MATH525 Crib Sheet
Julian Lore

$$V(aX+bY)=a^2V(X)+b^2V(Y)+2abCov(X,Y)$$

$$V(aX-bY)=a^2V(X)+b^2V(Y)-2abCov(X,Y) \qquad X\perp Y \implies Cov(X,Y)=0$$

$$V(Y)=E(Y^2)-(E(Y))^2=E[(Y-E[Y])^2]$$

$$MSE(\hat{Y})=E((\hat{Y}-Y)^2)=V(\hat{Y})+Bias(\hat{Y})^2$$

$$Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-E(X))(y_i-E(Y))$$

$$Cov(X,Y)=E\left[(X-E[X])(Y-E[Y])\right] \qquad V(X)=Cov(X,X)$$

$$Cov(X,Y)=E(XY)-E(X)E(Y) \qquad Bias(\hat{\theta})=E[\hat{\theta}]-\theta$$

$$V_W[W]=V_Y(E_{W|Y}[W\mid Y])+E_Y(V_{W|Y}[W\mid Y]) \qquad E_W[W]=E_Z[E_{W|Z}[W\mid Z]]$$

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

With replacement, select one unit with probability $\frac{1}{N}$ and repeat n times, getting $\pi_i=\frac{n}{N}$. Without replacement: Select one of all possible subsets of n population units. $\binom{N}{n}$ 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

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

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

Under SRS:

$$E(\hat{t}_s)=t,E(\overline{y}_S)=\overline{y}_U$$

$$V(\overline{y}_s)=\left(1-\frac{n}{N}\right)\frac{S^2}{n} \qquad \implies SE(\overline{y}_s)=\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$$

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

$$V(\hat{t}_s)=N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} \qquad \implies SE(\hat{t}_s)=N\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$$

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

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

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

Confidence interval for $\overline{y}_U$ $N\overline{y}_S\pm z_{\alpha/2}\widehat{SE}(\overline{y}_S)$
Finite population correction (fpc) $\left(1-\frac{n}{N}\right)$

Coefficient of variation (CV) $CV(\overline{y})=\frac{\sqrt{V(\overline{y})}}{E(\overline{y})}=\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}\overline{y}_U}$

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)

$$Z_i=\begin{cases}1 & \text{if unit } i \text{ is in the sample}\\ 0 & \text{otherwise}\end{cases} \qquad \overline{y}=\sum_{i\in\mathcal{S}}\frac{y_i}{n}=\sum_{i=1}^{N}Z_i\frac{y_i}{n}$$

$$\Pr(Z_i=1)=\frac{\binom{N-1}{n-1}}{\binom{N}{n}}=\frac{n}{N}\implies E[Z_i]=\frac{n}{N}$$

$$V(Z_i)=E(Z_i^2)-E(Z_i)^2=\frac{n}{N}-\left(\frac{n}{N}\right)^2=\frac{n}{N}\left(1-\frac{n}{N}\right)$$

$$E[Z_iZ_j]=\left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right),i\neq j$$

$$Cov(Z_i,Z_j)=E[Z_iZ_j]-E[Z_i]E[Z_j]=-\frac{1}{N-1}\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)$$

Sample Size Estimation

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

Relative error: $Pr\left(\frac{|\overline{y}_s-\overline{y}_U|}{|\overline{y}_u|}\leq r\right)=1-\alpha$

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

Naive sample size calc with no fpc (SRSWR): $n_0=\frac{S^2z_{\alpha/2}^2}{e^2}\implies n=\frac{n_0}{1+\frac{n_0}{N}}$

Weights

$$\hat{t}_S=\sum_{i\in S}\frac{N}{n}y_i=\sum_{i\in S}w_iy_i \qquad w_i=\frac{N}{n}=\frac{1}{\pi_i}=\frac{1}{Pr(z_i=1)} \qquad \overline{y}_S=\frac{\hat{t}}{N}=\frac{\sum_{i\in S}w_iy_i}{\sum_{i\in S}w_i}$$

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

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:

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

$$t_h=\sum_{j=1}^{N_h}y_{hj}=\text{ stratum } h \text{ total} \qquad t=\sum_{h=1}^{H}t_h=\text{ pop tot}$$

$$\overline{y}_{hU}=\frac{t_h}{N_h}=\text{ true stratum } h \text{ mean} \qquad \overline{y}_U=\frac{t}{N}=\text{ true pop mean}$$

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

Sample quantities:

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

$$\hat{t}_h=\frac{N_h}{n_h}\sum_{j\in S_n}y_{hj}=N_h\overline{y}_h \qquad s_h^2=\sum_{j\in S_h}\frac{(y_{jh}-\overline{y}_U)^2}{n_h-1}$$

$$\hat{t}_{str}=\sum_{h=1}^{H}\hat{t}_h=\sum_{h=1}^{H}N_h\overline{y}_h(\textbf{unbiased}) \qquad V(\hat{t}_{str})=\sum_{h=1}^{H}N_h^2\frac{S_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)$$

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

$$\overline{y}_{str}\pm z_{\alpha/2}SE(\overline{y}_{str}) \qquad \text{To est } V, \text{ use } s_h$$

Weights

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

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

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

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)
$V_{prop}(\hat{t}_{str})=\left(1-\frac{n}{N}\right)\frac{N}{n}\sum_{h=1}^{H}N_hS_h^2=\left(1-\frac{n}{N}\right)\frac{N}{n}(SSW+\sum_{h=1}^{H}S_h^2)$
$V(\hat{t}_{SRS})=\left(1-\frac{n}{N}\right)N^2\frac{S^2}{n}=\frac{N}{N-1}V_{prop}(\hat{t}_{str})+\frac{(N-h)N}{n(N-1)}(SSB-\sum_{h=1}^{H}s_h^2)$
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 (for stratified to beat SRS) s.t. sum of squares is large, variability within strata is smaller. **ANOVA** for stratified sampling:

Between	$df=H-1$	$SSB=\sum_{h=1}^{H}\sum_{j=1}^{N_h}\left(\overline{y}_{hU}-\overline{y}_U\right)^2=\sum_{h=1}^{H}N_h(\overline{y}_{hU}-\overline{y}_U)^2$
Within	$df=N-H$	$SSW=\sum_{h=1}^{H}\sum_{j=1}^{N_h}\left(y_{hj}-\overline{y}_{hU}\right)^2=\sum_{h=1}^{H}(N_h-1)S_h^2$
tot	$df=N-1$	$SSTO=\sum_{h=1}^{H}\sum_{j=1}^{N_h}\left(y_{hj}-\overline{y}_U\right)^2=(N-1)S^2$

Cost: $C=c_0+\sum_{h=1}^{H}c_hn_h$. Minimize $V(\hat{y}_{str})$ subject to the constraint that $C<C_{max}$.
Optimal allocation

$$n_h\propto\frac{N_hS_h}{\sqrt{c_h}}\implies\frac{n_h}{n}=\left(\frac{\frac{N_hS_h}{\sqrt{c_h}}}{\sum_{t=1}^{H}\frac{N_tS_t}{\sqrt{c_t}}}\right)$$

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

$$n_h\propto N_hS_h\implies\frac{n_h}{n}=\frac{N_hS_h}{\sum_{t=1}^{H}N_tS_t}$$

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. Prop better if S_h^2 relatively uniform, vary little. If S_h^2 vary lots, optimal alloc will result in smaller cost.

