Poststratification

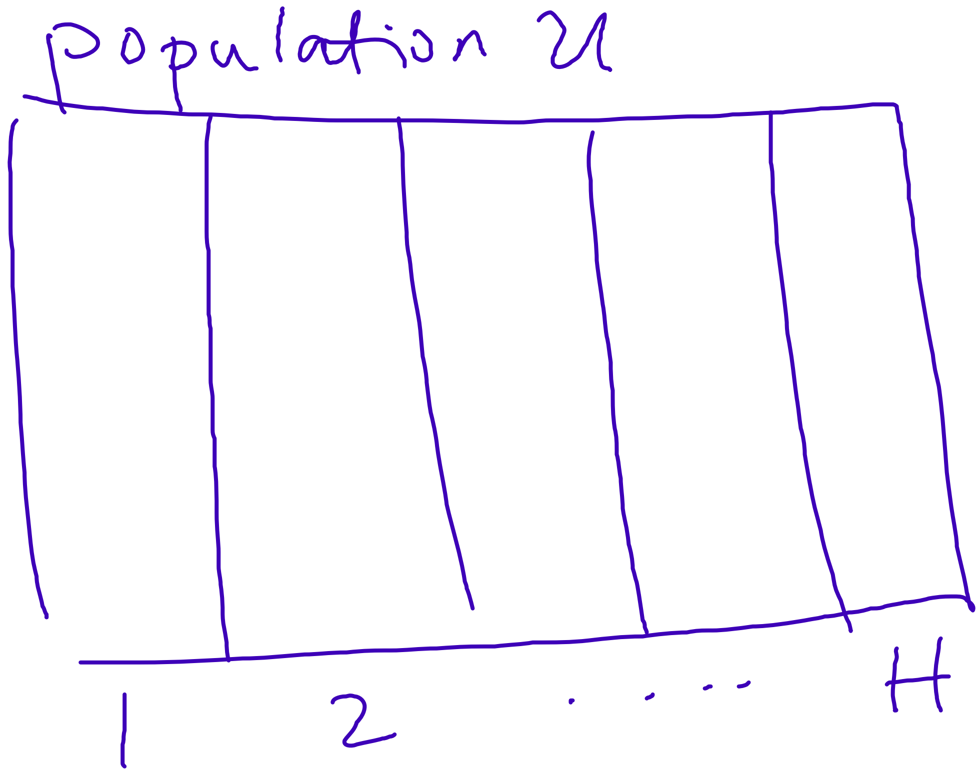
Week 5 (4.4)

Stat 260, St. Clair

Idea: Population

You have a population with **known** stratum fractions:

$$\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_H}{N}$$



Idea: Sample Design

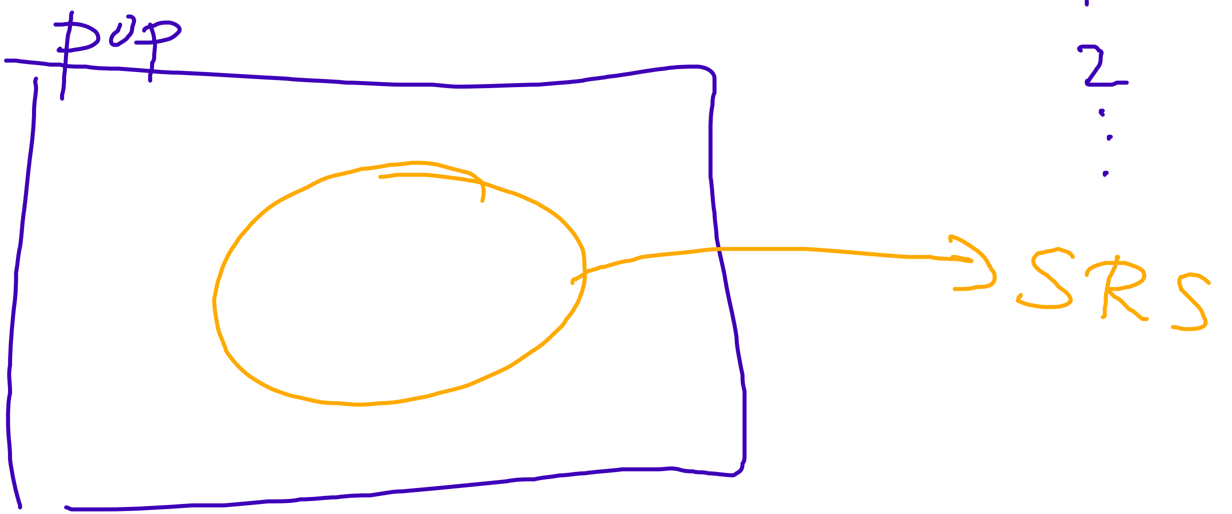
You **don't** have the stratification variable in the sampling frame.

