STAT 341: Tutorial 6 – Probabilistic Sampling

Friday February 28, 2020

Part 1: Horvitz-Thompson Estimation with SRSWOR

Question 1

Suppose that a sample $\mathcal{S}$ of size n is to be drawn from a population $\mathcal{P}$ of size N. Suppose that the units are selected at random and *without replacement.*

(a) Derive the marginal and joint inclusion probabilities π_u and π_{uv}.

(b) Provide an expression for the Horvitz-Thompson estimate.

(c) Provide an expression for the variance of the Horvitz-Thompson estimator.

(a) $$\pi_u = P(u \in S) = \frac{\#\text{ of samples containing } u}{\#\text{ of possible samples}} = \frac{1 \times \binom{N-1}{n-1}}{\binom{N}{n}}$$

$$= \frac{(N-1)!}{(n-1)!\,(\cancel{N-n})!} \div \frac{N!}{n!\,(\cancel{N-n})!}$$

$$= \frac{(\cancel{N-1})!}{(\cancel{n-1})!} \div \frac{N(\cancel{N-1})!}{n(\cancel{n-1})!}$$

$$= \frac{n}{N}$$

$$\pi_{uv} = P(u \in S \text{ and } v \in S) = \frac{\#\text{ of samples containing } (u,v)}{\#\text{ of possible samples}} = \frac{1 \times 1 \times \binom{N-2}{n-2}}{\binom{N}{n}}$$

$$= \frac{(\cancel{N-2})!}{(\cancel{n-2})!\,(\cancel{N-n})!} \div \frac{N(N-1)(\cancel{N-2})!}{n(n-1)(\cancel{n-2})!\,(\cancel{N-n})!}$$

$(N-n)!$

$$= \frac{n(n-1)}{N(N-1)}$$

(b) $$a_{HT}(S) = \sum_{u \in S} \frac{y_u}{\pi_u}$$

plug in $\frac{n}{N}$

(c) $$Var[\tilde{a}_{HT}(S)] = \sum_{u \in S} \sum_{v \in S} (\pi_{uv} - \pi_u \pi_v) \frac{y_u}{\pi_u} \frac{y_v}{\pi_v}$$

Plug in $\frac{n(n-1)}{N(N-1)}$

Plug in $\frac{n}{N}$

Question 2

(a) Load the `titanic` data and calculate the survival rate (i.e., the proportion of passengers that survived the disaster).

```
titanic <- read.csv("/Users/nstevens/Dropbox/Teaching/STAT_341/Tutorials/Tutorial 6/titanic.csv")
survRate <- mean(titanic$Survived)
print(survRate)
```

```
## [1] 0.323035
```

(b) Take a simple random sample without replacement of size $n = 100$.

```
n <- 100
N <- dim(titanic)[1]
set.seed(341)
indx_srswor <- sample(N, n, replace = FALSE)
titan_srswor <- titanic[indx_srswor,]
```

(c) Calculate the Horvitz-Thompson estimate of the survival rate, given the SRSWOR from (b).

```
pi_u <- rep(n/N, n) # marginal inclusion probabilities for units in the sample
y_u <- titan_srswor$Survived/N # variate values being summed in the sample

survRate_HT_srswor <- sum(y_u/pi_u)
print(survRate_HT_srswor)
```

```
## [1] 0.33
```

(d) Calculate the standard error of the HT estimate from (c).

To do this we will use a slightly modified version of the `estVarHT` function from the notes:

```
estVarHT <- function(y_u, pi_u, pi_uv){
  delta = pi_uv - outer(pi_u, pi_u)
  estimateVar =  sum( (delta/pi_uv) * outer(y_u/pi_u,y_u/pi_u) )
  return(estimateVar)
}
```

Now we simply need to calculate the joint inclusion probability matrix and plug everything into this function.

```
pi_uv <- matrix((n*(n-1)) / (N*(N-1)), nrow=n, ncol=n) # joint inclusion probabilities for units in the
diag(pi_uv) <- pi_u

var_HT_srswor <- estVarHT(y_u, pi_u, pi_uv)
se_HT_srswor <- sqrt(var_HT_srswor)
print(se_HT_srswor)
```

```
## [1] 0.04617212
```

(e) Calculate an approximate 95% confidence interval for the true survival rate.

