Chapter 1

Exercise 1.1.
c. It is a complete enumeration of the population

Exercise 1.2.
a. It is important to also ask people who do not go to the swimming pool. Also children belong to the target population.

Exercise 1.3.
b.

Exercise 1.4.
d.

Exercise 1.5.
c. Quota sampling did produce a sample that was reasonably representative.

Chapter 2

Exercise 2.1.
a. $\frac{N}{N-1}\sigma^2 = \frac{N}{N-1}\frac{1}{N}\sum_{k=1}^{N}(Y_k - \bar{Y})^2 = \frac{1}{N-1}\sum_{k=1}^{N}(Y_k - \bar{Y})^2 = S^2$

Exercise 2.2.
b.

Exercise 2.3.
a. Under-coverage (single households are missing) and over-coverage (there are multi-person households in the frame)

Exercise 2.4.
This is not a good way to obtain a good estimate. Forms will only be completed by employees reading the newsletter.

Exercise 2.5.
a.

Exercise 2.6.
$$\sum_{k=1}^{N}\pi_k = \sum_{k=1}^{N}E(a_k) = E\left(\sum_{k=1}^{N}a_k\right) = E(n) = n.$$

Exercise 2.7.
d.

Exercise 2.8.
c.

Chapter 3

Exercise 3.1.
- The answer will be affected by memory errors.
- It is a double negative question
- It is a double question

- It is a leading question

Exercise 3.2.
- The meaning of "how often" is not clear
- The meaning of "lately" is unclear
- There will be a memory error
- The meaning of "private pension insurance" will not clear

Exercise 3.3.
- Advantage of open question: respondents may overlook packages or mention wrong packages.
- Disadvantage of open question: respondents may overlook packages, difficult to analyze.
- Advantage of check-all-that-apply question: respondents do not overlook packages, easy to analyze.
- Disadvantage of check-all-that-apply question: The list my be very long.

Exercise 3.4.
"Regularly" can mean two things: (1) with a high frequency, and (2) at fixed points in time.

Exercise 3.5.
Offering a category "don't know" will give people the possibility to avoid giving an opinion. They take the easy way out. If "don't know" is not offered, people without an opinion cannot answer the question properly.

Exercise 3.6.

10. Have you drunk any alcoholic beverages in the last week?	O Yes O No	 Goto 13
11. Have you drunk any wine in the last week?	O Yes O No	 Goto 13
12. How many glasses of wine did you drink last week?	. . . . .	
13. Have you smoked ant cigarettes last week?	O Yes O Nee	

Exercise 3.7.
In case of an open question, people tend to forget some magazines (e.g. tv-guides).
The list of magazines for a closed may be very long. This makes it difficult to choose the write magazines.
It may be better to separate the magazines in different category, and to ask a closed question of each category.

Exercise 3.7.
Apparently, people choose the "don't know" option to avoid having to formulate an opinion.

Chapter 4

Exercise 4.1.
06 75 46 15 23 31 36 38 44 26 62 89 84 38 50 43 83 76 73 70 65 44 09 67 28
Only underscored numbers are eligible. The other numbers are either to large (over 80) or have already been drawn.

Exercise 4.2.a.

Formula 4.1.10: $V(p) = \frac{1-\frac{1}{20}}{50}\frac{1000}{999}(36)(64) = 43.81982 = (6.62)^2$

Exercise 4.2.b.

Formula 4.1.11: $v(p) = \frac{1-\frac{1}{20}}{49}(36)(64) = 44.669388 = (6.68)^2$

Exercise 4.2.c.

Formula 4.1.11: $v(p) = \frac{1-\frac{1}{20}}{49}\ (28)(72) = 39.085714 = (6.25)^2$

Exercise 4.2.d.

Formula 4.1.11: $v(p) = \frac{1-\frac{1}{20}}{49}\ (44)(56) = 47.771429 = (6.91)^2$

Exercise 4.3.a.

Sample mean: $\bar{y} = 77.75$

Estimated variance of the sample mean:

$v(\bar{y})$ = (1 / 20 – 1 / 10 000) 11.07332^2 = 6.118659

Estimated standard error:

$s(\bar{y})$ = $\sqrt{6.118659}$ = 2.4736

Exercise 4.3.b.

(77.75 – 1.96 × 2.4736 ; 77.75 + 1.96 × 2.4736) = (72.9 ; 82.6)

The mean satisfaction index could be 80 or higher, since the value of 80 is contained in the confidence interval.

Exercise 4.3.c.

Application of formula (4.1.18) gives:

$$n \geq \frac{1}{\left(\frac{M}{1.96\text{s}}\right)^2 + \frac{1}{N}} = \frac{1}{\left(\frac{2}{1.96 \times 11{,}07332}\right)^2 + \frac{1}{10\,000}} = 116.39206$$

So *n* must at least be equal to 117.

Exercise 4.4.

Formula 4.1.15: $n \geq 1 / \left(\frac{999}{1000} (\frac{1.5}{1.96})^2 \frac{1}{(30)(70)} + \frac{1}{N} \right) = 782.1$

The approximate formula 4.1.16 would produce: $n \geq (\frac{1.96}{1.5})^2 (30)(70) = 3585.5$.

This sample size is much larger than the population size! This answer is wrong. The formula cannot be applied because the population size is too small.

Exercise 4.5.

b. (selection probabilities may be based on other things than auxiliary variables)

Exercise 4.6.a.

The mean iss: $\bar{Y}$ = 24034 / 48 = 500.7.

The standard deviation is: *S* = √ 25145.7 = 158.6.

Exercise 4.6.b.

Selected elements are: 17, 42, 45, 43, 14, 40, 9, 26.

