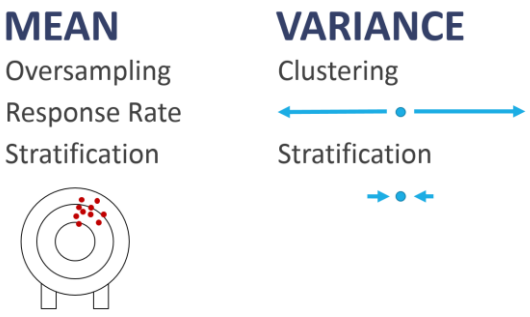
Methods: Complex sampling

Topics: Stratification, multistage clustering, oversampling, impact on mean and variance

1. Introduction to complex sampling

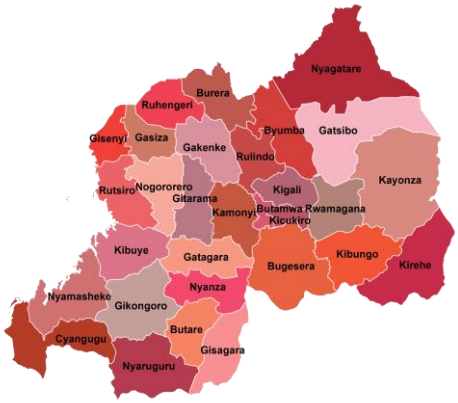
Most national household surveys have three complex sampling design characteristics:

1. Stratification – choosing an independent sample in subnational geographic areas
2. Multistage cluster sampling – First sampling communities, then sampling households, and sometimes sampling individuals within households
3. Oversampling – ensuring there are a sufficient number of rural and urban households to estimate national indicators.



Together, these design characteristics bias the mean estimates and introduce variance, which is why we have to account for the complex sample design when performing survey data analysis.

This video will review each of these sample design characteristics, and their impacts on the mean and variance estimators. Learn how to account for complex sample design characteristics in the video called “Survey Analysis in Stata”.

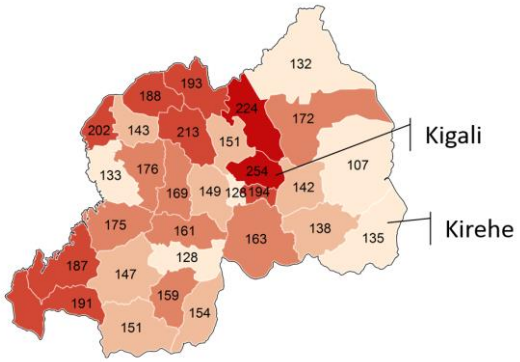


2. Stratification

Stratification means that the survey implementers divide the country into mutually exclusive, exhaustive subnational regions, and then select independent samples within each subnational region. The purpose of stratification is to produce representative statistics for subnational regions, and to be able to compare indicators across subnational regions.

3. Stratification: impact on mean

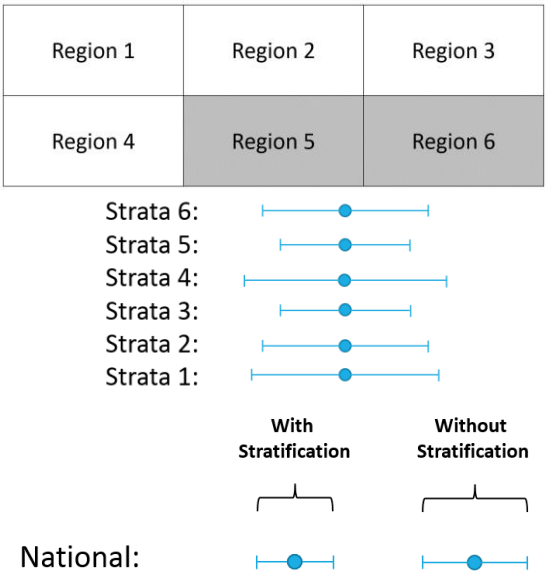
Stratification can bias national mean estimates when populations are different in each subnational region, and these subnational regions have different population sizes. Why? It has to do with the fact that approximately the same number of households are sampled from each stratum, but the stratum have different population sizes.



$$\frac{\text{Pop in stratum } j}{\text{Total population}} \div \frac{\text{Sample stratum } j}{\text{Total sample}}$$

Let us think about the impact of stratification through an example. Let us say we are going to select a stratified sample in Rwanda, and we stratify by Rwanda's 30 districts. Here are the total 2010 population figures (in 1000s) where darker colors represent larger populations. Kigali district has nearly twice the population as Kirehe. So if we sample the same number of people from each district, people from Kirehe would be over-represented in the national sample, and people from Kigali would be under-represented. We account for stratification by weighting the observations in a given district by the ratio of proportion of district population represented in the sample versus the underlying population.

