STA 522 HW4 (Stratified Sampling)

Daniel Truver

2/7/2018

(1) Golf Data

```
library(survey)
golfdata = read.csv("golfsrs.csv")
n.golf1 = 120
N.golf1 = 14938
fpc.golf1 = rep(N.golf1, n.golf1)
wts.golf1 = N.golf1/n.golf1
svy.golf1 = svydesign(~1, weights = wts.golf1, fpc = fpc.golf1, data = golfdata)
```

(1)(a) Back-tee Yardage

```
mean.golf1 = svymean(~backtee, svy.golf1)
confmean.golf1 = confint(svymean(~backtee, svy.golf1), level = 0.95)
```

Table 1: Estimate of Mean Back-tee Yardage on US Golf Courses

Estimate of Mean	2.5% Quantile	97.5% Quantile
637172	598474.9	675869.1

(b) Pros Available

```
library(dplyr)
confprop.golf1 = svyciprop(design = svy.golf1, level = 0.95, formula = ~pro)
proppro.golf1 = confprop.golf1[1]
propconfint.golf1 = confprop.golf1 %>%
  attr("ci")
```

Table 2: Estimate of Proportion of Golf Courses with Pros Available

Proportion Estimate	2.5% Quantile	97.5% Quantile
0.733	0.646	0.805

(c) Average Fee for 9 Weekday Holes

```
meanwkday9.golf1 = svymean(~wkday9, svy.golf1)
confwkday9.golf1 = confint(svymean(~wkday9, svy.golf1), level = .95)
```

Table 3: Estimate of Mean Weekday 9 Hole Green Fees

Mean Estimate	2.5% Quantile	97.5% Quantile
2418.4	2035.06	2801.74

(d)

```
diff.golf1 = svymean(~I(wkday9-wkend9), svy.golf1)
confdiff.golf1 = confint(diff.golf1, level = .95)
```

```
table_1d = data.frame(diff.golf1[1],
                      confdiff.golf1[1],
                      confdiff.golf1[2])
knitr::kable(table_1d,
             col.names = c("Mean Estimate", "2.5% Quantile", "97.5% Quantile"),
             caption = "Difference Between Weekday 9 Fees, Weekend 9 Fees (wkday - wkend)",
             digits = 2, row.names = FALSE)
```

Table 4: Difference Between Weekday 9 Fees, Weekend 9 Fees (wkday - wkend)

Mean Estimate	2.5% Quantile	97.5% Quantile
-10	-29.52	9.52

(2) Lohr, Chapter 3 Problem 8

(a) all households interviewed in person

We have a budget of $C = 20000$. After ruling out an investment in bitcoin, our fixed costs are $c_0 = 5000$. So, $D = 15000$. Let n_1, n_2 denote the sample size from houses with telephones and without telephones, respectively. Also let c_1, c_2 be the cost to sample an individual from each population. Using the formula from the lecture notes and $c_1 = c_2 = 30$, with the assumption that $S_1 \approx S_2$ we obtain

$$\begin{aligned} n_h &= D\left(\frac{N_h S_h/(N\sqrt{c_h})}{\frac{1}{N}\sum_{l=1}^{H} N_h S_h \sqrt{c_h}}\right) \quad \text{recall } N_1 = .90N \\ \implies n_1 &= 15000\left(\frac{.90S_1/\sqrt{30}}{.90S_1\sqrt{30} + .10S_2\sqrt{30}}\right) = 450 \\ \implies n_2 &= 15000\left(\frac{.10S_2/\sqrt{30}}{.90S_1\sqrt{30} + .10S_2\sqrt{30}}\right) = 50 \end{aligned}$$

(b) households with phone interviewed by phone, in-person for the rest

We use the same formula, but $c_1 = 10, c_2 = 40$.

$$\begin{aligned} n_1 &= 15000\left(\frac{.90S_1/\sqrt{10}}{.90S_1\sqrt{10} + .10S_2\sqrt{40}}\right) \approx 1227 \\ n_2 &= 15000\left(\frac{.10S_2/\sqrt{40}}{.90S_1\sqrt{10} + .10S_2\sqrt{40}}\right) \approx 68 \end{aligned}$$

We have rounded down the values above to be sure of our budget. This means we come in under budget. The remaining ten dollars we can then bet on red or spend to acquire one more sample by phone.

