# 11
# Theory of Sample Surveys with R: Solutions

*Solutions to exercises.*

## 11.1 Introduction

### 11.1.1 Exercises

Assume a sample survey shall be carried out to find out about how satisfied students are with their faculty.

1. How would you define the population?
2. Would you consider a census of all students or rather a sample survey? (Why?)
3. How would you operationalise "being satisfied with their faculty"?
4. What is a sampling frame and how could one be obtained in the example?
5. How could a random sample be obtained?
6. Would you consider stratified sampling? How would you define sensible strata? Would you expect the results to more or less efficient compared to simple random sampling?
7. Would you consider clustered sampling? How would you define sensible clusters? Would you expect the results to more or less efficient compared to simple random sampling?
8. How do you consider the idea of obtaining a sample from alumni?

### 11.1.2 Solutions

1. E.g. students enrolled in faculty ... at a specific date.
2. Consider size of faculty and costs of the survey/census.
3. Consider open and closed questiones, general and issue (teaching, facilities, and so on) specific opinions.
4. Sampling frame could be a list provided by the faculty administration inclding, name, subject, date of enrollment, and so on.
5. E.g. using pseudo random numbers generated with R.

6. Possible strata could be year of study, field of study, and so on. Depending on the variance (de)composition, stratified sampling may reduce the variance considerably.

7. Clusters may be generated according to attended courses. Depending on the homogeneity within clusters, the variance may be increased by clustering considerably.

8. Alumni present a selective subsample and results may be misleading.

## 11.2 Introduction to R

### 11.2.1 Exercises

1. Define a vector $x$ containing the values $\{1, 3, 5, 12\}$.
   a) Sort the $x$ using **`sort()`** and **`order()`** commands.
   b) Calculate $\sum_i x_i$.
   c) Let object $n$ contain the length of $x$.
   d) Calculate the variance of $x$ and compare your result with the result obtained from **`var(x)`**.
   e) Calculate the median of $x$ and compare your results with the results obtained from the following commands:

```
quantile(x,type=1,0.5)
quantile(x,type=7,0.5)
median(x)
```

2. Probability distributions
   a) Consider the exponential distribution with rate $\lambda = 0.2$.
      i. Calculate the value of the probability distribution for $x = 2.5$.
      ii. Calculate the value of the density distribution for $x = 0.8$.
      iii. Calculate the median.
      iv. Generate a realization with $n = 10$ from the exponential distribution (rate $\lambda = 0.2$).
   b) Consider the normally distributed random variate $X \sim \mathcal{N}(\mu = 2, \sigma^2 = 5)$ and calculate

$$\mathrm{P}(X < 3), \quad \mathrm{P}(X > 0.5), \quad \text{and} \quad \mathrm{P}(-2 < X < 3).$$

   c) Consider the Poisson distributed random variate $X \sim \mathrm{Pois}(\lambda = 4)$ and calculate

$$\mathrm{P}(X = 3), \quad \mathrm{P}(X \leq 2), \quad \text{and} \quad \mathrm{P}(1 < X \leq 5).$$

   d) Generate a realization with $n = 100$ from the normal distribution ($X \sim \mathcal{N}(\mu = 2, \sigma^2 = 5)$) and display the distribution using **`hist()`**.

3. Draw a sample of size $n = 20$ from $X$ and calculate the mean. Repeat this for $B = 1000$ samples, store the calculated means on display the distribution.

4. Define the following matrix using the command

```
x <- matrix(1:16,nrow=4,ncol=4,byrow=F)
```

$$x = \begin{pmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{pmatrix}.$$

   a) Select the vector containing the second column of $x$.

   b) Select the vector containing the third row of $x$.

   c) Select the $2 \times 2-$matrix containing the elements $x_{2,3}$, $x_{2,4}$, $x_{3,3}$, $x_{34}$.

5. Consider the following simple numerical example

$$y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}.$$

   and obtain the parameter vector $\hat{\beta}$ of the linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i.$$

### 11.2.2 Solutions

1. 
```
x <- c(1,3,4,12)
```

   a)
```
sort(x)
x[order(x)]
```

   b)
```
sum(x)
```

   c)
```
n <- length(x)
```

   d)
```
mean((x-mean(x))^2)
var(x)
# compare denominator
```

e)
```
?quantile # compare the definitions
```

2. a) i.
```
pexp(2.5,0.2)
```

ii.
```
dexp(0.8,0.2)
```

iii.
```
qexp(0.5,0.2)
```

iv.
```
rexp(10,0.2)
```

b)
```
pnorm(3,2,sqrt(5))
1-pnorm(0.5,2,sqrt(5))
pnorm(3,2,sqrt(5))-pnorm(-2,2,sqrt(5))
```

c)
```
dpois(3,4)
ppois(2,4)
ppois(5,4)-ppois(1,4)
```

d)
```
X <- rnorm(100,2,sqrt(5))
hist(X)
```

3.
```
x <- sample(X,20)
B <- 1000
e <- rep(NA,B)
for (i in 1:B) e[i] <- mean(sample(X,20))
hist(e)
```

4.
```
x[,2]
x[3,]
x[2:3,3:4]
```

5.
```
y <- c(3,1,8,3,5)
X <- cbind(1,c(3,1,5,2,4),c(5,4,6,4,6))
solve(t(X)%*%X)%*%t(X)%*%y
lm(y~X[,-1])
```

## 11.3 Inclusion probabilities

### 11.3.1 Exercises

1. A sample of size $n = 3$ shall be drawn from a population of size $N = 6$ by simple random sampling (without replacement). The number of combinations is given by

$$\binom{N}{n}.$$

   Derive this expression.

2. The inclusion indicator $I$ is a function of the random variate $S$. Please explain.

3. What is the meaning of $\pi_k$ and how can it be obtained?

4. What is the meaning of $\pi_{kl}$ and how can it be obtained?

5. Consider a population with four elements ($N = 4$)

$$U := \{u_1, u_2, u_3, u_4\}$$

   A sample of size $n = 2$ is to be drawn from the population.

   a) Consider for the moment the case of simple random sampling (SI).

      i. Obtain $M = |\mathcal{S}|$.

      ii. Obtain $\pi_k$.

      iii. Obtain the sum of all inclusion probabilities of elements $k \in U$.

   b) Consider the following sampling design: $\mathcal{S}_n = \{s_1, s_2, s_3\}$ where

$$s_1 = \{u_1, u_3\}, \quad s_2 = \{u_1, u_4\}, \quad s_3 = \{u_2, u_4\}.$$

      Assume the following probabilities for the samples:

$$p(s_1) = 0.1, \quad p(s_2) = 0.6, \quad p(s_3) = 0.3.$$

      i. Obtain all inclusion probabilities $\pi_k$.

      ii. Obtain numerically the sum of inclusion probabilities.

iii. Obtain all inclusion probabilities $\pi_{k,l}$

6. Solve exercise 5 using R.

7. Show that the covariance of the inclusion indicators $I_k$ and $I_l$ is negative in the case of simple random sampling (SI).