4. Stratification: impact on variance

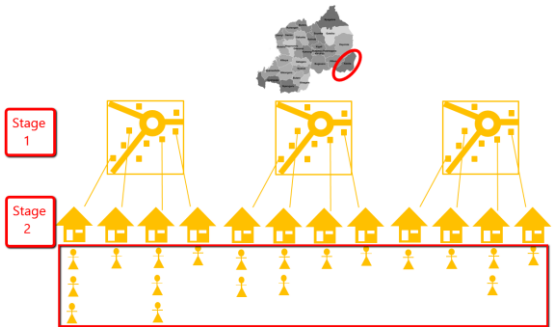


For national estimates, stratification can slightly improve the precision of estimates if the outcome follows a geographic pattern – for example, if we are estimating the percent of women in polygamous marriage, and polygamy is practiced mainly within one region. This is because there is less variation within a given region compared to the country overall. For example, say a country has six strata, and each strata has less variation within, than the overall sample, then we could generate an estimate at the national level with more certainty than if we had selected a simple random sample nationwide.

5. Multistage cluster sampling

Multistage cluster sampling means that we sample increasingly smaller, embedded units. The classic example is of students, embedded in (or "clustered in") classrooms, and classrooms which are clustered in schools. A multistage cluster sample of students might first list all of the schools in the study area, and randomly sample a few schools. Then within the sampled schools, list all classrooms and randomly sample classrooms. If we stopped here, and conducted a survey of all students in the sampled classrooms, then we would have performed a two-stage cluster sample with schools as stage one, and classrooms as stage two.

You might be wondering why this is not a three-stage cluster design since students are clustered in classrooms, and we ultimately survey students. The reason is, we do not sample the students; we take a census of all students in the sampled classrooms. Therefore, the sampling stops after two-stages.



A two-stage cluster sampling design is used to sample women in the Demographic and Health Surveys (DHSs), Multiple Indicator Cluster Surveys, and other similar household surveys. Within strata, all enumeration areas from the last census are listed and then sampled. In sampling terminology, we call the selected enumeration areas "clusters", or "primary sampling units (PSUs)" since they are units of the first-stage sample. Within PSUs, all households are listed and then sampled. We call these secondary sampling units (SSUs) since they are the units of the second-stage sample. Then a census of all eligible women is performed within the sampled households. Similar to our classroom example, we describe this design as a two-stage cluster sample with enumeration areas sampled in stage one, and households sampled in stage two. If we randomly sampled one woman per household rather than interview all eligible women, this would become a three-stage cluster sample design.

Here is another question you might ask: Most household surveys are stratified by district or province, resulting in non-overlapping geographic areas. Why doesn't the DHS consider strata as a level in the multistage cluster sampling frame? Think about how many strata get sampled. All of them, right? Since we have a census of strata, strata are not considered a stage in the multistage cluster design.

```
svyset [pweight=weight], psu(v021) strata(sdistr)
svy: tab v106
```

highest education al level	proportions
no educa	.155
primary	.6829
secondar	.1469
higher	.0152
Total	1

```
svyset [pweight=weight], strata(sdistr)
svy: tab v106
```

highest education al level	proportions
no educa	.155
primary	.6829
secondar	.1469
higher	.0152
Total	1

6. Multistage cluster sampling: impact on mean

How are means and proportions effected by multistage cluster sampling? They are not effected at all, actually. Clustering does not affect means, only variance. Let me demonstrate this. Let us say that we are estimating the highest level of education among women in the Rwanda 2010 DHS sample. Whether we correctly specify the two-stage cluster design with a svyset statement, or run

statistics without specifying the two-stage cluster design, the mean proportion of women at different levels of education remains the same.

7. Multistage cluster sampling: impact on variance

Cluster-sampling only effects variance estimates. This is because people from the same community are often similar in some way. For example, my neighbor is more likely to have a similar social economic status and occupation to me than a stranger living elsewhere in the country. Likewise, patients who attend the same clinic are often similar in terms of social economic status and health risks. And think about how similar children from the same classroom are in terms of age and social-economic status. When we sample multiple people from the same PSU, we are getting less information about the population, than if we had sampled people randomly from across the population. Does that make sense?