Ordered: 9, 14, 17, 26, 40, 42, 43, 45,

Lengths (process data row-wise): 640, 307, 663, 366, 344, 342, 627, 612

Mean: 3901 / 8 = 487.6

Standard deviation: √ 25446.0 = 159.5

Exercise 4.6.c.

Selected elementsn: 3, 9, 15, 21, 27, 33, 39, 45

Lengths (row-wise): 664, 640, 633, 675, 699, 698, 663, 612

Mean: 5254 / 8 = 656.75.

Standard deviation: $\sqrt{731.9} = 27.1$.

Exercise 4.6.d.
The results of the simple random sample are close to the population values. In case of the systematic sample something goes wrong. There is a repeating pattern in the population that corresponds to the step length. Therefore, only large values are selected. The sample mean is too large and the standard deviation too low.

Exercise 4.7.a.

$$V(\bar{y}) = (\frac{1}{n} - \frac{1}{N}) S^2 = (\frac{1}{8} - \frac{1}{20}) \times 1653.4211 = 124.00658$$

$$S(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{124.00658} = 11.13582$$

$$M = 1.96 \times S(\bar{y}) = 1.96 \times 11.13582 = 21.83$$

Exercise 4.7.b.
To apply the Cumulative Method, First the subtotals of the numbers of trucks have to be computed:

Company	Number of trucks	Cumulative
1	3	3
2	4	7
3	5	12
4	6	18
5	7	25
6	6	31
7	4	35
8	5	40
9	3	43
10	9	52

Company	Number of trucks	Cumulative
11	12	64
12	20	84
13	4	88
14	3	91
15	5	96
16	8	104
17	7	111
18	3	114
19	4	118
20	3	121

To draw the sample, recipe 4.3.1 is used. Random values are drawn from the interval $(0, T_N] = (0, 121]$. Such values are obtained by multiplying $(1 - u)$ by 121, where u is a random value from [0, 1):

Element	u	121 (1 – u)	Sequence number	y	x	z = $\frac{y}{x}$
1	0.,314	83.006	12	195	20	9.750
2	0.658	41.382	9	29	3	9.667
3	0.296	85.184	13	42	4	10.500
4	0.761	28.919	6	62	6	10.333
5	0.553	54.087	11	124	12	10.333
6	0.058	113.982	18	25	3	8.333
7	0.128	105.512	17	71	7	10.143
8	0.163	101.277	16	83	8	10.375

The estimate of the population mean is obtained by applying formula (4.3.13):
$\bar{y}_{OK} = \bar{X}\,\bar{z} = 6.05 \times 9.929 = 60.07$.

Exercise 4.7.c.
The estimate for the variance of the estimator is obtained by applying formula (4.3.16):

$$v(\bar{y}_{OK}) = \frac{\bar{X}^2}{n} s_z^2 = \frac{(6.05)^2}{8} 0.5064 = 2.317 .$$

$$s(\bar{y}_{OK}) = \sqrt{v(\bar{y}_{OK})} = \sqrt{2.317} = 1.522 .$$

The margin of the 95% confidence interval is: 1.96 × 1.522 = 2.98

Exercise 4.7.d.
The variance of the estimator in the unequal probability sample is much smaller. The is caused by the strong relationship between the amount of transported goods and the number of trucks.
The 95% confidence interval is (57.09 ; 63.06). This interval indeed contains the true value (60,5).

Exercise 4.8.a.
The step length is equal to $F = N/n = 9/3 = 3$. Three starting values are possible: b=1, b=2 or b=3. The three possible samples are (1, 1, 1), (2, 2, 2) en (3, 3, 3).
In all three cases: $\sigma_b^2 = 0$. So the variance is equal to σ^2. This value is equal to 6 / 9 = 0.667.

Exercise 4.8.b.
The variance is minimal if $\sigma_b^2 = \sigma$ holds for each sample. This is the case for the sequence 1, 1, 1, 2, 2, 2, 3, 3, 3. Then, the variance is 0.

Exercise 4.8.c.
Situation a: for each sample $s^2 = 0$. This creates an impression of a very precise estimator (which is not the case).
Situation b: for each sample $s^2 = 1$. Therefore, the variance of the estimator is estimated by:
(1/3 – 1/9) × 1 = 0.222. However, the real value of the variance is 0,667.
In both cases, a wrong conclusion is drawn.

Chapter 5

Exercise 5.1.
b. Note that d does not apply, because he did not apply random sampling**.**

Exercise 5.2.
c. See expression (5.1.19)

Exercise 5.3.
c.

Exercise 5.4.a.
Variance: $V(\bar{y}) = \frac{1-f}{n} S^2 = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 = \left(\frac{1}{400} - \frac{1}{25000}\right) \times 960000 = 2361.6$
Standard error: $S(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{2361.6} = 48.596296$
Margin: $1.96 \times S(\bar{y}) = 1.96 \times 48.596296 = 95.24874$

Exercise 5.4.b.
$n_1 = \frac{N_1}{N} n = \frac{15000}{25000} 400 = 240$; $n_2 = \frac{N_2}{N} n = \frac{10000}{25000} 400 = 160$
Furthermore: $f_1 = \frac{n_1}{N_1} = \frac{240}{15000} = 0.0016$; $f_2 = \frac{n_2}{N_2} = \frac{160}{10000} = 0.0016$
Expression (5.1.17):
$V(\bar{y}_S) = \left(\frac{15000}{25000}\right)^2 \frac{1-0.0016}{240} 40000 + \left(\frac{10000}{25000}\right)^2 \frac{1-0.0016}{160} 640000 = 698.88$
$S(\bar{y}_S) = \sqrt{V(\bar{y}_S)} = 26.245$
Margin M = 1.96 × 26.245 = 51.44

