Stratified sampling

Purpose

Suppose that, unlike in the previous example, we do not want to estimate prevalence in each of the three Territories, but rather we want to estimate the proportion of infected people in the whole of the Basque Autonomous Community (CAPV). Let's see how we could put to good use stratified sampling principles.

Notation

Let's call the respective proportions of infected people in the three Territories p_A, p_B and p_G: they are unknown, but on account of the aviaalable data believed to be wideliy different. The respective population sizes are $N_A = 331549$, $N_B = 1152651$ and $N_G = 723576$ persons, giving a total population for the CAPV of $N = 2207776$ persons. Hence, the (true, unknown) proportion of infected people in the CAPV is:

$$p = \frac{N_A}{N} \times p_A + \frac{N_B}{N} \times p_B + \frac{N_G}{N} \times p_G = 0.15017 \times p_A + 0.52209 \times p_B + 0.32774 \times p_G$$

Alternatives

Simple random sampling

We can of course take a single sample of size n, count the number of infected people I and estimate p by:

$$\hat{p} = \frac{I}{N}$$

The previoujs example shows us that if we want, for instance, a 95% confidence error not larger that 0.01 in $\hat{p}$, we would need a sample size of about $n = 9604$, the finite population correction being negligible here on account of the large N.

Worst case scenario

The variance of our $\hat{p}$ estimator would be in the worst scenario ($p = q = 0.50$) equal to:

$$\sigma^2 = \frac{0.25}{9604} = 2.60308 \times 10^{-5}$$

again neglecting finite population correction. If we introduce that correction, the variance would be:

```
 (0.25 / 9604) * (1 - 9604/2207776)
```

```
## [1] 2.591758e-05
```

With estimated rates of infected people

As of today (8 may 2020), the numbers of confirmed cases in the three Territories have been: 4480 (Araba), 9627 (Bizkaia) and 2852 (Gipuzkoa). The estimated proportions so far are:

```
cases <- c(4480, 9627, 2852)
pop   <- c(331549, 1152651, 723576)
names(pop) <- names(cases) <- c("Araba","Bizkaia","Gipuzkoa")
```

```
prop  <- cases / pop
prop
```

```
##       Araba     Bizkaia    Gipuzkoa
## 0.013512332 0.008352051 0.003941535
```

These will be different from the estimated proportions after we complete our survey, but give an idea of the relative incidence of the virus across Territories. We cannot know what the true proportion of infected people is until we finish the survey, but for the sake of the example, and given that only a small fraction of the population has been given tests, let us **assume** that the true proportion is 10 times as much:

```
prop <-10 *  prop
prop
```

```
##      Araba    Bizkaia   Gipuzkoa
## 0.13512332 0.08352051 0.03941535
```

which would give a value for the global p of

```
p <- sum(pop * prop) / sum(pop)
p
```

```
## [1] 0.07681486
```

We can then compute the appproximate variance of our estimator as:

$$\sigma^2 = \frac{\hat{p}\hat{q}}{n}$$

which evaluates to:

```
( p*(1-p) / 9604 ) * (1 - 9604/2207776)
```

```
## [1] 7.351713e-06
```

Proportional sampling

One improvement that we could make is to allocate the sample proportionally to the population sizes of the Territories. The reason for this is clear: with random sampling, we might end up with a sample in which some Territories were over-represented or under-represented, which would bias our estimate of p towards (or away from) the p_i $(i = A, B, G)$ of the over-represented (or under-represented) Territorie(s).