8. Obtain the covariance matrix of inclusion indicators for the example of exercise 5b using R.

### 11.3.2 Solutions

1. See section 7 in chapter 3.

2. $I_k$ takes values 0 or 1 depending on the sample $S$ as some samples contain $k$ and some do not.

3. $\pi_k$ is probability for $k$ being element of $S$, denoted $I_k = 1$. Sum of all $p(S)$ for samples containing $k$:

$$\pi_k = \mathrm{P}(k \in S) = \mathrm{P}(I_k = 1) = \sum_{s \ni k} p(s)$$

4. $\pi_{k,l}$ is probability for both elements $k$ and $l$ being element of $S$. Sum of all $p(S)$ for samples containing $k$ and $l$:

$$\pi_{k,l} = \mathrm{P}(k \,\&\, l \in S) = \mathrm{P}(I_k I_l = 1) = \sum_{s \ni k \& l} p(s)$$

5. a)

$$|\mathcal{S}_n| = \binom{N}{n} = \binom{4}{2} = \frac{4!}{2!\,2!} = \frac{4 \cdot 3}{2} = 6$$

$$\pi_k = \frac{n}{N} = 0.5$$

$$\sum_{k \in U} \pi_k = \sum_{k \in U} \sum_{s \ni k} p(s) = n = 4 \cdot 0.5 = 2$$

b) i.

$$\pi_1 = 0.7,\ \pi_2 = 0.3,\ \pi_3 = 0.1,\ \pi_4 = 0.9$$

$$\sum_{k \in U} \pi_k = 2$$

$$\pi_{1,2} = 0,\ \pi_{1,3} = 0.1,\ \pi_{1,4} = 0.6,$$

$$\pi_{2,3} = 0,\ \pi_{2,4} = 0.3,\ \pi_{3,4} = 0$$

6.
```
si <- cbind(c(1,3),c(1,4),c(2,4));si
p <- c(0.1,0.6,0.3)
pik <- apply(matrix(1:4),1,
             function(z) sum(t(si==z)*p))
sum(pik)
N <- 4
pikl <- matrix(NA,N,N)
for (k in 1:N){
for (l in 1:N) pikl[k,l] <- sum(
  apply(si,2,function(z)
    as.numeric(is.element(k,z)*is.element(l,z)))*p
  )
}
pikl
```

7. See chapter 3, section 8.4.

8.
```
ckl <- matrix(NA,N,N)
for (k in 1:N){
for (l in 1:N) ckl[k,l] <- pikl[k,l]-
                           pikl[k,k]*pikl[l,l]
}
ckl
```

## 11.4 Estimation

### 11.4.1 Exercises

1. How is the $\pi$-estimator ($\hat{t}_\pi$) of the population total defined?
2. Proof that $\hat{t}_\pi$ is an unbiased estimation function of the population total $t$.
3. Derive the variance of the $\pi$-estimator $\hat{t}_\pi$.
4. Proof that
$$\widehat{\mathrm{V}}(\hat{t}_\pi) = \sum\sum_s \breve{\Delta}_{kl}\breve{y}_k\breve{y}_l$$
is an unbiased estimation function of $\mathrm{V}(\hat{t}_\pi)$.
5. We consider a population with $N = 3$ elements
$$y_1 = 1, y_2 = 2, y_3 = 5$$
and want to estimate the total of the population based on a sample of size $n = 2$.

   The sampling design is characterized by the following probabilities $p(s)$:
$$p(s_1) = 0.5, \quad p(s_1) = 0.3, \quad p(s_1) = 0.2.$$
   a) Obtain the population total.
   b) Obtain the number $M = |\mathcal{S}|$ of possible samples.
   c) Obtain all elements of $\mathcal{S}_n$?
   d) Obtain the first order inclusion probabilities $\pi_k$ and the second order inclusion probabilities $\pi_{k,l}$.
   e) Obtain the covariances of the inclusion indicators $I_k$ and $I_l$.
   f) Obtain the numerical estimates $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$ for the total for each of the $M$ samples.
   g) Proof that $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$ is an unbiased estimation function for the total $t$.
   h) Derive the variance of $\hat{t}_\pi$.
   i) Obtain the numerical estimates of the variance estimators (for the $\pi$-estimator of the total) for each of the $M$ samples.

j) Show numerically that the estimation function $\widehat{\mathrm{V}}(\hat{t}_\pi)$ is an unbiased estimator of $\mathrm{V}(\hat{t}_\pi)$.

6. Solve exercise 5 using R.

### 11.4.2 Solutions

1. See text.
2. See text.
3. See text.
4. See text.
5. See 6.
6. 
```
y <- c(1,2,5)
p <- c(0.5,0.3,0.2)
# a
sum(y)
# b
choose(3,2)
# c
si <- combn(3,2);si
# d
pik <- apply(matrix(1:3),1,
             function(z) sum(t(si==z)*p))
N <- 3
pikl <- matrix(NA,N,N)
for (k in 1:N){
for (l in 1:N) pikl[k,l] <- sum(
  apply(si,2,function(z)
    as.numeric(is.element(k,z)*is.element(l,z)))*p)
}
pikl
# e
ckl <- matrix(NA,N,N)
for (k in 1:N){
for (l in 1:N) ckl[k,l] <- pikl[k,l]-
                           pikl[k,k]*pikl[l,l]
}
ckl
```

```
# f
that <- apply(si,2,function(z) sum(y[z]/pik[z]))
that
# g
sum(that*p)
sum(y)
# h
vkl <- matrix(NA,N,N)
for (k in 1:N){
for (l in 1:N) vkl[k,l] <- ckl[k,l]*
                           y[k]/pik[k]*y[l]/pik[l]
}
sum(vkl)
ty <- sum(y)
sum((that-ty)^2*p)
# i
M <- ncol(si)
vhat <- rep(NA,M)
for (i in 1:M){
  sii <- si[,i]
vkli <- matrix(0,N,N)
  for (k in sii){
  for (l in sii) vkli[k,l] <-
    ckl[k,l]/pikl[k,l]*y[k]/pik[k]*y[l]/pik[l]
  }
  vhat[i] <- sum(vkli)
}
# j
sum(vhat*p)
```