Exercise 5.4.c.
$N_1S_1 = 15\,000 \times 200 = 3\,000\,000$
$N_2S_2 = 10\,000 \times 800 = 8\,000\,000$
$N_1S_1 + N_2S_2 = 11\,000\,000$

$$n_1 = \frac{3000000}{11000000}\,400 = 109\,;\; n_2 = \frac{8000000}{11000000}\,400 = 291$$

Expression (5.1.17):

$$V(\bar{y}_S) = \left(\frac{15000}{25000}\right)^2 \frac{1-\frac{109}{15000}}{109}\,40000 + \left(\frac{10000}{25000}\right)^2 \frac{1-\frac{291}{10000}}{291}\,640000 = 472.8$$

$$S(\bar{y}_S) = \sqrt{V(\bar{y}_S)} = 21.744$$

Margin $M = 1.96 \times 21.744 = 42.62$

Exercise 5.4.d.
Simple random sample: $V = 2362$; $S = 49$; $M = 95$
Stratification, proportional allocation: $V = 689$; $S = 26$; $M = 51$
Stratification, optimal allocation: $V = 473$; $S = 22$; $M = 43$
Stratified sampling leads to a much more precise estimator than simple random sampling. Precision is higher for optimal allocation than proportional allocation. Apparently, not every stratum is as homogeneous. For optimal allocation, more observations are carried out in less homogeneous strata.

Exercise 5.5.a.
$n_1 = (10 / 35) \times 140 = 40$
$n_2 = (5 / 35) \times 140 = 20$
$n_3 = (20 / 35) \times 140 = 80$

Exercise 5.5.b.
$n_1 = (10{\times}10 / (100 + 35 + 60) \times 140 = (100 / 195) \times 140 = 72$
$n_2 = (5{\times}7 / (100 + 35 + 60) \times 140 = (35 / 195) \times 140 = 25$
$n_3 = (20{\times}3 / (100 + 35 + 60) \times 140 = (60 / 195) \times 140 = 43$

Exercise 5.6.a.

$$V(\bar{y}) = \left(\frac{1-\frac{1}{5}}{6}\right)\left(\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right) = \left(\frac{2}{15}\right)(6.258621) = 0.8345$$

Exercise 5.6.b.

$$Var(\bar{y}_S) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left(\frac{1-\frac{1}{N_h}}{1}\right) S_h^2 =$$

$$= \frac{1}{30^2}\left(3^2\left(\frac{1-\frac{1}{3}}{1}\right)(1) + 3^2\left(\frac{1-\frac{1}{3}}{1}\right)(4) + 9^2\left(\frac{1-\frac{1}{9}}{1}\right)(3) + 6^2\left(\frac{1-\frac{1}{6}}{1}\right)(0.8) + 6^2\left(\frac{1-\frac{1}{6}}{1}\right)(0.8) + 3^2\left(\frac{1-\frac{1}{3}}{1}\right)(9)\right)$$

$= (1 / 900)\,(6 + 24 + 216 + 24 + 24 + 54) = 348 / 900 = 0.3867$

Exercise 5.6.c.
It is not possible to estimate the stratum variances. At least two observations per stratum are needed for this. Expression (5.1.18) annot be used, because the s_h are not defined.

Exercise 5.6.d.

There are $\binom{6}{2} = 15$ possible samples of 2 clusters. The values of the estimator for each cluster is computed:

1.5 3.3 5.4 3.6 2.1 3.6 5.7 3.9 2.4 7.5 5.7 4.2 7.8 6.3 4.5 .

The expected value of the estimator is the average of these values: 4.5. The variance of the estimator is the variance of all these values: 3.3.

Expression (5.2.11): $V(\bar{y}_{CL}) = \left(\frac{M}{N}\right)^2 \frac{1-f}{m} S_C^2 = (\frac{6}{30})^2 \frac{1-\frac{2}{6}}{2}\ 247.5 = 3.3$

The value 247.5 is the adjusted population variance of the six cluster totals.

Exercise 5.7.a.

$V(\bar{y}) = \frac{1-f}{n} S^2 = \frac{1-\frac{12}{60}}{12} 424 = 28.2667$

Exercise 5.7.b.

Expression (5.2.11): $V(\bar{y}_{CL}) = \left(\frac{M}{N}\right)^2 \frac{1-f}{m} S_C^2 = \left(\frac{10}{60}\right)^2 \frac{1-\frac{2}{10}}{2} 64809.1 = 720.1$

Exercise 5.7.c.

Expression (5.2.18): $V(\bar{y}_{CL}) = \frac{1}{Nm} \sum_{h=1}^{M} N_h (\bar{Y}^{(h)} - \bar{Y})^2 = \frac{1}{60 \times 2} (23505.73) = 195.9$

Exercise 5.8.a.

Expression (5.3.16) contains the variance of the estimator. Since $S_{2,h}^2 = 0$ for each cluster, the second term vanishes. Only the first term remains.

Exercise 5.8.b.

The variance decreases as *m* increases. Consequently, as much as possible primary units must be selected and as few as possible secondary units per primary unit.

Exercise 5.9.a.

Expression (5.3.33) contains the variance of the estimator. Since the means of all primary units are the same, the variance between primary units vanishes. The first term in the variance expression is 0. Only the second term remains.

Exercise 5.9.b.

The variance decreases as *m* increases, and n_h increases. Assuming that n_h is very small with respect to N_h, the factor n_h / N_h disappears, and the magnitude of the variance is determined by *m*. Consequently, as much as possible primary units must be selected and as few as possible secondary units per primary unit.

Chapter 6

Exercise 6.1.

a.

Exercise 6.2.

a. (expression 6.2.12)

Exercise 6.3.a.

