MATH525 Crib Sheet

Julian Lore

$$V(aX+bY)=a^2V(X)+b^2V(Y)+2abCov(X,Y)$$
$$V(aX-bY)=a^2V(X)+b^2V(Y)-2abCov(X,Y)$$
$$X\perp Y \implies Cov(X,Y)=0$$
$$V(Y)=E(Y^2)-(E(Y))^2=E[(Y-E[Y])^2]$$
$$Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-E(X))(y_i-E(Y))$$
$$Cov(X,Y)=E\left[(X-E[X])\,(Y-E[Y])\right]$$
$$Cov(X,Y)=E(XY)-E(X)E(Y)$$
$$V(X)=Cov(X,X)$$
$$Bias(\hat\theta)=E[\hat\theta]-\theta$$

Key Definitions

census - measuring the quantity of interest exactly
observation unit/population unit - single member of the population
target population - set of observation units that we want to estimate the quantity for
sample - subset of population units that we will measure
sample population - set of population units who could ever be sampled
sampling unit - unit that can be selected for a sample
sampling frame - list of all sampling units in the sample population
We hope that everyone in sample pop belongs to target pop, but not always the case. Perfect world would be target pop = sample pop.
selection bias: part of target pop is not in sampled pop.
judgment sample: deliberately or purposely selecting representative sample, sample units that you judge are representative. **undercoverage**: failing to include all of target pop in sampling frame. **overcoverage**: including pop units in sampling frame that are not in target pop.
cluster sample: randomly sample strata and take SRS or census within each (can't sample all strata)
systematic sample: choose a random starting point and then take every k-th unit in the list
Probability sampling: Each unit has a known prob of being sampled.

Simple Random Sample

Select one of all possible subsets of n population units (this is without replacement, with replacement can also be done). $\binom{N}{n}$ possible samples, each equally likely. So $P(s)=\frac{1}{\binom{N}{n}}$ and $\pi_i=\frac{n}{N}$ (probability of including unit i).

Estimation

$$t=\sum_{i=1}^{N}y_i=N\overline{y}_U$$

$$\hat{t}_s=\frac{N\sum_{s\in S}y_s}{n}=N\overline{y}_S(\textbf{unbiased})$$
$$\overline{y}_u=\frac{\sum_{i=1}^{N}y_i}{N}=\frac{t}{n}$$
$$\overline{y}_S=\frac{\sum_{s\in S}y_s}{n}=\frac{\hat{t}_s}{N}(\textbf{unbiased})$$

Under SRS:

$$E(\hat{t}_s)=t,E(\overline{y}_S)=\overline{y}_U$$
$$V(\overline{y}_s)=\left(1-\frac{n}{N}\right)\frac{S^2}{n}\implies SE(\overline{y}_s)=\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$$
$$\hat{SE}(\overline{y}_S)=\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$$
$$V(\hat{t}_s)=N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}\implies SE(\hat{t}_s)=N\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$$

where

$$S^2=\frac{\sum_{i=1}^{N}(y_i-\overline{y}_u)^2}{N-1}=\frac{\sum_{i=1}^{N}y_i^2-\frac{1}{N}\left(\sum_{i=1}^{N}y_i\right)^2}{N-1}$$
$$=\frac{\sum_{i=1}^{N}y_i^2-N\overline{y}_u^2}{N-1}$$

Under SRS: . For fixed N, as $n\to N, V(\hat{t}_s)\to 0$.
Estimate S^2 with sample variance s^2:

$$s^2=\frac{\sum_{s\in S}(y_s-\overline{y}_s)^2}{n-1}=\frac{\sum_{s\in S}y_i^2-n\overline{y}_u}{n-1}$$

Confidence interval for $\overline{y}_U$ $N\overline{y}_S\pm z_{\alpha/2}\hat{SE}(\overline{y}_S)$
Finite population correction (fpc) $(1-\frac{n}{N})$
Hajek $N_\nu-n_\nu\to\infty$ then
$\hat{t}_S\sim N\left(t,N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}\right),\hat{t}_S\pm 1.96\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$ (95% CI)