## 11.5 Simple random sampling

### 11.5.1 Exercises

1. A sample of size $n$ shall be drawn from a population of size $N$.
   a) What is the probability $\mathrm{P}(S)$ for each $s \in \mathcal{S}$?
   b) Derive that inclusion probability $\pi_k$.
   c) Derive the second order inclusion probability $\pi_{kl}$.
2. Consider a population of size $N = 6$ and a variable $X$:
   $$x_1 = 1,\ x_2 = 2,\ x_3 = 5,\ x_4 = 10,\ x_5 = 14,\ x_6 = 28.$$
   a) Obtain the number $M = |\mathcal{S}|$ of possible samples for $n = 3$.
   b) What is the largest (smallest) sample mean that can occur?
   c) What can you say about the distribution of the sample mean?
3. Consider the general $\pi$-estimator of the total which is defined for all sampling designs.
   a) How is the estimation function of the total defined?
   b) How does the estimation function simplifies in the SI case?
   c) Show that the simplified (SI case) estimation function of the total is unbiased.
   d) How is the variance of $\pi$-estimator for the total defined. How does the variance simplifies in the SI case?
   e) Obtain the estimation function of the variance of the total-estimator based on a sample.
   f) Show that the estimation function of the variance of the $\pi$-estimator for the total is unbiased.
4. Consider the variable wage of PSID data. Obtain the approximate distribution ($B = 1000$) of the estimation function for the population mean and of the estimation function for the variance of sample mean for $n = 30$. Comment on your findings.

5. Consider the variable sector of PSID data. Obtain the approximate distribution ($B = 1000$) of the estimation function for proportion of persons working in the service sector (identified by variable sector having value 7) and of the estimation function for the variance of the proportion estimator for $n = 30$. Comment on your findings.

### 11.5.2 Solutions

1. a) See text.
   b) See text.
   c) See text.

2. Consider a population of size $N = 6$ and a variable $X$:

$$x_1 = 1,\ x_2 = 2,\ x_3 = 5,\ x_4 = 10,\ x_5 = 14,\ x_6 = 28.$$

   a)
```
y <- c(1,2,5,10,14,28)
N <- 6
n <- 3
M <- choose(6,3);M
```

   b)
```
mean(y[1:n])
mean(y[(N-n+1):N])
```

   c) Population specific, non-standard, exact distribution obtained by complete enumeration of sample space.

3. See text.

4.
```
d <- read.csv2("psid_2007_sp.csv")
Y <- d$wage
N <- nrow(d)
n <- 30
f <- n/N
B <- 1000
e <- rep(NA,B)
v <- rep(NA,B)
for (i in 1:B){
  y <- sample(Y,n)
  e[i] <- mean(y)
  v[i] <- (1-f)/n*var(y)
}
```

```
plot(density(e))
# right-skewed, non-normal
plot(density(v))
# right-skewed, bimodal, non-normal
```

5. 
```
Y <- d$sector==7
for (i in 1:B){
  y <- sample(Y,n)
  e[i] <- mean(y)
  v[i] <- (1-f)/n*var(y)
}
plot(density(e))
# almost normal
plot(density(v))
# left-skewed, non-normal
# max of possible estimators is
(1-f)/n*0.5^2*n/(n-1)
max(v)
```

## 11.6 Confidence intervals

### 11.6.1 Exercises

1. Derive the Chebichev-inequality.

2. Explain how you can obtain a conservative confidence interval for your estimator without any specific knowledge about the distribution of the estimating function.

3. Despite the fact that an interval based on the Chebichev-inequality is conservative, it may well be by far too small. Explain.

4. Explain why, in general, variance estimation functions for drawing with replacement are much simpler.

5. We consider sampling with replacement with individual one-draw probabilities $p_k$, $k = 1, \ldots, N$. Because of drawing with replacement, the individual draws are independent and the probabilities are $p_k$, $k = 1, \ldots, N$ in each draw. Hansen and Hurwitz [1943] proposed the following estimation function ($p-$expanded with replacement, in short 'pwr')

$$\hat{t}_{\text{pwr}} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{k_i}}{p_{k_i}} = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

$$\text{with} \quad Z_i = \frac{y_k}{p_k} \text{ if } p_{k_i} = p_k$$

and

$$P\left(Z_i = \frac{y_k}{p_k}\right) = p_k$$

$$\text{with} \quad k = 1, \ldots, N \quad \text{and} \quad \sum_{k=1}^{N} p_k = 1.$$

The variance $\mathrm{V}(\hat{t}_{\text{pwr}})$ of the estimating function $\hat{t}_{\text{pwr}}$ is

$$\mathrm{V}(\hat{t}_{\text{pwr}}) = \frac{1}{n} \sum_{U} \left(\frac{y_k}{p_k} - t\right)^2 p_k.$$

An estimating function for the variance $\mathrm{V}(\hat{t}_{\text{pwr}})$ is

$$\widehat{\mathrm{V}}(\hat{t}_{\text{pwr}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{\text{pwr}}\right)^2.$$

(Note: Use the random variate $Z_i$ to solve the exercises.)

a) Obtain the expected value of the estimation function $\hat{t}_{\text{pwr}}$.

b) Derive the variance of the estimation function $\hat{t}_{\text{pwr}}$.

c) Proof that the variance estimation function is unbiased.

d) An alternative estimation function is

$$\hat{t}_{\pi} = \sum_{s} \frac{y_k}{\pi_k} = \sum_{s} \check{y}_k,$$

$s$ denotes the set of **unique** elements in the sample.

i. Proof that this alternative estimation function is unbiased.

ii. Derive a variance estimation function for the alternative estimation function.

iii. Proof that the variance estimation function of this alternative estimation function is unbiased.

6. Show for the small numerical example that the estimation function proposed by Hansen and Hurwitz is unbiased

$$x_1 = 1, \quad x_2 = 3, \quad x_3 = 6$$
$$p_1 = 0.25, \quad p_2 = 0.35, \quad p_3 = 0.4.$$

7. Consider the case of simple random sampling without replacement. Obtain the relative bias of the variance estimation when counterfactually assuming drawing with replacement. Use as an numerical example a sample of size $n = 100$ and a population with $N = 2500$.

8. Consider the population of the PSID data and simple random sampling without replacement.

a) Calculate a symmetric interval based on the Chebichev-inequality for a given probability $1 - \alpha = 0.9$ for the average wage of a sample of size $n = 20$.

b) Calculate a confidence interval an analogue interval for the sample mean of the logarithmic wage. Which interval will be more conservative (why?)?

c) Approximately obtain the true probabilities of the two intervals which are supposed to have a probability of $1 - \alpha = 0.9$ by drawing $B = 10{,}000$ samples.