$v(\bar{y}) = \frac{1-f}{n} s^2 = \frac{1-\frac{100}{5000}}{100} 300^2 = 882$

$s(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{882} = 29.698$

Exercise 6.3.b.

$$b = \frac{c_{XY}}{s_X^2} = \frac{770000}{3200^2} = 0.0751953$$

$$\bar{y}_R = \bar{y} - b(\bar{x} - \bar{X}) = 500 - 0.0751953 \times (4900 - 5000) = 507.195313$$

Exercise 6.3.c.

$$r_{XY} = \frac{c_{XY}}{s_X s_Y} = \frac{770000}{3200 \times 300} = 0.8020833$$

$$v(\bar{y}_R) = \frac{1 - f}{n} s_Y^2 (1 - r_{XY}^2) = \frac{1 - 100/5000}{100} 300^2 (1 - 0{,}802033^2) = 314.5761759$$

$$s(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{314.5761759} = 17.36295$$

Exercise 6.3.c.

The standard error of the regression estimator as much smaller than that of the sample mean. The reason is a strong correlation between target variable and auxiliary variable: $r_{XY} = 0.8$.

Exercise 6.4.a.

$$\bar{y} = \frac{25 \times 348 + 75 \times 1692}{25 + 75} = \frac{135600}{100} = 1356$$

Exercise 6.4.b.

$$s(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{\frac{1 - f}{n} s^2} = \sqrt{\frac{1 - \frac{100}{3410}}{100} \times 895455} = \sqrt{8691{,}85} = 93.23$$

Exercise 6.4.c.

(1356 – 1,96 × 93.23 ; 1356 + 1,96 × 93.23) = (1173.3 ; 1538.7).

One can say with 95% confidence that the mean income in the population will be between 1173 en 1539.

Exercise 6.4.d.

$$\bar{y} = \frac{1210 \times 348 + 2200 \times 1692}{1210 + 2200} = \frac{44143480}{3410} = 1215.097$$

The estimate is smaller because low income households in Old North were under-represented. The post-stratification estimator corrects this.

Exercise 6.4.e.

Expression (6.6.11):

$$\sqrt{v(\bar{y})} = \sqrt{\frac{1 - \frac{100}{3410}}{100} \left(\frac{1210}{3410} \times 48218 + \frac{2200}{3410} \times 724649 \right) + \frac{1}{100^2} \left(\frac{2200}{3410} \times 48218 + \frac{1210}{3410} \times 724649 \right)} =$$

$$\sqrt{4732.956} = 68.8.$$

Exercise 6.4.f.

The standard error of the post-stratification estimator is smaller. Apparently, the two neighborhoods are more homogeneous than the town as a whole.

Exercise 6.5.a.

$\bar{x} = 92.25$ (floor surface, sample)

$\bar{X} = 103.7$ (floor surface, population)

$\bar{y} = 962.5$ (income, sample)

$$\bar{y}_R = 962.5 \times \frac{103.7}{92.25} = 1081.96$$

Exercise 6.5.b.

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\,\bar{x}\,\bar{y}}{\sum x_i^2 - n\,\bar{x}^2} = \frac{12387.5}{1106.75} = 11.19268$$

$$a = \bar{y} - b\bar{x} = 962.5 - 11.19268 \times 92.25 = -70.025$$

$$\bar{y}_{LR} = 962.5 - 11.19268 \times (92.25 - 103.7) = 1090.7$$

Exercise 6.6.a.

p = 1396 / 2000 × 100 = 69.8%.

$$v(p) = \frac{1 - \frac{2000}{40000}}{1999} \times 69.8 \times 30.2 = 1.0018 = 1.001^2$$

95% confidence interval: (69.8 − 1.96 × 1.001 ; 69.8 + 1.96 × 1.001) ≈ (67.8 ; 71.8)

Exercise 6.6.b.

$$\bar{y}_{PS} = \frac{30000}{40000}\left(\frac{1338}{1520}\right) + \frac{10000}{40000}\left(\frac{58}{480}\right) = 0.6904 = 69.04\%$$

$$s_1^2 = \left(\frac{1338}{1520} \times 100\right)\left(100 - \frac{1338}{1520} \times 100\right) = 1054$$

$$s_2^2 = \left(\frac{58}{480} \times 100\right)\left(100 - \frac{58}{480} \times 100\right) = 1062$$

$$v(p) = \frac{1 - \frac{2000}{40000}}{2000}\left(0.75 \times 1054 + 0.25 \times 1062\right) + \frac{1}{2000^2}\left((0.25 \times 1054 + 0.75 \times 1062\right) = 0.502$$

Exercise 6.6.c.

$$n_1 = K\frac{30000 \times \sqrt{900}}{\sqrt{16}} = K\frac{900000}{4} = 225000K;\ n_2 = K\frac{10000 \times \sqrt{1600}}{\sqrt{25}} = K\frac{400000}{5} = 80000K$$

$$K = \frac{20000}{30000 \times \sqrt{900}) \times \sqrt{16} + 10000 \times \sqrt{1600} \times \sqrt{25}} = \frac{20000}{5600000} = 0.003571429$$

n_1 = 225000×0.00357.. = 803.57, n_2 = 80000 × 0,00357.. = 285.71

Rounding: n_1 = 803, n_2 = 286

Chapter 7

Exercise 7.1.

c.

Exercise 7.2.

b.

Exercise 7.3.

c.

Exercise 7.4.

```
DATAMODEL Internet "The Internet Survey"

FIELDS
  PC      "Do you have a PC at home?”: (Yes, No)
  Net     "Is your PC connected to the Internet?: (Yes, No)
  Email   "Do you use your PC for sending and receiving e-mail?": (Yes, No)
  Surf    "Do you use the Internet for surfing on the world wide web?": (Yes, No)
  Browser "Which browser do you use for surfing the web?":
          (IEM "Internet Explorer van Microsoft",
           FFX "Firefox",
           OTH "Other browswer")

RULES
```

```
  PC
  IF PC = Yes THEN
     Net
     IF Net = Yes THEN
        Email Surf
        IF Surf = Yes THEN
           Browser
        ENDIF
     ENDIF
  ENDIF
ENDMODEL
```

Exercise 7.5.
b.