Determining Sample Size

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

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. Also, small n_h may make Hajek not apply.

Ratio Estimators

Auxiliary Variables Additional info on units used to improve precision. This is used after sampling (gotten from sampling), whereas stratification variables used before. y_i and x_i measured on unit i in pop.

$$t_y = \sum_{i=1}^{N} Y_i \qquad t_x = \sum_{i=1}^{N} X_i \qquad B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U} \qquad \hat{B} = \frac{\bar{y}_s}{\bar{x}_s}$$

Note that $\bar{y}_s$ and $\bar{x}_s$ are random, so $E(\hat{B}) \neq B$ unless $x_i = c, \forall i$, not unbiased
Correlation between X and Y is most important idea. Population correlation coefficient of x and y:

$$R = \sum_{i=1}^{M} \frac{(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}$$

R is a measure of linear association.
Why use Ratio Estimation? Estimate a ratio, estimate pop total but we don't know N, increase the precision of estimated means/totals, adjust estimates to reflect demographic totals or adjust for nonresponse.
e.g. Assume we want to know t_y, don't know N (pop total) but know t_x:

$$N = \frac{t_x}{\bar{x}_U} \qquad \hat{N} = \frac{t_x}{\bar{x}} \qquad \hat{t}_{yr} = \bar{y}\hat{N} = \bar{y}\frac{t_x}{\bar{x}} = \hat{B}t_x$$

$$\hat{\bar{y}}_r = \frac{\bar{x}_u}{\bar{x}}\bar{y} = \hat{B}\bar{x}_u \qquad E(\bar{y}_r) = E\left(\frac{\bar{x}_U}{\bar{x}}\bar{y}\right) \neq \bar{y}_U$$

Note that $E[\bar{y}_r - \bar{y}_U] = -E[\hat{B}(\bar{x} - \bar{x}_U)] = -Cov(\hat{B}, \bar{x})$. Because $\frac{\bar{x}-\bar{x}_U}{\bar{x}} \approx \frac{\bar{x}}{\bar{x}_U}$:

$$\begin{aligned} Bias(\hat{\bar{y}}_r) &= E(\hat{\bar{y}}_r - \bar{y}_U) = E\left(\frac{\bar{x}_U(\bar{y} - B\bar{x})}{\bar{x}}\right) = (\bar{y} - B\bar{x})\left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}}\right) \\ &\approx \frac{1}{\bar{x}_U}[BV(\bar{x}) - Cov(\bar{y}, \bar{x})] = \left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}_U}(BS_x^2 - RS_xS_y) \end{aligned}$$

This bias is small if n large, $\frac{n}{N}$ large, $\bar{x}_U$ large, S_x small, correlation close to 1. As noted above, if all $x_i = c$, then $S_x = 0$ so no bias. Alternatively, bias small if x does not depend too strongly on y, i.e. $Cov(\bar{x}, \bar{y})$ is small.

$$\begin{aligned} MSE[\hat{\bar{y}}_r] &= E[(\hat{\bar{y}}_r = \bar{y}_U)^2] = E\left[\left((\bar{y} - B\bar{x})\left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}}\right)\right)^2\right] \\ &= E\left[(\bar{y} - B\bar{x})^2 + (\bar{y} - B\bar{x})^2\left(\left(\frac{\bar{x} - \bar{x}_U}{\bar{x}}\right)^2 - 2\frac{(\bar{x} - \bar{x}_U)}{\bar{x}}\right)\right] \\ &\approx E[(\bar{y} - B\bar{x})^2] = V(\bar{y} - B\bar{x}) = V(\bar{y}) + B^2V(\bar{x}) - 2BCov(\bar{x}, \bar{y}) \\ &= \left(1 - \frac{n}{N}\right)\frac{S_y^2 + B^2S_x^2 - 2BRS_xS_y}{n} \end{aligned}$$

Small when n large, $\frac{n}{N}$ large, deviations $y_i - Bx_i$ small, R close to +1.

$$E[(\bar{y} - B\bar{x})^2] = V(\bar{d}) \qquad d_i = y_i - Bx_i \qquad \bar{d}_U = 0$$

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right)\left(\frac{\bar{x}_U}{\bar{x}}\right)^2\frac{s_e^2}{n} \qquad s_e^2 = \frac{1}{n-1}\sum_{i \in \mathcal{S}} e_i^2 \qquad e_i = y_i - \hat{B}x_i$$

$$\hat{V}(\hat{t}_{yr}) = \hat{V}(t_x\hat{B}) = \left(1 - \frac{n}{N}\right)\left(\frac{t_x}{\bar{x}}\right)^2\frac{s_e^2}{n} \qquad \hat{V}(\hat{B}) = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n\bar{x}^2}$$

Weights

$$\hat{t}_{yr} = \frac{t_x}{\hat{t}_x}\hat{t}_y = \frac{t_x}{\hat{t}_x}\sum_{i \in \mathcal{S}} w_i y_i$$

Modification in ratio estimation is like an adjustment to each weight

$$g_i = \frac{t_x}{\hat{t}_x}$$

$$\hat{t}_{yr} = \sum_{i \in \mathcal{S}} w_i g_i y_i$$

The weights are $w_i^* = w_i g_i$, but they depend on values from the sample. The weight adjustments g_i (**calibration factors**) **calibrate** estimates on x, i.e. $\sum_{i \in \mathcal{S}} w_i g_i x_i = t_x$.
Ratio estimation changes the weights from unbiased estimator. Re-weight such that $\hat{t}_x = t_x$
Advantages of Ratio Estimation If x and y perfectly correlated ($y_i = Bx_i, \forall i$), then $\hat{t}_{yr} = t_y$. In general, if y_i roughly proportion to x_i, MSE will be small. If deviations of y_i from $\hat{B}x_i$ smaller than deviations of y_i from $\bar{y}$, then $\hat{V}(\hat{\bar{y}}_r) \leq \hat{V}(\bar{y})$.

