DEPARTAMENTO DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA

UNIVERSIDAD DE GRANADA

TÉCNICAS CUANTITATIVAS III

Grado en Marketing e Investigación de Mercados

Teoría y ejercicios

ÍNDICE

1. Muestreo aleatorio simple.

1.0 Definiciones y conceptos básicos

Nuestro objetivo a lo largo de la asignatura será conocer o investigar alguna característica de una **población**, por ejemplo el consumo de determinados productos, la audiencia televisiva de un programa, la intención de voto,... Claramente la recogida de información sobre toda la población resultaría cara y lenta. Por ello es preferible utilizar un subconjunto pequeño de la población, la **muestra**.

La **muestra** debe ser **representativa**, es decir, una versión a escala reducida de la población que refleje las características de toda la población.

Para obtener una muestra representativa hay diferentes métodos. Los **métodos de muestreo** más utilizados son:

- **Muestreo aleatorio simple**.
- **Muestreo aleatorio estratificado**.
- **Muestreo sistemático.**
- **Muestreo por conglomerados**.
- **Muestreo con probabilidades desiguales**

El ***error de muestreo*** es el que surge al considerar una muestra y no examinar toda la población. **El error de muestreo puede ser controlado y medido** mediante el diseño de la muestra.

Otros errores, más difíciles de controlar, pueden ocurrir al estudiar una muestra. Estos otros errores se llaman ***errores de no muestreo.*** En muchas muestras, el error de muestreo cometido para esa muestra puede ser despreciable en comparación con los errores que no son de muestreo.

Los errores de no muestreo más comunes son:

- **Sesgo de selección**. Este error ocurre cuando alguna parte de la población objetivo no puede ser elegida como parte de la muestra. Por ejemplo, si hacemos una encuesta por los domicilios en horario de trabajo, estamos vetando que ciertos individuos puedan ser elementos de la muestra.
- **Sesgo de medición**. El sesgo de medición ocurre cuando los datos observados difieren del valor verdadero, por ejemplo:
  - Los individuos no reconocen la verdad porque pudiera estar mal visto.
  - No comprenden las preguntas.
  - La formulación y el orden de las preguntas pueden afectar a las respuestas obtenidas.
  - …
- **No respuesta**. La no respuesta de un individuo seleccionado para formar parte de la muestra puede causar un sesgo en los datos muestrales similar al sesgo de selección. Puede ocurrir que las personas que respondan no representen a la población objetivo.

Los errores de no muestreo deben controlarse con acciones como reentrevistas, verificación de los datos,…

Son muchas las **razones para el uso del muestreo**, entre otras destacamos:

- ***Evitar la destrucción de la población***. En algunos casos, por ejemplo en el control de calidad, la observación de los elementos lleva a su destrucción.
- ***Rapidez***. Los datos se pueden reunir más rápido, de modo que las estimaciones se pueden publicar de una manera programada. Por ejemplo las elecciones.
- ***Economía y precisión***. El muestreo puede proporcionar información fiable con costes mucho menores que los de un censo (toda la población). Un censo completo implica mucho trabajo en la recolección de los datos y debido a su complejidad se pueden cometer muchos errores. En una muestra, por su menor tamaño, se puede dedicar más atención a la calidad de los datos.

1.1 Selección de una muestra aleatoria simple. Números aleatorios. Rutas aleatorias.

Si cada muestra posible de tamaño n tiene la misma probabilidad de ser seleccionada, el procedimiento de muestreo se denomina muestreo aleatorio simple y a la muestra así seleccionada se le llama muestra aleatoria simple.

La condición de que cada muestra tenga la misma probabilidad de ser seleccionada equivale a que cada elemento de la población tenga la misma probabilidad de pertenecer a la muestra. Para ello la selección de cada elemento de la muestra se debe hacer con un sorteo completamente aleatorio. Para facilitar la obtención de los resultados de ese sorteo aleatorio existen lo que se conoce como ***tablas de números aleatorios*** que suelen aparecer en un apéndice al final de muchos libros de estadística. Cada vez más, estas tablas de números aleatorios son sustituidas por la ***generación de números aleatorios*** mediante programas de ordenador (Excel, SPSS, R,...). Para asociar el valor de esos números aleatorios con los elementos de la población necesitamos que ésta esté numerada, en caso contrario deberíamos formar una lista y numerarla. Esto último, en muchos casos, no es tan sencillo. Una alternativa a la formación de una lista numerada para la selección mediante números aleatorios de los elementos de la muestra es el método de las ***rutas aleatorias***. Según este método cada número aleatorio o grupo de números aleatorios describe el camino hasta el elemento de la muestra. Veamos cómo se aplicaría este método con un sencillo ejemplo:
Se ha seleccionado el número aleatorio 11071032, las dos primeras cifras (11) indican el distrito de la ciudad, las dos siguientes (07) la calle del distrito, las dos siguientes (10) el número de la calle, la siguiente (3) la planta del edificio y la última (2) la letra B de dicha planta.

EL NÚMERO TOTAL DE ELEMENTOS QUE FORMAN UNA MUESTRA TIENE MENOS IMPORTANCIA QUE EL PRINCIPIO DE SELECCIÓN ALEATORIA. Utilizar un método más sencillo para seleccionar la muestra, con el que fácilmente se obtengan muchas observaciones, no garantiza una mejor información que una muestra aleatoria simple con muchos menos datos.