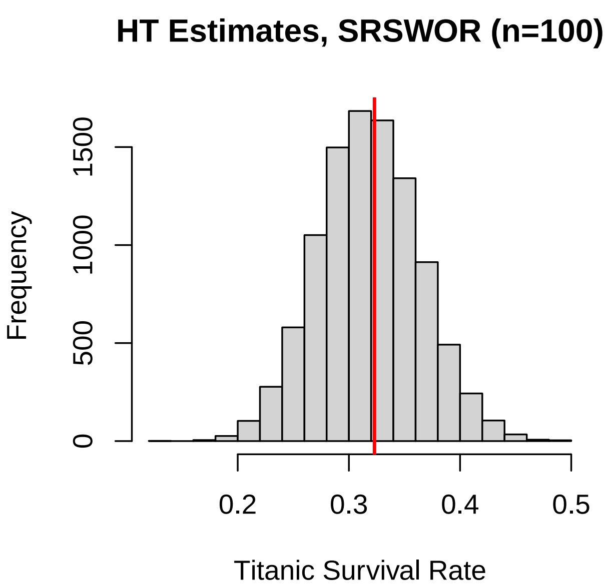
```
ci_srswor <- survRate_HT_srswor + 2*c(-1,1)*se_HT_srswor
print(ci_srswor)
```

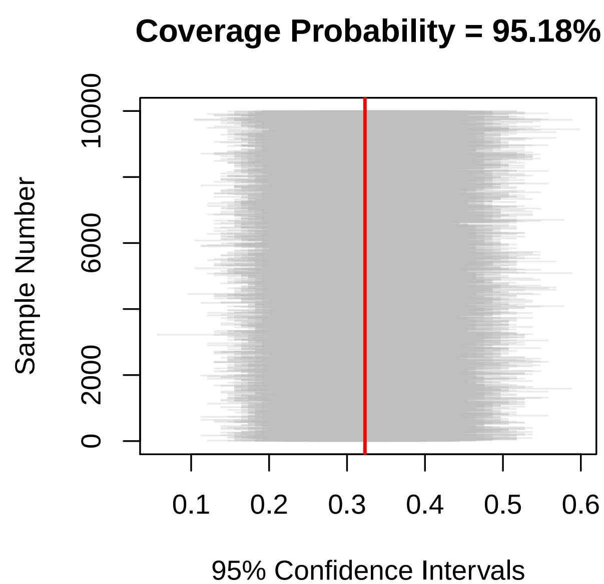
```
## [1] 0.2376558 0.4223442
```

(f) Draw 10,000 SRSWOR samples of size $n = 100$ from the `titanic` population. For each sample calculate the HT estimate and an approximate 95% confidence interval for the surival rate. Graphically summarize the sampling distribution of the HT estimator as well as the coverage of the corresponding confidence interval.

```r
# Run the simulations:
est <- rep(0, 10000)
ci <- matrix(0, nrow = 10000, ncol = 2)
for(i in 1:10000){
  samp <- sample(titanic$Survived, size = 100, replace = FALSE)
  y_u <- samp/N
  est[i] <- sum(y_u/pi_u)
  se <- sqrt(estVarHT(y_u, pi_u, pi_uv))
  ci[i,] <- sum(y_u/pi_u) + 2*c(-1,1)*se
}

# Construct the plots:
par(mfrow = c(1,2))
hist(est, col = "lightgrey", main = "HT Estimates, SRSWOR (n=100)", xlab = "Titanic Survival Rate")
abline(v = survRate, col ="red", lwd = 2)
coverage <- apply(X = ci, MARGIN = 1, FUN = function(u){survRate >= u[1] & survRate <= u[2]})
plot(0, type = "n", ylim = c(0,10000), xlim = c(min(ci[,1]),max(ci[,2])),
     xlab = "95% Confidence Intervals", ylab = "Sample Number",
     main = paste0("Coverage Probability = ", round(100*mean(coverage),2), "%"))
for(i in 1:10000){
  segments(x0 = ci[i,1], y0 = i, x1 = ci[i,2], y1 = i, col = adjustcolor("gray", alpha = 0.3))
}
abline(v = survRate, col ="red", lwd = 2)
```