Domain Estimation

Estimate totals within subgroups (i.e. strata) in population. Assume we have D domains, $d = 1, \ldots, D$. s_d is the set of indices of sample units belonging to D.

$$\bar{y}_{U_d} = \frac{\sum_{i \in U_d} y_i}{N_d} \qquad N = \sum_{d=1}^{D} N_d$$

Where U_d is collection of units in d in pop, N_d is # of units in d. Estimate $\bar{y}_{U_d}$ with (ratio estimator):

$$\bar{y}_d = \frac{\sum_{i \in S_d} y_i}{n_d \text{ (random)}} \qquad x_i = \begin{cases} 1 & \text{if } i \in U_d \\ 0 & \text{if } i \notin U_d \end{cases} \qquad u_i = x_i y_i = \begin{cases} y_i & \text{if } i \in U_d \\ 0 & \text{if } i \notin U_d \end{cases}$$

$$t_x = \sum_{i=1}^{N} x_i = N_d \qquad t_u = \sum_{i=1}^{N} u_i = \sum_{i \in U_d} y_i$$

$$\bar{y}_{U_d} = \frac{t_u}{t_x} \qquad \bar{x} = \frac{n_d}{n} \qquad \bar{y}_d = \hat{B} = \frac{\bar{u}}{\bar{x}} = \frac{\hat{t}_u}{\hat{t}_x}$$

$$\begin{aligned} SE(\bar{y}_d) &= \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}^2}\frac{\sum_{i \in S}(u_i - \hat{B}x_i)^2}{n-1}} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}^2}\frac{\sum_{i \in \mathcal{S}_d}(y_i - \hat{B}x_i)^2}{n-1}} \\ &= \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n_d^2}\frac{(n_d - 1)S_{y_d}^2}{n-1}} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{n}{n-1}\frac{n_d - 1}{n_d}\frac{s_{y_d}^2}{n_d}} \\ s_{yd}^2 &= \frac{\sum_{i \in \mathcal{S}_d}(y_i - \bar{y}_d)^2}{n_d - 1} \end{aligned}$$

If $E(n_d)$ large, then: $SE(\bar{y}_d) \approx \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_{y_d}^2}{n_d}}$
What about estimating totals?
If we know N_d, then $\hat{t}_u = N_d\bar{y}_d$. For SE, just multiply the above by N_d.

Otherwise if we don't know N_d, then $\hat{t}_u = N\bar{u}$ with $SE(\hat{t}_u) = NSE(\bar{u}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_u^2}{n}}$ (much worse than when we knew N_d, but unbiased)

Post stratification

If we know stratum membership before sampling, we should stratify. If we only know after: $\bar{y}_{str} = \sum_{h=1}^{H} \bar{y}_h \frac{N_h}{N}$ This is not unbiased, since $\bar{y}_h$ is a ratio estimator and n_hs are random instead of fixed like in stratified.
$\bar{y}_1, \ldots, \bar{y}_H$ are post-stratified means in sample, $n_1, \ldots, n_H$ are random # of units in each post stratum
Sample (y_i, x_i) pairs. If x_i is something you wished you could have stratified on before, stratify after:

$$x_{ih} = \begin{cases} 1 & \text{if } i \in \text{post stratum } h \\ 0 & \text{otherwise} \end{cases} \qquad x_{ih}y_i = u_{ih} = \begin{cases} y_i & \text{if } i \in \text{post stratum } h \\ 0 & \text{otherwise} \end{cases}$$

$$t_{xh} = \sum_{i=1}^{N} x_{ih} = N_h \qquad \hat{t}_{xh} = \frac{N}{n}\sum_{i \in \mathcal{S}} x_{ih} = \frac{N}{n}\underbrace{n_h}_{\text{random}} = \hat{N}_h$$

$$t_{uh} = \sum_{i=1}^{N} u_{ih} = t_{yh} \qquad \bar{y}_{uh} = \frac{\sum_{i=1}^{N} u_{ih}}{N_h} = \frac{t_{yh}}{t_{xh}} = B_h$$

$$\hat{t}_{uh} = \sum_{i \in \mathcal{S}} \frac{N}{n} u_{ih}$$

$$\hat{t}_{uhr} = \frac{t_{xh}}{\hat{t}_{xh}}\hat{t}_{uh} = \frac{N_h}{\hat{N}_h}\hat{t}_{uh} = N_h\bar{y}_h = \bar{y}_h t_{xh} = \frac{\sum_{i \in \mathcal{S}} y_i x_{ih}}{\sum_{i \in \mathcal{S}} x_{ih}} t_{xh} = \frac{\hat{t}_{uh}}{\hat{t}_{xh}} t_{xh}$$

$$\hat{t}_{y,post} = \sum_{h=1}^{H} \hat{t}_{uhr} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h}\hat{t}_{uh} = \sum_{h=1}^{H} N_h\bar{y}_h \qquad \bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h$$

$$V(\bar{y}_{post}) \approx \left(1 - \frac{n}{N}\right)\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\frac{s_h^2}{n} + \frac{1}{n}\left(\frac{N-n}{N-1}\right)\sum_{h=1}^{H}\left(1 - \frac{N_h}{N}\right)\frac{s_h^2}{n}$$

Ratio Estimation with Stratified Sampling

Combined ratio estimator: (apply ratio estimator on stratified sample) Assume fixed strata $h = 1, \ldots, H$

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

If we also have x, we can also use ratio estimation:

$$\hat{t}_{x,str} = \sum_{h=1}^{H} \hat{t}_{hx} = \sum_{h=1}^{H}\sum_{j \in \mathcal{S}_h} w_{hj} x_{hj} \qquad B = \frac{t_y}{t_X} \implies \hat{B} = \frac{\hat{t}_{y,str}}{\hat{t}_{x,str}}$$

$$t_y = Bt_x \implies \hat{t}_{yrc}\text{(c for combined)} = \hat{B}t_x$$

$$MSE(\hat{t}_{yrc}) \approx V(\hat{t}_{yrc} - B\hat{t}_{x,str}) = V\left[\sum_{h=1}^{H}\sum_{j \in \mathcal{S}_m} w_{hj}(y_{hj} - Bx_{hj})\right]$$

$$\begin{aligned} \hat{V}(\hat{t}_{yrc}) &= \left(\frac{t_{x,str}}{\hat{t}_{x,str}}\right)^2 \hat{V}\left(\sum_{h=1}^{H}\sum_{j \in \mathcal{S}_h} w_{hj} e_{hj}\right) \qquad e_{hj} = y_{hj} - \hat{B}_{x_{hj}} \\ &= \left(\frac{t_{x,str}}{\hat{t}_{x,str}}\right)^2 [\hat{V}(\hat{t}_{y,str}) + \hat{B}^2\hat{V}(\hat{t}_{x,str}) - 2\hat{B}\widehat{Cov}(\hat{t}_{y,str}, \hat{t}_{x,str})] \end{aligned}$$

Separate ratio estimator: (ratio estimation and then combine strata)

$$\hat{t}_{yrs} = \sum_{h=1}^{H} \hat{t}_{yhr} = \sum_{h=1}^{H} t_{xh}\frac{\hat{t}_{yh}}{\hat{t}_{xh}} \qquad V(\hat{t}_{yrs}) = \sum_{h=1}^{H} V(\hat{t}_{yhr})$$

Which is best? Depends.
Separate is good if ratios across strata $\left(\frac{t_{yh}}{t_{xh}}\right)$ vary a lot, but you are accumulating bias with h (each ratio adds bias)
Combined is good if ratios are constant across strata and less bias if some sample sizes in strata are small.