### 11.6.2 Solutions

1. See text.
2. Estimating mean and variance based on the sample and applying the Chebichev-inequality.
3. The estimate of the variance depends on the specific sample and can be way too small.
4. Indepence of draws results in covariances of 0.
5. Note: Use the random variate $Z_i$ to solve the exercises.

   a)

$$\hat{t}_{\text{pwr}} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{k_i}}{p_{k_i}} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \bar{Z}$$

$$\mathrm{E}(Z_i) = \mathrm{E}\left(\frac{y_{k_i}}{p_{k_i}}\right) = \sum_{U}\frac{y_k}{p_k}p_k = t$$

   b)

$$\mathrm{V}(Z_i) = \mathrm{E}\left[(Z_i - t)^2\right] = \sum_{U}\left(\frac{y_k}{p_k} - t\right)^2 p_k$$

$$\mathrm{V}(\hat{t}_{\text{pwr}}) = \mathrm{V}(\bar{Z}) = \mathrm{V}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \frac{1}{n}\mathrm{V}(Z_i)$$

$$= \frac{1}{n}\sum_{U}\left(\frac{y_k}{p_k} - t\right)^2 p_k$$

   Variance estimator

$$\widehat{\mathrm{V}}(\hat{t}_{\text{pwr}}) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{\text{pwr}}\right)^2$$

$$= \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2$$

   c) Because of independence of draws estimator

$$\widehat{\mathrm{V}}(Z_i) = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2$$

is unbiased

$$\mathrm{E}(\widehat{\mathrm{V}}(Z_i)) = \mathrm{V}(Z_i)$$

and therefore

$$\mathrm{E}(\widehat{\mathrm{V}}(\hat{t}_{\mathrm{pwr}})) = \mathrm{V}(\hat{t}_{\mathrm{pwr}})$$

d) Note that for the number of distinct sample elements $n_s$ it holds that $n_s \leq n$ and $n_s$ is random and

$$\pi_k = 1 - (1 - p_k)^n$$

i.

$$\begin{aligned}\mathrm{E}(\hat{t}_\pi) &= \mathrm{E}\left[\sum_U I_k \frac{y_k}{\pi_k}\right] = \sum_U \mathrm{E}(I_k)\frac{y_k}{\pi_k} \\ &= \sum_U \pi_k \frac{y_k}{\pi_k} = t\end{aligned}$$

ii. Note that covariances are 0 because of independent draws

$$\begin{aligned}\mathrm{V}(\hat{t}_\pi) &= \mathrm{V}\left(\sum_s \frac{y_k}{\pi_k}\right) = \mathrm{V}\left(\sum_U \frac{I_k}{\pi_k} y_k\right) \\ &= \mathrm{V}\left(\sum_U I_k \breve{y}_k\right) = \sum_U \mathrm{V}(I_k \breve{y}_k) \\ &= \sum_U \mathrm{V}(I_k)\breve{y}_k^2 = \sum_U \pi_k(1-\pi_k)\breve{y}_k^2\end{aligned}$$

Expansion using inclusion probabilities

$$\begin{aligned}\widehat{\mathrm{V}}(\hat{t}_\pi) &= \sum_s \frac{1}{\pi_k}\mathrm{V}(I_k)\breve{y}_k^2 = \sum_s \frac{1}{\pi_k}\pi_k\left(1-\pi_k\right)\breve{y}_k^2 \\ &= \sum_s \left(1-\pi_k\right)\breve{y}_k^2 = \sum_s \left(1-\pi_k\right)\frac{y_k^2}{\pi_k^2} \\ &= \sum_s \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k}\right) y_k^2\end{aligned}$$

iii.

$$\mathrm{E}\left[\widehat{\mathrm{V}}(\hat{t}_\pi)\right] = \mathrm{E}\left[\sum_s \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k}\right) y_k^2\right]$$

$$
\begin{aligned}
&= \mathrm{E}\left[\sum_U I_k \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k}\right) y_k^2\right] \\
&= \sum_U \pi_k \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k}\right) y_k^2 \\
&= \sum_U \pi_k (1 - \pi_k) \frac{y_k^2}{\pi_k^2}
\end{aligned}
$$

6. Show for the small numerical example that the estimation function proposed by Hansen and Hurwitz is unbiased

$$
\begin{aligned}
&x_1 = 1, \quad x_2 = 3, \quad x_3 = 6 \\
&p_1 = 0.25, \quad p_2 = 0.35, \quad p_3 = 0.4.
\end{aligned}
$$

```
N <- 3
n <- 2
# combinations with replacement
library(iterpc)
si <- t(getall(iterpc(N, n, replace=TRUE)))
si
M <- choose(N+n-1,n)
y <- c(1,3,6)
p <- c(0.25,0.35,0.4)
# sample probabilities
ps <- apply(si,2,
            function(z) prod(p[z])*
                        length(unique(z)))
# Hansen-Hurwitz
t.pwr <- 1/n*apply(si,2,
                   function(z) sum(y[z]/p[z]))
t.pwr
v.t.pwr <- 1/n/(n-1)*apply(si,2,function(z) {
  zi <- y[z]/p[z]
  sum((zi-mean(zi))^2)}
  )
v.t.pwr
# total estimator
sum(t.pwr*ps)
sum(y)
# variance estimator
sum(v.t.pwr*ps)
```

```
1/n*sum((y/p-sum(y))^2*p)
```

7. Consider the case of simple random sampling without replacement. Obtain the relative bias of the variance estimation when counterfactually assuming drawing with replacement. Use as an numerical example sample size $n = 100$ and a population of $N = 2500$.

```
N <- 2500
n <- 100
f <- n/N
f/(1-f)
```

8. a)
```
d <- read.csv2("psid_2007_sp.csv")
Y <- d$wage
N <- nrow(d)
n <- 20
f <- n/N
m <- mean(Y)
ml <- mean(log(Y))
# wage
s <- sqrt(1/n*(1-f)*var(Y))
a <- sqrt(1/(1-0.9));a
low <- m-a*s
high <- m+a*s
```

b)
```
# log-wage
sl <- sqrt(1/n*(1-f)*var(log(Y)))
lowl <- ml-a*sl
highl <- ml+a*sl
# more conservative, distribution
# closer to normal
```

c)
```
B <- 10000
e1 <- rep(NA,B)
e2 <- rep(NA,B)
for (i in 1:B){
  y <- sample(Y,n)
  e1[i] <- mean(y)
  e2[i] <- mean(log(y))
}
mean(e1>low & e1<high)
```

```
mean(e2>lowl & e2<highl)
```

## 11.7 Stratified sampling

### 11.7.1 Exercises

1. Consider a population with variable $Y$ and attribute $Z$ to be used for stratification.

| $Y$ : | 1 | 2 | 4 | 3 | 5 | 7 | 6 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $Z$ : | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 |

   a) Obtain the mean and the variance of $Y$.

   b) Obtain the means and variances for the $H = 3$ strata.

   c) Calculate the fraction of $\mathrm{V}(Y)$ that is explained by stratification variable $Z$.