Estimating variance: multistage cluster sampling

p = prevalence in the total population
Y_{ij} = indicator that observation j in cluster i has trait of interest
w_{ij} = sample probability weight of observation j in cluster i
m_i = sample size in cluster i
M_i = total population in cluster i
N = number of clusters
n = number of clusters sampled
N_{pop} = total population size

Simple Random Sampling:
Variance: $\widehat{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$

Variance: $\widehat{var}(\hat{p}) = \frac{1}{\hat{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{\hat{p}_i(1-\hat{p}_i)}{m_i-1}\right)\right]$

Where: $s_r^2 = \frac{\sum_{i=1}^{n}(M_i\hat{p}_i - M_i\hat{p})^2}{n-1}$

```
svyset [pweight=weight], psu(v021) strata(sdistr)

svy: tab v106, ci
```

highest education al level	proportions	lb	ub
no educa	.155	.1463	.164
primary	.6829	.6713	.6943
secondar	.1469	.1368	.1577
higher	.0152	.0119	.0193
Total	1		

```
svyset [pweight=weight], strata(sdistr)

svy: tab v106, ci
```

TYPE 1 ERROR

highest education al level	proportions	lb	ub
no educa	.155	.1489	.1613
primary	.6829	.675	.6908
secondar	.1469	.1411	.1529
higher	.0152	.0133	.0173
Total	1		

Remember from the Simple Random Sampling lecture, variance of a proportion is calculated as $\frac{\hat{p}(1-\hat{p})}{n-1}$. The variance estimates from a cluster sample are comprised of two pieces: variance within a cluster, and variance across clusters. As you can see, the calculation of the variance estimator for clustered samples is complex. Conceptually, this variance estimator is discounting information from each additional person that is sampled within a PSU because people in the same PSU are expected to be similar, and therefore contribute less new information to the sample, than if we had sampled a random person from the population. If you do not account for clustering in a multistage cluster sample, your confidence intervals will be incorrectly too narrow and you will risk finding differences among subgroups that are not real. We call this a type I error. You must absolutely account for clustering when performing descriptive statistics with survey data. There is some debate about whether it is necessary to account for clustering in multivariate analysis, which I will discuss when we cover logistic regression.

$$DE = \frac{var(survey\ design)}{var(SRS)}$$

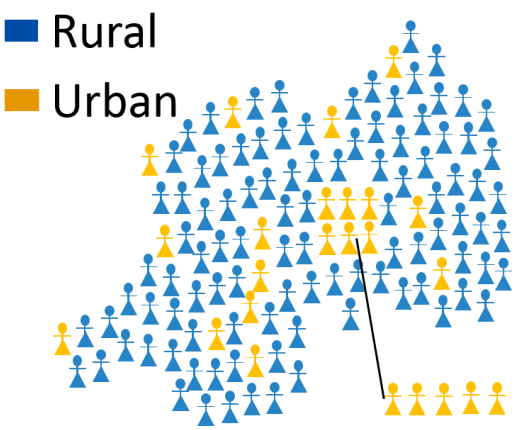
$DE > 1$: survey design has more variability (less precise)

$DE < 1$: survey design has less variability (more precise)

8. Design Effect

The design effect describes the effect of sampling on the variance estimate. The design effect compares the variance under one sample design with the variance of a simple random sample. To calculate design effect, simply divide the variance calculated with the survey design by variance assuming simple random sampling.

A design effect larger than 1 means that the variance under the sample design is larger than it would be under simple random sampling, and a design effect smaller than 1 means that the variance under the sample design is smaller than it would be under simple random sampling. The design effect of a complex survey design is usually reported by survey implementers for key indicator variables, and is helpful for sample size calculations in future surveys that use similar complex survey designs in that population.



9. Oversampling

In large household surveys, samples are often designed to ensure that a minimum number of respondents are selected from important subgroups so that robust estimates and comparisons can be made across subgroups. Urban and rural populations are usually considered the most important subgroups. In low-income countries where the overwhelming majority of the population is rural, oversampling is commonly performed in urban populations by sampling extra PSUs in cities.