(3) Lohr, Chapter 3 Problem 16

(a) Total Otter Dens in Shetland

Our opening move is to input the data into the `survey` package, a package which we now recognize as a divine gift to survey researchers.

```
N.otters = 237
n.otters = 82
N_1.otters = 89
N_2.otters = 61
N_3.otters = 40
N_4.otters = 47
n_1.otters = nrow(otters %>% filter(habitat == 1))
n_2.otters = nrow(otters %>% filter(habitat == 2))
n_3.otters = nrow(otters %>% filter(habitat == 3))
n_4.otters = nrow(otters %>% filter(habitat == 4))
wts.otters = rep(NA, n.otters)
wts.otters[otters$habitat == 1] = N_1.otters/n_1.otters
wts.otters[otters$habitat == 2] = N_2.otters/n_2.otters
wts.otters[otters$habitat == 3] = N_3.otters/n_3.otters
wts.otters[otters$habitat == 4] = N_4.otters/n_4.otters
fpc.otters = rep(NA, n.otters)
fpc.otters[otters$habitat == 1] = N_1.otters
fpc.otters[otters$habitat == 2] = N_2.otters
fpc.otters[otters$habitat == 3] = N_3.otters
fpc.otters[otters$habitat == 4] = N_4.otters
svy.otters = svydesign(~1, strata = ~habitat, data = otters,
                       weights = wts.otters, fpc = fpc.otters)
```

Now we proceed with the questions. Praise be to the `survey` package.

```
total.otters = svytotal(~holts, svy.otters)
knitr::kable(data.frame(total.otters), col.names = c("Estimated Total", "Standard Error"),
             row.names = FALSE, caption = "Total Number of Otter Dens on Schetland Coast",
             digits = 2)
```

Table 5: Total Number of Otter Dens on Schetland Coast

Estimated Total	Standard Error
984.71	73.92

(b) Possible Bias

For measurement bias, taking a 110-m-wide strip may not be wide enough; what defines 'along the coastline' is somewhat arbitrary. For selection bias, I don't know much about otters, but we might want to break down the strata further based on proximity to civilization. An over- or under-estimate could result if our sample contains disproportionately few or many areas of high human population.

(4) Lohr, Chapter 3 Problem 34 parts a, b, c

(a) Total Trucks in USA

Without the total size of the population, it's fortunate that we have the weights included in the data. We are going to add a variable called `existence` to the data, just so we can use the survey package to answer this

question. One can think of this variable as an answer to, 'does this truck exist?' Essentially, we want to accomplish the calculation

$$N = \sum_{h=1}^{H} N_h = \sum_{h=1}^{H} n_h \frac{N_h}{n_h} = \sum_{h=1}^{H} n_h w_h = \sum_{h=1}^{H} w_h \sum_{i=1}^{n_h} \text{existence}_i$$

```
wts.trucks = trucks$tabtrucks
trucks = trucks %>% mutate(existence = 1)
svy.trucks = svydesign(~1, strata = ~stratum, data = trucks,
                       weights = wts.trucks)
total.trucks = svytotal(~existence, design = svy.trucks)
```

Table 6: Total Number of Trucks in the USA

Estimated Total	Standard Error
85174776	0

The standard error is 0 since we are working with entirely fixed quanitities, n_h, N_h, and `existence`. There is not going to be a truck in the survey that does not exist.

(b) Total Truck Miles, 2002

```
miles_total.trucks = svytotal(~miles_annl, svy.trucks)
miles_confi.trucks = confint(miles_total.trucks)
```

Table 7: Total Truck Miles Driven in 2002

Estimated Total	2.5%	97.5%
1.114728e+12	1.102003e+12	1.127453e+12

(c) Total Miles by Truck Type

```
truckTypes = unique(trucks$trucktype)
total_byTruck = matrix(NA, nrow = length(truckTypes), ncol = 3)
for (type in truckTypes){
  svy.subtruck = subset(svy.trucks, trucktype == type)
  total_miles.trucktype = svytotal(~miles_life, svy.subtruck)
  confi_miles.trucktype = confint(total_miles.trucktype)
  total_byTruck[type,] = c(total_miles.trucktype[1],
                           confi_miles.trucktype[1],
                           confi_miles.trucktype[2])
}
total_byTruck = data.frame(total_byTruck)
```

Table 8: Total Lifetime Miles Driven by Type of Truck

	Estimated Total	2.5%	97.5&
1	3.952943e+12	3.892903e+12	4.012984e+12
2	3.607810e+12	3.550049e+12	3.665571e+12
3	5.191978e+11	5.098864e+11	5.285092e+11

	Estimated Total	2.5%	97.5&
4	3.269331e+11	3.219379e+11	3.319282e+11
5	5.528890e+11	5.460397e+11	5.597384e+11