Chapter 8

Exercise 8.1.
c.

Exercise 8.2.
c.

Exercise 8.3.
c.

Exercise 8.4.
a.

Exercise 8.5.
a.

Exercise 8.6.
Imputation of the mean: 3320174 / 8 = 415022.
No ok. All real values are close to 275000 or 550000. The impute values is far away from these values. The mean will not change.

Imputatie of the mean within regions: farm 5: 275848, farm 7: 554196.
Reasonable. The mean will not change.

Random imputation.
Not ok. Imputed value may come from a different region. Then the imputed value will be far away from real values. The mean will change. The expected value of the mean will not change.

Random imputation within regions.
Reasonable. Imputed vakues are close to real values. The mean will change. The expected value of the mean will not change.

Donor imputation. The value from the previous records is copied.
Farm 5: 253604, farm 7: 520534.
Not ok. This only works if donor record happens to come from the same group.

Donor imputation within groups.
Farm 1: 253604, farm 7: 520534.
Reasonable. The mean changes. If the order of the records is completely radom, the expected value does not change.

Use a model that assumes the manure production per pig to be approximately constant:
$Y_i = BX_i$, where Y is the manure production and X is the number of pigs.
B is estimated by $\overline{y} / \overline{x} = 1373$.

Farm 5: 208 × 1373 = 285584, farm 7: 435 × 1373 = 597255.
Reasonable. Both the mean and the expected value of the mean change.

Use a regression model: $Y_i = A+BX_i$, where Y is the manure production and X is the number of pigs.
B is estimated by 1431 and A by -17490.
Farm 5: −17490 + 208 × 1431 = 280158, farm 7: −17490 + 435 × 1431 = 604995.
Reasonable. Both the mean and the expected value of the mean change.

Exercise 8.7.a.
Sample mean for gas consumption: 962.5.
Sample mean for floor space: 92.25.
Ratio estimator: 962.5 × (103.7 / 92.25) = 1082

Exercise 8.7.b.
Estimate for B: 13.45.
Estimate for A: 189.8.
Regression equation: Electricity = 189.8 + 13.45 × Surface.
Imputed value: 189.8 + 13.45 × 99 = 1521.02.

Exercise 8.8.
Record 1: Rules 1 and 3 not satisfied.
One variable appears in both rules: I. Therefore, correct this variable: I = 230

Records 2 and 3 are OK.

Record 4: Rules 1 and 2 not satisfied.
Variables PC and OC appear in both rules.
Correction with rule 1:
Option 1: PC = 170. This is possible.
Option 2: OC = 160. Rules 2 not satisfied.

Exercise 8.9.a.

$$V(\bar{y}) = \frac{1-\frac{n}{N}}{n} S^2 = \frac{1-\frac{1000}{19000}}{1000} 600^2 = 341.053.\ S(\bar{y}) = \sqrt{341.053} = 18.468.$$

Exercise 8.9.b.
The number of available observations m is 90% of 1000 = 900.

$$V(\bar{y}) = \frac{1-\frac{m}{N}}{m} S^2 = \frac{1-\frac{900}{19000}}{900} 600^2 = 381.053.\ S(\bar{y}) = \sqrt{381.053} = 19.521.$$

Exercise 8.9.c.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{m-1}{n-1}\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2 = \frac{900-1}{1000-1}600 = 539.940.$$

The estimated variance of the sample mean becomes:

$$V(\bar{y}) = \frac{1-\frac{n}{N}}{n} s^2 = \frac{1-\frac{1000}{19000}}{1000} \times 539.940^2 = 276.191.\ S(\bar{y}) = \sqrt{276.191} = 16.612.$$

Exercise 8.9.d.
The estimate of the standard error is too low if imputation is not taken into account.

Chapter 9

Exercise 9.1.a.
Frame error. The company should be in the sampling frame at the time of sample selection.

Exercise 9.1.b.
Over-coverage. The company was correctly included in the sampling at the time of sample selection. The company did not belong to the target population on the reference date.

Exercise 9.1.c.
Frame error. The company should be in the sampling frame at the time of sample selection.

Exercise 9.1.d.
Over-coverage. The company was correctly included in the sampling at the time of sample selection. The company did not belong to the target population on the reference date.

Exercise 9.1.e.
Non-response. The company was correctly included in the sampling at the time of sample selection. The company did belong to the target population on the reference date. So the company should respond.

Exercise 9.2.a.
Percentage response: $100 \times 480 / 1000 = 48$ %

Exercise 9.2.b.
Percentage in favor: $100 \times (120 + 80) / (120 + 80 + 40 + 240) = 100$ x $200 / 480 = 41.7$ %

Exercise 9.2.c.
The lower bound is obtained by assuming that all non-respondents are opposed:
$100 \times (200 / 1000) = 20.0$ %
The lower bound is obtained by assuming that all non-respondents are in favor:
$100 \times ((200 + 520) / 1000) = 100 \times 720 / 1000 = 72.0$ %

Exercise 9.3.
Average amount for respondents: 1240.
Average amount for non-respondents: $1.1 \times 1240 = 1364$.
Estimate: $(600 \times 1240 + 200 \times 1364) / 800 = 1271$.

Exercise 9.4.a.
$15200 / (15200 + 3800) = 0.8$

Exercise 9.4.b.
$15200 / (15200 + 3800) – 7600 / (7600 + 11400) = 0.8 – 0.4 = 0.4$

Exercise 9.4.c.
$(19000 / 38000) \times 0.4 = 0.2$.