1.2 Muestreo aleatorio simple en poblaciones infinitas.

Supongamos que la característica en estudio de la población está representada por la variable Y (con media μ y varianza σ^2), una muestra aleatoria simple de tamaño *n* estará representada por *n* variables: $Y_1,..., Y_n$, independientes e idénticamente distribuidas (i.i.d.).
Nota: observaciones en poblaciones infinitas y también en poblaciones finitas, si se hacen con reemplazamiento, nos conducen a variables i.i.d.

1.2.1 Media, varianza y proporción muestrales: Propiedades. Error de estimación.

Como estimador de la media de la población, μ, se utiliza la media muestral, $\bar{y}$.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Un valor aislado $\bar{y}$ del estimador revela poco acerca de la media poblacional, deberíamos evaluar también su bondad.

Este estimador tiene propiedades deseables como ser insesgado y tener mínima varianza

$$E\left(\bar{y}\right) = \mu \qquad V\left(\bar{y}\right) = \frac{\sigma^2}{n}$$

Como estimador de la varianza de la población, σ^2, se utiliza la cuasivarianza muestral, S^2.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

que también tiene la propiedad de ser insesgado

$$E\left(S^2\right) = \sigma^2$$

de forma que la varianza de la media muestral se estima de forma insesgada por

$$\hat{V}(\bar{y}) = \frac{S^2}{n}$$

Cuando las variables Y, Y_1,..., Y_n son dicotómicas o binomiales, sólo toman dos valores (0 y 1), su media μ representa una proporción que se nota como p y el estimador de la misma, la proporción muestral, por $\hat{p}$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad\qquad y_i = 0, 1$$

Este estimador, como media muestral que es, tiene las mismas propiedades mencionadas antes.

La varianza de la población es en este caso $\sigma^2 = pq$, donde q=1-p. Como antes, el estimador insesgado de la varianza es la cuasivarianza muestral que para este tipo de variables es igual a

$$S^2 = \frac{n}{n-1}\hat{p}\hat{q}$$

y la varianza estimada de la proporción muestral es

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}$$

Si conocemos más características de las variables aleatorias Y, Y_1,..., Y_n, conoceremos más sobre el comportamiento de la media muestral, aparte de lo ya mencionado.

Si $Y \to N(\mu,\sigma^2)$ σ^2 *conocida* $\Rightarrow$ $\dfrac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$

Si $Y \to N(\mu,\sigma^2)$ σ^2 *desconocida* $\Rightarrow$ $\dfrac{\bar{y}-\mu}{\frac{S}{\sqrt{n}}} \to t_{n-1} \approx N(0,1)$

(*en la práctica para* $n > 30$, $t_{n-1} \approx N(0,1)$)

Si $Y \to$ *cualquier distribución* $\Rightarrow$ $\dfrac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx \dfrac{\bar{y}-\mu}{\frac{S}{\sqrt{n}}} \to N(0,1)$

(*por el Teorema Central del Límite cuando* $n \to \infty$, *en la práctica para n>30*)

Un caso particular del anterior es $Y \to B(1,p)$, *variable dicotómica, donde* $\mu = p$ $\bar{y} = \hat{p}$

$$\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \approx \frac{\hat{p}-p}{\sqrt{\frac{n}{n-1}\frac{\hat{p}\hat{q}}{n}}} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}\hat{q}}{n-1}}} \to N(0,1) \qquad \text{(en la práctica para } n>30\text{)}$$

Todo lo anterior puede resumirse diciendo que la media muestral (de variables numéricas, $\bar{y}$, o dicotómicas, $\hat{p}$) sigue una distribución Normal o se puede aproximar por ella si el tamaño de la muestra es suficientemente grande. De forma que podemos conocer la probabilidad de que dicha variable tome determinados valores, por ejemplo (tomando una de las anteriores expresiones de la media muestral tipificada, siendo válido lo que sigue también para las otras)

$$P\left[-1,96 \le \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1,96\right] = 0,95$$

o en un caso más general

$$P\left[-Z_{\alpha/2} \le \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \le Z_{\alpha/2}\right] = 1-\alpha$$

α=nivel de significación 1-α=nivel de confianza

Para un nivel de confianza del 95% (el más habitual) se suele redondear el anterior valor 1,96≈2.

En todos los apuntes que siguen trabajaremos con un nivel de confianza del 95% y con $Z_{\alpha/2}=2$. ***En el formulario consideraremos distintos niveles de confianza, por tanto distintos valores de*** $Z_{\alpha/2}$ ***que notaremos*** Z_c ***para simplificar la notación.***

De las probabilidades anteriores se puede hacer dos lecturas. **La primera:**

$$P\left[-2\frac{\sigma}{\sqrt{n}}\le \bar{y}-\mu\le 2\frac{\sigma}{\sqrt{n}}\right]=0,95 \quad\Rightarrow\quad P\left[\left|\bar{y}-\mu\right|\le 2\frac{\sigma}{\sqrt{n}}\right]=0,95$$

En esta expresión aparecen valores y expresiones fundamentales en las técnicas de estimación:

1-α=0,95= nivel de confianza del 95%.

$\left|\bar{y}-\mu\right|$= error de estimación o diferencia entre la estimación que hacemos, $\bar{y}$, y el verdadero valor del parámetro que se quiere estimar, μ.

$2\frac{\sigma}{\sqrt{n}}$= cota o límite para el error de estimación, es el máximo error de estimación que se puede estar cometiendo, con una confianza del 95%. En la práctica, σ es desconocida y se estima por S.