Cluster Sampling

Clusters are the primary sampling units (**psu**'s)
Units within clusters are secondary sampling units (**ssu**'s), units of the target pop of interest
N psu's indexed by i, M_i ssu's in i-th cluster, $M_0 = \sum_{i=1}^{N} M_i$ = # of ssu's in target population. **Why use cluster sampling**? Making sampling frame list of observations units can be difficult, expensive, impossible. Population may be widely distributed geographically, may occur in natural clusters (e.g. houses) and less expensive to sample clusters than SRS of individuals. However, unlike stratified sampling, cluster sampling tends to decrease precision.

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{psu } i \text{ total} \qquad t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \text{pop total}$$

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(t_i - \frac{t}{N}\right)^2 = \text{pop var of psu tots}$$

$$\bar{y}_U = \sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{y_{ij}}{M_0} = \frac{t}{M_0} = \text{pop mean} \qquad \bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i} = \text{pop } \mu \text{ in psu } i$$

$$S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{M_0 - 1} = \text{pop var at ssu level} \qquad S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1} = \text{var in psu } i$$

Assume sample of both levels. Let Ω be set of psu indices in sample and Ω_i be set of ssu indices for psu i in sample.

$$n = \text{# of psu's in sample} \qquad m_i = \text{# of ssu's from psu } i \text{ in sample}$$

$$\bar{y}_i = \sum_{j \in \Omega_i} \frac{y_{ij}}{m_i} = \text{sample mean for psu } i$$

$$\hat{t}_i = M_i\bar{y}_i = \sum_{j \in \Omega_i} \frac{M_i}{m_i} y_{ij} \qquad \hat{t} = \sum_{i \in \Omega} N\frac{\hat{t}_i}{n} = \sum_{i \in \Omega} \frac{N}{n}\hat{t}_i$$

$$s_t^2 = \frac{1}{n-1}\sum_{i \in \Omega}\left(\hat{t}_i - \frac{\hat{t}}{N}\right)^2 \qquad s_i^2 = \frac{1}{m_i - 1}\sum_{j \in \Omega_i}\left(y_{ij} - \bar{y}_i\right)^2$$

One-stage Sampling

All ssu's within sampled psu's are included in the sample (census of ssu's). Assume initially $M_i = m_i = M$.

$$\hat{t} = \frac{N}{n}\sum_{i\in\Omega} t_i \qquad V(\hat{t}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}$$

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

$$V(\hat{\bar{y}}) = \left(1-\frac{n}{N}\right)\frac{S_t^2}{nM^2} \qquad \widehat{SE}(\hat{\bar{y}}) = \frac{1}{M}\sqrt{\left(1-\frac{n}{N}\right)\frac{s_t^2}{N}}$$

Still self-weighting:

$$w_{ij} = \frac{1}{\Pr(\text{ssu } j \text{ in psu } i \text{ in sample})} = \frac{N}{n}$$

$$\hat{t} = \sum_{i\in\Omega}\frac{N}{n}t_i = \sum_{i\in\Omega}\sum_{j\in\Omega_i}\frac{N}{n}y_{ij} = \sum_{i\in N}\sum_{j=1}^{M}\left(\frac{N}{n}\right)y_{ij} = \sum_{i\in\Omega}\sum_{j=1}^{M} w_{ij}y_{ij}$$

$$\hat{\bar{y}} = \frac{\sum_{i\in\Omega}\sum_{j=1}^{M} w_{ij}y_{ij}}{\sum_{i\in\Omega}\sum_{j=1}^{M} w_{ij}} = \frac{\frac{N}{n}\sum_{i\in\Omega}\sum_{j=1}^{M} y_{ij}}{nM\frac{N}{n}} = \frac{\sum_{i\in\Omega}\sum_{j=1}^{M} y_{ij}}{nM}$$

Role of variability changes. In stratified, it is good, in cluster sampling, it is bad.

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_U\right)^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_{iU}\right)^2 + \sum_{i=1}^{N} M\left(\bar{y}_{iU}-\bar{y}_U\right)^2 = SSW + SSB$$

ANOVA for cluster sampling:

Between	$df = N-1$	$SSB = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\bar{y}_{iU}-\bar{y}_U\right)^2$	$MSB = \frac{SSB}{N-1}$
Within	$df = N(M-1)$	$SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_{iU}\right)^2$	$MSW = \frac{SSW}{N(M-1)}$
tot	$df = NM-1$	$SSTO = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_U\right)^2$ $= SSB + SSW$	$S^2 = \frac{SSTO}{NM-1}$

$$S_t^2 = \frac{\sum_{i=1}^{N}\left(t_i-\frac{t}{N}\right)^2}{N-1} = \frac{\sum_{i=1}^{N}\left(M\bar{y}_{iU}-M\bar{y}_U\right)^2}{N-1} = \frac{\sum_{i=1}^{N}M^2\left(\bar{y}_{iU}-\bar{y}_U\right)^2}{N-1}$$

$$= M\frac{SSB}{N-1} = M\times MSB$$

$$V(\hat{t}_{SRS}) = (NM)^2\left(1-\frac{nM}{NM}\right)\frac{s^2}{nM} = N^2\left(1-\frac{n}{N}\right)\frac{S^2 M}{n} = N^2\left(1-\frac{n}{N}\right)\frac{SSB+SSW}{NM-1}\frac{M}{n}$$

$$V(\hat{t}_{cluster}) = \left(1-\frac{n}{N}\right)N^2\frac{M\left(\frac{SSB}{N-1}\right)}{n} = N^2\left(1-\frac{n}{N}\right)\frac{M\times MSB}{n}$$