Exercise 9.5.a.
215 / 10 = 21,5 hours

Exercise 9.5.b.
Average household size in the sample: 70 / 20 = 3.5
Average household size in the respons: 43 / 10 = 4.3
Apparently, small households are under-represented in the response. It is not unlikely that there is a correlation between household size and hours spend on the Internet. So, the estimate may be too high.

Exercise 9.5.c.

X = Household size, Y = hours on the Internet; Ratio estimator: $\bar{y}_Q = \bar{y}_r \frac{\bar{x}}{\bar{x}_r} = 21.5 \times \frac{3.5}{4.3} = 17.5$

Exercise 9.5.d.
The estimate is smaller. That is correct, because the ratio estimator corrects for the under-representation of small households. These households spend less time on the Internet.

Exercise 9.6.a.

$p = 100 \times \frac{300}{1000} = 30\%$.

$$s(p) = \sqrt{\frac{1-f}{n-1} p\,(100-p)} = \sqrt{\frac{1-\frac{1000}{20000}}{999}\ 30 \times 70} = \sqrt{1.997} = 1.413$$

Margin of the 95% confidence interval: $m = 1.96 \times 1.413 = 2.77$.
Confidence interval: (30 – 2.77 ; 30 + 2.77) = (27.23 ; 32.77).

Exercise 9.6.b.
Method of Hansen and Hurvitz:

$$\bar{y}_{HH} = \frac{n\bar{y}_r + m\bar{y}_{nr}}{n + m} = \frac{1000 \times 30 + 1000 \times 10}{2000} = 20\%$$

Exercise 9.6.c.
The bias is more than three times the margin of the confidence interval. Hence, the confidence interval will not contain the true value in many cases. Therefore, the confidence level is very low.

Exercise 9.7.
a. (to avoid a learning effect)

Chapter 10

Exercise 10.1.
c.

Exercise 10.2.a.

$$n \geq \frac{1}{\frac{N-1}{N}\left(\frac{M}{1.96}\right)^2 \frac{1}{P(100-P)} + \frac{1}{N}} = \frac{1}{\frac{2499}{2500}\left(\frac{4}{1.96}\right)^2 \frac{1}{50(100-50)} + \frac{1}{2500}} = 484.2$$

The sample size should at least be equal to 483.

Exercise 10.2.b.
Percentage: 100 × 310 / 380 = 81.579%

Margin: $1.96 \times \sqrt{v(\bar{y})} = 1.96 \times \sqrt{\frac{1-f}{n-1} p(100-p)} = 1.96x\sqrt{\frac{1-\frac{380}{2500}}{379}} 81.6 \times 18.4 = 1{,}96 \times 1.833 = 3.592$

Confidence interval: (81.579 – 3.592 ; 81.579 + 3.592) = (77.987 ; 85.171)

Exercise 10.2.c.
Lower bound: all non-respondents unsatisfied: 100 × (310 + 0) / 500) = 62%
Upper bound: alle non-respondenten tevreden → 100 × (310 + 120) / 500 = 100 × 430 / 500 = 86%

Exercise 10.2.d.
Percentage low educated in the sample: 100 × 346 / 380 = 91.053%
Percentage high educated in the sample: 100 × 34 / 380 = 8,947%

Weight for low educated: 79 / 91.053 = 0.868
Weight for low educated: 21 / 8.947 = 2.347
Weighted percentage: 100 × (306 × 0.868 + 4 × 2.347) / 380 = 100 × 274.996 / 380 = 72.4%

Exercise 10.3.a.
Percentage respons is: 100 × 740 / 1000 = 74 %

Exercise 10.3.a.
Percentage in favor: 100 × (128 + 60) / (128 + 60 + 512 + 40) = 100 × 188 / 740 = 25.4 %

Exercise 10.3.c.
The lower bound is obtained i fit is assumed that all non-respondents are opposed:
100 × (188 / 1000) = 18.8 %
The lower bound is obtained i fit is assumed that all non-respondents are in favor:
100 × ((188 + 260) / 1000) = 100 × 448 / 1000 = 44.8 %

Exercise 10.3.d.
Percentage car owners in response: 100 × 640 / 740 = 86.486 %
Percentage car owners in de population: 80.000 %.
Weight for car owners: 80.000 / 86.486 = 0.925.

Percentage without car in response: 100 × 100 / 740 = 13.514 %
Percentage without car in population: 20.000 %.
Weight for without car: 20.000 / 13.514 = 1.480.

Exercise 10.3.e.

	In favor	Not in favor
Has car	128 × 0.925 = 118.4	512 × 0.925 = 473.6
Har no car	60 × 1.480 = 88.8	40 × 1.480 = 59.2

Percentage in favor: 100 × (118.4 + 88.8) / (118.4 + 88.8 + 473.6 + 59.2) = 100 × 207.2 / 740 = 28 %

Exercise 10.3.f.
The percentage ‘in favor’ increased. That is because those in favour were under-represented due to nonresponse.

Exercise 10.4.a.
130 / 21 = 6.19.

Exercise 10.4.b.
Respons distribution for experience:
Much: 100 × 10 / 21 = 47.619 %
Little: 100 × 11 / 21 = 52.381 %

Respons distribution for age:
Young: 100 × 7 / 21 = 33.333 %
Middle: 100 × 7 / 21 = 33.333 %
Old: 100 × 7 / 21 = 33.333 %

If the response distribution is compared to the population distribution, we see the largest difference for age. So the response is selective with respect to age. This variables should be used for weighting.

