The Demographic and Health Surveys (DHS) are large multi-stage surveys that are carried out globally. For this exercise, we are going to consider the DHS survey conducted in India in 2005-6. This survey contains information on 515,507 individuals from 109,041 sample households. Respondents provide information on a large number of demographic and health related areas (thus the name of the survey!).

You have access to a subset of this data for this exercise. (See data dictionary at the end of the assignment) We will focus on Tuberculosis (TB) and consider TB prevalence and correlates with TB in India. TB is currently the leading cause of death due to infectious disease globally and India is one of the countries with a particularly high TB burden. TB is primarily a respiratory infection, making smoking and cooking fuel, as well as ventilation of interest as potentially increasing the chance of disease in individuals.

Question 1

Read the sampling design section of Volume II of the Country report associated with this study (page 4 of Appendix C or page 599 of the pdf report).

PSU = primary sampling unit
SSU = secondary sampling units
TSU = tertiary sampling units
PPS = proportional to population size
CEB = census enumeration block

Rural Area Sampling – Stratified Sampling
1) creating strata of villages with probability PPS
2) systematic sampling in each strata

Urban Area Sampling – Cluster Sampling
1) PSU: PPS sampling of wards
2) SSU: one CEB per ward
3) TSU: households in each CEB

Urban Area 8 Cities – Cluster Sampling
1) PSU: slum and non-slum CEB
2) SSU: households in each CEB

Part A

Why were different sampling strategies employed for rural and urban areas? What is the difference between the sampling strategies?

Different sampling strategies were used for urban and rural areas because the 2001 census information for urban areas was not published at the time the DHS was conducted so listing all the households in each large urban area proved extremely difficult. Stratified sampling was used for rural areas, as each strata was systematic sampled. Cluster sampling was used for urban areas because at each stage, a simple random sampling was done on the primary and

secondary (and tertiary for the eight cities) units. Stratified and clustering sampling are different because in stratified sampling, every stratus is sampled but in clustering sampling, random sampling will determine which clusters are selected.

Part B

What was the sampling frame in rural areas (do not go in to detail of what was and was not included)?

The sampling frame of the rural areas was the people recorded in the 2001 Census list of Villages.

Part C

Why was stratification used in the rural sample? What was the first level of stratification done on?

Stratification was used in the rural sample to make sure villages with different socioeconomic statuses were represented in the sample. The first level of stratification was geographic.

Question 2 – Unweighted Analysis

Part A

What is your overall estimate of TB prevalence?

The overall estimate of TB prevalence is 2.20%.

Part B

Create a table for summary statistics of those with TB, without TB, and the overall sample.

Analysis not Accounting for Sampling Design

	Overall N=50,000	TB N=1,101	No TB N=48,899	P-value
Male (N, %)	25,253 (50.51%)	537 (48.77)	24,716 (50.55)	0.2451
Female (N, %)	24,747 (49.49%)	564 (51.23)	24,183 (49.45)	
Age (mean, 95% CI)	26.80 (26.63, 26.97)	26.03 (24.87, 27.17)	26.82 (26.65, 26.99)	0.1784
Age Category				
0-4	5,246 (10.49%)	116 (10.54%)	5,130 (10.49%)	0.3283
5-14	11,225 (22.45%)	272 (24.70%)	10,953 (22.40%)	
15-39	20,849 (41.70%)	442 (40.15%)	20,407 (25.37%)	
≥40	12,675 (25.4%)	271 (24.61%)	12,404 (25.37%)	
BMI (mean, 95% CI)	21.07 (20.99, 21.15)	20.23 (19.51, 20.94)	21.09 (21.01, 21.17)	0.0183
BMI category				
Malnourished	1,410 (12.40%)	49 (18.28%)	1,361 (12.25%)	0.0014
Underweight	1,916 (16.84%)	49 (18.28%)	1,867 (16.81%)	
Healthy	6,316 (55.53%)	147 (54.85%)	6,169 (55.54%)	
Overweight	1,317 (11.58%)	21 (7.84%)	1,296 (11.67%)	
Obese	416 (3.66%)	2 (0.75%)	414 (3.73%)	
Windows (N, %)	37,039 (74.12%)	681 (61.85%)	36,358 (74.40%)	<0.0001
Wood Fuel (N, %)	22690 (45.38%)	588 (53.41%)	22,102 (45.20%)	<0.0001

Part C

Create a table for the results of a multivariate logistic regression.