Intraclass correlation coefficient (ICC) tells us how similar elements in same cluster are, measure the extent to which clusters are homogeneous relative to overall variability. ICC is the Pearson correlation coefficient for the $NM(M-1)$ pairs (y_{ij}, y_{ik}) for $1 \le i \le N, j \ne k$ s.t.

$$ICC = 1 - \frac{M}{M-1}\times\frac{SSW}{SSB+SSW}$$

$$0 \le SSW/SSTO \le 1 \implies -\frac{1}{M-1} \le ICC \le 1$$

$$MSB = \frac{N(M-1)}{M(N-1)}S^2[1+(M-1)ICC]$$

$$\frac{V(\hat{t}_{cluster})}{V(\hat{t}_{SRS})} = \frac{MSB}{S^2} = \frac{N(M-1)}{M(N-1)}[1+(M-1)ICC]$$

If ICC is positive, cluster is worse. Also note that if clusters are perfectly homogeneous, then $SSW = 0 \implies ICC = 1$. Design effect (ratio of variances) is about 1 when $ICC \approx 1$.
ICC only works for clusters of equal sizes. Alternative measure of homogeneity is the **adjusted** R^2: $R_a^2 = 1 - \frac{MSW}{S^2}$

Unequal size clusters

What happens if M_i's differ?

$$\hat{t}_{unb} = \frac{N}{n}\sum_{i\in\Omega} t_i \qquad M_0 = \sum_{i=1}^{N} M_i \qquad V(\hat{t}_{unb}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} \qquad V(\hat{\bar{y}}_{unb}) = \frac{V(\hat{t}_{unb})}{N^2M_0^2}$$

Unbiased, but inefficient due to variability between M_i's. To reduce this variance, we use **ratio estimation**.

$$\bar{y}_U = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{M} M_i} = \frac{t}{M_0} \qquad \hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\frac{N}{n}\sum_{i\in\Omega} t_i}{\frac{N}{n}\sum_{i\in\Omega} M_i} = \frac{\sum_{i\in\Omega} t_i}{\sum_{i\in\Omega M_i}} = \hat{B}$$

Alternatively, with weights:

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\Omega} M_i\bar{y}_i}{\sum_{i\in\Omega} M_i} = \frac{\sum_{i\in\Omega}\sum_{j\in\Omega_i} y_{ij}}{\sum_{i\in\Omega}\sum_{j\in\Omega_i} 1} = \frac{\sum_{i\in\Omega}\sum_{j\in\Omega_i}\frac{N}{n}y_{ij}}{\sum_{i\in\Omega}\sum_{j\in\Omega_i}\frac{N}{n}} = \frac{\sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}y_{ij}}{\sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}}, w_{ij} = \frac{1}{\pi_{ij}}$$

$$SE(\hat{\bar{y}}_r) = \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\overline{M}^2}\sum_{i\in\Omega}\frac{(t_i-\hat{\bar{y}}_r M_i)^2}{n-1}} \qquad \overline{M} = \frac{\sum_{i\in\Omega} M_i}{n} \qquad \hat{t}_r = \hat{B}M_0 = \frac{\sum_{i\in\Omega} t_i}{\sum_{i\in\Omega} M_i}M_0$$

Two Stage Cluster Sampling

1. Select an SRS of size n of psu's (from N)
2. Select an SRS of size m_i of ssu's (from M_i) within psu i (note that if $m_i \ge M_i$, we just sample the whole M_i)

$$\hat{t}_i = \sum_{j\in\Omega_i}\frac{M_i}{m_i}y_{ij} = M_i\bar{y}_i$$

$$\frac{1}{w_{ij}} = \Pr(\text{ssu } j \text{ from psu } i \text{ in sample}) = \frac{n}{N}\times\frac{m_i}{M_i} = \Pr(\text{psu in samp})\Pr(\text{ssu in samp})$$

$$= \frac{nm_i}{NM_i} \implies w_{ij} = \frac{NM_i}{nm_i}$$

$$\hat{t}_{unb} = \sum_{i\in\Omega}\frac{N}{n}\hat{t}_i = \sum_{i\in\Omega}\frac{N}{n}M_i\bar{y}_i = \sum_{i\in\Omega}\frac{NM_i}{n}\sum_{j\in\Omega_i}\frac{1}{m_i}y_{ij}$$

$$= \sum_{i\in\Omega} w_{ij}\sum_{j\in\Omega_i} y_{ij} = \sum_{j\in\Omega_i} y_{ij} = \sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}y_{ij}$$

Not automatically a self weighted sample, only if $\frac{M_i}{m_i} = c, \forall i$

Two sources of variability: Between cluster variability (t_i's), $N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}$. Within cluster variability ($\hat{t}_i$), $\left(1-\frac{m_i}{M-i}\right)M_i^2\frac{S_i^2}{m_i}$

$$V(\hat{t}_{unb}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{M}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i} \qquad \hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0}$$

$$\hat{V}_{WR}(\hat{t}_{unb}) = N^2\frac{s_t^2}{n} \text{ (with replacement close to var)}$$

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\Omega}\hat{t}_i}{\sum_{i\in\Omega} M_i} = \frac{\sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}y_{ij}}{\sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}}$$

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\overline{M}^2}\sum_{i\in\Omega} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\Omega}(m_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2$$

Designing a Cluster Sample

Assume $M_i = M, m_i = m$

$$V(\hat{\bar{y}}_{unb}) = \left(1-\frac{n}{N}\right)\frac{MSB}{nM} + \left(1-\frac{m}{M}\right)\frac{MSW}{nM}$$

If $MSW = 0$, then choose $m = 1$ (all points equal).

c_1 = cost for sampling 1 psu

c_2 = cost for sampling 1 ssu

$C = c_1 n + c_2 nm$ = total cost of sampling

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}} \qquad m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$$

Choosing sample size (# psu's): If clusters same size and we ignore psu-level fpc

$$V(\hat{y}_{unb}) \le \frac{1}{n}\left[\frac{MSB}{M} + \left(1-\frac{m}{M}\right)\frac{MSW}{m}\right] = \frac{1}{n}v \qquad \hat{\bar{y}}_{unb} \pm z_{\alpha/2}\sqrt{\frac{1}{n}v} \qquad n = z_{\alpha/2}^2 v/e^2$$

Systematic Sampling

Is cluster sampling. Choose a starting point and then every i^{th} index after to make up the psu's. We get the mean of one psu (take SRS of one psu generated by the systematic counting above).

$$\bar{y}_i = \bar{y}_{iU} = \hat{\bar{y}}_{sys} \qquad E[\hat{y}_{sys}] = \bar{y}_U$$

$$V(\hat{y}_{sys}) = \left(1-\frac{1}{N}\right)\frac{S_t^2}{M^2} = \left(1-\frac{1}{N}\right)\frac{MSB}{M} \approx \frac{S^2}{M}[1+(M-1)ICC]$$