   From population $U$ with size $N = 9$ a simple random sample of size $n = 6$ is drawn.

   d) What is the estimating function for the population mean?

   e) Proof that this estimating function is unbiased.

   f) Provide an unbiased estimating function for the variance of the mean estimator.

   g) Calculate the numerical estimator of the variance of the mean estimator.

   h) What is the smallest numerical estimate possible? What is the largest?

2. Now consider stratified sampling with simple random sampling in the strata and identical sample size $n_h = 2/3$ in all strata.

   a) Calculate $|\mathcal{S}|$.

   b) How is the estimating function for the mean defined?

   c) Proof that the estimating function is unbiased.

   d) How is the estimating function for the mean estimator defined?

   e) Calculate the variance of the mean estimator for the population data given in 1.

f) What is the smallest numerical estimate possible? What is the largest?

g) Why would you prefer to use a stratification variable being highly correlated with $Y$?

h) How do you value the quality of the stratification variable $Z$ in the numerical example?

i) Obtain the optimal $Z$ for the numerical example.

3. Assume the specific stratified sample $s$ has been drawn:

| $Y$ : | 1 | 2 | 5 | 7 | 6 | 8 |
|---|---|---|---|---|---|---|
| $Z$ : | 1 | 1 | 2 | 2 | 3 | 3 |

a) Calculate the numerical estimator for the population mean.

b) Calculate the numerical estimator for the variance of the mean estimator.

4. Consider the PSID data set and the variables wage $Y$ and eduyears. Generate a classification variable $Z$ with three different values indicating whether the individual obtained less than 12 years of education, 12 but less than 16 years of education and 16 or more years. Use $Z$ to stratify $U$.

a) Obtain two vectors containing the size and the relative size of the strata.

b) Obtain the means and variances for the $H = 3$ strata.

c) Calculate the share of the variance of $Y$ that is explained by the stratification variable $Z$.

d) Calculate the variances for sample size $n = 120$ in the case of

i. Simple random sampling without stratification.

ii. Stratification and proportional sample size $n_h$.

iii. Stratification and optimal sample size $n_h$.

e) Generate one sample for each sampling design and calculate the estimator for the mean and the estimator for the variance of the mean.

f) Draw $B = 10{,}000$ samples for each design and obtain the approximate distributions of the three different estimating functions for the population mean. Comment on your results.

### 11.7.2 Solutions

1. 
```
Y <- c(1, 2, 4, 3, 5, 7, 6, 8, 9)
Z <- c(1, 1, 1, 2, 2, 2, 3, 3, 3)
N <- length(Y)
```

   a)
```
mY <- mean(Y)
vY <- var(Y)
```

   b)
```
Nh <- tapply(Y,Z,length);Nh
Wh <- Nh/N
mYh <- tapply(Y,Z,mean)
vYh <- tapply(Y,Z,var)
sYh <- sqrt(vYh)
```

   c)
```
anova(lm(Y~as.factor(Z)))
# Between strata
Vext <- sum((mYh-mY)^2*Nh)/(N-1);Vext
# Within strata
Vint <- sum(vYh*(Nh-1))/(N-1);Vint
# Fraction explained (%)
round(Vext/vY*100,1)
```

   d)
$$\hat{\bar{y}} = \frac{1}{n}\sum_{k=1}^{N} y_k I_k = \frac{1}{n}\sum_{k\in U} y_k I_k = \frac{1}{n}\sum_{k\in s} y_k$$

   e)
$$\begin{aligned}\mathrm{E}[\hat{\bar{y}}] &= \mathrm{E}\left[\frac{1}{n}\sum_{k=1}^{N} y_k I_k\right] = \frac{1}{n}\sum_{k=1}^{N} y_k \,\mathrm{E}\left[I_k\right] \\ &= \frac{1}{n}\sum_{k=1}^{N} y_k \frac{n}{N} = \frac{1}{N}\sum_{k=1}^{N} y_k = \bar{y}\end{aligned}$$

   f)
$$\mathrm{V}(\hat{\bar{y}}_\pi) = \frac{(1-f)}{n} S_y^2 \quad \text{with} \quad S_y^2 = \frac{1}{N-1}\sum_{U}\left(y_k - \bar{y}\right)^2 .$$

   g)
```
n <- 6
f <- n/N
(1-f)/n*vY
```

h)
```
Yo <- Y[order(Y)];Yo
mean(Yo[1:n])
mean(Yo[(N-n+1):N])
```

2. a)
```
choose(3,2)^3
```

b)

$$\hat{\bar{y}} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_{s_h} \quad \text{with} \quad \bar{y}_{s_h} = \frac{1}{n_h}\sum_{k=1}^{n_h} y_k$$

$$\hat{\bar{y}} = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{n_h} y_k = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} y_k I_{hk}$$

c)

$$\begin{aligned}
\mathrm{E}\left[\hat{\bar{y}}\right] &= \mathrm{E}\left[\frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} y_k I_{hk}\right] \\
&= \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} y_k \,\mathrm{E}\left[I_{hk}\right] \\
&= \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} y_k \pi_{hk} \\
&= \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} y_k \frac{n_h}{N_h} \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{N_h} y_k = \bar{y}
\end{aligned}$$

d)

$$\begin{aligned}
\mathrm{V}\left(\hat{\bar{y}}\right) = \mathrm{V}\left(\frac{\hat{t}_\pi}{N}\right) &= \frac{1}{N^2}\,\mathrm{V}\left(\hat{t}_\pi\right) = \frac{1}{N^2}\sum_{h=1}^{H} \mathrm{V}(\hat{t}_{h\pi}) \\
&= \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S^2_{yU_h}
\end{aligned}$$

$$S^2_{yU_h} = \frac{1}{N_h - 1}\sum_{k=1}^{N_h} (y_k - \bar{y}_{U_h})^2$$

e)
```
nh <- c(2,2,2)
fh <- nh/Nh
1/N^2*sum(Nh^2*(1-fh)/nh*vYh)
```

f)
```
1/N*sum(Nh*unlist(
  tapply(Y,Z,function(z) mean(sort(z)[1:2]))))
1/N*sum(Nh*unlist(
  tapply(Y,Z,function(z) mean(sort(z)[2:3]))))
```

g) Stratifying by highly correlated variable will reduce variance in the strata and therefore reduce variance through the sampling and estimating effect.

