IHSD 7440 - Homework #3 2024

Name__Paige Zeltzer___________________________

Background

The Roll Back Malaria (RBM) initiative began in 1998 by WHO as an international effort to halve the 2000 levels of malaria morbidity and mortality by 2010 and to reduce this malaria burden by a further 50 percent by 2015. One of RBM's core indicators is the proportion of households with at least one insecticide treated net (ITN). ITNs are a key tool in reduction of malaria transmission and subsequent reduction in child and adult morbidity and mortality. An ITN is defined as a mosquito net treated with a long-lasting insecticide or a mosquito net that has been dipped in insecticide within the past 12 months. Efforts to scale up ITN coverage are underway in most African countries. Nationally representative population-based surveys such as the DHS are the data collection methods preferred to measure RBM indicators including proportion of households with at least one ITN. More information on RBM can be found at http://www.rbm.who.int.

Assignment

In this assignment, we are interested in the following indicators at the household and child level:

Among households:

1) Proportion of households with at least 1 ITN

Among children:

1) Proportion of children under the age of 5 that used an ITN the previous night

We will look at overall estimates for these two indicators as well as by the following factors:

1. Residence: urban/rural (all indicators)
2. Household socioeconomic status: wealth quintile (all indicators)
3. Household head education (for the household ITN possession indicators)
4. Child's age (for the ITN use analysis among children)
5. Mothers education (for the ITN use analysis among children)

We will be using subsets of the household and child-level dataset for Zambia 2007. Both are available for download on Canvas under Assignments, Homework #3, as well as in the IHSD 7440 HH Sampling GitHub repository.

These data were collected using the standard DHS sampling protocol, which consists of a 2-stage cluster design with first stage selection of primary sampling units selected proportional to their size (PPS). All women of reproductive age were asked for information of their children. Independent samples were selected within survey domains at the regional level. Within each survey domain, data were collected using a proportional stratification system to improve the precision of the estimates. For all household and child level data, sample weights were created based relative strata sizes, and on the difference between estimated cluster size (M) and actual cluster size (B).

Please note variable **HV005 is the sampling weight**, **variable HV021, labeled as Primary Sampling Unit, provides the cluster number for the analysis**, and **variable HV022, labeled Sample stratum number, provides the strata number for the analysis**.

Problem 1: HH-level analysis

Using the ***2009_Zambia_HH_2024.csv*** dataset located in the IHSD 7440 GitHub repository, please answer the following questions. You can use either R Studio or STATA to complete this exercise, but R Studio is recommended.

a) What is your element in this analysis and how many are there (n)?
The element in this analysis is households. There are 6439 households in this dataset.

b) How many clusters are there in this sample?
There are 285 clusters in this sample.

c) How many households were selected in each cluster / PSU?
Between 6 and 48 households were selected in each cluster/PSU.

d) How many survey domains are there in this dataset?

There are 16 survey domains in this dataset.

e) How many strata are there in this dataset?
There are 16 strata in this dataset.

Now complete the following tables. The tables show the proportion of households that own at least 1 ITN. For column d, you need to analyze the data appropriately, taking into account the following: 1) the use of a 2-stage cluster design that results in correlated data at the cluster / PSU level; 2) adjustment for differences in the ultimate probability of selection through sampling weights; and 3) uses the strata information to improve the precision of your estimates.

In each cell, include the proportion and the standard error (round to 3 decimal points)

