SRS: estimation details

Week 2 (2.1, 2.3-2.6)

Stat 260, St. Clair

Design: Simple Random Sample (SRS)

Defined: Each sample of size n units is equally likely

- Assumption: sampling unit = observation unit
- Assumption: done *without replacement*
- **implies** that each *sampling unit* is equally likely (reverse is *not* true)

Inclusion probabilities: SRS

What is the probability that unit i is selected in a SRS of size n from a population of N units?

"N choose n"

Math fact: $\binom{N}{n} = \dfrac{N!}{n!(N-n)!}$ counts the number of samples of size n

$\binom{N}{n}$ = # of SRS's possible

→ how many contain unit 1?

how many way?

how many → $\binom{N-1}{n-1}$

$S = \{1, \ldots\ldots\}$

n−1 more units

$$\pi_1 = \frac{\#\text{ of sample that look like}}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!\,(N-n)!}}{\frac{N!}{n!\,(N-n)!}} = \frac{n}{N}$$

Sampling weights: SRS

What is the sampling weight for unit i under a SRS?

$$w_i = \frac{1}{\pi_i} = \frac{1}{n/N} = \frac{N}{n}$$

e.g. D3 SRS $N=3$ $n=2$ $\pi_i = \frac{2}{3}$ $w_i = \frac{3}{2} = 1.5$

* a design with sampling weights that are the same for every unit is called a <u>self-weighting</u> design.

Estimation plan: SRS

- Use a Horvitz-Thompson estimator to estimate the **population total**

$$\hat{t}_{HT} = \sum_{\text{sampled units}} w_i y_i = \sum_{\text{sampled units}} \frac{y_i}{\pi_i} \underset{\text{WLOG}}{=} \sum_{i=1}^{n} w_i y_i$$

$$\underline{\underline{\text{SRS:}}} \quad w_i = \frac{N}{n}$$

$$\hat{t} = \sum_{i=1}^{n} \left(\frac{N}{n}\right) y_i = \frac{N}{n} \sum y_i = \boxed{N\bar{y}}$$

$$\bar{y} = \text{sample mean}$$

- Divide by population size to estimate a **population mean**
- Define a **binary** indicator (1=yes, 0=no) as the response and use **mean** results to estimate a **population proportion**

Population Total: SRS

- Parameter: $t = \sum_{i=1}^{N} y_i$

- Estimator (unbiased)

$$\hat{t} = N\bar{y}$$

  where $\bar{y}$ is the sample mean response.

- Standard error (estimated variation in $\hat{t}$)

$$SE(\hat{t}) = N \times SE(\bar{y}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

  where s is the sample standard deviation.

Population Mean: SRS

$\bar{y}_{\mathcal{U}}$ = pop. mean

$\mathcal{U}$ = "universe"

$\hat{t} = N\bar{y}$

- Parameter: $\bar{y}_{\mathcal{U}} = \frac{t}{N}$

- Estimator (unbiased)

$$\hat{\bar{y}}_{\mathcal{U}} = \frac{\hat{t}}{N} = \bar{y}$$

- Standard error (estimated variation in $\bar{y}$)

$$SE(\hat{\bar{y}}) = \frac{SE(\hat{t})}{N} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

Population Proportion of "successes": SRS

$$y_i = \begin{cases} 1 & , \text{succes} \\ 0 & , \text{failure} \end{cases}$$

- $y_i = 1$ if unit i's response is a "success" and 0 otherwise
- Parameter: $p = \frac{t}{N} = \frac{\text{number of successes in pop.}}{\text{pop. size}}$ $= \frac{\sum_{i=1}^{N} y_i}{N}$
- Estimator (unbiased)

$$\hat{p} = \bar{y} = \text{sample proportion of successes}$$

- Standard error (estimated variation in $\hat{p}$)

$$SE(\hat{p}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}}$$

end of slides

SE comments

big reduction is SE when n gets close to N

- $\left(1 - \frac{n}{N}\right)$ is the **finite population correction** (fpc) that we get from sampling **without replacement**
  - omit the FPC if N unknown

$SE(\bar{y}) \approx \frac{s}{\sqrt{n}}$

- Why all the SE details?
  - Need to understand SE's to compare competing designs!