h)
```
cor(Y,Z)
Vext/vY
```

i)
```
o <- order(Y)
Z[o]
cor(Y,Z[o])
```

3.
```
y <- c(1, 2, 5, 7, 6, 8)
z <- c(1, 1, 2, 2, 3, 3)
```

a)
```
1/N*sum(Nh*unlist(tapply(y,z,mean)))
```

b)
```
1/N^2*sum(Nh^2*(1-fh)/
             nh*unlist(tapply(y,z,var)))
```

4.
```
d <- read.csv2("psid_2007_sp.csv")
Y <- d$wage
N <- length(Y)
E <- d$eduyears
Z <- cut(E,c(0,12,16,100),right=F,labels=1:3)
```

a)
```
Nh <- unlist(tapply(Y,Z,length));Nh
Wh <- Nh/N;Wh
```

b)
```
mYh <- unlist(tapply(Y,Z,mean));mYh
vYh <- unlist(tapply(Y,Z,var));vYh
```

c)
```
anova(lm(Y~Z))
# var
vY <- var(Y)
```

```
# Between strata
Vext <- sum((mYh-mY)^2*Nh)/(N-1);Vext
# Within strata
Vint <- sum(vYh*(Nh-1))/(N-1);Vint
# Fraction explained (%)
round(Vext/vY*100,1)
```

d) i.
```
n <- 120
f <- n/N
(1-f)/n*vY
```

ii.
```
nh <- round(Wh*n);sum(nh)
fh <- nh/Nh
1/N^2*sum(Nh^2*(1-fh)/nh*vYh)
```

iii.
```
nho <- round(sqrt(vYh)*Nh/
                sum(sqrt(vYh)*Nh)*n*0.999)
sum(nho)
fho <- nho/Nh
1/N^2*sum(Nh^2*(1-fho)/nho*vYh)
```

e)
```
# simple random sampling
y <- sample(Y,n)
mean(y)
(1-f)/n*var(y)
# stratified, prop.
y <- list()
H <- 3
for (i in 1:H) y[[i]] <- sample(Y[Z==i],nh[i])
sum(Wh*unlist(lapply(y,mean)))
1/N^2*sum(Nh^2*(1-fh)/nh*unlist(lapply(y,var)))
# stratified, opt
y <- list()
for (i in 1:H) y[[i]] <- sample(Y[Z==i],nho[i])
sum(Wh*unlist(lapply(y,mean)))
1/N^2*sum(Nh^2*(1-fho)/
            nho*unlist(lapply(y,var)))
```

f)
```
set.seed(1)
B <- 10000
e <- matrix(NA,B,6)
for (b in 1:B){
```

```
  ysi <- sample(Y,n)
  e[b,1] <- mean(ysi)
  e[b,4] <- (1-f)/n*var(ysi)
  ypr <- list()
  for (i in 1:H) ypr[[i]] <- sample(Y[Z==i],
                                    nh[i])
  e[b,2] <- sum(Wh*unlist(lapply(ypr,mean)))
  e[b,5] <- 1/N^2*
            sum(Nh^2*(1-fh)/
                  nh*unlist(lapply(ypr,var)))
  yop <- list()
  for (i in 1:H) yop[[i]] <- sample(Y[Z==i],
                                    nho[i])
  e[b,3] <- sum(Wh*unlist(lapply(yop,mean)))
  e[b,6] <- 1/N^2*
            sum(Nh^2*(1-fho)/
                  nho*unlist(lapply(yop,var)))
  }
colMeans(e)/1000
plot(density(e[,3]))
lines(density(e[,2]),col=3)
lines(density(e[,1]),col=2)
plot(density(e[,6]))
lines(density(e[,5]),col=3)
lines(density(e[,4]),col=2)
```

## 11.8 Cluster sampling

### 11.8.1 Exercises

1. Consider the following population with variable $Y$ and cluster indicator $Z$:

| $Y$ : | 1 | 2 | 4 | 3 | 5 | 7 | 6 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $Z$ : | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 |

   The sample will consist of $n_I = 2$ clusters drawn randomly with identical inclusion probabilities.

   a) Despite fixed $n_I$ in practice sample size $n$ is often a random variate depending on the specific sample chosen. Is $n$ random in this small example?

   b) Obtain the total of the population and the three possible numeric estimators of the population total based on samples with $n_I = 2$.

   c) Obtain the variance of the total estimator based on the complete enumeration as well as analytically.

   d) Compare the variance of the three esimators for the population mean numerically.

      - Simple random sampling with sample size $n = 6$
      - Stratified sampling with identical sampling fraction $f_h$ in all strata and sample size $n = 6$. $Z$ indicates the partition of $U$ into $H$ strata.
      - Cluster sampling with $n_I = 2$ and identical inclusion probabilities for all clusters.

2. Assume the following cluster sample has been drawn:

| $Y$ : | 1 | 2 | 4 | 3 | 5 | 7 |
|---|---|---|---|---|---|---|
| $Z$ : | 1 | 1 | 1 | 2 | 2 | 2 |

   a) Calculate the numeric estimator of the population mean.

   b) Calculate the variance estimator of the mean estimator.

   c) Explain the variance decomposition in the context of cluster sampling. What is the share of the between cluster (within cluster) variance of the total variance for the population given in exercise 1?

d) Do you regard the clustering in the example as optimal?

e) How would you optimally aggregate units towards clusters?

3. Consider the PSID data set. Sort the data according to sector, years of education and age. Build $n_I = 200$ clusters of identical cluster size $n_i = 5$.

a) Obtain a vector of length $N_I$ containing the cluster totals.

b) Draw a cluster sample of size $n_I = 20$ and calculate the numeric estimator of the population mean and of the estimator of its variance.

c) Obtain the approximate distributions of the estimator of the population mean and of the estimator of its variance. Describe both distributions using descriptive statistics.

d) Compare the distributions with the corresponding distributions for simple random sampling with identical sample size $n = 100$.

### 11.8.2 Solutions

1.
```
Y <- c(1, 2, 4, 3, 5, 7, 6, 8, 9)
Z <- c(1, 1, 1, 2, 2, 2, 3, 3, 3)
N <- length(Y)
NI <- 3
nI <- 2
fI <- nI/NI
```

a) Identical size of clusters, $n$ is fixed.

b)
```
tY <- sum(Y)
si <- combn(3,2);si
htY <- NI/nI*apply(si,2,
                   function(z) sum(Y[Z%in%z]))
htY
```

c)
```
mean(htY)
var(htY)*(NI-1)/NI
ti <- tapply(Y,Z,sum);ti
vhtY <- Ni^2*(1-fI)/nI*var(ti);vhtY
```

d)

- ```
  n <- 6
  f <- n/N
  (1-f)/n*var(Y)
  ```

- ```
  Nh <- tapply(Y,Z,length)
  Wh <- Nh/N
  vYh <- tapply(Y,Z,var)
  nh <- round(Wh*n)
  fh <- nh/Nh
  1/N^2*sum(Nh^2*(1-fh)/nh*vYh)
  ```

- ```
  1/N^2*NI^2*(1-fI)/nI*var(ti)
  ```