Sample Size Estimation

Absolute error: $Pr(|\overline{y}_s-\overline{y}_U\le e|)=1-\alpha$, e is called the **margin of error**
Relative error: $Pr\left(\frac{|\overline{y}_s-\overline{y}_U|}{|\overline{y}_u|}\le r\right)=1-\alpha$

$$n=\frac{S^2z_{\alpha/2}^2}{e^2+\frac{S^2z_{\alpha/2}^2}{N}}\text{ or }n=\frac{z_{\alpha/2}^2S^2}{(r\overline{y}_U)^2+\frac{z_{\alpha/2}^2S^2}{N}}$$

Naive sample size calc with no fpc:

$$n_0=\frac{S^2z_{\alpha/2}^2}{e^2}\implies n=\frac{n_0}{1+\frac{n_0}{N}}$$

Weights

$$\hat{t}_S=\sum_{i\in S}\frac{N}{n}y_i=\sum_{i\in S}w_iy_i$$

where

$$w_i=\frac{N}{n}=\frac{1}{\pi_i}=\frac{1}{Pr(z_i=1)}$$

and z_i is indicator function

$$\overline{y}_S=\frac{\hat{t}}{N}=\frac{\sum_{i\in S}w_iy_i}{\sum_{i\in S}w_i}$$

All weights are same in SRS. A sample in which every unit has same sampling weight is called a **self-weighting** sample

Model-based perspective

Assume model for $Y_1,\dots,Y_n$, these are now random (instead of only Z_i being random). $Y_1,\dots,Y_n\overset{iid}{\sim}f_Y(y|\theta)$, i.e.
$E_f(Y_j)=\mu,V_f(Y_j)=\sigma^2\ \forall j$
Want to estimate $T=\sum_{i=1}^{N}Y_i$
Need to **predict** values for y_i not in the sample.
Best linear unbiased predictor:

$$\hat{T}=\sum_{i\in S}Y_i+\sum_{i\notin S}\hat{Y}_i=\sum_{i\in S}Y_i+\frac{N-n}{n}\sum_{i\in S}Y_i=\frac{N}{n}\sum_{i\in S}Y_i$$

(because of common mean)
This is model-unbiased (if mean and var not different, i.e. assumptions made under model are correct):

$$E_f(\hat{T}-T)=\frac{N}{n}\sum_{i\in S}E_f(Y_i)-\sum_{i=1}^{N}E_f(Y_i)=\frac{N}{n}(n\mu)-N\mu=0$$
$$E_f[(\hat{T}-T)^2]=N^2\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}$$

CLT applies to $\underline{Y}_s$ because of model assumptions
$\overline{Y}_s\sim N\left(\mu,\frac{\sigma^2}{n}\right),\hat{T}\pm z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\frac{\sigma^2}{n}\right)}$

When to use SRS

Do not use if a controlled experiment is better (i.e. is this brand of bath oil an effective mosquito repellent), do not have a list of obs units/expensive to take an SRS or have extra information to make a more cost-effective scheme.
Good for little extra info available or interested in multivariate relationships and no need to take stratified/cluster sample. Easier to perform.

Stratified Sampling

1. Divide pop units into H subpops or strata (requires additional info)
2. Take a probability sample within each stratum (independently)
3. Make inference about target param within each stratum, then pool results together

Strata should be disjoint and partition the sampling frame.

Stratified Random Sampling

SRS within each stratum. Divide pop of N sampling units into H strata, N_h units in strata h. Membership of strata must be mutually exclusive. Must know $N_1, \ldots, N_H$ s.t. $N = \sum_{h=1}^{H} N_h$. Stratified with equal strata size is not the same as SRS because you are forcing to have samples in each strata.
Independently take SRS of size n_h from each stratum: $n = \sum_{h=1}^{H} n_h$
Population quantities:

$$y_{hj} = \text{ val of } j^{th} \text{ unit in stratum } h$$

$$t_h = \sum_{j=1}^{N_h} y_{hj} = \text{ stratum } h \text{ total}$$

$$t = \sum_{h=1}^{H} t_h = \text{ pop tot}$$

$$\overline{y}_{hU} = \frac{t_h}{N_h} = \text{ true stratum } h \text{ mean}$$

$$\overline{y}_U = \frac{t}{N} = \text{ true pop mean}$$

$$S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj} - \overline{y}_{hU})^2}{N_h - 1} = \text{ stratum } h \text{ pop var}$$

$$S^2 = \frac{\sum_{h=1}^{H} \sum_{j=1}^{N_h} (y_{jh} - \overline{y}_U)^2}{N - 1} = \text{ pop var}$$