Coefficient of Variation

Definition The estimated coefficient of variation for an unbiased estimator is

$$CV = \frac{\text{SE of estimate}}{\text{estimate}}$$

- allows you to compare estimator precision across measurements of different magnitude
  - e.g. compare the precision of estimates of mean monthly housing expenditures between residents of NYC and Kansas City

Confidence intervals: SRS

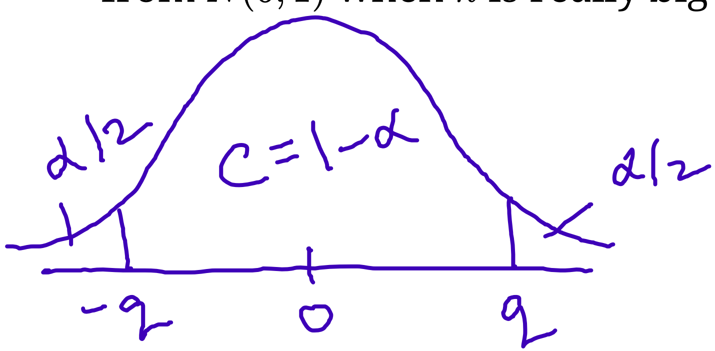
- When n, N and $N - n$ are "large", our SRS estimators have a roughly **normally distributed sampling distribution**
  - A Central Limit Theorem for "finite" populations

- A "C"% confidence interval for our parameters looks like:

$$\boxed{\text{estimate} \pm q \times SE}$$

$q \times SE =$ margin of error

- q is a quantile that is determined by the confidence level $C = 1 - \alpha$ and is either
  - from the t distribution with $n - 1$ degrees of freedom
  - from $N(0, 1)$ when n is really big (or not yet determined)



$-q = \frac{\alpha}{2}$ quantile

$q =$ qnorm(.975)   95%

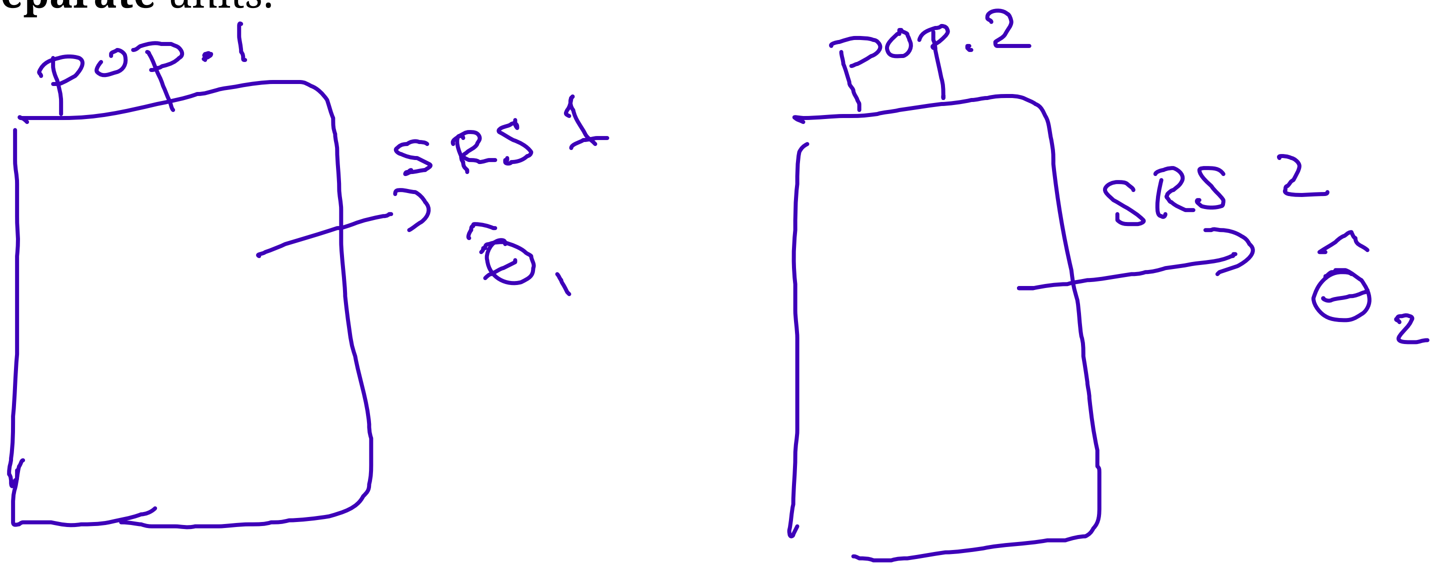
qt(.975, df = n-1)

Confidence interval for a difference: SRS

A CI for the difference of two population parameters $\theta_1 - \theta_2$ looks like

$$\hat{\theta}_1 - \hat{\theta}_2 \pm q \times \sqrt{SE_1^2 + SE_2^2}$$

- Condition: We have two **separate** SRS so $\hat{\theta}_1$ and $\hat{\theta}_2$ are computed from **separate** units.