Unequal Probability Sampling

$$w_i = \frac{1}{\pi_i} \qquad w_{ij} = \frac{1}{\pi_{ij}} \qquad \hat{t} = \sum_{i\in\Omega} w_i y_i \qquad \bar{y}_r = \frac{\sum_{i\in\Omega} w_i y_i}{\sum_{i\in\Omega} w_i} \qquad \bar{y}_{unb} = \frac{\sum_{i\in\Omega} w_i y_i}{N}$$

Start with sampling one unit, $n = 1$: $\psi_i = \pi_i = \Pr(i \text{ is selected}), y_i = t_i$

$$\hat{t}_\psi = w_i y_i = \frac{1}{\psi_i}y_i, \text{ where } i \in \Omega$$

$$E(\hat{t}_\psi) = E\left(\sum_{i\in\Omega} w_i y_i\right) = E\left(\sum_{j=1}^{N} w_j y_j z_j\right) = \sum_{j=1}^{N}\frac{1}{\psi_j}y_j E(Z_j) = \sum_{j=1}^{N}\frac{1}{\psi_j}y_j\psi_j = t$$

$$V(\hat{t}_\psi) = E((\hat{t}_\psi - t)^2) = \sum_{\text{possible samples }\Omega}\Pr(\Omega)(\hat{t}_{\psi\Omega} - t)^2 = \sum_{i=1}^{N}\psi_i\left(\frac{y_i}{\psi_i} - t\right)^2$$

With ψ proportional to pop size, we get smaller var here than for SRS, which makes sense since we use auxiliary information to sample, which we believe to be correlated to what we're measuring.

One Stage Sampling

$$\hat{t}_\psi = \sum_{i\in\Omega} w_i t_i = \sum_{i\in\Omega}\frac{1}{\psi_i}t_i \qquad E(\hat{t}_\psi) = t \qquad V(\hat{t}_\psi) = \sum_{i=1}^{N}\psi\left(\frac{t_i}{\psi_i} - t\right)^2 \qquad \pi_i = 1 - (1-\psi_i)^n$$

We take a sample of n psu's with replacement, straightforward

if we know all psu sizes $\implies \psi_i = \frac{M_i}{M_0}, M_0 = \sum_{i=1}^{N} M_i$

To sample with only ssu's: Take a sample of size 1 from list of ssu's by SRS. Take all ssu's from the psu of the one ssu sampled $\implies \psi_i \propto M_i$.

Lahiri's Algorithm

N = # of psu's in pop. Let $\max\{M_i\}$ = max psu size. **Lahiri's method** is as follows:

1. Draw # i from 1 to N uniformly (with replacement)
2. Draw second random number k between 1 and $\max\{M_i\}$. If $k \le M_i$ for psu in step 1, select psu i, otherwise go back to 1.
3. Repeat until n psu's selected.

This is an example of **rejection sampling**. Sample with replacement with $\psi_i = \frac{M_i}{\sum_{i=1}^{N} M_i}$. Note that the prob for selection for slot $j = \frac{1}{N}\frac{M_i}{\max\{M_i\}} \propto M_i$.

$\mathcal{R}$ denotes units included in the sample, including repeats.

$$\hat{t}_\psi = \frac{\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i}}{n} = \frac{1}{n}\sum_{i\in\mathcal{R}} u_i = \bar{u}, u_i = \frac{t_i}{\psi_i}$$

$$Q \sim Mult(n, \psi_1, \ldots, \psi_n), Q_i = \text{\# times } i \text{ appears in samp}$$

$$E(\hat{t}_\psi) = E\left(\frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i}\right) = \frac{1}{n}E\left(\sum_{i=1}^{N} Q_i\frac{t_i}{\psi_i}\right) = \frac{1}{n}\sum_{i=1}^{N} n\psi_i\frac{t_i}{\psi_i} = t$$

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N} Q_i\frac{t_i}{\psi_i} = \frac{1}{n}\sum_{i=1}^{N}\left[\sum_{j=1} q_{ij}\right]\frac{t_i}{\psi_i} = \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i=1}^{N} q_{ij}\right]\frac{t_i}{\psi_i} \implies$$

$$V(\hat{t}_\psi) = \frac{1}{n^2}\sum_{j=1}^{n} V(\underbrace{\sum_{i=1}^{N} q_{ij}}_{\text{only one of these} \ne 0}\frac{t_i}{\psi_i}) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2$$

$$\hat{V}(\hat{t}_\psi) = \frac{s_u^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}(u_i - \bar{u})^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2$$

$$E(\hat{V}(\hat{t}_\psi)) = V(\hat{t}_\psi)$$

$$\hat{\overline{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}} \qquad \hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{M_i}{\psi_i} \qquad \hat{V}(\hat{\overline{y}}_\psi) = \frac{1}{(\hat{M}_0\psi)^2}\frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \frac{\hat{\overline{y}}_\psi M_i}{\psi_i}\right)^2$$

If N is small or some ψ_i are very large, possible that the sample will be one psu sampled n times, estimated variance will be 0. Better to use sampling without replacement in this case.

Choosing ψ_i Optimal $\psi_i = \frac{t_i}{t}$, but this requires t and t_i. Take **probability proportional to size** (pps) sampling.

$$\psi_i = \frac{M_i}{M_0} \qquad \frac{t_i}{\psi_i} = t_i\frac{M_0}{M_i} = M_0\overline{y}_i$$

$$\hat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}} M_0\overline{y}_i \qquad \hat{\overline{y}}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\overline{y}_i$$

$$\hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{M_i}{\psi_i} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{M_i}{M_i/M_0} = \frac{1}{n}\sum_{i\in\mathcal{R}} M_0 = M_0$$

$$\hat{V}(\hat{\overline{y}}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}(\overline{y}_i - \hat{\overline{y}}_\psi)^2$$

Weights without replacement, $w_i = \frac{1}{E(Z_i)}$. With replacement:

$$w_{ij} = w_i = \frac{1}{E(Q_i)} = \frac{1}{n\psi_i} \qquad \hat{t}_\psi = \sum_{i\in\mathcal{R}}\sum_{j=1}^{M_i} w_{ij}y_{ij} \qquad \hat{\overline{y}}_\psi = \frac{\sum_{i\in\mathcal{R}}\sum_{j=1}^{M_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{R}}\sum_{j=1}^{M_i} w_{ij}}$$

If ψ_i are unequal, not self-weighting. In one stage pps, elements in larger psu's have smaller weights than elements in small psu's.