2. ```
   y <- c(1, 2, 4, 3, 5, 7)
   z <- c(1, 1, 1, 2, 2, 2)
   ```

   a) ```
      Ni/ni*sum(y)
      ```

   b) ```
      Ni^2*(1-fi)/ni*var(tapply(y,z,sum))
      ```

   c) ```
      anova(lm(Y~Z))
      # var
      vY <- var(Y)
      # Between cluster
      mYh <- tapply(Y,Z,mean)
      mY <- mean(Y)
      Vext <- sum((mYh-mY)^2*Nh)/(N-1);Vext
      # Within cluster
      Vint <- sum(vYh*(Nh-1))/(N-1);Vint
      # Fraction explained (%)
      round(Vext/vY*100,1)
      ```

      Only variance between clusters is relevant for variance.

   d) Not optimal, main fraction of variance is between clusters.

   e) Clusters with equal totals.

3. ```
   d <- read.csv2("psid_2007_sp.csv")
   N <- nrow(d)
   o <- order(d$sector,d$eduyears,d$age)
   Y <- d$wage[o]
   NI <- 200
   ```

```
Z <- rep(1:NI,each=5)
```

a)
```
ti <- tapply(Y,Z,sum)
```

b)
```
nI <- 20
sti <- sample(ti,nI)
1/N*NI/nI*sum(sti)
1/N^2*NI^2*(1-fI)/nI*var(sti)
```

c)
```
B <- 10000
e <- rep(NA,B)
v <- e
for (i in 1:B){
  sti <- sample(ti,nI)
  e[i] <- 1/N*NI/nI*sum(sti)
  v[i] <- 1/N^2*NI^2*(1-fI)/nI*var(sti)
  }
plot(density(e))
plot(density(v))
summary(e)
summary(v)
```

d)
```
e2 <- rep(NA,B)
v2 <- e2
n <- 100
f <- n/N
for (i in 1:B){
  yi <- sample(Y,n)
  e2[i] <- mean(yi)
  v2[i] <- (1-f)/n*var(yi)
  }
plot(density(e),ylim=c(0,8e-05))
lines(density(e2),col=2)
plot(density(v))
lines(density(v2),col=2)
```

## 11.9 Auxiliary variables

### 11.9.1 Exercises

1. Consider the following small population with variable $Y$ and auxiliary variable $X$:

| $Y$ : | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $X$ : | 1 | 2 | 4 | 3 |

   a) Calculate $t_y$, $t_x$, $r = t_y/t_x$, $\mathrm{Cov}(Y, X)$ and $\mathrm{Cor}(Y, X)$.

   The population total $t_y$ shall be estimated based on a simple random sample of size $n = 2$.

   b) Obtain the number of possible samples $|\mathcal{S}| = M$.

   c) Obtain for all $M$ possible samples the free estimator $\hat{t}$ and the ratio estimator $\hat{t}_r = \hat{t}_y \frac{t_x}{\hat{t}_x}$.

   d) Calculate the expected value of $\hat{t}$ and $\hat{t}_r$ by complete enumeration.

   e) Calculate the variance of $\hat{t}$ and $\hat{t}_r$ by complete enumeration.

   f) Obtain for all $M$ possible samples the estimator of the variance of $\hat{t}$ and the estimator of the approximate variance of (obtained by first order Taylor series approximation) $\hat{t}_r$.

   g) Obtain the mean square error of both estimating functions for the population total. Which estimating function would you prefer?

   h) Obtain the variance of the free total estimator $\hat{t}$ according to the following expression

$$\mathrm{V}\left(\hat{t}_\pi\right) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_{y,U}^2$$

   and the approximate variance of the ration estimator according to

$$\mathrm{AV}\left(\hat{t}_{\mathrm{ra}}\right) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left(S_{y,U}^2 + r^2 S_{x,U}^2 - 2r S_{xy,U}\right)$$

2. Consider the PSID data set. Use wage as variable $Y$ and age as auxiliary variable $X$.

a) Do you expect the ratio estimator of the population total $t_y$ to be more efficient than the free estimator of the total?

b) Draw $B = 10{,}000$ simple random samples from the population of size $n = 50$ and calculate for all samples both estimators $\hat{t}_r$ and $\hat{t}$.

c) Compare the approximate distributions of both estimating function graphically.

d) Which estimating function has smaller variance?

### 11.9.2 Solutions

1. 
```
Y <- c(2, 4, 6)
X <- c(1, 2, 4)
```

a)
```
tY <- sum(Y)
tX <- sum(X)
r <- tY/tX
cXY <- cov(X,Y)
rXY <- cor(X,Y)
```

b)
```
N <- 3
n <- 2
M <- choose(N,n)
```

c) Obtain for all $M$ possible samples the free estimator $\hat{t}$ and the ratio estimator $\hat{t}_r = \hat{t}_y \frac{t_x}{\hat{t}_x}$.

```
si <- combn(N,n)
htY <- N*apply(si,2,function(z) mean(Y[z]));htY
htX <- N*apply(si,2,function(z) mean(X[z]));htX
htYr <- htY*tX/htX;htYr
```

d)
```
mean(htY)
mean(htYr)
```

e)
```
vhtX <- var(htX)*(M-1)/M
vhtY <- var(htY)*(M-1)/M
vhtYr <- var(htYr)*(M-1)/M
chtXY <- mean((htY-tY)*(htX-tX))
```

f)
```
mean((htY-tY)^2) # equal to variance, no bias
mean((htYr-tY)^2) # greater than variance
```

```
vhtYr+(mean(htYr)-tY)^2 # mse=var+bias^2
```

g)
```
# SI
f <- n/N
N^2*(1-f)/n*var(Y)
# ratio
vhtY+r^2*vhtX-2*r*chtXY
N^2*(1/n-1/N)*(var(Y)+r^2*var(X)-2*r*cov(X,Y))
```

h)
```
# for all samples
v <- rep(NA,M)
vr <- v
for (i in 1:M){
  sii <- si[,i]
  y <- Y[sii]
  x <- X[sii]
  hty <- N*mean(y)
  htx <- N*mean(x)
  hr <- hty/htx
  v[i] <- N^2*(1-f)/n*var(y)
  vr[i] <- N^2*(1/n-1/N)*
           (var(y)+hr^2*var(x)-2*hr*cov(x,y))
}
v
vr
mean(v) # unbiased
mean(vr) # biased
```

2.
```
d <- read.csv2("psid_2007_sp.csv")
Y <- d$wage
X <- d$age
```

a)
```
cor(X,Y)
0.5*(sd(X)/mean(X))/(sd(Y)/mean(Y))
```

Ratio estimator is expected to be slightly more efficient.

b)
```
B <- 10000
t <- rep(NA,B)
tr <- t
N <- nrow(d)
n <- 50
```

```
U <- 1:N
tX <- sum(X)
for (i in 1:B){
  s <- sample(U,n)
  y <- Y[s]
  x <- X[s]
  hty <- N*mean(y)
  htx <- N*mean(x)
  t[i] <- hty
  tr[i] <- hty*tX/htx
}
```

c)
```
plot(density(tr))
lines(density(t),col=2)
```

d)
```
var(t)
var(tr)
```

## 11.10 Regression

### 11.10.1 Exercises

1. Consider a population with two variables $X$ and $Y$:

| | | | | | |
|---|---|---|---|---|---|
| $X:$ | 6 | 8 | 9 | 10 | 17 |
| $Y:$ | 4 | 8 | 7 | 10 | 11 |

   a) Calculate $t_y$, $t_x$, $s_X^2$, $s_Y^2$, $\text{Cov}(Y,X)$, $\text{Cor}(Y,X)$.

   b) Calculate the population parameters $B_1$ and $B_2$.

   c) Calculate population residuals $E_k$.

