



PHI Applied Research Fellows 2021: Survey Statistics

Jessica Godwin

July 7, 2021

Sampling Schemes

Design-based Estimation

Model-based estimation

References

- ▶ Statisticians often choose models and likelihoods based on a combination of:
 - how closely they reflect the true data-generating process
 - the mathematical and statistical properties
 - ► (hopefully) the principle of parsimony
- ▶ If our outcome is binary:

- Statisticians often choose models and likelihoods based on a combination of:
 - how closely they reflect the true data-generating process
 - the mathematical and statistical properties
 - (hopefully) the principle of parsimony
- ▶ If our outcome is binary:
 - Flipping a coin? or
 - Drawing from an urn?

- Statisticians often choose models and likelihoods based on a combination of:
 - how closely they reflect the true data-generating process
 - the mathematical and statistical properties
 - (hopefully) the principle of parsimony
- ► If our outcome is binary:
 - Flipping a coin? or
 - Drawing from an urn?
 - ▶ What does flipping a coin or drawing from an urn have to do with surveys with human respondents?

Finite vs. superpopulations

For observations i = 1, ..., n, let

- ▶ Superpopulation: If $y \sim \text{Bernoulli}(p)$

 - ▶ Finite population: If $y \sim \text{Hypergeometric}(N, K)$

Finite vs. superpopulations, cont'd

If $y_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$,

If $y_i \stackrel{\text{iid}}{\sim} \text{Hypergeometric}(N, K, n)$,

$$\widehat{p} = \frac{\sum_{i=1}^{n} y_{i}}{n},$$

$$\widehat{p} = \frac{\sum_{i=1}^{n} y_{i}}{n} = \frac{k}{n}$$

$$\widehat{Var}(\widehat{p}) \neq \frac{\widehat{p}(1-\widehat{p})}{n}.$$

$$\widehat{Var}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n} \times (1-\frac{n}{N}).$$

How do we say what \hat{p} means in either case? Is it the same?

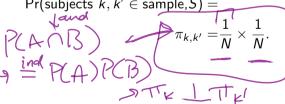
Sampling Schemes

▶ Under an **SRS** of *n* observations



$$\boxed{\pi_{k}} = \frac{1}{N}$$

 $Pr(subjects k, k' \in sample, S) =$



▶ Under an **SRSWOR** of *n* observation

 $Pr(subject k \in sample, S) =$

$$\pi_k = \frac{n}{N} = {}^{\dagger}k!$$

 $Pr(subjects k, k' \in sample, S) =$

$$\tau_{k,k'} = \frac{N}{N} \times \frac{N}{N-1}$$



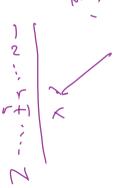
Jessica Godwin

References

Systematic sampling

 \triangleright Select every r^{th} sampling unit from the sampling frame of length

 $N: r \times n \leq N < r \times (n+1)$ What is π_k for individual k = r? k = r + 1?



 \triangleright Select every r^{th} sampling unit from the sampling frame of length

$$N: r \times n \leq N < r \times (n+1)$$

- What is π_k for individual k = r? k = r + 1?
- ► Can a systematic sample be implemented so that it is the equivalent of an SRS?

 \triangleright Select every r^{th} sampling unit from the sampling frame of length

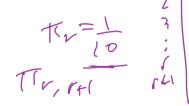
$$N: r \times n \leq N < r \times (n+1)$$

- What is π_k for individual k = r? k = r + 1?
- ▶ Can a systematic sample be implemented so that it is the equivalent of an SRS?
- ▶ What is $\pi_{r,r+1}$?

 \triangleright Select every r^{th} sampling unit from the sampling frame of length

 $N: r \times n \leq N \leq r \times (n+1)$

- What is π_k for individual k = r? k = r + 1?
- ▶ Can a systematic sample be implemented so that it is the equivalent of an SRS?
- What is $\pi_{r,r+1}$?
- ightharpoonup Random single start \rightarrow what changes?