La segunda lectura:

$$P\left[\bar{y}-2\frac{\sigma}{\sqrt{n}}\le \mu\le \bar{y}+2\frac{\sigma}{\sqrt{n}}\right]=0,95$$

expresa la confianza que tenemos de que el verdadero valor del parámetro μ se encuentre entre los extremos del intervalo $\left(\bar{y}-2\frac{\sigma}{\sqrt{n}},\bar{y}+2\frac{\sigma}{\sqrt{n}}\right)$.

Todo lo anterior se puede asegurar si el estimador sigue una distribución Normal (si el tamaño de la muestra es suficientemente grande, n>30, está garantizado). Pero qué ocurre si no es así. En ese caso la desigualdad de Tchebychev nos da la respuesta.

La desigualdad de Tchebychev asegura que si X es una variable aleatoria con media $E(X)=\mu$ y varianza $V(X)=\sigma^2$, sea cual sea su distribución de probabilidad

$$P\left[\left|X-\mu\right| \le k\sigma\right] \ge 1-\frac{1}{k^2}$$

Aplicando lo anterior a la media muestral para k=2 se obtiene

$$P\left[\left|\bar{y}-\mu\right| \le 2\frac{\sigma}{\sqrt{n}}\right] \ge 1-\frac{1}{4} = 0,75$$

resultado parecido al que obteníamos anteriormente

$$P\left[\left|\bar{y}-\mu\right| \le 2\frac{\sigma}{\sqrt{n}}\right] = 0,95$$

salvo que en este caso lo más que podemos asegurar es que dicha probabilidad es mayor que 0,75.

1.2.2 Estimación puntual. Intervalos de confianza. Contrastes de hipótesis.

Cuando estimamos el valor de un parámetro poblacional con el valor que ha presentado en una determinada muestra el estimador asociado, hacemos una **estimación puntual**.

Si dicha estimación puntual se acompaña de un margen de error (límite para el error de estimación) y de una medida de la certidumbre que se tiene en tal estimación (nivel de confianza), hablamos de **intervalo de confianza**. Por ejemplo, utilizando muestras grandes, el intervalo de confianza para la media poblacional μ con un nivel de confianza del 95% es

$$\left(\bar{y}-2\frac{S}{\sqrt{n}}, \bar{y}+2\frac{S}{\sqrt{n}}\right)$$

En ocasiones se quiere contrastar con los valores observados en una muestra la posibilidad de que el verdadero valor de un parámetro de la población sea un valor dado, por ejemplo, se quiere **contrastar la hipótesis** nula $H_0 : \mu = \mu_0$ con un nivel de significación del 5%. Lo anterior equivale a comprobar si

$$\mu_0 \in \left(\bar{y}-2\frac{S}{\sqrt{n}}, \bar{y}+2\frac{S}{\sqrt{n}}\right)$$

en cuyo caso se aceptaría la hipótesis nula, rechazándose en caso contrario.

1.2.3 Determinación del tamaño muestral.

Si se fija de antemano el máximo error de estimación que estamos dispuestos a aceptar en una estimación, $2\frac{\sigma}{\sqrt{n}} = B$, la cantidad de información necesaria para conseguirlo depende del tamaño de la muestra según la siguiente expresión

$$4\frac{\sigma^2}{n} = B^2 \Rightarrow n = \frac{\sigma^2}{\frac{B^2}{4}} = \frac{\sigma^2}{D} \quad , \quad D = \frac{B^2}{4}$$

En la práctica la varianza poblacional σ^2 es desconocida. Si disponemos de S^2 de un estudio anterior podemos obtener el valor de *n* sustituyendo en la anterior expresión σ^2 por S^2.

Si no se dispone de información previa para estimar la varianza podemos usar que en variables Normales el rango es aproximadamente cuatro veces su desviación típica

$$\sigma \cong \frac{R}{4} \quad \Leftrightarrow \quad \sigma^2 \cong \frac{R^2}{16}$$

La proporción poblacional p es la media μ de una variable dicotómica ($Y \sim B(1, p)$, $E(Y) = p$, $V(Y) = pq$), luego el problema de determinar el tamaño muestral se hace de forma análoga sustituyendo σ^2 por pq

$$n = \frac{pq}{D}, \qquad D = \frac{B^2}{4}$$

En la práctica p se desconoce. Una aproximación se obtiene reemplazándolo por el valor estimado $\hat{p}$ obtenido en muestras preliminares. Si no se cuenta con información anterior, suponiendo $p = \frac{1}{2}$ se obtiene un tamaño muestral conservador (mayor que el requerido para obtener la cota del error de estimación fijada).

Ejemplo 1.1. Un hipermercado desea estimar la proporción de compras que los clientes pagan con su "Tarjeta de Compras". Durante una semana observaron al azar 200 compras de las cuales 35 fueron pagadas con la tarjeta.

a) Estime con un intervalo de confianza la proporción de compras pagadas con dicha tarjeta.

b) ¿Cuántas compras deberían observarse para estimar, con un error inferior al 3%, la proporción de compras pagadas con la tarjeta? (Consideren los datos anteriores como una muestra previa)

c) Si no se tuviera ninguna información acerca de los clientes que utilizan la tarjeta, cuántas compras deberíamos observar para asegurar que la anterior estimación se realiza con un error inferior al 3%.

d) Este mismo hipermercado desea estimar también el valor medio de las compras realizadas con su "Tarjeta de Compras". Basándose en los anteriores datos se observa que

el valor total de las compras hechas con la tarjeta fue de 5600€ (siendo la cuasivarianza de los datos 625). Estime el valor medio de las compras pagadas con la tarjeta y el error de estimación asociado.

