Stratified Simple Random Sampling: Quality Control

Darryl V. Creel, RTI International

October 24, 2018

```
## -- Attaching packages ----------------------------------------------------------------------- tidyver
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts -------------------------------------------------------------------------- tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Outline

Quality control at each each stage in the process.

1. Sampling Frame
2. Sample Size Data Set
3. Check Sampling Strata Across the Sampling Frame and Sample Size Data Set
4. Probability of Selection
5. Sample Indicator
6. Design Weight
7. Sample Selection Summary
8. Session Information
9. References

For this study, the sampling unit is the physician.

1. Sampling Frame

At the time of sampling, the frame consists of eligible sampling units.

```
## Parsed with column specification:
## cols(
##   id = col_integer(),
##   samplingStratum = col_integer(),
##   probabilityOfSelection = col_double(),
##   sampleIndicator = col_integer(),
##   designWeight = col_double()
## )
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1200 obs. of  5 variables:
##  $ id                    : int  1 2 3 3 NA 6 7 8 9 9 ...
##  $ samplingStratum       : int  1 1 1 1 1 1 1 NA NA 1 ...
##  $ probabilityOfSelection: num  0.25 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 ...
##  $ sampleIndicator       : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ designWeight          : num  6 6.67 6.67 6.67 6.67 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 5
##   .. ..$ id                    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ samplingStratum       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ probabilityOfSelection: list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ sampleIndicator       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ designWeight          : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

1.1. Check for Missing IDs

```
## Fail: Missing identifier value.
## [1] 3
```

1.2. Check for Duplicate IDs

Fail: Duplicate unique identifier.

Table 1: Frame Duplicate IDs

id	n
3	2
9	3
	3

1.3. Check for Missing Stratification Values

```
## Fail: Missing sampling stratum value.
## [1] 2
```

2. Sample Size Data Set

```
## Parsed with column specification:
## cols(
##   samplingStratum = col_integer(),
##   populationCount = col_integer(),
##   sampleSize = col_integer()
## )
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    4 obs. of  5 variables:
##  $ samplingStratum: int  1 2 3 4
##  $ populationCount: int  100 200 300 600
##  $ sampleSize     : int  15 20 25 100
##  $ posPop         : num  0.15 0.1 0.0833 0.1667
##  $ dw              : num  6.67 10 12 6
```

Pass: All observations have a value for the sampling stratum.

2.1. Check for Duplicate Strata

Pass: All observations have a unique sampling stratum value.

2.2. Check for Strata with Sample Size Less Than Two

Pass: All sampling strata have at least two sampling units.

3. Check Sampling Strata Across the Sampling Frame and Sample Size Data Set

Fail: At least one samling stratum does not have the expected number of sampling units.

Table 2: Sampling Stratum Does Not Have Expected Number of Sampling Units

samplingStratum	populationCount	frameCount	countDiff
1	100	98	2.00
		2	

4. Probability of Selection

4.1. Check Probability of Selection Values

In a sampling stratum, all sampling uints should have the same probability of selection.

Fail: At least one samling stratum does not have all the sampling units with the same value for the probability of selection.

Table 3: Probabilities of Selection Do Not Have the Same Value

samplingStratum	posPop	psVal	diff
2	0.10		
		0.15	

4.2. Check Sum of Probabilities of Selection

In a sampling stratum, the sample size should equal the sum of the probabilities of selection. For the h^{th} sampling stratum, the sample size, n_h, should equal the sum of the probabilities of selection, p_{hi}. That is, in the h^{th} sampling stratum, the check to ensure that the probability of selection was calculated correctly is

$$n_h = \sum_{i=1}^{N_h} p_{hi}.$$

Fail: At least one samling stratum does not have the sum of the probabilities of selection equal to the sample size.

Table 4: Sum Probabilites of Selection not Equal Sample Size

samplingStratum	sampleSize	psSum	diff
2	20		
		0.30	

5. Sample Indicator

In a sampling stratum, the sample size should equal the sum of the sample indicators. For the h^{th} sampling stratum, the sample size, n_h, should equal the sum of the sample indicators, s_{hi}. That is, in the h^{th} sampling stratum, the check to ensure that the sample indicators were calculated correctly is

$$n_h = \sum_{i=1}^{N_h} s_{hi}.$$

Fail: At least one samling stratum does not have the sum of the sample indicators equal to the sample size.