Two Stage Sampling

With replacement:

1. Take sample of psu's with replacement choosing psu i with prob ψ_i at slow j out of n.
2. Take probability sample of m_i ssu's from psu i at each replication.
3. If psu i appears Q_i times, $\hat{t}_{i1},\dots,\hat{t}_{iQ_i}$ are different.

Conditions: Different subsamples within psu i must be done independently and with same sampling scheme. Also, j^{th} subsample from psu i is selected s.t. $E[\hat{t}_{ij}] = t_i, j = 1,\dots,Q_i$. Using the same procedure each time means $V[\hat{t}_{ij}] = V_i, \forall j$

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i}$$

$$E(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N} E_{Q_i}\left(E_{\hat{t}_{ij}}\left(\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i} \mid Q_i\right)\right) = \frac{1}{n}\sum_{i=1}^{N} E_{Q_i}\left(\frac{1}{\psi_i}\sum_{j=1}^{Q_i} E\left(\hat{t}_{ij} \mid Q_i\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{N}\frac{t_i}{\psi_i}E_{Q_i}[Q_i] = t$$

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\left(\frac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi\right)^2$$

$$\hat{\overline{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}}, \hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{M_i}{\psi_i}$$

For an SRS of size m_i at stage 2, $w_{ij} = \frac{1}{n\psi_i}\frac{M_i}{m_i}$.
If $\psi_i = \frac{M_i}{M_0}$ (**Probability proportional to size** - PPS), $w_{ij} = \frac{M_0}{nM_i}\frac{M_i}{m_i} = \frac{M_0}{nm_i}$. If $m_i = m \implies w_{ij} = \frac{M_0}{nm}$, self-weighted

Without Replacement

Probability of being chosen second is different from being chosen first. For $n = 2$ we have:

$$\Pr(i \text{ chosen } 1^{st}, \text{ then } j\ 2^{nd}) = \psi_i\frac{\psi_j}{1-\psi_i} \neq \psi_j\frac{\psi_i}{1-\psi_j} = \Pr(j\ 1^{st}, i\ 2^{nd})$$

$$\pi_{ij} = \Pr(i \text{ and } j \text{ in sample}) = \psi_i\frac{\psi_j}{1-\psi_i} + \psi_j\frac{\psi_i}{1-\psi_j}$$

In general:

$$\pi_i = E(Z_i) \implies \sum_{i=1}^{N} Z_i = n$$

$$\pi_{ij} = E(Z_iZ_j) = \sum_{\substack{j=1\\ j\neq i}}^{N}\pi_{ik} = \sum_{\substack{j=1\\ j\neq i}}^{N} E(Z_iZ_j) = E\left[\sum_{j=1}^{N} Z_iZ_j - Z_i^2\right] = E\left[\sum_{j=1}^{N} Z_iZ_k - Z_i\right]$$

$$= E[Z_in - Z_i] = (n-1)E(Z_i) = (n-1)\pi_i$$

$\frac{\pi_i}{n}$ is average prob of selection for each draw

Horvitz-Thompson Estimator for One Stage

$$\hat{t}_{HT} = \sum_{i\in\Omega}\frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i\frac{t_i}{\pi_i} \qquad Cov(Z_i,Z_i) = \pi_i(1-\pi_i) \qquad Cov(Z_i,Z_k) = \pi_{ik} - \pi_i\pi_k, i\neq k$$

Assume subsampling independent between psu's, $\hat{t}_i \perp Z_1,\dots,Z_n$ and $E(\hat{t}_i \mid Z_1,\dots,Z_n) = t_i$

$$E(\hat{t}_i \mid Z_1,\dots,Z_n) = t_i$$
$$V(\hat{t}_i \mid Z_1,\dots,Z_n) = v_i$$

$$E[\hat{t}_{HT}] = E\left[\sum_{i=1}^{N} Z_i\frac{\hat{t}_i}{\pi_i}\right] = E_{Z_1,\dots,Z_N}\left[E_{\hat{t}_1,\dots,\hat{t}_N|Z_1,\dots,Z_n}\left[\sum_{i=1}^{N} Z_i\frac{\hat{t}_i}{\pi_i} \mid Z_1,\dots,Z_n\right]\right]$$

$$= E_{Z_1,\dots,Z_n}\left[\sum_{i=1}^{N} Z_iE_{\hat{t}_i|Z_i}\left[\frac{\hat{t}_i}{\pi_i}\right]\right] = E_{Z_1,\dots,Z_n}\left[\sum_{i=1}^{N} Z_i\frac{t_i}{\pi_i}\right] = \sum_{i=1}^{N} t_i = t$$

$$V(\hat{t}_{HT}) = V_{\vec{Z}}(E_{\vec{\hat{t}}|\vec{Z}}(\hat{t}_{HT} \mid Z_1,\dots,Z_n)) + E_{\vec{Z}}(V_{\vec{\hat{t}}|\vec{Z}}(\hat{t}_{HT} \mid Z_1,\dots,Z_N))$$

$$= V\left[\sum_{i=1}^{N} Z_iE\left[\frac{\hat{t}_i}{\pi_i}\right]\right] + E\left[\sum_{i=1}^{N} Z_i^2V\left[\frac{\hat{t}_i}{\pi_i}\right]\right]$$

$$= V\left[\sum_{i=1}^{N} Z_i\frac{t_i}{\pi_i}\right] + E\left[\sum_{i=1}^{N} Z_i^2\frac{v_i}{\pi_i}\right] = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{t_it_k}{\pi_i\pi_k}Cov(Z_i,Z_k) + \sum_{i=1}^{N}\pi_i\frac{v_i}{\pi_i^2}$$

$$= \sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{t_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{i\neq k}(\pi_{ik} - \pi_i\pi_k)\frac{t_it_k}{\pi_i\pi_k} + \underbrace{\sum_{i=1}^{N}\frac{v_i}{\pi_i}}_{\text{ssu var}}$$

For one stage, $V(\hat{t}_i) = 0$ for $i\in\Omega \implies$

$$V_{HT}(\hat{t}_{HT}) = \sum_{i=1}^{N}(1-\pi_i)\frac{t_i^2}{\pi_i} + \sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}(\pi_{ik} - \pi_i\pi_k)\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}(\pi_i\pi_k - \pi_{ik})\left(\frac{t_i}{\pi} - \frac{t_k}{\pi_k}\right)^2 = V_{SYG}(\hat{t}_{HT})$$