Solución:

a)

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{35}{200} = 0{,}175 \qquad n = 200 \qquad \hat{q} = 1 - 0{,}175 = 0{,}825 \qquad \hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1} = 0{,}000726$$

$$B = 2\sqrt{\hat{V}(\hat{p})} = 0{,}0539 \qquad \boxed{p \in (12{,}11\% \, , \, 22{,}89\%)}$$

b)

$$B = 0{,}03 \qquad D = \frac{B^2}{4} = 0{,}000225 \qquad \boxed{n = \frac{pq}{D} = 641{,}\hat{6} \approx 642}$$

c)

$$B = 0{,}03 \qquad D = \frac{B^2}{4} = 0{,}000225 \qquad p = q = 0{,}5 \qquad \boxed{n = \frac{pq}{D} = 1111{,}\hat{1} \approx 1112}$$

d)

$$n = 35 \qquad \boxed{\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{5600}{35} = 160€}$$

$$S^2 = 625 \qquad \hat{V}(\bar{y}) = \frac{S^2}{n} = \frac{625}{35} = 17{,}8571 \qquad \boxed{B = 2\sqrt{\hat{V}(\bar{y})} = 8{,}45€}$$ ■

1.3 Muestreo aleatorio simple en poblaciones finitas.

Suponemos que la población es finita, tiene N elementos, y además que la muestra se selecciona ***sin reemplazamiento*** (*lo que suele ser habitual, en caso contrario estaríamos ante el mismo modelo que el muestreo aleatorio simple en poblaciones infinitas con variables i.i.d.*)

1.3.1 Estimación de la media, proporción y total poblacionales.

Estimación de la media poblacional.

Para estimar la media poblacional, μ, se utiliza la media muestral

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Este estimador es insesgado y su varianza decrece conforme crece el tamaño de la muestra

$$E\left(\bar{y}\right)=\mu \qquad V\left(\bar{y}\right)=\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

En este tipo de muestreo la cuasivarianza muestral no es un estimador insesgado de la varianza de la población

$$E\left(S^2\right)=\frac{N}{N-1}\sigma^2 \qquad E\left(\frac{N-1}{N}S^2\right)=\sigma^2$$

De lo anterior se sigue que la varianza de la media muestral puede ser estimada insesgadamente por

$$\hat{V}\left(\bar{y}\right)=\left(\frac{N-1}{N}S^2\right)\frac{1}{n}\left(\frac{N-n}{N-1}\right)=\frac{S^2}{n}\left(\frac{N-n}{N}\right)$$

expresión igual a la del caso de poblaciones infinitas, $\left(\hat{V}(\bar{y})=\frac{S^2}{n}\right)$, salvo el coeficiente $\left(\frac{N-n}{N}\right)$ que se denomina coeficiente corrector para poblaciones finitas (c.p.f.).

En la práctica el coeficiente c.p.f. suele despreciarse si está próximo a 1, $\left(\frac{N-n}{N}\right)\geq 0{,}95$ o lo que es equivalente si $n\leq 5\%N$. En muchos casos N no está claramente definido o se desconoce, pero si N se supone suficientemente grande el c.p.f. se omite, $\left(\frac{N-n}{N}\right)\cong 1$.

Para calcular el límite para el error de estimación, con un 95% de confianza, se halla $2\sqrt{\hat{V}(\bar{y})}$. Igual que en el caso de poblaciones infinitas, se habla de un nivel de confianza del 95% cuando trabajamos con el coeficiente 1,96≈2. Pero en algunos casos, según la desigualdad de Tchevychev, sólo se puede asegurar que este nivel es mayor que un 75%.

Estimación del total poblacional.

Para estimar el total poblacional, τ, dado que $\mu=\frac{\tau}{N} \Leftrightarrow \tau=N\mu$ utilizaremos el estimador

$$\hat{\tau}=N\bar{y}=\frac{N}{n}\sum_{i=1}^{n}y_i\ .$$

Para hallar su varianza usamos que $V(kX)=k^2V(X)$, por tanto:

Varianza estimada de $\hat{\tau}$

$$\hat{V}(\hat{\tau})=\hat{V}(N\bar{y})=N^2\hat{V}(\bar{y})=N^2\frac{S^2}{n}\frac{N-n}{N}=N(N-n)\frac{S^2}{n}$$

Como en el caso de la media, el límite para el error de estimación con una confianza del 95% está dado por $2\sqrt{\hat{V}(\hat{\tau})}$. Valiendo comentarios análogos a los hechos anteriormente.

En lo sucesivo se dará solamente el valor de la varianza del estimador para los distintos tipos de muestreo, omitiéndose el límite para el error de estimación.

Ejemplo 1.2. Un auditor examina las cuentas abiertas con diferentes clientes de una empresa. Suponga que existen 1000 cuentas de las cuales se examinan 300. La media muestral de las cuentas fue $\bar{y} = 1040€$ y la cuasivarianza muestral es S^2=45000€^{2}. Estime el promedio de la deuda y el total de la deuda por cobrar para las 1000 cuentas abiertas con un intervalo de confianza al 95%.