Exercise 10.4.c.
Weights for age:
Young: 22 / 33.333 = 0.660
Middle: 30 / 33.333 = 0.900

Old: 48 / 33.333 = 1.440

Weights for experience:
Much: 48 / 47.619 = 1.008
Liitle: 52 / 52.381 = 0.993

Exercise 10.4.d.
Weighting by age: $(0.660 \times 25 + 0.900 \times 42 + 1.440 \times 63) / 21 = 145.02 / 21 = 6.901.$
Weighting by experience: $(1.008 \times 63 + 0.993 \times 67) / 21 = 130.035 / 21 = 6.192$

Exercise 10.4.e.
The response distribution for experience resembles the population distribution. The response is not selective with respect to experience. Weighting does not correct the estimator.
The response distribution for age does not resemble the population distribution. The response is selective with respect to age. Weighting does correct the estimator.

Chapter 11

Exercise 11.1.
b. (no interviewer guidance)

Exercise 11.2
d.

Exercise 11.3.a.
$0.4 \times (5 - 0) = 2.$

Exercise 11.3.b.
$0.6 \times 5 + 0.4 \times 0 = 3.$

Exercise 11.4.a.

$$s(p) = \sqrt{\frac{1-f}{n-1}p(1-p)} = \sqrt{\frac{1-\frac{10000}{1000000}}{9999}60 \times 40} = \sqrt{0.2376} = 0.4875.$$

$$s(p) = \sqrt{\frac{1-f}{n-1}p(100-p)} = \sqrt{\frac{1-\frac{10000}{1000000}}{9999}60 \times 40} = \sqrt{0{,}2376} = 0{,}4875$$

Margin $m = 1.96 \times 0.4875 = 0.96$
Confidence interval: $(60 - 1 ; 60 + 1) = (59 ; 61)$

Exercise 11.4.b.
$p = (3 / 10) \times 40 + (7 / 10) \times 60 = 54\%$

Exercise 11.4.c.

Internet stratum: $v(p) = \frac{1-\frac{10000}{700000}}{9999}60 \times 40 = 0.2366$

Non-internet-stratum: $v(p) = \frac{1-\frac{100}{300000}}{99}40 \times 60 = 24.2343.$

Total: $v(p) = \left(\frac{7}{10}\right)^2 \times 0.2366 + \left(\frac{3}{10}\right)^2 \times 24.2343 = 2.297.$

$s(p)$ = 1.516. M = 2.97. Confidence interval: (51.7 ; 56.3)

Exercise 11.4.d.
The sample form the non-Internet population makes it possible to correct the bias. The estimate is reduced from 60% tot 54%. Apparently people without Internet do less voluntary work. The second confidence interval is wider. This is a consequence of the small sample size in the non-Internet stratum.

Exercise 11.5.a.
Table with marginal totals:

Politician	Votes				
	CDA	VVD	SP	Other	Total
Jan-Peter Balkenende	1980	254	38	218	2490
Jan Marijnissen	135	97	3006	2080	5318
Rita Verdonk	385	1000	183	866	2434
Other politicians	1427	1644	1540	6685	11296
Total	3927	2995	4767	9849	21538

Percentage Balkenende: 100 × 2490 / 21538 = 11.6%
Percentage Marijnissen = 100 × 5318 / 21538 = 24.7%
Percentage Verdonk = 100 × 2434 / 21538 = 11.3%
Marijnissen is the beste politician.

Exercise 11.5.c.
Percentage CDA = 100 × 3927 / 21538 = 18.2%. Weight for CDA = 19.4 / 18.2 = 1.066
Percentage VVD = 100 × 2995 / 21538 = 13.9%. Weight for VVD = 10.7 / 13.9 = 0.770
Percentage SP = 100 × 4767 / 21538 = 22.1%. Weight for SP = 12.1 / 22.1 = 0.548
Percentage Anders = 100 × 9849 / 21 538 = 45.7%. Weight for other = 57.8 / 45.7 = 1.264

Exercise 11.5.d.
Table with weighted frequencies:

Politician	Votes				
	CDA	VVD	SP	Other	Total
Jan-Peter Balkenende	2111	196	21	276	2604
Jan Marijnissen	144	75	1647	2629	4495
Rita Verdonk	410	770	100	1095	2375
Other politician	1521	1266	844	8450	12081
Total	4186	2307	2612	12450	21555

Exercise 11.5.e.
Percentage Balkenende: 100 × 2604 / 21555 = 12.1%
Percentage Marijnissen = 100 × 4495 / 21555 = 20.9%
Percentage Verdonk = 100 × 2375 / 21555 = 11.0%
Marijnissen is still the best politician. His score is a bit lower, because the outcomes are corrected for the over-representation of voters for his party.

Chapter 12

Exercise 12.1.
a.

Exercise 12.2.a.

$$v(p) = \frac{1-f}{n} s^2 = \frac{1-\frac{5}{20}}{5} 2.5 = 0.375.$$

Exercise 12.2.b.

$$v(p) = \frac{1}{n} s^2 = \frac{1}{5} 2.5 = 0.500.$$

Exercise 12.2.c.
The sample in b is with replacement. This allows for more variation in the possible values of the estimator.

Exercise 12.3.a.
p = (3/4) × 40 + (1/4) × 20 = 35%
rhinegate: $v(p) = 40 \times 60 / 499 = 4.810$
Millwood: $v(p) = 20 \times 80 / 499 = 3.206$
Total: $v(p) = (3/4)^2 \times 4.810 + (1/4)^2 \times 3.206 = 2.906$.
$s(p) = 1.705$. $m = 3.341$. CI = (32.7 ; 38.3)

Exercise 12.3.b.
p = (40 + 20) / 2 = 30%.
$v(p) = 30 \times 70 / 999 = 2.102$. $s(p) = 1.405$. $m = 2.842$. CI = (27.2 ; 32.8)

Exercise 12.3.c.
The estimator in 12.3.b does not take the sampling design into account. Therefore a wrong conclusion is drawn. Note that the confidence interval almost do not overlap.

Exercise 12.4.
c.

Exercise 12.5.
b.

Exercise 12.6.
a.

