Simple example on sampling

It is commonly said that an image is worth a thousand words, so here is an example to help you with problems in Handout 14. I will use a a test case the ongoing study on Covid-19 prevalence: the goal is to determine the proportion of persons who have developed antigens, because they have been in contact with the virus, knowingly or unknowingly. Thus, we want to estimate **a proportion.**

The study is based in 90.000 tests, of which about 60.000 are expected to be made (people may refuse or not be available, for various reasons). The following questions and answers may help you to understand the computations required in simple, non-stratified sampling.

What standard error of estimation can be expected from a sample size $n = 60.000$?

If p is totally unknown, we must make provisión for the worst case scenario of $p = q = 0.50$. If T is the total number of positives, our estimate of p will be:

$$\hat{p} = \frac{T}{n}$$

with variance

$$\frac{pq}{n} \leq 0.25/n$$

If we replace n by 60.000 we get a variance of $0.25/60000 = 4.166667 \times 10^{-6}$ which translates to an standard error of ≈ 0.002041.

What would be a 95% confidence interval for p with said sample?

It can be computed as,

$$\hat{p} \pm 1.96 \times 0.002041 = \hat{p} \pm 0.004001$$

Arrived at this point you may ask yourself: do we really need a precision of under one half of a percentage point in the estimation of p? The answer is that probably not, if all we want is the proportion for the whole of Spain; but see below on what are the needs if we want to have results at the provincial level.

Would the results above change much if we account for the finite population of Spain?

Assuming the total population is 47 million people, the variance would be reduced by a factor of

$$\left(1 - \frac{n}{N}\right) = 1 - \frac{60000}{47000000} = 0.9987$$

and the standard deviation by the square root of that: $\sqrt{0.9987} \approx 0.9993615$. Hardly worth considering.

If all we want is an estimation error under 1% with confidence 95%, what sample size is needed?

We would obtain it from:

$$0.01 = 1.96\sqrt{\frac{0.25}{n}}$$

giving

$$n = \frac{1.96^2 \times 0.25}{0.01^2} = 9604$$

We make no correction for a finite population as we have seen it is negligible.

If we knew that in no event more than 10% of the population can possibly be infected, how would the previous sample size change?

The upper bound for pq would now be $0.10 \times 0.90 = 0.09$; then, a sample of

$$n = \frac{1.96^2 \times 0.09}{0.01^2} = 3457$$

would be enough.

We make no correction for a finite population as we have seen it is negligible.

If we want a 95% estimation error at the province level, what sample size do we need to use?