Solución:

$$\hat{V}(\bar{y}) = \frac{S_{n-1}^2}{n}\frac{N-n}{N} = \frac{45000}{300}\ \frac{1000-300}{1000} = 105 \qquad 2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{105} = 20,49€$$

$$(1040 \mp 20,49) = (1019,51\ ,\ 1060,49)$$

$$\hat{\tau} = N\bar{y} = 1000 \times 1040 = 1040000€$$

$$2\sqrt{\hat{V}(\hat{\tau})} = N2\sqrt{\hat{V}(\bar{y})} = 1000 \times 20,49 = 20490€ \quad (\textit{valor exacto} \quad 20493,9)$$

$$(1040000 \mp 20490) = (1019510\ ,1060490) \qquad \blacksquare$$

Estimación de la proporción poblacional.

Para estimar la proporción poblacional p, dado que se trata de una media, usaremos la media muestral con la siguiente notación en este caso

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad y_i = 0,1$$

su varianza estimada, teniendo en cuenta que $S^2 = \frac{n\hat{p}\hat{q}}{n-1}$, es igual a

$$\hat{V}(\hat{p}) = \frac{S^2}{n}\frac{N-n}{N} = \frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N}$$

Para estimar el total poblacional de una variable dicotómica usamos

$$\hat{\tau} = N\hat{p} \qquad \hat{V}(\hat{\tau}) = \hat{V}(N\hat{p}) = N^2\hat{V}(\hat{p}) = N(N-n)\frac{\hat{p}\hat{q}}{n-1}$$

Ejemplo 1.3. Se toma una muestra aleatoria simple de 100 estudiantes de un centro con 900 estudiantes para estimar

- La proporción que votarán a un determinado representante de centro.
- La proporción de ellos que tienen algún tipo de trabajo.

Sean $y_i, z_i \quad (i = 1,...,100)$ las respuestas del i-ésimo estudiante seleccionado ($y_i = 0$ cuando responden NO, $y_i = 1$ cuando responden SI, análogamente para z_i).

Según la muestra $\sum_{i=1}^{100} y_i = 70 \qquad \sum_{i=1}^{100} z_i = 25$

Usando los datos de la muestra, estime p_1 (proporción de estudiantes que votarán a un determinado representante), p_2 (proporción de estudiantes con algún tipo de trabajo), τ_2 (número de estudiantes con algún tipo de trabajo) y los límites para los errores de estimación correspondientes.

Solución:

$$\hat{p}_1 = \frac{\sum_{i=1}^{100} y_i}{100} = 0,70 \quad (70\%) \qquad \hat{p}_2 = \frac{\sum_{i=1}^{100} z_i}{100} = 0,25 \quad (25\%)$$

$$\hat{V}(\hat{p}_1) = \frac{\hat{p}_1\hat{q}_1}{n-1}\frac{N-n}{N} = 0,0018855 \qquad \hat{V}(\hat{p}_2) = \frac{\hat{p}_2\hat{q}_2}{n-1}\frac{N-n}{N} = 0,0016835$$

$$2\sqrt{\hat{V}(\hat{p}_1)} = 0,0868 \quad (8,68\%) \qquad 2\sqrt{\hat{V}(\hat{p}_2)} = 0,0821 \quad (8,21\%)$$

$$\hat{\tau}_2 = N\hat{p}_2 = 900 \times 0,25 = 225 \qquad 2\sqrt{\hat{V}(\hat{\tau}_2)} = 900 \times 0,0821 = 73,89$$ ■

1.3.2 Determinación del tamaño muestral.

El número de observaciones necesarias para estimar μ con un límite para el error de estimación de magnitud B se obtiene resolviendo $2\sqrt{V(\bar{y})} = B$

$$2\sqrt{V(\bar{y})} = B \quad \Leftrightarrow \quad V\left(\bar{y}\right) = \frac{B^2}{4} = D$$

$$V(\bar{y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = D \quad \Rightarrow \quad \frac{\sigma^2 N}{n} - \frac{\sigma^2 n}{n} = (N-1)D \quad \Rightarrow \quad \frac{\sigma^2 N}{n} = (N-1)D + \sigma^2$$

$$\frac{N\sigma^2}{(N-1)D + \sigma^2} = n$$

Para estimar el total poblacional con un límite para el error de estimación B, dado que $2\sqrt{\hat{V}(\hat{\tau})} = N2\sqrt{\hat{V}(\bar{y})} = B$, se llega a la misma expresión de *n* pero con $D = \frac{B^2}{4N^2}$

Habitualmente la varianza poblacional σ^2 es desconocida. Si disponemos de S^2 de un estudio anterior podemos obtener el valor de *n* sustituyendo en la anterior expresión σ^2 por S^2.
Si no se dispone de información previa para estimar la varianza podemos usar que en variables Normales el rango es aproximadamente cuatro veces su desviación típica

$$\sigma \cong \frac{R}{4} \quad \Leftrightarrow \quad \sigma^2 \cong \frac{R^2}{16}$$

La proporción poblacional p es la media μ de una variable dicotómica ($Y \sim B(1, p)$, $E(Y) = p$, $V(Y) = pq$), luego el problema de determinar el tamaño muestral se hace de forma análoga sustituyendo σ^2 por *pq*, obteniéndose

$$n = \frac{Npq}{(N-1)D + pq} \qquad D = \frac{B^2}{4} \text{ (para la proporción)} \qquad D = \frac{B^2}{4N^2} \text{ (para el total)}$$

En la práctica p se desconoce. Una aproximación al mismo se obtiene reemplazándolo por el valor estimado $\hat{p}$ obtenido en encuestas preliminares. Si no se cuenta con información anterior, suponiendo $p = \frac{1}{2}$ se obtiene un tamaño muestral conservador (mayor que el requerido para obtener la cota del error de estimación fijada).