You take a **SRS** of size n and observed

- response y_i
- stratification variable

e.g. Census ⇒ age demo. in Northfield are known

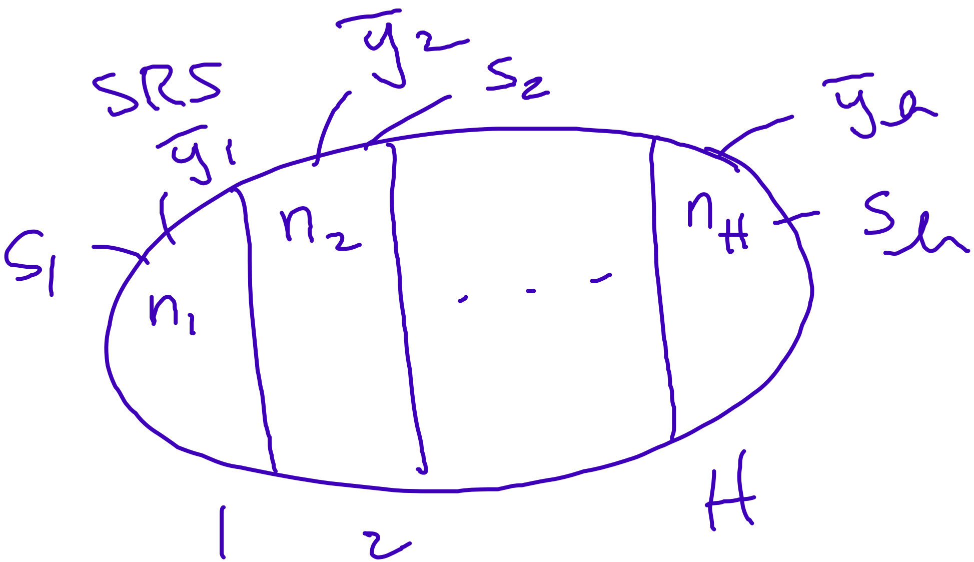
Stratum	Age	N_h/N
1	0–18	.20
2	19–29	.10
⋮	⋮	⋮



Idea: Poststratification

After the SRS, divide the SRS into strata and calibrate estimates of $t_y, \bar{y}_{\mathcal{U}}, P$ based on known population stratum characteristics

- divide responses by stratum
- get *domain* estimates for each stratum
- combine domain estimates like you would in a stratified sample



$n_1, n_2, \ldots, n_H$ vary from sample to sample

Benefits

- Nonresponse: if related to stratum, poststratification can reduce non-response bias IF
  - within stratum: non-respondents ≈ respondents

e.g. strata = age groups

y_i = presidential preference

Non-response:
older ⇒ higher response rate
younger ⇒ lower " "

- If n is large, can increase precision over SRS if stratification *would* have been beneficial

Poststratification estimation: mean

$\bar{y}_h$ = domain est. of the domain/stratum h mean.

- **Parameter** $\bar{y}_{\mathcal{U}}$
- **Estimator**

same as strat. est. of mean

$$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

- **SE:**

Not be strat. SE

$$SE(\bar{y}_{post}) \approx \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \underbrace{\left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left(\frac{n_h - 1}{n_h}\right)\frac{s_h^2}{n_h}}_{SE(\bar{y}_h)^2 \text{ for domain estimation}}}$$

indep. of strata est.

$$\approx \sqrt{\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} \frac{N_h}{N}\frac{s_h^2}{n}}$$

when n & the n_h are big

① $\left(\frac{n}{n-1}\right)\left(\frac{n_h-1}{n_h}\right) \approx 1$

② $n_h \approx n\left(\frac{N_h}{N}\right)$

Poststratification estimation: proportion

- **Parameter** p
- **Estimator**

$$\hat{p}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$

- **SE**: same as mean but with $s_h^2 = \dfrac{n_h}{n_h - 1} \hat{p}_h (1 - \hat{p}_h)$

Poststratification estimation: total

- **Parameter** $t = N\bar{y}_{\mathcal{U}}$
- **Estimator**

$$\hat{t}_{post} = \underline{N\bar{y}_{post}} = \sum_{h=1}^{H} N_h \bar{y}_h$$

- **SE**:

$$SE(\hat{t}_{post}) = NSE(\bar{y}_{post})$$

Example

The population of N=1,500 students. We have a SRS of 200 students and obtained 135 responses to the question *How many hours per week do you devote to study outside of regular class time?*

	Female	Male	all
n_h	104	31	135
$\bar{y}_h$	12.38	8.03	11.38
s_h	6.68	5.50	6.68

SRS
$n = 135$
$\bar{y} = 11.38$
$s = 6.68$

SRS estimate/SE of mean study hours per week?

SRS

$\bar{y} = 11.38$ hours

$$SE = \sqrt{\left(1 - \frac{135}{1500}\right)\frac{6.68^2}{135}} = .55$$

Example

	Female	Male	all
N_h	900	600	1500
n_h	104	31	135
$\bar{y}_h$	12.38	8.03	11.38
s_h	6.68	5.50	6.68

$$\frac{N_F}{N} = \frac{900}{1500} = .60$$

$$\frac{N_M}{N} = \frac{600}{1500} = .40$$

Poststratified estimate/SE of mean study hours per week?

$$\bar{y}_{post} = \sum_{h=1}^{2} \frac{N_h}{N} \bar{y}_h = (.60)(12.38) + (.40)8.03 = \boxed{10.64}$$

$$SE(\bar{y}_{post}) \approx \sqrt{\left(1 - \frac{135}{1500}\right)\left((.6)\frac{6.68^2}{135} + (.4)\frac{5.5^2}{135}\right)}$$

$$\approx .51$$

Example

	Female	Male	all
N_h	900	600	1500
n_h	104	31	135
$\bar{y}_h$	12.38	8.03	11.38
s_h	6.68	5.50	6.68

→ Female 60% of pop. but $\frac{104}{135} \approx 77\%$ of the SRS.

↓ suggesting F higher response rate

$$\boxed{\begin{aligned} \bar{y} &= 11.38 \\ \bar{y}_{post} &= 10.64 \end{aligned}} \quad \begin{aligned} SE(\bar{y}) &= 0.55 \\ SE(\bar{y}_{post}) &= 0.51 \end{aligned}$$

Why is the poststratified estimator of mean study hours lower than the SRS?

→ Response related to Strat: F had higher study hours

SRS

$\left.\begin{array}{l}\text{F overrep. +} \\ \text{F higher response}\end{array}\right\}$ $\bar{y}_{SRS} > \bar{y}_{post}$.