

# Sampling

Brad McNeney

2019-01-28

# Introduction to Sampling: Definitions

Reference: Sampling: Design and Analysis, by S. Lohr (1999).

- ▶ **Observational unit** An object on which a measurement is taken.
- ▶ **Target population** The complete collection of observations that we want to study.
- ▶ **Sample** A subset of the population
- ▶ **Sampled population** The population from which the sample was taken.
- ▶ **Sampling unit** The unit we actually sample.
  - ▶ E.G., we may want to study individuals, but do not have a list of individuals in the target population. Instead we sample households. The observational units are the individuals living in the households.
- ▶ **Sampling frame** The list of sampling units.

# Selection Bias

- ▶ Selection bias occurs when some part of the target population is not in the sampled population.
  - ▶ For example, if a survey of household income omits transient people, the estimates of average household income are likely too large.
- ▶ Convenience samples are often biased, since the units that are easiest to select or are most likely to respond are usually not representative.
  - ▶ Mall surveys, mail-in surveys, web surveys.
  - ▶ Nonresponders often differ from responders.

# Probability Sampling

- ▶ In a probability sample, each unit in the population has a known probability of selection.
- ▶ In a **simple random sample** (SRS), every unit in the population has the same probability of being included in the sample.
- ▶ In a **stratified random sample** the population is divided into subgroups, or strata, and an SRS is taken from each stratum.
- ▶ In a **cluster sample**, observation units are aggregated into larger sampling units called clusters. We draw a sample of clusters and then subsample all or some observational units within the sampled clusters.

# Stratified Random Sampling

- ▶ One example is case-control sampling, where we draw separate samples from the population of cases and controls.
  - ▶ The motivation is our need for a representative sample of a rare segment (cases) of the population.
- ▶ Other reasons to draw a random sample include protecting against a bad SRS, lower cost of administering the survey, and ensuring comparable precision of estimates (e.g., of means) within strata.

# Cluster Sampling

- ▶ We aggregate the observational units into clusters, sample clusters, and then sample all or some observational units within clusters.
- ▶ The clusters are called primary sampling units (psu's) and the samples within clusters are called secondary sampling units (ssu's).
- ▶ The primary motivation for cluster sampling is to save money when sampling units from a population that is geographically dispersed (like Canada) or naturally grouped into clusters (like schools, or hospitals).

# Complex Surveys

- ▶ Complex surveys may employ a combination of clustering and stratification.
- ▶ Example: The Canadian Community Health Survey – Healthy Aging.
- ▶ See the documentation available on canvas.