Planning a survey: SRS

M.E for est. mean

$$e = z \times \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

solve for n!

e_t was M.E. for est. total

$$e = \frac{e_t}{N}$$

What should n be to achieve

- to estimate the population **mean**/prop
- a fixed margin of error e
- with confidence level "C"?

N **known**: $n = \dfrac{n_0}{1 + n_0/N}$

N **unknown**: $n = n_0$

where

- $n_0 = \left(\dfrac{sz}{e}\right)^2$

95% $\Rightarrow$ $z = 1.96$

- s is our "best guess" at the SD of our quantitative response
  - for proportion: $s \approx \sqrt{p(1-p)}$

use $p = \frac{1}{2}$ for conservative

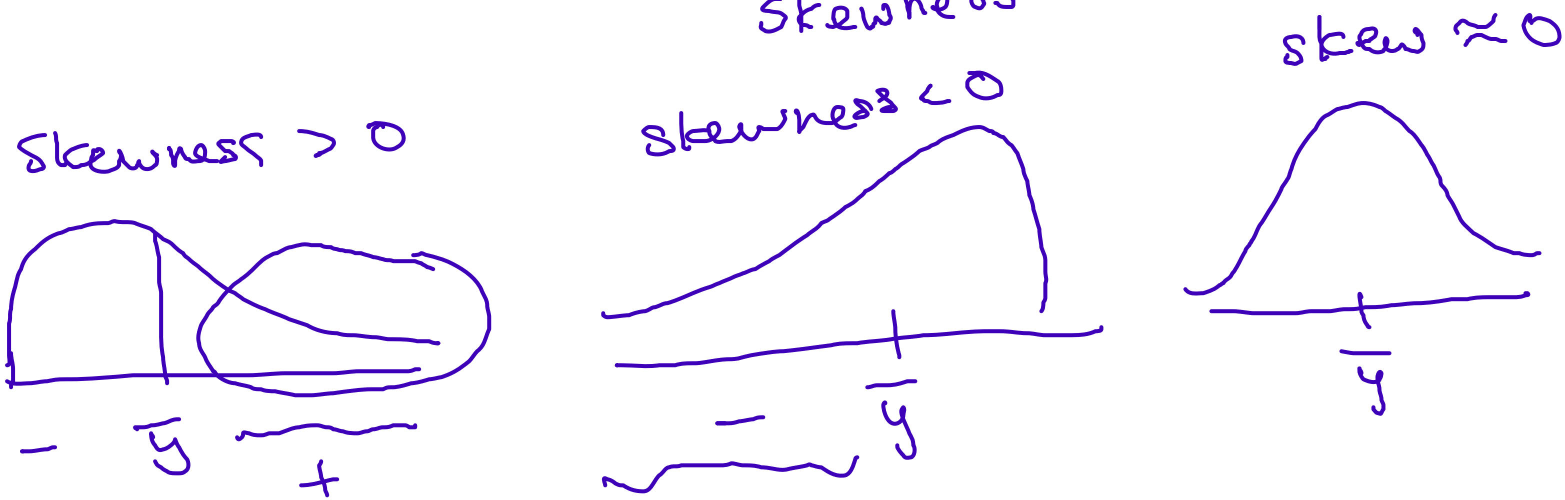
- z is a $N(0,1)$ quantile for "C" level of confidence

Planning a survey: SRS

What should n be to trust the "large enough" sample size condition for CI?

- Sugden et al. suggests

$$n_{min} \approx 28 + 25\left(\underbrace{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{ns^3}}_{\text{skewness}}\right)^2$$



For fun:

why $n - 1$ in the $SE(\hat{p})$?

$$SE(\hat{p}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}}$$

$$SE(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}$$

sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{(y_i - \bar{y})^2}$$

what does this simplify to when y_i's are 0/1 ?

$y_i - \bar{y}$ $\Rightarrow$ either $0 - \hat{p}$ or $1 - \hat{p}$

Also know: $n\hat{p}$ = # of successes / # of 1's

$n(1-\hat{p})$ = # of 0's

$$S^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{1}{n-1}\Big[\underbrace{n(1-\hat{p})(0-\hat{p})^2}_{y=0} + \underbrace{n\hat{p}(1-\hat{p})^2}_{y=1}\Big]$$

$$= \frac{1}{n-1}\left[n\hat{p}^2(1-\hat{p}) + n\hat{p}(1-\hat{p})^2\right]$$

$$= \frac{1}{n-1}\left[n\hat{p}(1-\hat{p})\right]\underbrace{\left[\hat{p}+1-\hat{p}\right]}_{1}$$

$$S^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

$$SE(\hat{p}) = \sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}} = \sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} \quad \checkmark$$