2. Consider the following sampling design:

$$s_1 : \{u_1, u_2, u_3\}\ p(s_1) = 0.5$$
$$s_2 : \{u_1, u_2, u_5\}\ p(s_2) = 0.3$$
$$s_3 : \{u_3, u_4, u_5\}\ p(s_3) = 0.2$$

   a) Obtain the first order inclusion probabilities $\pi_k$ and second order inclusion probabilities $\pi_{k,l}$.

   b) Obtain the covariances $\Delta_{kl}$?

   c) Estimate based on sample $s_1$ the $\pi$-estimators $\widehat{B}_1$ and $\widehat{B}_2$.

   d) Obtain for sample $s_1$ the estimated residuals $e_k$.

   e) Estimate the variance estimator $\widehat{\text{V}}(\widehat{B}_2)$ based on sample $s_1$.

3. Now consider the case of simple random sampling with sample size $n = 3$.

   a) Obtain the number of possible samples $|\mathcal{S}| = M$.

   b) Obtain the inclusion probabilities $\pi_k$ and $\pi_{kl}$?

   c) Obtain the covariances $\Delta_{kl}$?

   d) Consider the specific sample $s = \{u_1, u_2, u_3\}$ and obtain for the SI-design the $\pi$-estimators $\widehat{B}_1$ and $\widehat{B}_2$.

   e) Estimate the residuals $e_k$ for the specific sample.

   f) Obtain the variance estimator $\widehat{\text{V}}(\widehat{B}_2)$ for the specific example.

4. Use the PSID data file and consider the variables wage $Y$ and age $X$.

   a) Obtain the population regression parameters $B_1$ and $B_2$ for the linear regression of wages on age.

   b) Draw $B = 1000$ samples of size $n = \{20, 50, 100, 500\}$ and plot the approximate distribution of slope estimator $\widehat{B}_2$ for the different sample sizes.

### 11.10.2 Solutions

1. 
```
X <- c(6, 8, 9, 10, 17)
Y <- c(4, 8, 7, 10, 11)
```

   a)
```
tY <- sum(Y)
tX <- sum(X)
vY <- var(Y)
vX <- var(X)
cXY <- cov(X,Y)
rXY <- cor(X,Y)
```

   b)
```
reg <- lm(Y~X)
B <- reg$coef
```

   c)
```
E <- reg$resid
```

2. 
```
N <- length(Y)
U <- 1:N
si <- cbind(c(1,2,3),c(1,2,5),c(3,4,5))
M <- ncol(si)
p <- c(0.5,0.3,0.2)
```

   a)
```
I.k.s <- function(k,s) as.numeric(k %in% s)
D <- apply(si,2,function(s) I.k.s(U,s));D
pi.k <- as.vector(D%*%p);pi.k
D2 <- array(NA,dim=c(N,N,M))
for(i in 1:M) D2[,,i] <- D[,i]%o%D[,i]
pa <- array(rep(p,each=N*N),c(N,N,M))
pi.kl <- apply(D2*pa,c(1,2),sum)
round(pi.kl,2)
```

b)
```
D.kl <- matrix(NA,N,N)
for (i in 1:N){
  for (j in 1:N){
  D.kl[i,j] <- pi.kl[i,j]-pi.k[i]*pi.k[j]
  }}
round(D.kl,2)
```

c)
```
s1 <- si[,1]
y <- Y[s1]
x <- cbind(1,X[s1])
Ipii <- solve(diag(pi.k[s1]))
hB <- solve(t(x)%*%Ipii%*%x)%*%t(x)%*%Ipii%*%y
```

d)
```
e <- y-x%*%hB
```

e)
```
q <- ncol(x)
V1 <- matrix(NA,q,q)
pi.k1 <- pi.k[s1]
D.kl1 <- D.kl[s1,s1]
for (i in 1:q){
  for (j in 1:q){
    xepi.i <- as.vector(x[,i]*e/pi.k1)
    xepi.j <- as.vector(x[,j]*e/pi.k1)
    V1[i,j] <- sum(D.kl1*(xepi.i%o%xepi.j))
  }
  }
round(V1,2)
Ti <- solve(t(x)%*%x)
AV1.B <- Ti%*%V1%*%Ti
round(AV1.B,3)
```

3. a)
```
n <- 3
M <- choose(N,n)
```

b)
```
pik <- n/N
pikl <- n/N*(n-1)/(N-1)
```

c)
```
Dkl <- pikl-pik^2
```

d)
```
s1 <- c(1:3)
y <- Y[s1]
x <- X[s1]
```

```
hB2 <- sum((y-mean(y))*(x-mean(x)))/
       sum((x-mean(x))^2)
hB1 <- mean(y)-mean(x)*hB2
```

e)
```
e <- y-hB1-x*hB2
```

f)
```
f <- n/N
(1-f)*n/(n-1)*sum((x-mean(x))^2*e^2)/
sum((x-mean(x))^2)^2
```

4.
```
d <- read.csv2("psid_2007_sp.csv")
Y <- d$wage
X <- d$age
N <- length(Y)
```

a)
```
lm(Y~X)$coef
```

b)
```
B <- 1000
e <- matrix(NA,B,4)
for (i in 1:B){
  s20 <- sample(N,20)
  s50 <- sample(N,50)
  s100 <- sample(N,100)
  s500 <- sample(N,500)
  e[i,1] <- cov(Y[s20],X[s20])/var(X[s20])
  e[i,2] <- cov(Y[s50],X[s50])/var(X[s50])
  e[i,3] <- cov(Y[s100],X[s100])/var(X[s100])
  e[i,4] <- cov(Y[s500],X[s500])/var(X[s500])
 }
colMeans(e)
plot(density(e[,4]),xlim=c(-1000,3000))
for (i in 1:3) lines(density(e[,i]),col=(i+1))
```