 \triangleright Select every r^{th} sampling unit from the sampling frame of length

$$N: r \times n \leq N < r \times (n+1)$$

- What is π_k for individual k = r? k = r + 1?
- Can a systematic sample be implemented so that it is the equivalent of an SRS?
- What is $\pi_{r,r+1}$?
- ightharpoonup Random single start \rightarrow what changes?
- Multiple starts
 - No individual sampling probabilities are 0 or 1
 - ► Joint sampling probabilities defined

Stratified simple random sampling (strSRS)

▶ Consider h = 1, ..., H strata from each of which you want to sample n_h individuals.

Pr(subject
$$k \in S_h$$
) = $\pi_k = \frac{n_h}{N_h}$

Pr(subjects $k, k' \in S_h$) = $\pi_{k,k'} = \frac{n_h}{N_h} \times \frac{n_h - 1}{N_h - 1}$

Pr(subjects $k \in S_h, k' \in S_{h'}$) = $\pi_{k,k'} = \frac{n_h}{N_h} \times \frac{n_{h'}}{N_{h'}}$.

▶ Why stratify? Why not an SRS or SRSWOR?

strSRS. cont'd

- ▶ Why stratify? Why not an SRS or SRSWOR?
 - Availability of sampling frame
 - Cost. convenience, speed
 - \triangleright N_1, \ldots, N_h vary widely $\not \leftarrow$
 - Rare outcomes within certain strata
 - \blacktriangleright We know strata are related to outcome of interest \rightarrow precision gains!
- ▶ What happens if we ignore the stratification?
 - Waste a lot of folks' money!!
 - ▶ Implicit assumption that outcome of interest doesn't differ by strata
 - \rightarrow obscure differences in outcomes by strata
 - → OVERESTIMATE variance/standard errors
 - → worsens variability in outcomes between strata grows and within strata shrinks
 - \rightarrow worsens as variability in $\pi_{k \in S_k}$ between strata grows

Cluster sampling

Consider sampling c = 1, ..., C clusters or **primary sampling units (PSU)** from your population of N_C clusters and N units.

Indivuals k are the **observation units** contained within clusters on which we will make measurements.

One-stage cluster sampling

$$\Pr(\mathsf{PSU}\ c \in S) = \frac{C}{N_c}$$

$$\pi_{k \in Sc} = \begin{cases} 1, & \mathsf{PSU}\ c \in S, \\ 0, & \mathsf{otherwise}. \end{cases}$$

Two-stage cluster sampling Sample m_c from M_c units in cluster c.

$$\Pr(\mathsf{PSU}\ c \in S) = \frac{C}{N_c}$$

$$\pi_{k \in S_c} = \begin{cases} \frac{m_c}{M_c}, & \mathsf{PSU}\ c \in S, \\ 0, & \mathsf{otherwise}. \end{cases}$$

Cluster sampling, cont'd

- ▶ Probability proportional to size (PPS) sampling
 - $\rightarrow \pi_c \propto M_c$
 - When does this make sense?
- Why implement a cluster sample?
 - ▶ The only sampling frame we have is a list of groups of observation units
 - Cost and convenience
- ▶ What happens if we ignore clustering in our sample?



- \triangleright The m_c observation units sampled in cluster c are **not** independent samples
- \rightarrow we have LESS information than m_c observations from an SRS
- ▶ → we will UNDERESTIMATE variances and standard errors if we ignore this dependence
- ▶ → this underestimation worsens as the correlation between outcomes from individuals in a cluster increases

Multi-stage sampling

- **Example:** DHS (among others) stratify clusters by administrative divisions × urban/rural \rightarrow select women within households within clusters within strata
- Stratified two-stage cluster sampling
- ightharpoonup PSUs ightharpoonup secondary sampling units (SSUs) ightharpoonup observation units
- One could stratify within clusters if a sampling frame necessitates (never encountered this vet)

Multi-phase sampling

- Fancy term for trying again to reach non-respondents!!
- Sub-sample (perhaps fully) your nonrespondents in attempts to get a response.