(5) Another Friend Does Something Wrong

(a) Inclusion Probability

For individual i in stratum h, the inclusion probability is

$$Pr(I_{hi} = 1) = \frac{n_h}{N_h}.$$

(b) Expected Value of Friend's Estimator

$$\begin{aligned}
E(\tilde{y}) &= E\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{y_{hi}}{n}\right) \\
&= \frac{1}{n}\sum_{h=1}^{H} E\left(\sum_{i=1}^{n_h} y_{hi}\right) \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}E(I_{hi}) \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}\frac{n_h}{N_h} \\
&= \frac{1}{n}\sum_{h=1}^{H}\frac{n_h}{N_h}\sum_{i=1}^{N_h} y_{hi} \\
&= \frac{1}{n}\sum_{h=1}^{H} n_h\bar{Y_h}
\end{aligned}$$

(c) Bias of Estimator

The estimator $\tilde{y}$ is generally biased. It's bias is given by

$$\text{Bias}(\tilde{y}) = \frac{1}{n}\sum_{h=1}^{H} n_h\bar{Y_h} - \frac{1}{N}\sum_{h=1}^{H} N_h\bar{Y_h}.$$

This tells us, however, that $\tilde{y}$ will be unbiased in the special case that $n_h/n = N_h/N$. That is, it will be unbiased when the stratum sample is proportional in size to the size of the stratum.

(6) Simulation Study

(a) Growing the Population

```
H = 3
N_1 = 50000
group1 = rnorm(N_1, mean = 10, sd = 5)
N_2 = 35000
group2 = rnorm(N_2, mean = 40, sd = 2)
N_3 = 15000
group3 = rnorm(N_3, mean = 100, sd = 20)
population = data.frame(
  strat = c(rep(1, N_1), rep(2,N_2), rep(3,N_3)),
  value = c(group1, group2, group3)
  )
N = sum(N_1, N_2, N_3)
```

(b) Optimal Allocation of Sample Size

We will use the allocation formula from class with $n = 200$ and the values given above for each S_h.

$$n_h = n\frac{N_h S_h/N}{\sum_{l=1}^{3} N_l S_l/N}$$

When we perform the calculation below, we get non-integer results. Unfortunately, all of them would round up, which would put us at a sample size of 201. Instead, we round up only the two that are closest to the next greatest integer.

```
n = 200
denom = sum(N_1*5/N, N_2*2/N, N_3*20/N)
n_1 = ceiling(n*(N_1*5/N)/denom)
n_2 = floor(n*(N_2*2/N)/denom)
n_3 = ceiling(n*(N_3*20/N)/denom)
cat("n_1 =",n_1,
    "\nn_2 =", n_2,
    "\nn_3 =", n_3)
```

```
## n_1 = 81
## n_2 = 22
## n_3 = 97
```

(c) Running the Simulation

```
# first, some QOL enhancements
population = population %>%
  mutate(p_inclusion = ifelse(strat == 1, n_1/N_1,
                              ifelse(strat == 2, n_2/N_2, n_3/N_3))) %>%
  mutate(weight = 1/p_inclusion)
# now onto the calculation
set.seed(2018)
nsim = 1000
res.sim = data.frame(HT = rep(NA, nsim), AVG = rep(NA, nsim))
for (t in 1:nsim){
  samp1 = sample_n(population %>% filter(strat == 1), n_1)
  samp2 = sample_n(population %>% filter(strat == 2), n_2)
  samp3 = sample_n(population %>% filter(strat == 3), n_3)
  horvitz.thompson = sum(samp1$value * samp1$weight,
                         samp2$value * samp2$weight,
```

```
                        samp3$value * samp3$weight) * (1/N)
  average = mean(c(samp1$value,
                   samp2$value,
                   samp3$value))
  res.sim$HT[t] = horvitz.thompson
  res.sim$AVG[t] = average
}
res = apply(res.sim, 2, mean)
knitr::kable(data.frame(res["HT"], res["AVG"], mean(population$value)),
             col.names = c("Avg. Horvitz-Thompson", "Avg. Sample Mean", "Real Population Mean"),
             digits = 2, row.names = FALSE)
```

Avg. Horvitz-Thompson	Avg. Sample Mean	Real Population Mean
33.98	56.89	33.98

(d) Empirical Results

The sample mean is a poor estimator of the population mean, by the results for this survey. Bias in the sample mean is possible (shown analytically in part (5) and empirically here), and we should rely first on the Horvitz-Thompson estimator.