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum_{i=1}^{N} Z_i(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N} R_{ik}\left(\frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}}\right)\frac{\hat{t}_i\hat{t}_k}{\pi_i\pi_k}$$

$$R_{ik} = \begin{cases} 1 & \text{if } Z_i = Z_k = 1\\ 0 & \text{otherwise}\end{cases}$$

$$\hat{V}_{SYG}(\hat{t}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N} Z_iZ_k\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 \neq \hat{V}_{HT}(\hat{t}_{HT})$$

However, both are unbiased. $\hat{V}_{SYG}$ usually preferred. Since π_{ik} can be hard to compute, we can still estimate the variance by pretending it was taken with replacement, $\psi_i = \frac{\pi_i}{n}$:

$$\hat{V}_{WR}(\hat{t}_{HT}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\Omega}\left(\frac{t_i}{\psi_i} - \hat{t}_{HT}\right) = \frac{n}{n-1}\sum_{i\in\Omega}\left(\frac{t_i}{\pi_i} - \frac{\hat{t}_{HT}}{n}\right)^2$$

Two Stage Horvitz-Thompson

$$\hat{t}_{HT} = \sum_{i\in\Omega}\frac{\hat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i\frac{\hat{t}_i}{\pi_i}$$

$$V(\hat{t}_{HT}) = \sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{t_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{i\neq k}(\pi_{ik} - \pi_i\pi_k)\frac{t_it_k}{\pi_i\pi_k} + \sum_{i=1}^{N}\frac{v_i}{\pi_i}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N}(\pi_i\pi_k - \pi_{ik})\left(\frac{t_i}{\pi} - \frac{t_k}{\pi_k}\right)^2 + \sum_{i=1}^{N}\frac{v_i}{\pi_i}$$

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum_{i=1}^{N} Z_i(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N} R_{ik}\left(\frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}}\right)\frac{\hat{t}_i\hat{t}_k}{\pi_i\pi_k} + \sum_{i=1}^{N} Z_i\frac{\hat{v}_i}{\pi_i}$$

$$\hat{V}_{SYG}(\hat{t}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N} Z_iZ_k\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 + \sum_{i=1}^{N} Z_i\frac{\hat{v}_i}{\pi_i} \neq \hat{V}_{HT}(\hat{t}_{HT})$$

$$\hat{V}_{WR}(\hat{t}_{HT}) = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\Omega}\left(\frac{n\hat{t}_i}{\pi_i} - \hat{t}_{HT}\right)^2 = \frac{n}{n-1}\sum_{i\in\Omega}\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_{HT}}{n}\right)^2$$

Weights

One Stage $\hat{t}_{HT} = \sum_{i\in\Omega} w_i\hat{t}_i, w_i = \frac{1}{\pi_i}$

Two Stage with SRS at stage 2:

$$\hat{t}_i = \sum_{j\in\Omega_i}\frac{y_{ij}}{\pi_{j|i}} \qquad \pi_{j|i} = \frac{m_i}{M_i} \qquad w_{ij} = \frac{1}{\pi_i\pi_{j|i}}$$

$$\hat{t}_{HT} = \sum_{i\in\Omega} w_i\hat{t}_i = \sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}y_{ij} \qquad \hat{\overline{y}}_{HT} = \frac{\sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}y_{ij}}{\sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}} \qquad \hat{M}_0 = \sum_{i\in\Omega}\sum_{j\in\Omega_i} w_{ij}$$

Residual from the fact that $\hat{\overline{y}}$ is ratio estimator: $\hat{e}_i = \hat{t}_i - \hat{\overline{y}}_{HT}\hat{M}_i, \hat{M}_i = \sum_{j\in\Omega_i}\frac{1}{\pi_{j|i}}$ we then get:

$$\hat{V}_{WR}(\hat{\overline{y}}_{HT}) = \frac{n}{n-1}\sum_{i\in\Omega}\left(\frac{\hat{e}_i}{\hat{M}_0\pi_i}\right) = \frac{n}{n-1}\sum_{i\in\Omega}\left(\frac{\sum_{j\in\Omega_i} w_{ij}(y_{ij} - \overline{y}_{HT})}{\sum_{k\in\Omega}\sum_{j\in\Omega_i} w_{kj}}\right)^2$$

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
$S_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_{i,Dec} - y_{i,Nov} - [\overline{y}_{u,Dec} - \overline{y}_{u,Nov}])^2 = V(Y_{Dec}) + V(Y_{Nov}) - 2Cov(Y_{Dec}, Y_{Nov})$.
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: $Cov(\overline{y}_n, \overline{y}_m) = Cov\left(\sum_{i\in S_1}\frac{y_i}{n}, \sum_{j\in S_2}\frac{y_j}{m}\right) = Cov\left(\sum_{i=1}^{N} Z_i\frac{y_i}{n}, \sum_{j=1}^{N} Z_j'\frac{y_j}{m}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} y_iy_j\frac{1}{n}\frac{1}{m}\underbrace{Cov(Z_i, Z_j')}_{E(Z_iZ_j') - E(Z_i)E(Z_j')}$. Use

$E(Z_iZ_j') = Pr(Z_i = 1, Z_j' = 1) = Pr(Z_j' = 1 \mid Z_i = 1)Pr(Z_i = 1)$

Design Effect of $\hat{t}$ relative to $\hat{t}_2$ is $\frac{V(\hat{t})}{V(\hat{t}_2)}$

Lagrange: Set constraint = 0 (e.g. $C = c_0 + c_1 \to 0 = c_0 + c_1 - c^*$), then partial derive wrt to each variable (one at a time) of $f - \lambda(constraint)$, e.g. $f - \lambda(c_0 + c_1 - c^*)$ and derive wrt to n_h. If second deriv is positive, then min.
Two-phase, sample of size n from N and each unit has category. N_g be # units in pop with category g and n_g be # units in samp with cat g. 2nd phase, sample m units from each category.

Inclusion prob, $\pi^* = \frac{n}{N}\times\sum_{g=1}^{G} x_{ig}\min\left(\frac{m}{n_g}, 1\right)$, where $x_{ig} = 1$ if i in cat g.

Write $n_g = \sum_{i\in S_1} x_{ig}$ for S_1 being units sampled in phase 1. $E(n_g) = \frac{N_g}{N}n$. Simplify π_i^*, gets rid of max for large n and converges to inclusion for stratified with m from each strata.
When deriving var, try to get var in terms of an expr we already know var of.
Cannot estimate var between if only one cluster in samp, so cannot estimate var.
If one cluster has higher var than all the others in two stage, we can sample more from that cluster to decrease variance.
Multinomial: $p(x_1,\dots,x_k) = \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}, E(X_i) = np_i, V(X_i) = np_i(1-p_i), Cov(X_i, X_j) = -np_ip_j(i\neq j)$. SRSWR is multinomial.