Handout 14

1. A sample survey is carried out to estimate the proportion p of families possessing a certain characteristic. The value of such proportion p is expected to lie between 20 and 60 per cent. Find the sample size necessary to estimate p with an standard error not exceeding 0.02 if:

   - The population is very large.
   - The population is known to consist of 3800 families.

   How would your answer change if you were certain that p does not exceed 0.30?

2. We want to estimate the unemployment rate in the three Territories of the CAPV, with sampling error (at 95% confidence) no larger than 0.02 in each Territory and 0.02 overall.

   We assume the active population sizes equal to 600000, 400000 and 200000 respectively, for Bizkaia, Gipuzkoa and Araba. Unemployment is expected to be below the 22% mark in all cases, with no large differences among the three Territories.

   Compute the sample size needed for each Territory. How does this compare with the sample size required if our only requirement were an overall unemployment rate for the CAPV?

3. A sample of 315 households was drawn from a city area of 15.762 households. Each family was asked about whether it owned or rented the house and also whether it had water supply system. Results were as follows: It is required to estimate:

   Table 1: Availability of piped water versus tenancy status

	Owned	Rented
With piped water	153	121
Without piped water	10	31

   (a) The total number of renting families in the area with piped water and the standard error of estimation.

   (b) A 95% confidence interval for the total number of renting families in the area with piped water.

   Compare results using independent and non-independent (finite population) sampling.

4. A sample survey is carried out to estimate in a very large population the proportion of families in different income classes: under 3000$, between 3000$ and 8000$, and over 8000$. A previous study has shown that the proportions were respectively around 65%, 25% and 10%, and are now expected not to deviate much from such values. Find,

   (a) The sample size necessary to estimate such proportions with an standard error not greater than 0.02.

   (b) The sample size to estimate such proportions so that the 95% confidence error is under 0.03.

5. We want to estimate the total daily market size for liquors in a population made of three ethnic groups, A, B, C, whose consumption is quite different. The percentage of people in the three groups is 50%, 40% and 10%, and the population size is $N = 9000$. Previous surveys have shown average daily intakes for the three groups of (respectively) 0.3, 0.2 and 0.01 deciliters (dl) with estimated quasi-variances of 0.25, 0.10 and 0.30 (dl^2). We decide we can afford a $n = 400$ sample.

   (a) What would be the estimated variance of the mean liquor intake per person if we take a sample of $n = 400$ persons and we neglect finite population correction?

   (b) What would be the estimated variance of the mean if we account for the finite population size?

   (c) What would be the estimated variance of the mean if we use stratified sampling with proportional samples?

   (d) What would be the optimal allocation of the sample and the estimated variance of the mean obtained with such allocation?

Sampling is almost absent in the manuals recommended in the syllabus. For a much more detailed treatment than we can afford in this course, you may turn to Cochran (1977) or Pérez-López (2006).

References

W. G. Cochran. *Sampling techniques.* Wiley, 1977.

C. Pérez-López. *Muestreo Estadistico - Conceptos y Problemas Resueltos.* Pearson Educacion, 2006.

Respuestas

1. Se trata de estimar una proporción. El rango de valores en que puede estar dicha proporción (de 0.20 a 0.60) incluye el caso más desfavorable desde el punto de vista de varianza que es $p = 0.50$. Haremos los cálculos conservadoramente, asumiendo este escenario más desfavorable.

   - Cuando la población es muy grande, la dependencia entre observaciones será despreciable y la aproximación de muestreo aleatorio simple adecuada. Para tamaños realistas de la muestra, la frecuencia relativa se distribuirá aproximadamente como:

     $$\hat{p} \sim N(p, \sigma^2 = pq/n);$$

     en el caso más desfavorable, $pq = 0.25$ y la desviación standard será no mayor que $\sqrt{0.25/n}$. Si queremos que no exceda de 0.02, tenemos que tomar n tal que

     $$\sqrt{\frac{0.25}{n}} \leq 0.02,$$

     de donde obtenemos:

     $$n \geq \frac{0.25}{(0.02)^2} = 625$$

     Bastaría encuestar a 625 familias.

   - Cuando la población es de 3800 familias, el tamaño muestral ya es una fracción apreciable del tamaño de la población y hemos de aplicar los resultados para poblaciones finitas. En la fórmula

     $$\mathrm{Var}(\hat{p}) = \frac{\tilde{\sigma}^2}{n}\left(1 - \frac{n}{N}\right)$$

     reemplazamos una cota superior de la varianza, el valor deseado de la desviación típica, el tamaño de la población y resolvemos para n:

     $$(0.02)^2 = \frac{0.25}{n}\left(1 - \frac{n}{3800}\right)$$

     Despejando n,

     $$n = \frac{0.25}{(0.02)^2 + 0.25/3800} = 536.72,$$

     de modo que un tamaño de muestra de 537 familias bastaría.