# Demographic and HUI Variables

- Illustrate survey information with a selection of demographic and health-utility-index (HUI) variables.

```
uu<-url("http://people.stat.sfu.ca/~mcneney/Teaching/Stat305/Data/HUI.csv")
hui <- read.csv(uu)
head(hui,n=3)
```

```
##      GEO_PRV  GEOGCMA2      DHHGAGE DHH_SEX      HUIDCOG
## 1      ONT  NON - CMA 45 TO 49 YEARS  FEMALE COG. ATT. LEVE 1
## 2      ONT      CMA 55 TO 59 YEARS    MALE COG. ATT. LEVE 1
## 3      ONT      CMA 75 TO 79 YEARS    MALE COG. ATT. LEVE 1
##              HUIGDEX      HUIDEMO      HUIGHER HUIDHSI      HUIGMOB
## 1 USE OF HANDS/F. EMOT. ATT. LEV.1 NO PROBLEMS    0.838 NO PROBLEMS
## 2 USE OF HANDS/F. EMOT. ATT. LEV.1 NO PROBLEMS    0.973 NO PROBLEMS
## 3 USE OF HANDS/F. EMOT. ATT. LEV.1 NO PROBLEMS    0.973 NO PROBLEMS
##              HUIGSPE      HUIGVIS      WTS_M
## 1 NO PROBLEMS VISUAL PROB. COR 1026.07
## 2 NO PROBLEMS VISUAL PROB. COR 1987.81
## 3 NO PROBLEMS VISUAL PROB. COR  343.27
```



```

##      GEO_PRV      GEOGCM2      DHHGAGE      DHH_SEX
## ONT      :6525   CMA      :17335   55 TO 59 YEARS:4788   FEMALE:17568
## QUE      :5217   NON - CMA:13530   60 TO 64 YEARS:4542   MALE  :13297
## BC       :3860
## AB       :2735
## NS       :2282
## NB       :2225
## (Other):8021
##      (Other)      :7519
##      HUIDCOG      HUIDDEX      HUIDEMO
## COG. ATT. LEVE 1:21495   LIM. HANDS/F : 405   EMOT. ATT. LEV.1:22980
## COG. ATT. LEVE 2: 765   USE OF HANDS/F.:30446   EMOT. ATT. LEV.2: 6342
## COG. ATT. LEVE 3: 5896   NA's      : 14   EMOT. ATT. LEV.3: 1136
## COG. ATT. LEVE 4: 1938
## COG. ATT. LEVE 5: 641
## COG. ATT. LEVE 6: 94
## NA's      : 36
##      HUIGHER      HUIDHSI      HUIGMOB
## NO PROBLEMS :26706   Min. : -0.3170   NEED MECH. SUPP: 2454
## PROB./CORR. : 2490   1st Qu.: 0.7270   NO AID REQUIRED: 487
## PROB./NOT CORR.: 1221   Median : 0.9050   NO PROBLEMS :27007
## NA's      : 448   Mean : 0.8057   REQUIRES HELP : 891
##      :      3rd Qu.: 0.9730   NA's      : 26
##      :      Max. : 1.0000
##      :      NA's :759
##      HUIGSPE      HUIGVIS      WTS_M
## NO PROBLEMS :30605   NO PROBLEMS : 6473   Min. : 10.00
## PARTIAL/NOT UND.: 244   VISUAL P. UNCOR.: 1020   1st Qu.: 91.21
## NA's      : 16   VISUAL PROB. COR:23144   Median : 228.41
##      :      NA's : 228   Mean : 441.78
##      :      3rd Qu.: 514.93
##      :      Max. :23740.26
##

```

# Sampling Weights

- ▶ The sampling weight for a sample member is the number of units in the population represented by the sample member.
- ▶ For example, if the population has 1600 men and 400 women, and a stratified sample is of 200 men and 200 women, then each sampled man represents 8 and each sampled woman represents 2.
- ▶ For complex surveys, determining sampling weights requires specialized expertise.

# Sampling Weights for CCHS-HA

```
library(dplyr)
hui %>% group_by(GEO_PRV, GEOGCMA2) %>%
  summarize(Q1=quantile(WTS_M,.25),
            Q2=quantile(WTS_M,.5),
            Q3=quantile(WTS_M,.75))
```

```
## # A tibble: 19 x 5
## # Groups:   GEO_PRV [?]
##   GEO_PRV      GEOGCMA2      Q1      Q2      Q3
##   <fct>        <fct>    <dbl> <dbl> <dbl>
## 1 AB          CMA        146.  300.  525.
## 2 AB          NON - CMA    156.  339.  618.
## 3 BC          CMA        180.  308.  439.
## 4 BC          NON - CMA    197.  386.  650.
## 5 MB          CMA        88.0  152.  226.
## 6 MB          NON - CMA    87.5  156.  263.
## 7 NB          CMA        55.5  104.  158.
## 8 NB          NON - CMA    60.1  111.  171.
## 9 NFLD & LAB. CMA        43.8  81.8  114.
## 10 NFLD & LAB. NON - CMA    48.7  101.  133.
## 11 NS          CMA        71.5  124.  187.
## 12 NS          NON - CMA    78.7  135.  200.
## 13 ONT         CMA        290.  536.  792.
## 14 ONT         NON - CMA    314.  570.  903.
## 15 PEI         NON - CMA    13.5  26.5  44.4
```

# Weighted Means

- The estimate of the population mean from a weighted sample  $y_1, \dots, y_n$  with weights  $w_1, \dots, w_n$  is

$$\frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$

```
hui %>% group_by(DHH_SEX) %>%  
  summarize(unwtd.mean=mean(HUIDHSI,na.rm=TRUE),  
            wtd.mean=weighted.mean(HUIDHSI,w=WTS_M,na.rm=TRUE))
```

```
## # A tibble: 2 x 3  
##   DHH_SEX unwtd.mean wtd.mean  
##   <fct>      <dbl>    <dbl>  
## 1 FEMALE    0.794    0.838  
## 2 MALE     0.822    0.857
```