Sample quantities:

$$\overline{y}_h = \frac{\sum_{j \in S_h} y_{nj}}{n_h}$$

$$\hat{t}_h = \frac{N_h}{n_h} \sum_{j \in S_n} y_{hj} = N_h \overline{y}_h$$

$$s_h^2 = \sum_{j \in S_h} \frac{(y_{jh} - \overline{y}_U)^2}{n_h - 1}$$

$$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \overline{y}_h (\textbf{unbiased})$$

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

$$\overline{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h (\textbf{unbiased})$$

$$V(\overline{y}_{str}) = \frac{1}{N^2} V(\hat{t}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}$$

$$\overline{y}_{str} \pm z_{\alpha/2} SE(\overline{y}_{str})$$

Weights

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h \overline{y}_h = \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj} = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}$$

where

$$w_{hj} = \frac{N_h}{n_h} = \frac{1}{Pr(Z_{hj} = 1)} = \frac{1}{\pi_{hj}}$$

because $\pi_{hj} = \frac{n_h}{N_h}$

$$\overline{y}_{str} = \frac{\hat{t}_{str}}{N} = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj}}$$

Allocating Observations

Proportional allocation:

$$n_h \propto N_h \implies n_h = \left(\frac{N_h}{N}\right) n \implies \pi_{hj} = \frac{n_h}{N_h} = \frac{n}{N}$$

$\hat{t}_{str} = \frac{N}{n} \sum_{h=1}^{H} \sum_{j \in S_h} y_{hj}$ (self-weighting sample)
$(N-1)S^2 = \left[\sum_{h=1}^{H} (N_h - 1) S_h^2\right] + \sum_{h=1}^{H} N_h (\overline{y}_{hU} - \overline{y}_U) \implies$ $TSS = SSW + SSB$ (total sum of squares = sum of squares within + sum of squares between)
$V_{prop}(\hat{t}_{str}) = \left(1 - \frac{n}{N}\right) \frac{N}{n} \sum_{h=1}^{H} N_h S_h^2 =$
$\left(1 - \frac{n}{N}\right) \frac{N}{n} (SSW + \sum_{h=1}^{H} S_h^2)$
$V(\hat{t}_{SRS}) = \left(1 - \frac{n}{N}\right) N^2 \frac{S^2}{n} =$
$\frac{N}{N-1} V_{prop}(\hat{t}_{str}) + \frac{(N-h)N}{n(N-1)} (SSB - \sum_{h=1}^{H} s_h^2)$
So $\hat{t}_{SRS}$ will have larger variance than $\hat{t}_{str}$ unless
$SSB < \sum_{h=1}^{H} S_h^2 \implies \sum_{h=1}^{H} N_h (\overline{y}_h - \overline{y}_U)^2 < \sum_{h=1}^{H} S_h^2$
(variability between clusters is smaller than variability of each strata). Don't want all strata to have same mean, or else variability between will be small, making prop alloc worse. We want stratum means to differ a lot s.t. sum of squares is large, variability within strata is smaller.
Cost: $C = c_0 + \sum_{h=1}^{H} c_h n_h$. Minimize $V(\hat{y}_{str})$ subject to the constraint that $C < C_{max}$. **Optimal allocation**

$$n_h \propto \frac{N_h S_h}{\sqrt{c_h}} \implies \frac{n_h}{n} = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{t=1}^{H} \frac{N_t S_t}{\sqrt{c_t}}} \right)$$

If costs are equal across strata $c_1 = c_2 = \ldots = c_h$, then

$$n_h \propto N_h S_h \implies \frac{n_h}{n} = \frac{N_h S_h}{\sum_{t=1}^{H} N_t S_t}$$

this is **Neyman allocation**. If S_h is known for all h, then Neyman beats prop alloc. With no max cost, Neyman gives optimal alloc.