Exercise 12.7.
It is difficult to establish the height of a vertical bar.
It is difficult to compare values at the vertical scale.
It is difficult to compare sectors of pie charts.

Exercise 12.8.
Non-equidistant x/y-scale.
Do not start scales at origin.
Use different y-scales in same plot.
Use symbols the size of which do not reflect true value.
Change the colors of the bars in comparable chart.
Use three-dimensional graphs
Add a lot of chart junk.

Chapter 13

Exercise 13.1.
d.

Exercise 13.2.
c.

Exercise 13.3.a.
The population is divided into 18 groups:

$$\left.\begin{array}{l}
40000\left\{\begin{array}{l}
12000\left\{\begin{array}{l}6000\\6000\end{array}\right.\\
16000\left\{\begin{array}{l}8000\\8000\end{array}\right.\\
12000\left\{\begin{array}{l}6000\\6000\end{array}\right.
\end{array}\right.\\
70000\left\{\begin{array}{l}
20000\left\{\begin{array}{l}
6000\left\{\begin{array}{l}3000\\3000\end{array}\right.\\
8000\left\{\begin{array}{l}4000\\4000\end{array}\right.\\
6000\left\{\begin{array}{l}3000\\3000\end{array}\right.
\end{array}\right.\\
10000\left\{\begin{array}{l}
3000\left\{\begin{array}{l}1500\\1500\end{array}\right.\\
4000\left\{\begin{array}{l}2000\\2000\end{array}\right.\\
3000\left\{\begin{array}{ll}1500\\1500 & \pi_{18}=1500/70000=1.5/70\end{array}\right.
\end{array}\right.
\end{array}\right.
\end{array}\right. \qquad \pi_1 = 6000/70000 = 6/70$$

$$\sum_{i=1}^{18}\pi_i^2 = 4\times(\frac{6}{70})^2 + 2\times(\frac{8}{70})^2 + 4\times(\frac{3}{70})^2 + 2\times(\frac{4}{70})^2 + 4\times(\frac{1{,}5}{70})^2 + 2\times(\frac{2}{70})^2 = \frac{51}{700} = 0.0728571$$

R = 700/51 = 13.7255

Exercise 13.3.b.

$$R = \frac{1}{18\times(\frac{1}{18})^2} = 18$$

Exercise 13.4.

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

The maximum value is obtained if the denominator is minimal. This is the case if F=N/2. Then R=2.
The minimum value is obtained if the denominator is maximal. This is the case if F=1.
Then $R = \frac{N^2}{2-2N+N^2} = \frac{N^2}{(N-1)^2+1}$.

Exercise 13.5.
b.

Exercise 13.6.a.
$K = 2 \times 12 \times 6 \times 10 \times 7 = 10080$. Each F_i has a Poisson distribution with μ = 100800 / 10080 = 10.
Therefore $P(F = 0) = e^{-10} = 0.0000454$ and $P(F = 1) = 10e^{-10} = 0.000454$.
The number of key values with $F_i = 0$ has a binomial distribution with n = 10080 and p = 0.0000454. The expected value is 0.46. The expected number if key values with $F_i > 0$ is 10080 – 0.46 = 10079.54.

Exercise 13.6.b.
The number of key values with $F_i = 1$ has a binomial distribution with $n = 10080$ and $p = 0.000454$. The expected value is 4.6.

Exercise 13.7.a.
File 1:
$K = 600 \times 2 \times 10 \times 7 \times 12 = 1008000$.
$U_p = N \exp(-N/K) = 7000000 \exp(-6.944444) = 6748$.
$U_p / N = 6748 / 7000000 (= e^{-6.944444}) = 0.000964 < 0.001$. The file is OK.

File 2:
$K = 12 \times 2 \times 20 \times 7 \times 12 \times 2 \times 13 = 1048320$.
$U_p = N \exp(-N/K) = 7000000 \exp(-6.677350) = 8814$.
$U_p / N = 8814 / 7000000 (= e^{-6.67735}) = 0.001259 > 0.001$. File is not OK.

Exercise 13.7.b.
$K = 1000000$, $a = 0.00005$, $b = 0.02$.
$U_p = N (1 + Nb)^{-(1+a)} = 7000000 (1 + 7000000 \times 0.02)^{-(1+0.00005)} = 7000000 \times (140001)^{-1.00005} = 50$.
Since $50 / 7000000 < 0.001$, the file is OK.

Exercise 13.7.c.
$N_C = (C^{-1/(1+a)} - 1) / b = (0.001^{-1/1.00005} - 1) / 0.02 = 49933$
The file can only be disseminated if regions have at least 49933 inhabitants.

Exercise 13.8.a.
$K = 2 \times 4 \times 20 \times 40 \times 15 = 96000$.
$R = K = 96000$.

Exercise 13.8.b.
$U_p = Ne^{-N/K} = 600000e^{-600000/96000} = 1158.$
Fraction uniques: $1158 / 600000 = 0.00193 > 0.001$. File is not OK.

Exercise 13.8.c.
$(1 + nb)\frac{n}{K} = 2.1875$
Substitution of $K = 96000$ and $n = 10000$, and solving for b gives $b = 0.002$.
$a = \frac{1}{bK}$. Substitution of $b = 0.002$ and $K = 96000$ gives $a = 0.0052$.
$U_p = N(1 + Nb)^{-(1+a)} = 600000(1 + 600000 \times 0.002)^{-1.0052} = 482.$
Estimated fraction uniques: $482 / 600000 = 0.0008 < 0.001$. The file is OK.

Exercise 13.8.d.
$N_c = \frac{1}{b}(C^{-1/(1+a)} - 1) = \frac{1}{0.002}(0.001^{-1/1.0052} - 1) = 481948.$
Neighborhoods have much less inhabitants. Publication at neighborhood level is not possible.