
+                    fpc = ~pii,
+                    data = agpps)
> svytotal(~acres92, ag_pps, deff=T)  # H-T SE
             total         SE   DEff
acres92 992665088  11508289 0.0023
```

omit pps