10. Oversampling: impact on mean

Oversampling changes the probability of selection. If urban areas are oversampled, then individuals living in urban areas have a higher probability of selection than individuals living in rural areas. Oversampling biases the mean estimates. For example, access to formal education is usually higher in urban areas. If we oversample urban people, then the proportion of urban people in the sample will be greater than the proportion of urban people in the population, and our estimates will be biased; they will not represent the population. Going back to Stata, if we estimate highest level of education among women

```
svyset [pweight=weight], psu(v021) strata(sdistr)

svy: tab v106
```

highest education al level	proportions
no educa	.155
primary	.6829
secondar	.1469
higher	.0152
Total	1

```
svyset , psu(v021) strata(sdistr)

svy: tab v106
```

highest education al level	proportions
no educa	.1508
primary	.6786
secondar	.1529
higher	.0178
Total	1

in the Rwanda 2010 DHS by correctly adjusting for sampling weights, and then we estimate the same variable without sampling weights, the unweighted estimates incorrectly show women to have higher access to education than actually exists.

A key principle of sampling is that everyone in the study population has an equal probability of selection. When there is unequal probability of selection, we must (a) know the probability of selection among each respondent, and (b) incorporate these probabilities into our analysis using sampling probability weights. As with other aspects of complex sample design, probability weights *must* be used in descriptive analysis of means and proportions, and there is debate about whether oversampling must be accounted for in multivariate analysis.

11. Oversampling: impact on variance

Oversampling in household surveys usually entails the addition of a few extra PSUs. If these PSUs are sampled randomly – in other words, everyone within the subgroup has an equal probability of selection – then oversampling does not affect variance estimates. The only way that variance could be effected is if the overall sample size is increased by adding PSUs, in which case the effect on variance is the result of increased sample size; not of oversampling.

12. Response rates

Response rates can bias mean estimates if certain kinds of people systematically refuse, or are not available, to participate in the household surveys. If, for example, interviewing is performed at a time of day when women farmers tend to be away from the house working in the fields, and interviewers are more likely to encounter sick women than healthy women, then the sample is not representative of the population. Although we do not know the characteristics of the women who were not interviewed, we can calculate the probability of selection among those women who were interviewed, and adjust the contribution of their information to our mean estimate.

Weighted estimator for unequal probability of selection

p = prevalence in the total population

Y_j = indicator that observation j has trait of interest

w_j = sampling weight

n = sample size

N = population size

Estimator: $\hat{p} = \frac{\sum w_j Y_j}{N} = \frac{\sum \left(\frac{N}{n}\right) Y_j}{N} = \frac{\left(\frac{N}{n}\right) \sum Y_j}{N} = \frac{1}{n} \sum Y_j = \frac{\sum Y_j}{n}$

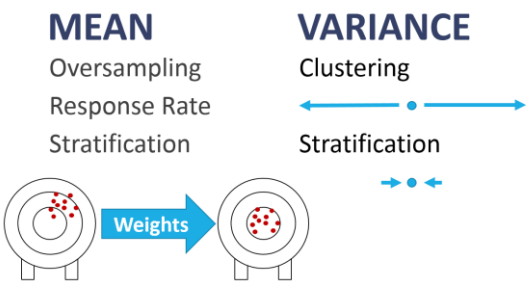
13. Sampling probability weights

Sampling probability weights adjust for unequal probability of selection introduced to the sample by stratification, oversampling, and unequal response rates. If you are analyzing a national household survey like the DHS, then sampling probability weights will have been calculated for you and provided in the datasets.

The sampling weight is equal to 1 over the probability of being sampled, or $\frac{1}{n/N}$. You will remember from math class that this can be re-written as N/n. The sampling weight is applied to each observation when estimating prevalence.

Notice that the denominator of the weighted estimator is N, the total population. This is different from the unweighted estimator presented in the Simple Random Sampling video where sample size, n, was in the denominator. Another way to think about the unweighted estimator is that it has a constant weight – an equal probability of selection for each person in the population. The estimators are equivalent, and this is how we get from the weighted estimator equation, to the unweighted – or constant weight – estimator equation: the weight, $\frac{N}{n}$, can be pulled out of the sum operator because the weights are equal to 1 for all individuals. Then the Ns cancel out, and we are left with $\frac{\sum Y_j}{n}$.

14. Summary



Several characteristics of complex survey design can bias mean and variance estimates. Any survey design characteristic which effects the probability of selection - including stratification, oversampling, and response rates – must be accounted for with the application of sampling probability weights in descriptive data analysis. Descriptive data analysis must also adjust for clustering by widen the variance estimates to avoid making type I errors. Accounting for stratification can slightly narrow confidence intervals in analyses of multiple strata, but the effect is usually negligible, and so the effect of stratification on variance can be ignored in descriptive data analysis.

Acknowledgement: This presentation was based off of learning material by Dr. Bethany Hedt-Gauthier.