Regression not Accounting for Study Design

	OR	95% CI	P-value
Female vs Male	1.073	0.952, 1.209	0.2496
Age Category			
0-4	0.973	0.781, 1.213	0.8808
5-14	1.074	0.905, 1.274	0.8678
15-39	0.987	0.847, 1.150	0.4118
≥40	-	-	-
Windows	0.591	0.520, 0.672	<0.0001
Wood Fuel	1.230	1.087, 1.392	0.0011

Question 3 – Weighted Analysis

Part A

What is your overall estimate of TB prevalence?

The overall estimate of TB prevalence is 2.31%.

Part B & C

Create a table for summary statistics of those with TB, without TB, and the overall sample using the weighted frequencies.

Analysis Accounting for Sampling Design

	Overall N= 512333	TB N= 11,845	No TB N= 500,488	P-value
Male (N, %)	256,480 (50.06%)	5596 (47.24%)	250,884 (50.12)	0.1399
Female (N, %)	255,853 (49.94%)	6249 (52.76%)	249,604 (49.87)	
Age (mean, 95% CI)	26.18 (25.94, 26.43)	25.47 (23.97, 26.97)	26.21 (25.96, 26.45)	0.3450
Age Category				
0-4	58,365 (11.39)	1515 (12.79)	56,850 (11.36)	0.5800
5-14	121,878 (23.79)	2893 (24.42)	118,985 (23.78)	
15-39	205,376 (40.09)	4485 (37.86)	200,892 (40.4)	
≥40	126,663 (24.73)	2952 (24.92)	123,711 (24.72)	
BMI (mean, 95% CI)	20.52 (20.42, 20.63)	19.54 (18.94, 20.15)	20.54 (20.44, 20.65)	0.0014
BMI Category				
Malnourished	17,296 (15.07)	647.28 (22.63%)	16,649 (14.89%)	0.0152
Underweight	22,560 (19.66)	661.19 (23.12%)	21,899 (19.57%)	
Healthy	60,822 (53.00)	1359 (47.52%)	59,463 (53.14)	
Overweight	10,763 (9.37)	168.95 (5.91%)	10,594 (9.47%)	
Obese	3,320 (2.89)	23.58 (0.82%)	3,296 (2.95%)	
Windows (N, %)	343,541 (67.09)	6330 (53.44)	337,210 (67.41)	<0.0001
Wood Fuel (N, %)	253,289 (49.44)	6039 (50.98)	247,250 (49.40)	0.5591

Part D

Create a table for the results of a multivariate logistic regression.

Regression Accounting for Study Design

	OR	95% CI	P-value
Female vs Male	1.122	0.963, 1.307	0.1399
Age Category			
0-4	1.062	0.801, 1.407	0.6767
5-14	0.972	0.776, 1.219	0.8085
15-39	0.933	0.765, 1.138	0.4938
≥40	-	-	-
Windows	0.552	0.45, 0.68	<0.001
Wood Fuel	0.947	0.758, 1.182	0.6299

Question 4

What appears to be associated with TB disease? How might the sample size impact your conclusions?

Having windows in the home is associated with TB. Using survey weights, individuals with TB only made up 2.31% of the total sample, so there may not be enough power to detect potential associations with TB.

Data Dictionary

Variable	Description	Coding
Gender	Gender of the individual	M, F
TB	Indicator of whether individual has TB disease	0: no disease 1: has TB disease
Age		
Age_cat		0: 0-4 1: 5-14 2: 15-39 3: >40
BMI	BMI quantitative value	
BMI_cat	BMI in categories established by WHO	Malnourished (BMI<17) Underweight (17 ≤ BMI < 18.5) Healthy (18.5 ≤ BMI < 25) Overweight (25 ≤ BMI < 30) Obese (BMI ≥ 30)
Wood_fuel	Indicator of whether wood fuel is used in cooking	0: no wood fuel 1: use wood fuel
Smoker	Indicator of whether the respondent is a smoker	0: not a smoker 1: smoker
Windows	Indicator of whether the home has windows	0: no windows 1: windows
Wgtdhs	Weighting variable for each individual	
Sho21	Cluster variable for sampling design	
hv022	Stratum variable for sampling design	