Table 1: Proportion of households that own at least 1 ITN

n = 6439	(a) Assuming SRS	(b) Assuming SRS, with weights	(c) 2-stage cluster sampling with weights	(d) 2-stage cluster sampling with weights and stratification
Residence				
Urban	Proportion=0.540 Se=0.010	Proportion=0.520 Se=0.012	Proportion=0.520 Se=0.018	Proportion=0.520 Se=0.018
Rural	Proportion=0.572 Se=0.008	Proportion=0.556 Se=.008	Proportion=0.556 Se=0.019	Proportion=0.556 Se=0.018
SES				
Poorest	Proportion=0.506 Se=0.015	Proportion=0.481 Se=0.016	Proportion=0.481 Se=0.028	Proportion=0.481 Se=0.027
Poorer	Proportion=0.560 Se=0.013	Proportion=0.549 Se=.014	Proportion=0.549 Se=.025	Proportion=0.549 Se=.023
Middle	Proportion=0.582 Se=.013	**Proportion=.585** Se=.014	Proportion=.585 Se=.022	Proportion=.585 Se=.020
Richer	Proportion=.529 Se=.013	Proportion=.510 Se=.015	Proportion=.510 Se=.023	Proportion=.510 Se=.022
Richest	Proportion=.624 Se=.015	Proportion=.589 Se=.017	Proportion=.589 Se=.021	Proportion=.589 Se=.020
HH head education				
None	Proportion=.421 Se=.016	Proportion=.405 Se=.017	Proportion=.405 Se=.026	Proportion=.405 Se=.025
Primary	Proportion=.546 Se=.009	Proportion=.528 Se=.010	Proportion=.528 Se=.018	Proportion=.528 Se=.017
Secondary	Proportion=.612 se=.011	Proportion=.599 Se=.013	Proportion=.599 Se=.018	Proportion=.599 Se=.016
Higher	Proportion=.704 Se=.020	Proportion=.681 **Se=.023**	Proportion=.681 Se=.022	Proportion=.681 Se=.022
All Households	**Proportion=.560 Se=.006**	**Proportion=.543 Se=.007**	**Proportion=.543 Se=.014**	**Proportion=.543 Se=.013**

f) What is the effect of sample weights on point estimates and standard errors?

Sample weights generally lower the point estimates (except for middle SES households) and slightly increase standard errors by .001-.003. Also, once weighting was applied, the point estimates remained the same regardless of how the data was analyzed.

g) What is the effect of the cluster sampling design (i.e. use of clusters at first stage) on the standard errors (i.e. when using the Huber-White Sandwich estimator in SAS or STATA- e.g. using a cluster command)?

The cluster sampling design generally increased the standard error for each parameter. While adding weighting to an SRS analysis caused a slight increase in the standard error (by 0.001-0.003), accounting for 2-stage cluster sampling resulted in a larger increase. For most parameters, analyzing the data with 2-stage cluster sampling with weights resulted in the highest standard errors. Accounting for stratification generally lowered the standard error slightly compared to estimate c (by 0.001-0.002), but it typically remained higher than when analyzing the data assuming SRS, with or without weighting.

However, for households where the head of household completed higher education, the standard errors were as follows: 0.020 assuming SRS without weighting, 0.023 assuming SRS with weighting, and 0.022 assuming 2-stage cluster sampling with weighting (both with and without stratification), a slight decrease of .001.

h) What is the design effect for this 2-stage cluster sampling design for the proportion of households with at least 1 ITN, with sample weights and stratification included in the analysis?

DEFT=SE(estimate d)/SE(SRS)
=SE(estimate d)/SE(estimate a)
=.013/.006
=2.166667 ->2.167

i) How does household residence – urban versus rural - affect the proportion of households with at least 1 ITN?

Rural households have a higher proportion of households with at least 1 ITN. The proportion of rural households with at least 1 ITN is .572 unweighted and .540 weighted. The proportion of urban households with at least 1 ITN is .556 unweighted and .520 weighted.

j) Which of the four estimates (a, b, c, d in table 1 above) provides the least biased point estimates and standard errors of the ITN household possession estimates, and why?

Estimate d provides the least biased point estimate and standard errors because the data is analyzed correctly. The survey design follows a 2-stage cluster sampling approach, with the first stage involving the selection of primary sampling units proportional to their size (PPS) and strata were used. Estimate d is represents the point estimate and standard error using 2-stage cluster sampling PPS accounting for weights and stratification.

Problem 2: Individual-level analysis

Using the ***2009_Zambia_child_2023.csv*** dataset located in the IHSD 7440 GitHub Repository, please answer the following questions.

a) What is your element in this analysis and how many are there (n)?
The element in this analysis is the individual child. There are 5194 children in this analysis.

b) How many clusters are there in this sample?
There are 285 clusters in this sample.