This is tedious to answer by hand, so we can make use of R. A file with the provinces populations has been obtained from INE, that we read here:

```
pop <- read.csv("2852bsc.csv", sep=";", )
pop
```

```
##                   Provincias  Sexo Periodo    Total
## 1                      Total Total    2019 47026208
## 2                 02 Albacete Total    2019   388167
## 3          03 Alicante/Alacant Total    2019  1858683
## 4                  04 Almería Total    2019   716820
## 5              01 Araba/Álava Total    2019   331549
## 6                 33 Asturias Total    2019  1022800
## 7                    05 Ávila Total    2019   157640
## 8                  06 Badajoz Total    2019   673559
## 9            07 Balears, Illes Total    2019  1149460
## 10                08 Barcelona Total    2019  5664579
## 11                  48 Bizkaia Total    2019  1152651
## 12                   09 Burgos Total    2019   356958
## 13                  10 Cáceres Total    2019   394151
## 14                    11 Cádiz Total    2019  1240155
## 15                39 Cantabria Total    2019   581078
## 16     12 Castellón/Castelló Total    2019   579962
## 17              13 Ciudad Real Total    2019   495761
## 18                  14 Córdoba Total    2019   782979
## 19                15 Coruña, A Total    2019  1119596
## 20                   16 Cuenca Total    2019   196329
## 21                  20 Gipuzkoa Total    2019   723576
## 22                   17 Girona Total    2019   771044
## 23                  18 Granada Total    2019   914678
## 24              19 Guadalajara Total    2019   257762
## 25                   21 Huelva Total    2019   521870
## 26                   22 Huesca Total    2019   220461
## 27                     23 Jaén Total    2019   633564
## 28                     24 León Total    2019   460001
## 29                   25 Lleida Total    2019   434930
## 30                     27 Lugo Total    2019   329587
```

```
## 31                 28 Madrid Total    2019  6663394
## 32                 29 Málaga Total    2019  1661785
## 33                 30 Murcia Total    2019  1493898
## 34                31 Navarra Total    2019   654214
## 35                32 Ourense Total    2019   307651
## 36               34 Palencia Total    2019   160980
## 37            35 Palmas, Las Total    2019  1120406
## 38             36 Pontevedra Total    2019   942665
## 39              26 Rioja, La Total    2019   316798
## 40             37 Salamanca Total    2019   330119
## 41 38 Santa Cruz de Tenerife Total    2019  1032983
## 42                40 Segovia Total    2019   153129
## 43                41 Sevilla Total    2019  1942389
## 44                  42 Soria Total    2019    88636
## 45              43 Tarragona Total    2019   804664
## 46                 44 Teruel Total    2019   134137
## 47                 45 Toledo Total    2019   694844
## 48      46 Valencia/València Total    2019  2565124
## 49              47 Valladolid Total    2019   519546
## 50                 49 Zamora Total    2019   172539
## 51               50 Zaragoza Total    2019   964693
## 52                 51 Ceuta Total    2019    84777
## 53                52 Melilla Total    2019    86487
```

We omit the first row and second and third columns:

```
pop <- pop[-1,-(2:3)]
pop
```

```
##                   Provincias   Total
## 2                02 Albacete  388167
## 3         03 Alicante/Alacant 1858683
## 4                 04 Almería  716820
## 5              01 Araba/Álava  331549
## 6                 33 Asturias 1022800
## 7                   05 Ávila  157640
## 8                 06 Badajoz  673559
## 9           07 Balears, Illes 1149460
## 10               08 Barcelona 5664579
## 11                48 Bizkaia 1152651
## 12                 09 Burgos  356958
## 13                10 Cáceres  394151
## 14                  11 Cádiz 1240155
## 15               39 Cantabria  581078
## 16     12 Castellón/Castelló  579962
## 17             13 Ciudad Real  495761
## 18                14 Córdoba  782979
## 19               15 Coruña, A 1119596
## 20                  16 Cuenca  196329
## 21                20 Gipuzkoa  723576
## 22                  17 Girona  771044
## 23                 18 Granada  914678
## 24             19 Guadalajara  257762
## 25                  21 Huelva  521870
## 26                  22 Huesca  220461
```

```
## 27                   23 Jaén  633564
## 28                   24 León  460001
## 29                 25 Lleida  434930
## 30                   27 Lugo  329587
## 31                 28 Madrid 6663394
## 32                 29 Málaga 1661785
## 33                 30 Murcia 1493898
## 34                31 Navarra  654214
## 35                32 Ourense  307651
## 36              34 Palencia  160980
## 37            35 Palmas, Las 1120406
## 38             36 Pontevedra  942665
## 39              26 Rioja, La  316798
## 40             37 Salamanca  330119
## 41 38 Santa Cruz de Tenerife 1032983
## 42                40 Segovia  153129
## 43                41 Sevilla 1942389
## 44                  42 Soria   88636
## 45              43 Tarragona  804664
## 46                 44 Teruel  134137
## 47                 45 Toledo  694844
## 48      46 Valencia/València 2565124
## 49             47 Valladolid  519546
## 50                 49 Zamora  172539
## 51               50 Zaragoza  964693
## 52                  51 Ceuta   84777
## 53                52 Melilla   86487
```

We will complete this data frame with two columns giving the sample sizes required for the desired precision with and without finite population correction.

```
pop <- cbind(pop, n.inf=NA, n.fin=NA)
pop[,"n.inf"] <- round( (1.96^2) * 0.25 / 0.01^2 )
pop[,"n.fin"] <- round( ((1.96)^2 * 0.25) / (0.01^2 + 0.25/pop[,"Total"]) )
```

(We have made the simplification $N = N - 1$.)