Ejemplo 1.4. Encuentre el tamaño de la muestra necesario para estimar el valor total de 1000 cuentas por cobrar con un límite para el error de estimación de 10000€. Aunque no se cuenta con datos anteriores para estimar la varianza poblacional, se sabe que la mayoría de las cuentas caen dentro del intervalo (600, 1400).

Solución:

$$D = \frac{B^2}{4N^2} = \frac{10000^2}{4 \times 1000^2} = 25 \qquad 4\sigma \cong 800 \quad \Rightarrow \quad \sigma \cong 200 \quad \Rightarrow \quad \sigma^2 \cong 40000$$

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = 615,62 \approx 616 \qquad \blacksquare$$

Si se realizan dos mediciones (o más) sobre cada elemento de la muestra, se calcularán los tamaños muestrales que satisfacen los límites para el error de estimación fijados para cada estimación y finalmente el mayor de los dos será el tamaño de la muestra que satisface ambos límites.

Ejemplo 1.5. Los alumnos de TC3 de una facultad con 3000 estudiantes desean realizar una encuesta para determinar la proporción de estudiantes que están a favor de hacer los exámenes en sábado con un límite para el error de estimación del 10%. La información previa disponible indica que el 60% preferían los exámenes en sábado. También se quiere estimar la proporción de estudiantes que apoyan al equipo decanal con un error de estimación del 5%. Determínese el tamaño muestral que se requiere para estimar ambas proporciones con los límites de error especificados.

Solución:

$p_1 =$ proporción de estudiantes que prefieren los exámenes en sábado.

$$D_1 = \frac{B_1^2}{4} = \frac{(0,10)^2}{4} = 0,0025$$

$$n_1 = \frac{Np_1q_1}{(N-1)D_1 + p_1q_1} = \frac{3000 \times 0,60 \times 0,40}{(2999 \times 0,0025) + (0,60 \times 0,40)} = 93,05 \approx 94$$

$p_2 =$ proporción de estudiantes que apoyan al equipo decanal.

$$D_2 = \frac{B_2^2}{4} = \frac{(0,05)^2}{4} = 0,000625$$

$$n_2 = \frac{Np_2q_2}{(N-1)D_2 + p_2q_2} = \frac{3000 \times 0,50 \times 0,50}{(2999 \times 0,000625) + (0,50 \times 0,50)} = 353,04 \approx 354$$

para cumplir con ambos objetivos habría que tomar n=354 con lo que el límite para el error de la estimación de p_1 disminuiría (con un 95% de confianza) hasta:

$$2\sqrt{\hat{V}(\hat{p}_1)} = 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n-1}\frac{N-n}{N}} = 2\sqrt{\frac{0,60 \times 0,40}{353}\frac{3000-354}{3000}} = 0,0489 \quad (\cong 4,9\%)$$

o bien la cota del error de estimación del 10% se tiene con un nivel de confianza mucho mayor

$$z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_1)} = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n-1}\frac{N-n}{N}} = z_{\alpha/2}\sqrt{\frac{0,60 \times 0,40}{353}\frac{3000-354}{3000}} = 0,10$$

$$z_{\alpha/2}\,0,02445 = 0,10 \quad \Rightarrow \quad z_{\alpha/2} = 4,09$$

buscando en la tabla de la Normal (o con ayuda de la hoja de cálculo Excel,...) la probabilidad comprendida entre (-4,09 , 4,09), se obtiene 0,99995684, es decir, prácticamente un nivel de confianza del 100%. ■

APÉNDICE: Estudio empírico de todas las posibles muestras en un muestreo aleatorio simple sobre una población finita.

Habitualmente el tamaño de la población es grande y no es posible comprobar empíricamente las propiedades del muestreo aleatorio simple que hemos estudiado en este tema. A continuación vamos a suponer una población con sólo 6 elementos y vamos obtener todas las posibles muestras de tamaños 4 y 5 para estudiar las propiedades de los estimadores de la media poblacional y proporción poblacional.

En la siguiente tabla se recogen las características de la población que vamos a considerar:

①②③ ⓫⓬⓭	Población finita	Proporción poblacional, p (*de negras*)	Media poblacional, μ	Varianza poblacional, σ^2
	N=6	$p=\frac{3}{6}=0,50$	$\mu=\frac{1+2+3+11+12+13}{6}=7$	25,6667 **(#)**

$$\frac{(1-7)^2+(2-7)^2+(3-7)^2+(11-7)^2+(12-7)^2+(13-7)^2}{6}=25,6667 \quad \textbf{(\#)}$$

Tomamos todas las muestras aleatorias simples de tamaño 4 para estimar tanto la proporción como la media poblacional usando la proporción y media muestral, comprobando que son estimadores insesgados. Asimismo calculamos la cuasivarianza muestral comprobando que en el caso de poblaciones finitas no es un estimador insesgado de la varianza de la población, pero sí lo es: $\frac{N-1}{N}S^2$, $\frac{5}{6}S^2$ en nuestro ejemplo.

$$\frac{(3-9,75)^2+(11-9,75)^2+(12-9,75)^2+(13-9,75)^2}{4-1}=20,92 \quad \textbf{(\#)}$$

De igual forma se calculan el resto de cuasivarianzas muestrales.