It makes more sense to allocate our $n = 9604$ subjects proportionally to the size of the respective populations:

```
pop / sum(pop)
```

```
##     Araba   Bizkaia  Gipuzkoa
## 0.1501733 0.5220869 0.3277398
```

```
n.i <- round( 9604 * (pop / sum(pop)) )
n.i
```

```
##    Araba  Bizkaia Gipuzkoa
##     1442     5014     3148
```

```
sum(n.i)
```

```
## [1] 9604
```

Worst case scenario

Doing that would lead to a variance of our estimator in the worst case scenario ($p = q = 0.50$ within each Territory) of:

$$\sigma^2 = \left(\frac{N_A}{N}\right)^2 \times \frac{0.25}{1442} + \left(\frac{N_B}{N}\right)^2 \times \frac{0.25}{5014} + \left(\frac{N_C}{N}\right)^2 \times \frac{0.25}{3148}$$

or introducing the finite population correction,

$$\sigma^2 = \left(\frac{N_A}{N}\right)^2 \frac{0.25}{1442}\left(1 - \frac{1442}{331549}\right) + \left(\frac{N_B}{N}\right)^2 \frac{0.25}{5014}\left(1 - \frac{501}{1152651}\right) + \left(\frac{N_C}{N}\right)^2 \frac{0.25}{3148}\left(1 - \frac{3148}{723576}\right)$$

which when computed, happens to be:

```
sum(  (pop/sum(pop))^2 * (0.25 / n.i  ) * (1 - n.i/pop) )
```

```
## [1] 2.591758e-05
```

We have no gain from proportional allocation in this case in terms of variance, as would be expected, for in the worse case scenario all strata are exactly alike.

With estimated rates of infected people

In this case, the estimated variance would be:

```
sum(  (pop/sum(pop))^2 * (prop*(1-prop) / n.i  ) * (1 - n.i/pop) )
```

```
## [1] 7.249102e-06
```

We see a marginal improvement over what we would achieve taking a single sample.

With optimally allcoated sample

Now we will allocate the sample optimally, which means in proportion to

$$\frac{N_i \sigma_i}{c_i}$$

where c_i is the cost of sampling stratum i (we assume the cost is the same in all strata). Therefore, we can compute these factors

```
f <- pop * sqrt(prop*(1-prop))
f <- f / sum(f)
f
```

```
##     Araba   Bizkaia  Gipuzkoa
## 0.1977914 0.5565103 0.2456983
```

and allocate the sample in proportion:

```
strat.n.i <- round(9604 * f)
strat.n.i
```

```
##    Araba  Bizkaia Gipuzkoa
##     1900     5345     2360
```

We see that the optimal sample in Araba is 1950, nearly as much as that in Gipuzkoa, which is more than twice as big: the reason is the greater variance in the former Territory. Bizkaia has the largest sample, but not as large in comparison with Araba as its size would suggest, again an effect of the larger Araba variance.

With these sample sizes, the variance of the stratified sampling would be estimaed at

```
sum((pop/sum(pop))^2 * (prop*(1-prop) / strat.n.i  ) * (1 - strat.n.i/pop))
```

```
## [1] 6.982199e-06
```

Compare the share of population in each Territory,

```
pop / sum(pop)
```

```
##     Araba   Bizkaia  Gipuzkoa
## 0.1501733 0.5220869 0.3277398
```

with the share of sample size

```
f
```

```
##     Araba   Bizkaia  Gipuzkoa
## 0.1977914 0.5565103 0.2456983
```

We see that Araba has been oversampled with respecto to its size, on account of the higher incidence of the virus, and Gipuzkoa has been heavily subsampled.

Remarks

It is useful to compare the estimated variances with the hypothesized rates of incidence:

Method	Variance
Simple sampling	7.35171×10^{-6}
Proportional sampling	7.24910×10^{-6}
Optimal stratification	6.98219×10^{-6}

The reduction in variance is not that spectacular, because the strata are not very different (except Gipuzkoa).

It is also worth noting that we have computed these variances with **hypothesized** values of the respective proportions of people infected in the three Territories: the true variance can onlhy be estimated after the survey is done, replacing in the formulae our best estimated values for p_A, p_B and p_G.

However even if the hypothesized proportions were grossly wrong, our allocation of the sample would still be right as long as the relative values of intra-stratum variance are right.

In other words, we have assumed that the true prevalences of infection are 10 times as large as the present proportion of positives with respect to the population. We might be wrong and the true prevalences be 6 or 13 times as much. As long as the values of the σ_i ($i = A, B, G$) are in about the same proportion, nothing much happens: the allocation of the sample remains largely unaffected.