Determining Sample Size

$V(\overline{y}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} \leq$
$\frac{1}{n} \sum_{h=1}^{H} \frac{n}{n_h} \left(\frac{N_h}{N}\right)^2 S_h^2 = \frac{\nu}{n}$, where $\nu = \sum_{h=1}^{H} \left(\frac{n}{n_h}\right) \left(\frac{N_h}{N}\right)^2 S_h^2$.
So $\overline{y}_{str} \pm z_{\alpha/2} \sqrt{\nu/n}$
$n = z_{\alpha/2}^2 \nu / e^2$ for margin of error e

Model-Based

$Y_{hj} = \mu_h + \varepsilon_{hj}$ where
$\varepsilon_{hj} \sim f(\varepsilon), E_f(\varepsilon_{hj}) = 0, V_f(\varepsilon_{hj}) = \sigma_h^2, \varepsilon_{hj} \perp \varepsilon_{hi}$ for $i \neq j$ and $\varepsilon_{hj} \perp \varepsilon_{ki}$ for $h \neq k$.

$$T_h = \sum_{j=1}^{N_h} Y_{hj}, T = \sum_{h=1}^{H} T_h$$

$$\hat{T}_h = \frac{N_h}{n_h} \sum_{j \in S_h} Y_{hj}, T = \sum_{h=1}^{H} \hat{T}_h$$

$$E_f[\hat{T}_h - T_h] = 0 \text{ (similar to SRS model-based)}$$

$$\begin{aligned} E_f[(\hat{T} - T)^2] &= E_f \left[\left(\sum_{h=1}^{H} \hat{T}_h - T_h \right)^2 \right] \\ &= E_f \left[\sum_{h=1}^{H} (\hat{T}_h - T_h)^2 + \sum_{h=1}^{H} \sum_{k \neq h} (\hat{T}_h - T_h)(\hat{T}_k - T_k) \right] \\ &= E_f[\sum_{h=1}^{H} (\hat{T}_h - T_h)^2] = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} \end{aligned}$$

When not to Stratify

Expensive/impossible to identify the strata, need to know $N_1, \ldots, N_H$, need to choose strata and defend why it's good.

Example Problems

Gas stations, consider gas prices in November and December, want to estimate pop diff in average gas prices between two months. Design 1: SRS n gas stations in November and then SRS n gas stations in December (independently). Good estimator is $\hat{\overline{d}} = \overline{y}_{Dec} - \overline{y}_{Nov}$, sample avgs unbiased for $\overline{y}_{u,Month}$. so difference is unbiased. $V[\hat{\overline{d}}] = V(\overline{y}_{Dec}) + V(\overline{y}_{Nov})$
Design 2: SRS of n gas stations in November and then same stations in December. $\hat{\overline{d^*}} = \overline{y}_{Dec} - \overline{y}_{Nov}$, unbiased because sample means are unbiased. For var:
$\hat{\overline{d^*}} = \frac{1}{n} \sum_{i \in S} (y_{i,Dec} - y_{i,Nov}) = \frac{1}{n} \sum_{i \in S} d_i \implies V[\hat{\overline{d^*}}] =$
$\left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}$ where
$S_d^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_{i,Dec} - u_{i,Nov} - [\overline{y}_{u,Dec} - \overline{y}_{u,Nov}])^2 =$
$V(Y_{Dec}) + V(Y_{Nov}) - 2Cov(Y_{Dec}, Y_{Nov})$. Design 2 is better if covariance or correlation is positive (lower var).
Given margin of error for SRS total and want to compute n, convert margin of error to mean and then use formula for n for mean.
Cov: Design based:
$Cov(\overline{y}_n, \overline{y}_m) = Cov\left(\sum_{i \in S_1} \frac{y_i}{n}, \sum_{j \in S_2} \frac{y_j}{m}\right) =$
$Cov\left(\sum_{i=1}^{N} Z_i \frac{y_i}{n}, \sum_{j=1}^{N} Z'_j \frac{y_j}{m}\right) =$
$\sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \frac{1}{n} \frac{1}{m} \underbrace{Cov(Z_i, Z'_j)}_{E(Z_i Z'_j) - E(Z_i)E(Z'_j)}$. Use $E(Z_i Z'_j) =$
$Pr(Z_i = 1, Z'_j = 1) = Pr(Z'_j = 1 \mid Z_i = 1) Pr(Z_i = 1)$
Model based: $Cov(\overline{y}_n, \overline{y}_m) = \sum_{i \in S_1} \sum_{j \in S_2} \frac{1}{n} \frac{1}{m} Cov(Y_i, Y_j) =$
$\sum_{i \in S_1} \sum_{j \in S_2} \frac{1}{n} \frac{1}{m} Cov(\epsilon_i, \epsilon_j) = \sum_{i \in S_1} \sum_{j \in S_2} \frac{1}{n} \frac{1}{m} \times 0 = 0$