Sampling R Demo

Brad McNeney

2019-01-28

The survey package

- Computing standard errors, confidence intervals and test statistics from complex survey data is tricky and requires information on clusters and strata.
  - This information is not included in the CCHS-HA data files.
- However, we can use the survey weights to take the unequal sampling probabilities into account and get some "quick and dirty" CIs and tests.
  - A useful tool for doing so is the `survey` package.
- Use the `Tools -> Install Packages` menu in RStudio to install.

```
uu<-url("http://people.stat.sfu.ca/~mcneney/Teaching/Stat305/Data/HUI.csv")
hui <- read.csv(uu)
library(survey)
```

Specifying the Design in survey

- Details of the survey design and the data are encapsulated in a `survey.design` object by the `svydesign()` function.
- We pass `svydesign()` the variables in the data frame that identify clusters and strata, if any, and the survey weights.
  - Argument `id` specifies cluster IDs, `strata` the stratum variables, and `weights` the sampling weights.
  - For our CCHS-HA data we don't know the clusters or strata, and pass only the weights.

```
dd <- svydesign(id=~1,strata=NULL,weights=~WTS_M,data=hui)
```

Software Notes

- Cluster ID, stratum and weight variables are specified as one-sided "formulas", of the form `~variable`.
  - We saw two-sided formulas for relationships between a response and grouping variable in `t.test()`.

Means and SEs

- Use `svymean()` to calculate means and SEs that account for the design.
  - Remove missing values from the calculation with `na.rm=TRUE`
- Use `svyby()` to stratify means on a grouping variable.

```
svymean(~HUIDHSI,design=dd,na.rm=TRUE)
```

```
##             mean     SE
## HUIDHSI 0.84723 0.0022
```

```
svyby(~HUIDHSI,by=~DHH_SEX,design=dd,FUN=svymean,na.rm=TRUE)
```

```
##        DHH_SEX   HUIDHSI          se
## FEMALE  FEMALE 0.8378109 0.002921952
## MALE      MALE 0.8574178 0.003233301
```

Confidence Intervals

```
confint(svymean(~HUIDHSI,design=dd,na.rm=TRUE))
```

```
##              2.5 %    97.5 %
## HUIDHSI 0.8429717 0.8514957
```

```
confint(svyby(~HUIDHSI,by=~DHH_SEX,design=dd,FUN=svymean,na.rm=TRUE))
```

```
##              2.5 %    97.5 %
## FEMALE 0.8320840 0.8435378
## MALE   0.8510806 0.8637549
```

t-tests

- svyttest() does t-tests.

```
svyttest(HUIDHSI~DHH_SEX,design=dd,na.rm=TRUE)
```

```
##
##  Design-based t-test
##
## data:  HUIDHSI ~ DHH_SEX
## t = 4.4991, df = 30104, p-value = 6.85e-06
## alternative hypothesis: true difference in mean is not e
## 95 percent confidence interval:
##  0.01106542 0.02814841
## sample estimates:
## difference in mean
##         0.01960691
```