4

Now complete the following tables. The tables show the proportion of children that slept under an ITN the previous night. For column d, you need to analyze the data appropriately, taking into account the following: 1) the use of a 2-stage cluster design that results in correlated data at the cluster / PSU level; 2) adjustment for differences in the ultimate probability of selection through sampling weights; and 3) uses the strata information to improve the precision of your estimates. Please note variable **V005 is the sampling weight**, **variable V021, labeled as Primary Sampling Unit, provides the cluster number** for the analysis, and **variable V022, labeled Sample stratum number**, provides the strata number for the analysis.

In each cell, include the proportion and the standard error (round to 3 decimal points)

Table 2: Proportion of children that slept under an ITN the previous night, among all households

n = 5194	(a) Assuming SRS	(b) Assuming SRS, with weights	(c) 2-stage cluster sampling with weights	(d) 2-stage cluster sampling with weights and stratification
Child age				

0	Proportion=.375
	Se=.014
		Proportion=.360
		Se=.015
			Proportion=.360
			Se=.021
				Proportion=.360
				Se=.020

1	Proportion=.349
	Se=.014
		Proportion=.334
		Se=.015
			Proportion=.334
			Se=.020
				Proportion=.334
				Se=.019

2	Proportion=.285
	Se=.014
		Proportion=.270
		Se=.015
			Proportion=.270
			Se=.018
				Proportion=.270
				Se=.017

3	Proportion=.244
	Se=.014
		Proportion=.231
		Se=.015
			Proportion=.231
			Se=.018
				Proportion=.231
				Se=.017

4	Proportion=.207
	Se=.013
		Proportion=.192
		Se=.013
			Proportion=.192
			Se=.015
				Proportion=.192
				Se=.014

Residence				

Urban	Proportion=.288
	Se=.011
		Proportion=.264
		Se=.013
			Proportion=.264
			Se=.020
				Proportion=.264
				Se=.019

Rural	Proportion=.301	Proportion=.290	Proportion=.290	Proportion=.290

	Se=.008	Se=.008	Se=.017	Se=.015
SES				

Poorest	Proportion=.216
	Se=.012
		Proportion=.199
		Se=.012
			Proportion=.199
			Se=.019
				Proportion=.199
				Se=.018

Poorer	Proportion=.328
	Se=.014
		Proportion=.325
		Se=.014
			Proportion=.325
			Se=.023
				Proportion=.325
				Se=.022

Middle	Proportion=.325
	Se=.014
		Proportion=.334
		Se=.015
			Proportion=.334
			Se=.023
				Proportion=.334
				Se=.022

Richer	Proportion=.301
	Se=.014
		Proportion=.287
		Se=.016
			Proportion=.287
			Se=.024
				Proportion=.287
				Se=.024

Richest	Proportion=.317
	Se=.018
		Proportion=.270
		Se=.019
			Proportion=.270
			Se=.028
				Proportion=.270
				Se=.027

Mother's education				

None	Proportion=.230
	Se=.016
		Proportion=.228
		Se=.017
			Proportion=.228
			Se=.025
				Proportion=.228
				Se=.025

Primary	Proportion=.29
	Se=.008
		Proportion=.281
		Se=.008
			Proportion=.281
			Se=.016
				Proportion=.281
				Se=.014

Secondary	Proportion=.33
	Se=.014
		Proportion=.308
		Se=.015
			Proportion=.308
			Se=.021
				Proportion=.308
				Se=.020

Higher	Proportion=.436Se=.048	Proportion=.414
		Se=.054
			Proportion=.414
			Se=.057
				Proportion=.414
				Se=.058

All Households	Proportion=.296Se=.006	Proportion=.282
		Se=.007
			Proportion=.282
			Se=.013
				Proportion=.282
				Se=.012

c) How did the inclusion of the sampling strata in the analysis affect the precision of the estimates?

Compared to assuming SRS, analyzing the data assuming 2-stage cluster sampling with weights and stratification increased the standard errors and decreased the precision of the estimates.

