You should know...

- **Final Exam: December 15, 9 - 11 a.m., EX 200 (255 McCaul Street)**
- Test 2: November 20, 1.10 - 2 p.m.
- cluster sampling
- psu, ssu
- one stage and two stage
- reasons for cluster sampling
- difference from stratified sampling
- equal size clusters: $\hat{t} = (N/n)\sum t_i$; $\hat{\bar{y}} = \hat{t}/(NM)$
- $SE(\hat{t}) = ...,$ $SE(\hat{\bar{y}}) = ...$
- $s_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \hat{t}/N)^2$
- *SSB*, *SSW*, *SSTO*, anova table for population and sample
- §5.0, 5.1, 5.2.1, 5.2.2, 5.2.3.1
- HW: Ex. 5.9.1, 5.9.2, 5.9.3, 5.9.4, 5.9.5, 5.9.6

In the News

- "Obama love-in has little impact on how Canadians view Americans, poll finds" http://ca.news.yahoo.com/s/capress/091101/national/poll_cda_us
- The Historica-Dominion Institute is a national charitable organization focused on promoting a greater understanding of Canada's history and the rights and responsibilities of citizenship.
- http://www.historica-dominion.ca/en/news/7155_american-myths-revisited-barack-obama-has-not-fundamentally-changed-how-canadians-see-
- "5 College Majors That Can Help You Get a Job" http://www.smartmoney.com/personal-finance/college-planning/5-college-majors-that-can-help-you-land-a-job/

... in the News

- "Swine flu: A researcher's perspective – Hype can make us all ill"

`http://www.cbc.ca/canada/story/2009/11/02/f-viewpoint-cassels.html`

Hundreds of people wait in line outside a health clinic in Elmsdale, N.S. for their turn to be injected with the H1N1 flu vaccine. (Andrew Vaughan/Canadian Press)

... in the News

- "Q: Does the vaccine work?"
- "A: It depends on your definition of work. It works in terms of helping people develop antibodies to that particular virus. But are those antibodies enough to keep you from getting sick?
  Often people who get the flu shot still get the flu. And we know there are many other circulating viruses that could still make you sick.
  When people tell you the flu vaccine "reduces mortality by 50 per cent" you need to know that these stats come from cohort studies, which compare death rates in vaccinated people versus non-vaccinated people. The truth is, those two groups may be very fundamentally different to start with and the vaccine might have had nothing to do with the observed outcomes.

  This "healthy-user bias," as it is called, is rampant in vaccine studies. Without randomization and a true

  control group to compare, we don't really know for sure if flu campaigns achieve their intended outcomes. "
- "Q: Isn't it public spirited to get vaccinated, so you won't spread the virus to others?"
- "A: That sounds plausible, but is that recommendation evidence-based? Researchers who have combed through hundreds of flu-vaccine studies find very little evidence that suggests a vaccine will prevent the spread of the virus in the general population.
  Of the hundreds of studies on flu immunization campaigns, only about four are of sufficient rigour to say anything definitive. And two of those studies show the vaccine in question to be useless.

  Basing public health policy on only two quality studies doesn't seem sound to me."
- "Does the vaccine matter" (Atlantic Monthly)

  `http://www.theatlantic.com/doc/200911/brownlee-h1n1`

Back to reality: Ch. 5

- two-stage cluster sampling: N psu's (clusters), M_i ssu's in each cluster
- sample n psu's, and then m_i ssu's in each cluster
- see Figure 5.2
- special case: $m_i = M_i$ – one stage cluster sampling
- Estimate total t: $\hat{t}_{unb} = ...$
- Summary formulas: p.168, §5.8; 5.21, 5.25, 5.28, 5.29
- ?Huh? ratio estimation? p.148. Needed in two-stage if we don't know $K = \sum_{i=1}^{N} M_i$