   Si tuviéramos la certeza de que $p \leq 0.30$, la cota máxima de la varianza pq sería $0.3 \times 0.7 = 0.21$, que sería reemplazada en los lugares en que hemos empleado $pq = 0.25$. Rehaciendo los cálculos obtendríamos para población muy grande y para población de 3800 familias los siguientes tamaños muestrales necesarios:

```
> 0.21 / 0.02^2                    # población infinita

[1] 525

> 0.21 / ( 0.02^2 + 0.21/3800 )  # población con N=3800

[1] 461.2717
```

2. Para calcular los tamaños muestrales para cada Territorio emplearíamos la fórmula de poblaciones infinitas, puesto que las poblaciones son grandes. El error máximo con confianza 95% es la semi-amplitud del intervalo de confianza. Por tanto,

$$\delta = 0.02 = 1.96\sqrt{\frac{0.22 \times 0.78}{n}}$$

de donde despejamos

$$n = \frac{1.96^2 \times 0.22 \times 0.78}{0.02^2}.$$

Realizando los cálculos,

```
> ( n <- 1.96^2 * 0.22 * 0.78 / 0.02^2 )

[1] 1648.046
```

La fórmula para poblaciones finitas

$$n = \frac{N z_{\alpha/2}^2 \sigma^2}{(N-1)\delta^2 + \sigma^2 z_{\alpha/2}^2}$$

da prácticamente el mismo resultado. Como queremos que el error (con confianza 95%) no exceda de 0.02 en cada Territorio necesitaremos $z_{\alpha/2} = 1.96$ y tamaños muestrales respectivos:

```
> 600000*1.96^2*0.22*0.78 /
+     (599999*0.02^2 + 0.22*0.78*1.96^2)  # Bizkaia

[1] 1643.535

> 400000*1.96^2*0.22*0.78 /
+     (399999*0.02^2 + 0.22*0.78*1.96^2)  # Gipuzkoa

[1] 1641.288

> 200000*1.96^2*0.22*0.78 /
+     (199999*0.02^2 + 0.22*0.78*1.96^2)  # Araba
```

```
[1] 1634.585
```

Nótese que como las poblaciones son tan grandes, hace falta prácticamente la misma muestra en los tres Territorios, aunque sus tamaños son muy diferentes.

Si quisiéramos un error de muestreo con confianza 95% de 0.02 sólo para la CAPV, el tamaño muestral sería, utilizando la aproximación de poblaciones infinitas, el mismo que hemos calculado para cada Territorio (≈ 1649 personas). Utilizando la fórmula para poblaciones finitas tendríamos casi el mismo valor:

```
> 1200000*1.96^2*0.22*0.78 /
+     (1199999*0.02^2 + 0.22*0.78*1.96^2)
```

```
[1] 1645.787
```

Con cualquiera de las aproximaciones vemos que el requerimiento de precisión en cada Territorio es caro: prácticamente supone triplicar la encuestación.

3. La proporción de familias en alquiler y con agua corriente se estima por punto por: $\hat{p} = 121/(121 + 153 + 10 + 31)$:

```
> N <- 15762
> n <- (121+153+10+31)
> ( p <- 121/n )
```

```
[1] 0.384127
```

La desviación standard se estimaría por:

```
> ( s <-  sqrt(p*(1-p) / n) )
```

```
[1] 0.02740487
```

El factor de corrección por población finita introduce una corrección irrelevante:

```
> 1 - n/N
```

```
[1] 0.9800152
```

El TOTAL de personas en alquiler y con agua corriente se estima por:

```
> N * p
```

```
[1] 6054.61
```

y su desviación standard se estima por:

```
> sqrt( N^2 * s^2)

[1] 431.9555
```

o, si se desea introducir la corrección por población finita:

```
> sqrt( N^2 * s^2 * (1 - n/N))

[1] 427.6175
```

Intervalos de confianza para la magnitud deseada sin y con corrección por población finita vienen dados respectivamente por,

```
> c(N*p-1.96*sqrt( N^2 * s^2),
+   N*p+1.96*sqrt( N^2 * s^2))

[1] 5207.977 6901.242
```

y

```
> c(N*p-1.96*sqrt( N^2 * s^2 * (1-n/N) ),
+   N*p+1.96*sqrt( N^2 * s^2 * (1-n/N)) )

[1] 5216.479 6892.740
```

4. La mención de que la población es "muy grande" sugiere emplear las fórmulas de muestreo aleatorio simple.

   (a) Los grupos de renta son de tamaños variables. Para estimar la proporción que cada uno de ellos representa con desviación standard no mayor que 0.02, harían falta tamaños diferentes en cada caso: calcularemos cuales son, y tomaremos el mayor, que garantizará la precisión deseada en los tres casos.