In general, compared to analyzing the data assuming 2-stage cluster sampling with weighting (without accounting for stratification), including the sampling strata in the analysis slightly decreased the standard errors(by 0.001-0.002), thus increasing the precision of the estimates. However, for children whose mothers had higher education, including the sampling strata increased the standard error by 0.001 and decreased the precision compared to 2-stage cluster sampling without stratification.

d) What is the **design effect** for this 2-stage cluster sampling design for the proportion of children that slept under an ITN, with sample weights and stratification included in the analysis?

DEFT=SE(estimate d)/SE(SRS)
=SE(estimate d)/SE(estimate a)
=.012/.006
=2

e) How does child age affect the use of ITNs among children?

The proportion of children age 0 who slept under an ITN is .375 unweighted and .360 weighted
The proportion of children age 1 who slept under an ITN is .349 unweighted and .334 weighted
The proportion of children age 2 who slept under an ITN is .285 unweighted and .270 weighted
The proportion of children age 3 who slept under an ITN is .244 unweighted and .231 weighted
The proportion of children age 4 who slept under an ITN is .207 unweighted and .192 weighted

As the child's age increases, the proportion of children that slept under an ITN decreases. Therefore, younger children may be more likely to sleep under an ITN than older children.

f) How would you explain the level of children that slept under an ITN to the Ministry of Health- is it high or low, and how would you interpret the point estimate taking total survey error into account?

The proportion of children under 5 who slept under an ITN in the past night is .296 unweighted and .282 weighted. Both values suggest low ITN use among children under 5. In comparison, data from the bednet survey shows the proportion of children under 5 who slept under an ITN as 0.7655 unweighted and 0.7605 weighted, while the proportion of households that own at least one ITN is 0.8287 unweighted and 0.8217 weighted. There is not a large discrepancy between possession and usage in the bednet survey. In our data, the proportion of households that own at least 1 ITN is .560 unweighted and .543 weighted, indicating a higher discrepancy between ITN possession and usage in our data.

According to the WHO and the Malaria Journal, the minimum target for universal coverage is 80% for the possession and use of ITN in malaria-endemic areas (Koenker et al., 2018). The usage proportion of .282-.296 and the possession proportion of .543-.560 in our data suggest large coverage gaps. ITN use is low among older children, children in the poorest SES group, and children whose mothers had no education, so it may be a good idea to conduct targeted interventions to improve coverage among these groups.

The total survey error is the combination of sampling error and non-sampling error. Sampling error is measured by the standard error of the estimate and the 95% confidence intervals. The point estimate for the proportion of all children under 5 that slept under an ITN, .296 unweighted and .282 weighted, is accompanied by a standard error (ranging from .006 to .013 depending on how the data was analyzed). This suggests some sampling variation. The true population proportion of children under 5 who slept under an ITN could be within a broader range. For example, assuming for SRS unweighted, the 95% confidence interval for the proportion of children under 5 who slept under an ITN is .296+/-(1.96)(.006)= (.284, .308). We are 95% confident that the true proportion of children under 5 who slept under an ITN lies somewhere between .284 and .308. If we take multiple samples of children under 5 in Zambia, 95% of the samples will have an estimated proportion of children under 5 who slept under an ITN in the range of .284 to .308.

The other component of total survey error is non-sampling error. There are several types of non sampling error that can lead to the underreporting or overreporting of data. For example, one potential source is information error, where respondents may misunderstand the question and answer anyway, leading to inaccurate responses. Many respondents may not know what an ITN is and will overreport possession and usage. Information bias, specifically social desirability bias, may also lead to over reporting ITN use because respondents may feel socially obligated to respond positively to questions about recent ITN use. Selection bias, specifically coverage bias, arises from failing to include eligible respondents in the sampling frame, failing to capture the variability in the population. Finally, non sampling error can come from non-response bias when certain households or respondents may be unavailable during data collection or refuse to respond.

https://malariajournal.biomedcentral.com/articles/10.1186/s12936-018-2505-0#:~:text=In%20line%20with%20the%20definition,needed%20%5B3%2C%205%5D