(g) Using the 10,000 estimates from (f) estimate the sampling bias, variance and MSE of this HT estimator.

```r
bias_srswor <- mean(est - survRate)
variance_srswor <- var(est)
MSE_srswor <- mean((est - survRate)^2)
kable(data.frame(bias = bias_srswor, variance = variance_srswor, MSE = MSE_srswor))
```

bias	variance	MSE
-6e-06	0.0020905	0.0020903

Part 2: Horvitz-Thompson Estimation with SRSWR

Question 1

Suppose that a sample $\mathcal{S}$ of size n is to be drawn from a population $\mathcal{P}$ of size N. Suppose that the units are selected at random and *with replacement.*

(a) Derive the marginal and joint inclusion probabilities π_u and π_{uv}.

(b) Provide an expression for the Horvitz-Thompson estimate.

(c) Provide an expression for the variance of the Horvitz-Thompson estimator.

$$\pi_u = P(u \in S) = 1 - P(u \notin S)$$

$$= 1 - P(u \text{ not selected } 1^{st} \text{ and } \ldots \text{ and } u \text{ not selected } n^{th})$$

$$= 1 - \prod_{i=1}^{n} P(u \text{ not selected } i^{th})$$

$$= 1 - \prod_{i=1}^{n} \left(1 - \frac{1}{N}\right)$$

$$= 1 - \left(\frac{N-1}{N}\right)^n$$

$$\pi_{uv} = P(u \in S \text{ and } v \in S) = 1 - P(u \notin S \text{ or } v \notin S)$$

$$= 1 - \left[P(u \notin S) + P(v \notin S) - P(u \notin S \text{ and } v \notin S)\right]$$

$$= 1 - \left[\left(\frac{N-1}{N}\right)^n + \left(\frac{N-1}{N}\right)^n - \prod_{i=1}^{n} P(u \text{ and } v \text{ are not selected})\right]$$

$$= 1 - \left[2\left(\frac{N-1}{N}\right)^n - \prod_{i=1}^{n}\left(1 - \frac{2}{N}\right)\right]$$

$$= 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n$$

(b) and (c) Plug in π_u and π_{uv} (from above) into:

$$a_{HT}(S) = \sum_{u \in S} \frac{y_u}{\pi_u} \qquad Var\left[\tilde{a}_{HT}(S)\right] = \sum_{u \in S}\sum_{v \in S} (\pi_{uv} - \pi_u \pi_v) \frac{y_u}{\pi_u} \frac{y_v}{\pi_v}$$

Question 2

(a) Using the `titanic` data, take a simple random sample without replacement of size $n = 100$.

```
set.seed(341)
indx_srswr <- sample(N, n, replace = TRUE)
titan_srswr <- titanic[indx_srswr,]
```

(b) Calculate the Horvitz-Thompson estimate of the survival rate, given the SRSWOR from (b).

```
pi_u <- rep(1-((N-1)/N)^n, n) # marginal inclusion probabilities for units in the sample
y_u <- titan_srswr$Survived/N # variate values being summed in the sample

survRate_HT_srswr <- sum(y_u/pi_u)
print(survRate_HT_srswr)
```

```
## [1] 0.3374784
```

(c) Calculate the standard error of the HT estimate from (b).

We simply need to calculate the joint inclusion probability matrix and plug everything into the `estVarHT` function.

```
pi_uv <- matrix(1 - 2*((N-1)/N)^n + ((N-2)/N)^n, nrow=n, ncol=n) # joint inclusion probabilities for un
diag(pi_uv) <- pi_u

var_HT_srswr <- estVarHT(y_u, pi_u, pi_uv)
se_HT_srswr <- sqrt(var_HT_srswr)
print(se_HT_srswr)
```

```
## [1] 0.04724532
```

(d) Calculate an approximate 95% confidence interval for the true survival rate.

```
ci_srswr <- survRate_HT_srswr + 2*c(-1,1)*se_HT_srswr
print(ci_srswr)
```