Design-based Estimation

$$\frac{1}{2}$$
weight $w_k = -\frac{1}{2}$

- ► Each individual k has their responses weighted by their **sampling weight** $w_k = \frac{1}{\pi_k}$
 - lacktriangleright i.e. an individual with low chance of being sampled $o \pi_k$ small $o w_k$ big
 - w_k can be interpreted as number of individuals in the finite population that individual k's response represents
 - Caveat: nonresponse
- Average or arithmetic mean

$$\frac{\sum_{k=1}^{n} y_{k}}{n} \stackrel{?}{=} \frac{\sum_{k=1}^{n} w_{k} y_{k}}{\sum_{k=1}^{n} w_{k}} = \frac{\sum_{k=1}^{n} \frac{N}{n} y_{k}}{\sum_{k=1}^{n} \frac{N}{n}} \\
= \frac{\frac{N}{n} \sum_{k=1}^{n} y_{k}}{\frac{N}{n} \sum_{k=1}^{n} 1} = \frac{N}{n} \left(\frac{\sum_{k=1}^{n} y_{k}}{\frac{N}{n} \times n} \right) = \frac{N}{n} \left(\frac{\sum_{k=1}^{n} y_{k}}{N} \right) = \frac{N}{n} \left(\frac{\sum_{k=1}^{n} y_{k}}{N} \right)$$

- **Each** individual k has their responses weighted by their **sampling weight** $w_k = \frac{1}{k}$
 - i.e. an individual with low chance of being sampled $\to \pi_k$ small $\to w_k$ big
 - w_k can be interpreted as number of individuals in the finite population that individual k's response represents
 - Caveat: nonresponse
- Weighted average

$$\sum_{k=1}^n w_k y_k$$
 such that $w_k \in [0,1]$ and $\sum_k^n w_k = 1$

- ightharpoonup Consider a population of size N, a sample of size n, where each individual has outcome Y_k
- $ightharpoonup Y_k$ is **not** random, but Z_k is

$$Z_k = \begin{cases} 1, & k \in S \\ 0, & \text{otherwise.} \end{cases}$$

▶ Once sample taken $y_k = Y_k \times Z_k$ denotes an individual's observed response (may contain measurement error)

$$E[y_k] = E[Y_k \times Z_k] = Y_k E[Z_k] = Y_k \times \pi_k$$

▶ The population total of outcomes Y is

$$T = \sum_{k=1}^{N} Y_k$$

$$\widehat{T} = \sum_{k=1}^{n} \underbrace{w_k y_k} = \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

$$\widehat{Var}(\widehat{T}) = \sum_{k,k'} \frac{y_k y_{k'}}{\pi_k \pi_{k'}} - \frac{y_k y_{k'}}{\pi_k \kappa_{k'}}$$

▶ The population mean of outcomes Y is

$$\overline{Y} = \frac{\sum_{k=1}^{N} Y_k}{N}$$

$$\widehat{\overline{Y}} = \frac{\sum_{k=1}^{N} \frac{W_k y_k}{N}}{N} = \frac{1}{N} \sum_{k=1}^{N} \frac{y_k}{\pi}$$

$$\widehat{Var}(\widehat{\overline{Y}}) = \frac{\widehat{Var}(\widehat{T})}{N^2}$$

▶ The population mean of binary outcomes Y or **prevalence** is

$$P = \frac{\sum_{k=1}^{N} Y_k}{N}$$

$$\widehat{P} = \frac{\sum_{k=1}^{n} w_k y_k}{N} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

$$\widehat{Var}(\widehat{P}) = \frac{\widehat{Var}(\widehat{T})}{N^2}$$

Stratified sampling

$$\widehat{T} = \sum_{h=1}^{H} \widehat{T}_h = \sum_{h=1}^{H} \sum_{k=1}^{n_h} w_{hk} y_{hk},$$

$$\widehat{Var}(\widehat{T}) = \sum_{h=1}^{H} \widehat{Var}(\widehat{T}_h) = \sum_{h=1}^{H} \sum_{k,k'} \frac{y_{hk} y_{hk'}}{\pi_{hk} \pi_{hk'}} - \frac{y_{hk} y_{hk'}}{\pi_{hkk'}},$$

Calculate variance in terms of each individual's difference from their respective strata total.