Two-stage cluster sampling

- See Example 5.6
- Exercise 5.9.6, data in `books.dat`
- also see R code from October 30
- $N = 44$ shelves of books; $n = 12$ shelves selected at random; $M_1 = 26, M_2 = 52, \ldots, M_{12} = 27$ books on each of the sampled shelves (see Table 5.5)
- sample $m_i = 5$ books from each of the 12 shelves, look up the replacement values in *Books in Print*
- (a): estimate total replacement cost

$$\hat{t}_{unb} = \sum M_i \bar{y}_i = 32637.73$$

- and standard error:

$$\widehat{V}(\hat{t}_{unb}) = 44^2(1 - \frac{12}{44})\frac{s_t^2}{12} + \frac{N}{n}\sum_i (1 - \frac{5}{M_i})M_i^2\frac{s_i^2}{5}$$

$$= 31,507,635 + 1,365,699 = (5733)^2$$

- HW: Exercise 5.9.14a, 5.9.15a, 5.9.17

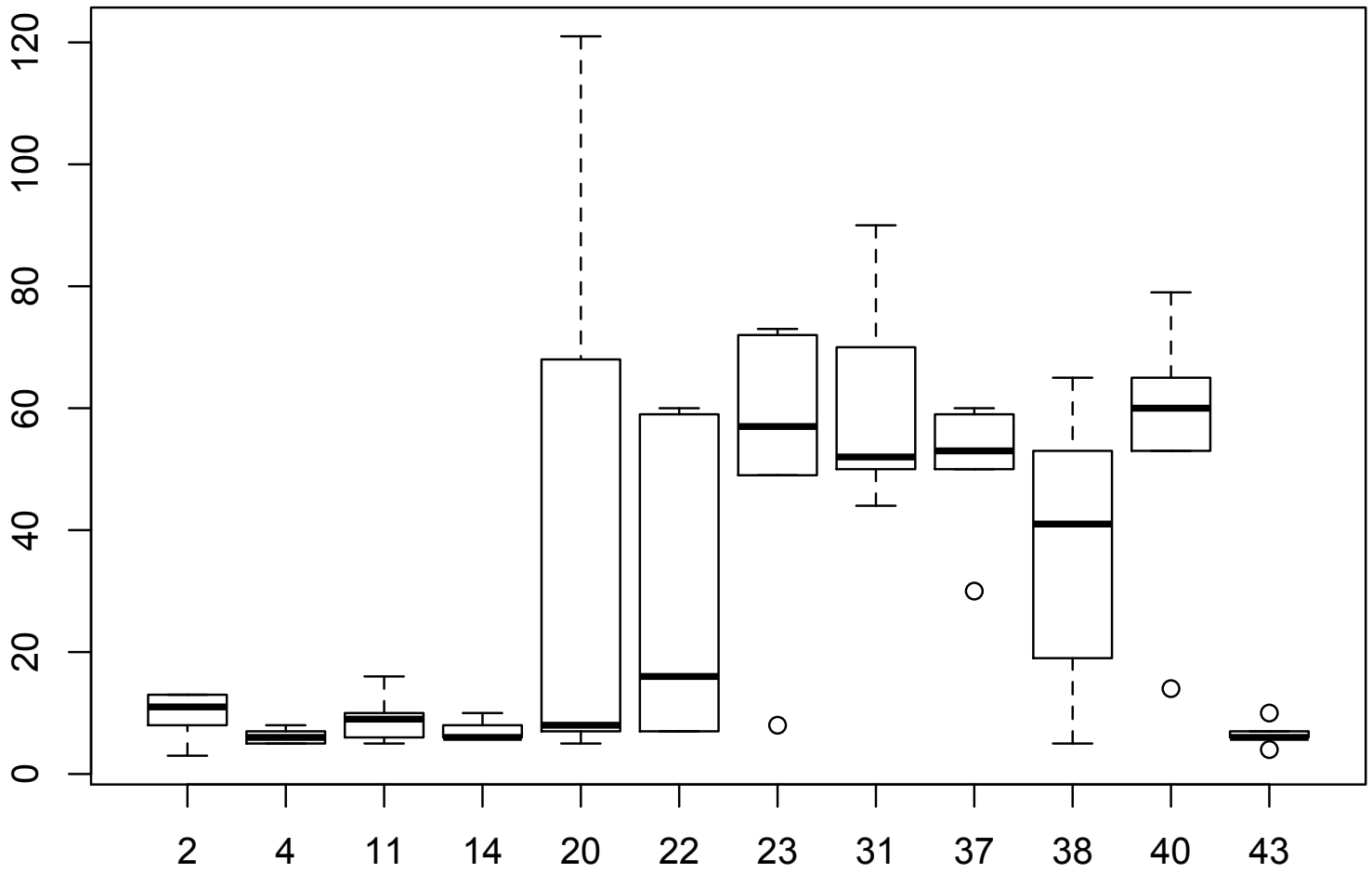
... two-stage cluster sampling

- (b): estimate average replacement cost and standard error
- $\widehat{\bar{y}}_r = \sum_{i\in\mathcal{S}} M_i\bar{y}_i / \sum_{i\in\mathcal{S}} M_i = 23.61$

- $$\widehat{V}(\widehat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left[(1-\frac{12}{44}\frac{s_r^2}{n} + \frac{1}{12\times 44}\sum_{i\in\mathcal{S}}(1-\frac{5}{M_i})M_i^2\frac{s_i^2}{5}\right]$$
$$= \frac{1}{31.416667^2}\left[(1-\frac{12}{44})\frac{476700.3}{12} + \frac{1}{12\times 44}372463.4\right]$$
$$= \frac{1}{31.416667^2}[28890.3 + 705.4228] = (5.48)^2$$

- $$s_r^2 = \frac{\sum_{i\in\mathcal{S}}[M_i(\bar{y}_i - \widehat{\bar{y}}_r)]^2}{n-1}$$

... two-stage cluster sampling



Another look at point estimators: §5.4

- $$\hat{t}_{unb} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$
- $$\widehat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}$$