Results can be seen next:

```
pop
```

```
##                   Provincias   Total n.inf n.fin
## 2                02 Albacete  388167  9604  9543
## 3         03 Alicante/Alacant 1858683  9604  9591
## 4                  04 Almería  716820  9604  9571
## 5              01 Araba/Álava  331549  9604  9532
## 6                 33 Asturias 1022800  9604  9581
## 7                    05 Ávila  157640  9604  9454
## 8                  06 Badajoz  673559  9604  9568
## 9           07 Balears, Illes 1149460  9604  9583
## 10               08 Barcelona 5664579  9604  9600
## 11                 48 Bizkaia 1152651  9604  9583
## 12                  09 Burgos  356958  9604  9537
## 13                 10 Cáceres  394151  9604  9543
## 14                   11 Cádiz 1240155  9604  9585
## 15              39 Cantabria  581078  9604  9563
## 16     12 Castellón/Castelló  579962  9604  9563
```

```
## 17            13 Ciudad Real  495761  9604  9556
## 18                14 Córdoba  782979  9604  9573
## 19              15 Coruña, A 1119596  9604  9583
## 20                 16 Cuenca  196329  9604  9483
## 21               20 Gipuzkoa  723576  9604  9571
## 22                 17 Girona  771044  9604  9573
## 23                18 Granada  914678  9604  9578
## 24            19 Guadalajara  257762  9604  9512
## 25                 21 Huelva  521870  9604  9558
## 26                 22 Huesca  220461  9604  9496
## 27                   23 Jaén  633564  9604  9566
## 28                   24 León  460001  9604  9552
## 29                 25 Lleida  434930  9604  9549
## 30                   27 Lugo  329587  9604  9532
## 31                 28 Madrid 6663394  9604  9600
## 32                 29 Málaga 1661785  9604  9590
## 33                 30 Murcia 1493898  9604  9588
## 34                31 Navarra  654214  9604  9567
## 35                32 Ourense  307651  9604  9527
## 36               34 Palencia  160980  9604  9457
## 37            35 Palmas, Las 1120406  9604  9583
## 38             36 Pontevedra  942665  9604  9579
## 39              26 Rioja, La  316798  9604  9529
## 40             37 Salamanca  330119  9604  9532
## 41 38 Santa Cruz de Tenerife 1032983  9604  9581
## 42                40 Segovia  153129  9604  9450
## 43                41 Sevilla 1942389  9604  9592
## 44                  42 Soria   88636  9604  9341
## 45              43 Tarragona  804664  9604  9574
## 46                 44 Teruel  134137  9604  9428
## 47                 45 Toledo  694844  9604  9570
## 48      46 Valencia/València 2565124  9604  9595
## 49             47 Valladolid  519546  9604  9558
## 50                 49 Zamora  172539  9604  9467
## 51               50 Zaragoza  964693  9604  9579
## 52                  51 Ceuta   84777  9604  9329
## 53                52 Melilla   86487  9604  9334
```

We see two things: that the finite population correction hardly matters for large provinces, but it becomes significant for Soria, Ceuta, Melilla, where the sample is a non negligible portion of the total population. If we add up the provincial sample sizes, we get a whooping total sample size of nearly half a million, about nine times as much as the projected 60.000 used in the study, which we now realize is by no means excessive.

```
colSums(pop[,3:4])
```

```
##  n.inf  n.fin
## 499408 496029
```

This shows clearly how expensive is the detail: 9604 persons were enough for the estimation of a global p with 95% confidence error of 1%, but if we want the same precision at the province level, 52 times as much is necessary (if you do not account for finite population size) and nearly as much if you make the finite population correction.