► Cluster sampling

$$\widehat{T} = \sum_{c=1}^{C} T_c = \sum_{c=1}^{C} \sum_{k=1}^{N_c} w_{ck} y_{ck} = \sum_{c=1}^{C} w_c \sum_{k=1}^{N_c} y_{ck},$$

▶ Calculate the variance in terms of each cluster total's difference from the overall population total

Stratified two-stage cluster sampling

$$\widehat{T} = \sum_{h=1}^{H} \widehat{T}_{h} = \sum_{h=1}^{H} \sum_{c_{1}=1}^{C_{1h}} \widehat{T}_{h[c_{1}]}$$

$$= \sum_{h=1}^{H} \sum_{c_{1}=1}^{C_{1h}} \sum_{c_{2}=1}^{C_{2h}} \widehat{T}_{h[c_{1}:c_{2}]} = \sum_{h=1}^{H} \sum_{c_{1}=1}^{C_{1h}} \sum_{c_{2}=1}^{C_{2h}} \sum_{k=1}^{n_{c_{2}}} w_{h[c_{1}:c_{2}]k} y_{h[c_{1}:c_{2}]k}$$

$$\widehat{Var}(\widehat{T}) = \sum_{h=1}^{H} \widehat{Var}(\widehat{T}_{h}).$$

Apply methods from previous two in appropriate summation order

▶ What if we don't know *N*?

$$\widehat{P} = \frac{\sum_{k=1}^{n} w_k y_k}{N} \approx \frac{\sum_{k=1}^{n} w_k y_k}{\widehat{N}} = \underbrace{\frac{\sum_{k=1}^{n} w_k y_k}{\sum_{k=1}^{n} w_k}}_{\sum_{k=1}^{n} w_k}$$

$$\widehat{Var}(\widehat{P}) = \widehat{Var}\left(\frac{\widehat{T}}{\widehat{N}}\right) = \frac{\widehat{??}}{\widehat{??}}$$

- **Linearization:** Use Taylor series expansions to approximate the variance
 - survey package in R

Linear regression mean model

$$E[Y|\theta,X]=X\beta,$$

The likelihood

$$L(\theta|y,\mathbf{X}) = \prod_{k=1}^{n} L(\theta|y_k,\mathbf{x_k})$$

The log-likelihood

$$I(\theta|y, \mathbf{X}) = \log L(\theta|y, \mathbf{X}) = \sum_{k=1}^{n} \log L(\theta|y_k, \mathbf{x_k})$$

► The score function

$$\nabla I(\theta|y,\mathbf{X}) = \begin{bmatrix} \frac{\partial I}{\partial \beta_0} & \dots & \frac{\partial I}{\partial \beta_p} \end{bmatrix}$$

is set equal to 0 to estimate $\widehat{\beta}$ in maximum likelihood estimation

- Incorporates sampling weights in pseudolikelihood method by weighting each observation unit's contribution to the score function by w_{ν}
- survey::svyglm function

Model-based estimation

Model-based estimation

Design-based methods have nice properties, but

- what if sampling weights not provided?
- ightharpoonup small sample sizes ightharpoonup design-based standard errors too large
- especially a concern in small area estimation

Fixed effects for strata \rightarrow different means for strata

Random effects for cluster \rightarrow account for dependence between observations within cluster

$$y_{h[c_1:c_2]k} = \mu_h + \mathbf{b}_{c_1} + \mathbf{b}_{c_2}$$

Is there enough information to estimate \mathbf{b}_{c_2} 's or all μ_h if H is large or $m_{c_1:c_2}$ is small compared to $m_{\rm Cl}$'s?

References

References

- ▶ Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replace- ment from a finite universe. Journal of the American statistical Association, 47(260):663–685.
- ▶ Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51:279–292.
- ► Lohr, S. (2010). Sampling: Design and Analysis, Second Edition. Brooks/Cole Cengage Learning, Boston.
- ▶ Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software, 9:1–19.
- ▶ Lumley, T. (2010). Complex Surveys: A Guide to Analysis using R. John Wiley and Sons, Hoboken, Jersey.

Maximum Likelihood Estimation Examples, if needed

- $k = 1, \ldots, n, v_k \stackrel{iid}{\sim} Bernoulii(p)$
- $k = 1, \ldots, n, v_k \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$