1. SRS: $\hat{\mu} = \bar{y}, \quad \hat{\tau} = N\bar{y}, \quad \hat{p} = \bar{y}$

$$\widehat{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}, \quad s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2, \quad \widehat{V}(\hat{\tau}) = N^2\widehat{V}(\hat{\mu}), \quad \widehat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$$

2. StRS: $\hat{\mu}_{st} = \frac{1}{N}\sum_{\ell=1}^{L} N_\ell \bar{y}_\ell, \quad \hat{\tau}_{st} = N\hat{\mu}_{st}, \quad \hat{p}_{st} = \frac{1}{N}\sum_{\ell=1}^{L} N_\ell \hat{p}_\ell$

$$\widehat{V}(\hat{\mu}_{st}) = \frac{1}{N^2}\sum_{\ell=1}^{L} N_\ell^2 \left(1 - \frac{n_\ell}{N_\ell}\right)\frac{s_\ell^2}{n_\ell}, \quad \widehat{V}(\hat{\tau}_{st}) = N^2\widehat{V}(\hat{\mu}_{st}), \quad \widehat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{\ell=1}^{L} N_\ell^2 \left(1 - \frac{n_\ell}{N_\ell}\right)\frac{\hat{p}_\ell(1-\hat{p}_\ell)}{n_\ell - 1}$$

3. Auxiliary variable x_i:

(i) $\hat{\mu}_{rat} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\mu_x = r\mu_x, \quad \hat{\tau}_{rat} = N\hat{\mu}_{rat}$

$$\widehat{V}(\hat{\mu}_{rat}) = \mu_x^2\widehat{V}(r), \quad \widehat{V}(r) = \frac{1}{\mu_x^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}, \quad s_r^2 = \sum_{i=1}^{n}(y_i - rx_i)^2/(n-1), \quad \widehat{V}(\hat{\tau}_{rat}) = N^2\mu_x^2\widehat{V}(r)$$

(ii) $\hat{\mu}_{reg} = \bar{y} + b(\mu_x - \bar{x}), \quad b = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) / \sum_{i=1}^{n}(x_i - \bar{x})^2,$

$$\widehat{V}(\hat{\mu}_{reg})\text{[1]} = \left(1 - \frac{n}{N}\right)\frac{MSE}{n}, \quad MSE = \sum_{i=1}^{n}(y_i - \bar{y} - b(x_i - \bar{x}))^2/(n-2)$$

4. One-stage cluster sampling:

(i)ratio: $\hat{\mu}_r = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} m_i, \quad y_i$ cluster total , $\quad m_i$ cluster size, $\quad \hat{\tau}_r = M\hat{\mu}_r$

$$\widehat{V}(\hat{\mu}_r) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}, \quad s_r^2 = \frac{\sum_{i=1}^{n}(y_i - m_i\hat{\mu}_r)^2}{n-1}, \quad \widehat{V}(\hat{\tau}_r) = M^2\widehat{V}(\hat{\mu}_r), \quad M = \sum_{i=1}^{N} M_i, \bar{M} = \frac{M}{N}$$

– if necessary, can use $\bar{m} = \sum_{i=1}^{n} m_i/n$ in place of $\bar{M}$

(ii)SRS/unbiased: $\hat{\tau} = N\bar{y}_t = N\dfrac{\sum_{i=1}^{n} y_i}{n}, \quad \hat{\mu} = \dfrac{\hat{\tau}}{M}$ $\quad M$ is the number of units in the population

$$\widehat{V}(\hat{\tau}) = N^2\widehat{V}(\bar{y}_t) = N^2\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n}, \quad s_t^2 = \sum_{i=1}^{n}(y_i - \bar{y}_t)^2/(n-1)$$

5. Two-stage cluster sampling:

(i)SRS/unbiased:

$$\hat{\mu}_{unb} = \frac{N}{M}\left(\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{n}\right), \quad \bar{M} = \sum_{i=1}^{N} M_i/N, \quad \hat{\tau}_{unb} = M\hat{\mu}_{unb}$$

$\widehat{V}(\hat{\mu}_{unb}) = T1 + T2, \quad T1$ involves between cluster variance, $\quad T2$ involves within cluster variance

[1] In regression the prediction variance is $\sigma^2\{(1/n) + (x^0 - \bar{x})^2/s_{xx}\}$;, here the 2nd term has been conveniently dropped

(ii)ratio:

$$\hat{\mu}_r = \frac{\sum_{i=1}^n M_i \bar{y}_{i.}}{\sum_{i=1}^n M_i}, \quad \bar{y}_{i.} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}, \quad m_i \text{ sample size in cluster}$$

$$\widehat{V}(\hat{\mu}_r) = T1 + T2, \quad T1 \text{ involves between cluster variance}, \quad T2 \text{ involves within cluster variance}$$

6. Interesting facts: A systematic sample is like one-stage cluster sampling with only one cluster – once an item has been chosen at random, the sample is determined. This means that we can't assess the variance properly (using between cluster variation), so we have to fake it with the SRS variance.

Cluster sampling can be compared to SRS sampling by looking at the analysis of variance table. If MSB is a large component of the overall variance, then cluster sampling is less efficient. If it is small, then cluster sampling is more efficient. Large and small are ascertained by comparing MSB to an estimate of σ^2, which itself is motivated by the *population* anova table. The algebra is easier if all stratum sizes are the same, in which case both estimates of μ are the same. The unbiased estimates, using principles of SRS, in cluster sampling are usually presented in terms of estimating the total τ, whereas the ratio estimates are usually presented in terms of estimating the mean μ. This sort of makes sense, as we rarely know M.

Ratio estimation is the same as least squares estimation in the model $y_i = \beta x_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2 x_i)$, whereas regression estimation is the same as least squares estimation in the mode $y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$. Ratio estimates are biased, but the bias is small compared to the variance, so the efficiency of ratio estimation compared to SRS (unbiased) estimation can be assess by comparing the variances of the two estimates. Usually we compare the estimates of the variances $\widehat{V}$ instead.

We can determine optimal choices for stratum sizes by comparing the variances within strata, and trading this off against the cost. A similar analysis can be done for cluster sampling, but is not covered in this course or in our text. See Lohr, §5.5. Section 4.6 on comparing two estimates has nothing to do with finite population sampling, which is why we skipped it. Post-stratification can be used to reduce variance, by dividing the sample into strata after the fact and using the usual formula. This can be a form of cheating, though, like looking for the largest difference among several groups and then pretending that this was the difference you were interested in all along. Post-stratification strategies planned ahead of the survey, typically to handle non-response, are considered legitimate.