- $$w_{ij} = \frac{NM_i}{nm_i} = \frac{1}{\mathrm{P}(\ j\text{th ssu in } i\text{th psu is selected})}$$
- Exercise 5.9 (books)
  $M = (26, 52, 70, 47, 5, 28, 27, 29, 21, 31, 14, 27); \quad m_i = 5$
- $$w_{ij} = \frac{N}{12}\frac{M_i}{5}$$
- relative weights are simply M_i, or $M_i/5$
- See Example 5.8 where $m_i \equiv 2$

Choosing sample sizes

- if variability within a cluster is large, then cluster sampling is efficient (R_a^2 is close to 1)
- that's good
- large psu's usually have more variability (extreme case ...)
- but if a psu is too large then it's expensive to sample from
- see Example 5.9, where data is collected from 5 sampling plans, with psu's of size 1 through 5; and 5 is best
- what about number m_i of ssu's?
- formulas in §5.5.2, but you don't need to use these
- see Figures 5.6 and 5.7

Systematic sampling: §5.6

- choose a sample of size 3 from $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
- choose a number from 1 to 4 as a random starting point
- gives 4 possible clusters: $\{1, 5, 9\}$, ...
- we only see 1
- the estimator of the population mean is still unbiased
- the variance might even be smaller (if ICC ...)
- but we can't estimate the variance, because we only have one psu

... systematic sampling

- the list is in random order; treat the sample like an SRS
- the sampling frame is increasing or decreasing order (e.g. auditing)
- systematic sample will have larger variance than a SRS
- SRS formula will over-estimate the variance
- sampling frame is periodic
- systematic sampling is dangerous
- see Figure 5.8; sampling for hazardous waste sites
- choose a point at random, construct a grid with points equi-distant
- grid points make a systematic random sample