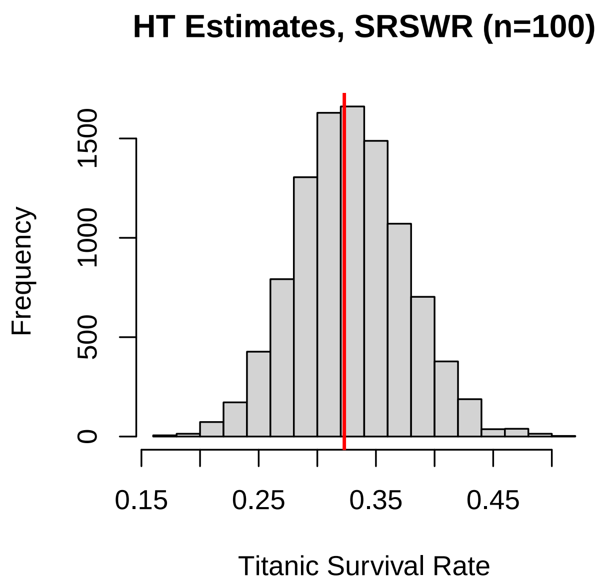
```
## [1] 0.2429878 0.4319690
```

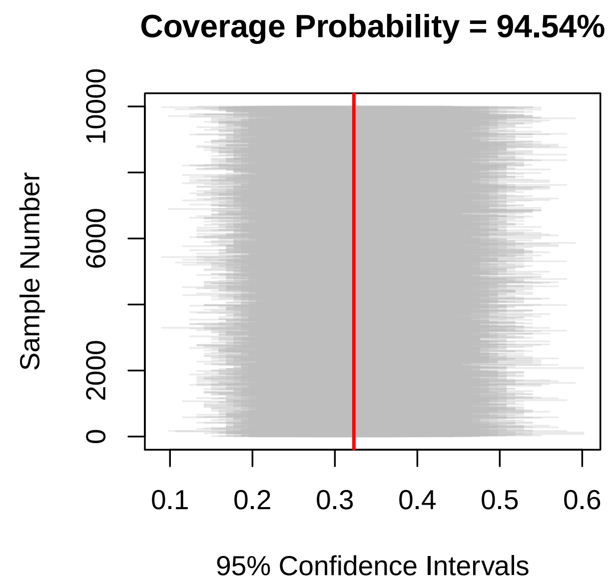
(e) Draw 10,000 SRSWR samples of size $n = 100$ from the `titanic` population. For each sample calculate the HT estimate and an approximate 95% confidence interval for the surival rate. Graphically summarize the sampling distribution of the HT estimator as well as the coverage of the corresponding confidence interval.

```
# Run the simulations:
est <- rep(0, 10000)
ci <- matrix(0, nrow = 10000, ncol = 2)
for(i in 1:10000){
  samp <- sample(titanic$Survived, size = 100, replace = TRUE)
  y_u <- samp/N
  est[i] <- sum(y_u/pi_u)
  se <- sqrt(estVarHT(y_u, pi_u, pi_uv))
  ci[i,] <- sum(y_u/pi_u) + 2*c(-1,1)*se
}

# Construct the plots:
par(mfrow = c(1,2))
hist(est, col = "lightgrey", main = "HT Estimates, SRSWR (n=100)", xlab = "Titanic Survival Rate")
abline(v = survRate, col ="red", lwd = 2)
coverage <- apply(X = ci, MARGIN = 1, FUN = function(u){survRate >= u[1] & survRate <= u[2]})
```

```r
plot(0, type = "n", ylim = c(0,10000), xlim = c(min(ci[,1]),max(ci[,2])),
     xlab = "95% Confidence Intervals", ylab = "Sample Number",
     main = paste0("Coverage Probability = ", round(100*mean(coverage),2), "%"))
for(i in 1:10000){
  segments(x0 = ci[i,1], y0 = i, x1 = ci[i,2], y1 = i, col = adjustcolor("gray", alpha = 0.3))
}
abline(v = survRate, col ="red", lwd = 2)
```





(f) Using the 10,000 estimates from (f) estimate the sampling bias, variance and MSE of this HT estimator.

```r
bias_srswr <- mean(est - survRate)
variance_srswr <- var(est)
MSE_srswr <- mean((est - survRate)^2)
kable(data.frame(bias = bias_srswr, variance = variance_srswr, MSE = MSE_srswr))
```

bias	variance	MSE
0.0073523	0.0022858	0.0023396