Table 5: Sum Sample Indicators not Equal Sample Size

samplingStratum	sampleSize	siSum	diff
1	15	12	-3.00
2	20		
4	100	101	1.00
		2	

6. Design Weight

6.1. Check Design Weight Values

In a sampling stratum, all the sampled sampling units should have the same design weight, and all non-sampled sampling units should have a design weight of zero.

Fail: At least one samling stratum, when the design weight is greater than zero, does not have all the sampling units with the same value for the design weight.

Table 6: Design Weights Do Not Have the Same Value

samplingStratum	dw	dwVal	diff
		6.67	

6.2. Check Sum of the Design Weights

In a sampling stratum, the population count should equal the sum of the design weights. For the h^{th} sampling stratum, the population count, N_h, should equal the sum of the design weights, d_{hi}. That is, in the h^{th} sampling stratum, the check to ensure that the design weights were calculated correctly is

$$N_h = \sum_{i=1}^{N_h} d_{hi}.$$

Fail: At least one samling stratum does not have the sum of the design weights equal to the population size.

Table 7: Sum Design Weights not Equal Population Size

samplingStratum	populationCount	dwSum	diff
1	100	86.00	-14.00
2	200		
4	600	606.00	6.00
		13.33	

7. Quality Control Summary

```
## Fail: At least one samling stratum does not have all the sampling units with the same value for the p

## Fail: At least one samling stratum does not have the sum of the probabilities of selection equal to t

## Fail: At least one samling stratum does not have the sum of the sample indicators equal to the sample

## Fail: At least one samling stratum, when the design weight is greater than zero, does not have all th

## Fail: At least one samling stratum does not have the sum of the design weights equal to the populatio
```

Table 8: Sampling Summary Table

samplingStratum	populationCount	sampleSize	probabilityOfSelection	sampleIndicator	designWeight	n
1	100	15	0.15	0	0.00	85
1	100	15	0.15	1	6.67	12
1	100	15	0.25	0	6.00	1
2	200	20	0.10	0	0.00	180
2	200	20	0.10	1	10.00	17
2	200	20	0.10	1		1
2	200	20	0.10		10.00	1
2	200	20		1	10.00	1
3	300	25	0.08	0	0.00	275
3	300	25	0.08	1	12.00	25
4	600	100	0.17	0	0.00	499
4	600	100	0.17	1	6.00	101
			0.15	1	6.67	2

8. Table, Document, and Session Information

The tables in this document were created using the xtable package (Dahl et al. 2018). This package was used in the R language and environment for statistical computing (R Core Team 2018). This document was created using the knitr package (Xie 2018) in RStudio (RStudio).

Session informaiton:

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] bindrcpp_0.2.2  xtable_1.8-3    forcats_0.3.0   stringr_1.3.1
##  [5] dplyr_0.7.6     purrr_0.2.5     readr_1.1.1     tidyr_0.8.1
##  [9] tibble_1.4.2    ggplot2_3.0.0   tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.19     cellranger_1.1.0 pillar_1.3.0     compiler_3.5.1
##  [5] plyr_1.8.4       bindr_0.1.1      tools_3.5.1      digest_0.6.18
##  [9] lubridate_1.7.4  jsonlite_1.5     evaluate_0.12    nlme_3.1-137
## [13] gtable_0.2.0     lattice_0.20-35  pkgconfig_2.0.2  rlang_0.2.2
## [17] cli_1.0.1        rstudioapi_0.8   yaml_2.2.0       haven_1.1.2
## [21] withr_2.1.2      xml2_1.2.0       httr_1.3.1       knitr_1.20
## [25] hms_0.4.2        rprojroot_1.3-2  grid_3.5.1       tidyselect_0.2.5
## [29] glue_1.3.0       R6_2.3.0         readxl_1.1.0     rmarkdown_1.10
## [33] modelr_0.1.2     magrittr_1.5     backports_1.1.2  scales_1.0.0
## [37] htmltools_0.3.6  rvest_0.3.2      assertthat_0.2.0 colorspace_1.3-2
## [41] stringi_1.1.7    lazyeval_0.2.1   munsell_0.5.0    broom_0.5.0
## [45] crayon_1.3.4
```

9. References

Dahl, David B., David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. 2018. *Xtable: Export Tables to Latex or Html*. https://CRAN.R-project.org/package=xtable.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Xie, Yihui. 2018. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.