PHI Applied Research Fellows 2021: Survey Statistics

Jessica Godwin

July 7, 2021

Data-generating Processes

Data-generating Processes

- 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:

Data-generating Processes

- 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?

Data-generating Processes

- 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

$E[y_i] = \cancel{0 \cdot P(y_i=0)} + 1\,P(y=1) = p$

For observations $i = 1, \ldots, n$, let $= \sum_{i=1}^{n} E[y_i] = \sum_{i=1}^{n} p = np$

$Z = \sum_{i=1}^{n} y_i \qquad E[\sum_i y_i]$

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

$Z \sim Bin(n, p)$

flipping coin

- Superpopulation: If $y \sim$ Bernoulli(p),
  - $E[y] = p$ $Var(y) = p(1-p)$.
- Finite population: If $y \sim$ Hypergeometric(N, K, n), drawing from an urn
  - $E[y] = \frac{K}{N}$ $Var(y) = \frac{K}{N}\left(1 - \frac{K}{N}\right)\left(1 - \frac{n}{N}\right)$. How do we say what $\widehat{p}$ means in either case? Is it the same?

p $\qquad p(1-p)$

$\frac{n}{N} \to 1 \quad n \to N$

K	N−K
green	white

N

Finite vs. superpopulations, cont'd

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

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

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

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

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

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

$\boxed{Var(\hat{p}) \to SE(\hat{p})} = \sqrt{Var(\hat{p})}$

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

Sampling Schemes

Simple random sampling(SRS)

- Under an **SRS** of n observations

$$\Pr(\text{subject } k \in \text{sample}, S) = \boxed{\pi_k} = \frac{1}{N}$$

$$\Pr(\text{subjects } k, k' \in \text{sample}, S) = \pi_{k,k'} = \frac{1}{N} \times \frac{1}{N}.$$

and

$P(A \cap B) \overset{ind}{=} P(A)P(B)$

$\rightarrow \pi_k \perp \pi_{k'}$

- Under an **SRSWOR** of n observation

w/out replacement

$$\Pr(\text{subject } k \in \text{sample}, S) = \pi_k = \frac{n}{N} = \pi_{k'}$$

$$\Pr(\text{subjects } k, k' \in \text{sample}, S) = \pi_{k,k'} = \frac{n}{N} \times \frac{n-1}{N-1}.$$

π_k $\pi_{k'|k}$

$P(A \cap B) = P(A|B)P(B)$

$\pi_{k'}$ ✗

Systematic sampling

$r = 2$

1
2
3
4
$N = 5$

- 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 $k = r$? $k = r + 1$?

$E[X+Y]$

1
2
...
r
$r+1$
...
N

$\pi_r = 1$

$\pi_{r,r+1} = 0 \quad \pi_1 = \quad = \quad \pi_{r+1} = 0$

Systematic sampling

- 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?

Systematic sampling

- 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 $\pi_{r,r+1}$?

Systematic sampling

- 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 $\pi_{r,r+1}$?
- Random single start $\rightarrow$ what changes?

$1 \leq r \leq 10$

$\pi_r = \frac{1}{10}$

$\pi_{r,r+1}$

1
2
3
⋮
r
r+1

Systematic sampling

- 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 $\pi_{r,r+1}$?
- Random single start $\rightarrow$ what changes?
- Multiple starts
  - No individual sampling probabilities are 0 or 1
  - Joint sampling probabilities defined

$1 \leq r_1, r_2 \leq 10$

Stratified simple random sampling (strSRS)

- Consider $h = 1, \ldots, H$ strata from each of which you want to sample n_h individuals.

$N = N_1 + \ldots + N_H$

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

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

SRSWOR

$\pi_k \pi_{k'|k}$

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

strSRS, cont'd

- Why stratify? Why not an SRS or SRSWOR?

strSRS, cont'd

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Why stratify? Why not an SRS or SRSWOR?
  - Availability of **sampling frame**
  - Cost, convenience, speed
  - $N_1, \ldots, N_h$ vary widely
  - Rare outcomes within certain strata
  - We know strata are related to outcome of interest → 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
  - → obscure differences in outcomes by strata
  - → OVERESTIMATE variance/standard errors
  - → worsens variability in outcomes between strata grows and within strata shrinks
  - → worsens as variability in $\pi_{k \in S_h}$ between strata grows

$-S \quad 0 \quad S$

$S_h^2 \leftrightarrow S_h^2$

Cluster sampling

Consider sampling $c = 1, \dots, C$ 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

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

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

Two-stage cluster sampling

Sample m_c from M_c units in cluster c.

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

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

Cluster sampling, cont'd

- Probability proportional to size (PPS) sampling
  - $\pi_c \propto M_c$
  - 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

$n = \sum_{c=1}^{C} m_c$

- What happens if we ignore clustering in our sample?
  - 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
  - $\rightarrow$ we will UNDERESTIMATE variances and standard errors if we ignore this dependence
  - $\rightarrow$ this underestimation worsens as the correlation between outcomes from individuals in a cluster increases

Complex surveys

Multi-stage sampling

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

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

Horvitz-Thompson estimators

$\frac{n}{N}$ $\frac{N}{n}$

- Each individual k has their responses weighted by their **sampling weight** $w_k = \dfrac{1}{\pi_k}$
  - i.e. an individual with low chance of being sampled $\rightarrow \pi_k$ small $\rightarrow w_k$ 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**

$\frac{1}{n}$ $\Sigma w_k y_k$

$$\frac{\sum_{k=1}^{n} y_k}{n} \stackrel{?}{=} \frac{\sum_{k=1}^{n} w_k y_k}{\sum_{k=1}^{n} w_k} = \frac{\sum_{k=1}^{n} \frac{N}{n} y_k}{\sum_{k=1}^{n} \frac{N}{n}}$$

$$= \frac{\frac{N}{n}\sum_{k=1}^{n} y_k}{\frac{N}{n}\sum_{k=1}^{n} 1} = \frac{N}{n}\left(\frac{\sum_{k=1}^{n} y_k}{\frac{N}{n} \times n}\right) = \frac{N}{n}\left(\frac{\sum_{k=1}^{n} y_k}{N}\right) = \boxed{\frac{\sum_{k=1}^{n} y_k}{n}}$$

Horvitz-Thompson estimators

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

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

Horvitz-Thompson estimators

- Consider a population of size N, a sample of size n, where each individual has outcome Y_k
- 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)
  - $E[y_k] = E[Y_k \times Z_k] = Y_k E[Z_k] = Y_k \times \pi_k$

Horvitz-Thompson estimators

- The population total of outcomes Y is

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

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

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

Horvitz-Thompson estimators

- The population mean of outcomes Y is

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

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

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

$Var(aX) = a^2 Var(X)$

Horvitz-Thompson estimators

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

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

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

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

Horvitz-Thompson estimators

- 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.

Horvitz-Thompson estimators

- 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

Horvitz-Thompson estimators

- 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

Horvitz-Thompson estimators

- What if we don't know N?

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

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

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

Binder (1983) and regression

- 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**

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

Binder (1983) and regression, cont'd

- The **score function**

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

  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_k
- `survey::svyglm` function

Model-based estimation

Model-based estimation

Design-based methods have nice properties, but

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

Fixed effects for strata → different means for strata

Random effects for cluster → 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_{c_1}'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

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