	MUESTRAS (*n*=4)	Proporción muestral, $\widehat{p}$ (*de negras*)	Media muestral, $\bar{y}$	Cuasivarianza muestral, S^2_{n-1}
1	③⓫⓬⓭	$\frac{3}{4}=0,75$	$\frac{3+11+12+13}{4}=9,75$	20,92 **(#)**
2	②⓫⓬⓭	$\frac{3}{4}=0,75$	$\frac{2+11+12+13}{4}=9,5$	25,67
3	②③⓬⓭	$\frac{2}{4}=0,50$	$\frac{2+3+12+13}{4}=7,5$	33,67
4	②③⓫⓭	$\frac{2}{4}=0,50$	$\frac{2+3+11+13}{4}=7,25$	30,92
5	②③⓫⓬	$\frac{2}{4}=0,50$	$\frac{2+3+11+12}{4}=7$	27,33

6	①⓫⓬⓭	$\frac{3}{4}=0,75$	$\frac{1+11+12+13}{4}=9,25$	30,92
7	①③⓬⓭	$\frac{2}{4}=0,50$	$\frac{1+3+12+13}{4}=7,25$	37,58
8	①③⓫⓭	$\frac{2}{4}=0,50$	$\frac{1+3+11+13}{4}=7$	34,67
9	①③⓫⓬	$\frac{2}{4}=0,50$	$\frac{1+3+11+12}{4}=6,75$	30,92
10	①②⓬⓭	$\frac{2}{4}=0,50$	$\frac{1+2+12+13}{4}=7$	40,67
11	①②⓫⓭	$\frac{2}{4}=0,50$	$\frac{1+2+11+13}{4}=6,75$	37,58
12	①②⓫⓬	$\frac{2}{4}=0,50$	$\frac{1+2+11+12}{4}=6,5$	33,67
13	①②③⓭	$\frac{1}{4}=0,25$	$\frac{1+2+3+13}{4}=4,75$	30,92
14	①②③⓬	$\frac{1}{4}=0,25$	$\frac{1+2+3+12}{4}=4,5$	25,67
15	①②③⓫	$\frac{1}{4}=0,25$	$\frac{1+2+3+11}{4}=4,25$	20,92
	TOTAL:	7,50	105	462

MUESTRAS (*n*=4)	Proporción muestral, $\hat{p}$ (*de negras*)	Media muestral, $\bar{y}$	Cuasivarianza muestral, S^2_{n-1}
MEDIA:	$E\left[\hat{p}\right]=\frac{7,50}{15}=0,50$	$E\left[\bar{y}\right]=\frac{105}{15}=7$	$E\left[S^2_{n-1}\right]=\frac{462}{15}=30,8$ (*)
VARIANZA:	$V\left[\hat{p}\right]=0,0250$	$V\left[\bar{y}\right]=2,5667$	

(*) $E\left[S^2_{n-1}\right]=\frac{462}{15}=30,8\neq 25,6667=\sigma^2$, S^2_{n-1} no es un estimador insesgado de la varianza poblacional en el muestreo aleatorio simple en poblaciones finitas (pero sí lo es en el muestreo aleatorio simple en poblaciones infinitas). Sin embargo, $\frac{N-1}{N}S^2_{n-1}$ sí lo es: $E\left[\frac{N-1}{N}S^2_{n-1}\right]=\frac{N-1}{N}E\left[S^2_{n-1}\right]=\frac{5}{6}30,8=25,\widehat{6}=\sigma^2$.

Volvemos a tomar todas las muestras aleatorias simples, esta vez de tamaño 5, para estimar tanto la proporción como la media poblacional usando la proporción y media muestral, comprobando de nuevo que son estimadores insesgados. Observando que para un tamaño mayor de la muestra, 5 en lugar de 4, los valores de los estimadores presentan menos

dispersión, como puede confirmarse comparando el valor de sus varianzas con el valor obtenido en muestras de tamaño 4.

$$\frac{(2-8,2)^2+(3-8,2)^2+(11-8,2)^2+(12-8,2)^2+(13-8,2)^2}{5-1}=27,7 \quad \textbf{(\#)}$$

De igual forma se calculan el resto de cuasivarianzas muestrales.

	MUESTRAS (*n*=5)	Proporción muestral, $\hat{p}$ (*de negras*)	Media muestral, $\bar{y}$	Cuasivarianza muestral, S_{n-1}^2
1	②③⓫⓬⓭	$\frac{3}{5}=0,60$	$\frac{2+3+11+12+13}{5}=8,2$	27,7 **(#)**
2	①③⓫⓬⓭	$\frac{3}{5}=0,60$	$\frac{1+3+11+12+13}{5}=8$	31,0
3	①②⓫⓬⓭	$\frac{3}{5}=0,60$	$\frac{1+2+11+12+13}{5}=7,8$	33,7
4	①②③⓬⓭	$\frac{2}{5}=0,40$	$\frac{1+2+3+12+13}{5}=6,2$	33,7
5	①②③⓫⓭	$\frac{2}{5}=0,40$	$\frac{1+2+3+11+13}{5}=6$	31,0
6	①②③⓫⓬	$\frac{2}{5}=0,40$	$\frac{1+2+3+11+12}{5}=5,8$	27,7
	TOTAL:	3	42	184,8