   ```
   > ( n1 <- (0.65*0.35) / 0.02^2 )

   [1] 568.75

   > ( n2 <- (0.25*0.75) / 0.02^2 )

   [1] 468.75

   > ( n3 <- (0.10*0.90) / 0.02^2 )

   [1] 225
   ```

   Tomaremos pues una muestra de 569 familias.

   (b) Análogamente si lo que queremos es un error 95% no mayor que 0.03:

```
> ( n1 <- 1.96^2 * (0.65*0.35) / 0.03^2 )
[1] 971.0711
> ( n2 <- 1.96^2 * (0.25*0.75) / 0.03^2 )
[1] 800.3333
> ( n3 <- 1.96^2 * (0.10*0.90) / 0.03^2 )
[1] 384.16
```

Tomaremos pues una muestra de 972 familias.

5. (a) La ingesta media de alcohol es:

```
> 0.5*0.3 + 0.4*0.2 + 0.1*0.01
[1] 0.231
```

La dispersión en torno a 0.231 tiene dos componentes: la dispersión de los valores medios en cada estrato respecto a 0.231 y la dispersión dentro de cada estrato. Por tanto,

```
> sigma2 <- 0.5 * ( (0.3-0.231)^2 + 0.25) +
+           0.4*((0.2-0.231)^2 + 0.10) +
+           0.1 * ((0.01-0.231)^2 + 0.30)
```

Por tanto, tomando una muestra de $n = 400$ y prescindiendo de la corrección por población finita obtenemos una varianza en la estimación de la ingesta media por persona de:

```
> ( varest1 <- sigma2 / 400 )
[1] 0.0005066225
```

(b) Si tomamos en cuenta el hecho de que la población es finita ($N = 9000$) tenemos en cambio la siguiente estimación de la varianza:

```
> ( varest2 <- ( sigma2 / 400 ) * (1 - 400/9000) )
[1] 0.0004841059
```

(c) Si realizamos un reparto de la muestra proporcional al tamaño de los estratos, obtenemos las siguientes estimaciones de las varianzas en cada estrato:

```
> ( var.str1 <- ( 0.25 / 200 ) * ( 1 - 200/4500) )
[1] 0.001194444
> ( var.str2 <- ( 0.10 / 160 ) * ( 1 - 160/3600) )
[1] 0.0005972222
> ( var.str3 <- ( 0.30 /  40 ) * ( 1 -  40/900)  )
[1] 0.007166667
```

El estimador de la media global es

$$\hat{m} = 0.5\hat{m}_1 + 0.4\hat{m}_2 + 0.1\hat{m}_3$$

y su varianza:

```
> var.str  <- 0.5^2 * var.str1 +
+             0.4^2 * var.str2 +
+             0.10^2 * var.str3
```

(d) La afijación óptima de la muestra tiene en cuenta tanto los tamaños como las varianzas intraestrato (y en su caso, los costes de muestreo respectivos, que aquí se suponen iguales). La muestra n_i asignada al estrato i-ésimo es:

$$n_i \propto \frac{N_i \tilde{\sigma}_i}{N}$$

Efectuando los cálculos,

```
> t1 <- 0.5 * sqrt(0.25)
> t2 <- 0.4 * sqrt(0.10)
> t3 <- 0.1 * sqrt(0.30)
> den <- t1 + t2 + t3
> ( n1 <- 400 * t1 / den )
[1] 231.8769
> ( n2 <- 400 * t2 / den )
[1] 117.3214
> ( n3 <- 400 * t3 / den )
[1] 50.80168
> n1+n2+n3
[1] 400
```

(Naturalmente, los tamaños respectivos los redondeamos a enteros.) Con esta asignación de la muestra, la varianza estimada de la media es:

```
> var.str.opt <- 0.5^2 * (0.25 / n1) * (1- n1/4500) +
+                0.4^2 * (0.10 / n2) * (1- n2/3600)  +
+                0.1^2 * (0.30 / n3) * (1- n3/900)
```

La ganancia en términos de reducción de varianza lograda por el muestreo estratificado es:

```
> var.str.opt - varest2
[1] -4.080239e-05
```

o en términos de mejora de la desviación típica:

```
> sqrt(var.str.opt) - sqrt(varest2)
[1] -0.0009476327
```