MEDIA:	$E\left[\hat{p}\right]=\frac{3}{6}=0,50$	$E\left[\bar{y}\right]=\frac{42}{6}=7$	$E\left[S_{n-1}^2\right]=\frac{184,8}{6}=30,8$ (*)
VARIANZA:	$V\left[\hat{p}\right]=0,0100$	$V\left[\bar{y}\right]=1,0267$	

(*) $E\left[S_{n-1}^2\right]=\frac{184,8}{6}=30,8\neq 25,6667=\sigma^2$, S_{n-1}^2 no es un estimador insesgado de la varianza poblacional en el muestreo aleatorio simple en poblaciones finitas

EJERCICIOS RESUELTOS

1. Se selecciona una m.a.s. de 9 compras de clientes de un centro comercial para estimar el valor medio de las compras por cliente.

VALOR en €	33,5	32	52	43	40	41	45	42,5	39

   a) Obtener un intervalo de confianza para el valor medio de las compras.
   b) ¿Podemos aceptar que la compra media es de 45€?

c) ¿Qué tamaño muestral deberíamos tomar para que el LEE sea de 2€?

SOLUCIÓN:

a) $\widehat{\mu} = \overline{y} = \dfrac{33,5 + ... + 39}{9} = 40,89\ €$

$$S^2 = \frac{1}{9-1}\left((33,5 - 40,89)^2 + ... + (39 - 40,89)^2 \right) = 35,67$$

$$\widehat{V}\left(\overline{y}\right) = \frac{S^2}{n} = 3,963 \qquad B = 2\sqrt{\widehat{V}\left(\overline{y}\right)} = 3,98\ €$$

$$\left(40,89 - 3,98\,;\, 40,89 + 3,98\right) = \left(36,91\,;\, 44,87\right)$$

b) No, porque $45 \notin \left(36,91\,;\, 44,87\right)$

c) $n = \dfrac{\sigma^2}{\dfrac{B^2}{4}} \cong \dfrac{S^2}{\dfrac{B^2}{4}} = \dfrac{35,67}{1} = 35,67 \approx 36\ \textit{compras}$

2. Se han entrevistado 1000 vecinos, elegidos aleatoriamente entre los más de cien mil habitantes de una ciudad para conocer su opinión sobre los nuevos impuestos municipales. 655 manifestaron su opinión desfavorable. Estime la proporción de vecinos que están en contra de los nuevos impuestos y establezca el límite para el error de estimación. ¿Se puede afirmar que la mayoría de los habitantes están en contra?

SOLUCIÓN:

$$\widehat{p} = \frac{655}{1000} = 0,655 \quad \Rightarrow \quad \widehat{p} = 65,5\%$$

$$\widehat{V}(\widehat{p}) = \frac{\widehat{p}\widehat{q}}{n-1} = \frac{0,655 \times (1 - 0,655)}{999} = 0,0002262012$$

$$2\sqrt{\widehat{V}(\widehat{p})} = 0,0301 \quad \Rightarrow \quad 3,01\%$$

$$(65,5\% - 3,01\%\,,\, 65,5\% + 3,01\%) = (62,49\%\,,\, 68,51\%)$$

$p \in (62,49\%\,,\, 68,51\%) \quad \Rightarrow \quad p > 50\% \quad \Rightarrow$ *sí se puede afirmar que la mayoría de los habitantes están en contra*

3. El Centro de Estadística desea estimar el salario medio de los trabajadores de los invernaderos de una región. Se decide clasificarlos en dos estratos, los que poseen contrato fijo y los que tienen un contrato temporal. El salario de los contratos fijos está comprendido entre los 1200 y 2200 euros mensuales, el salario de los contratos temporales está comprendido entre 500 y 1700 euros mensuales. ¿Cuál debe ser el tamaño muestral

total y su asignación para que se estime el salario medio de los contratos fijos con un error inferior a 100€ y el salario medio de los contratos temporales con un error inferior a 120€?

SOLUCIÓN:

R_i	$\frac{R_i}{4} \approx \sigma_i$	σ_i^2
2200-1200=1000	250	62500
1700-500=1200	300	90000

$$n_1 = \frac{\sigma_1^2}{D_1} = \frac{\sigma_1^2}{\frac{B_1^2}{4}} = \frac{62500}{\frac{100^2}{4}} = \frac{62500}{\frac{10000}{4}} = 25$$

$$n_2 = \frac{\sigma_2^2}{D_2} = \frac{\sigma_2^2}{\frac{B_2^2}{4}} = \frac{90000}{\frac{120^2}{4}} = \frac{90000}{\frac{14400}{4}} = 25$$

$$n = n_1 + n_2 = 50$$

4. Entre todas las oficinas bancarias de una pequeña ciudad se tienen concedidos 2000 préstamos hipotecarios. Existen razones para pensar que el préstamo hipotecario de menor cuantía es de algo más de 1200 euros, siendo de casi 11000 euros el de mayor cuantía. ¿cuál es el tamaño muestral necesario para estimar estos dos parámetros:
   - la cuantía media de los prestamos cometiendo un error de estimación menor de 400 euros y
   - la proporción de préstamos pendientes de amortizar más de la mitad de la deuda cometiendo un error máximo del 5%?

SOLUCIÓN:

$$N = 2000$$

$$R = 11000 - 1200 = 9800 \quad \Rightarrow \quad \sigma \cong \frac{R}{4} = 2450 \quad \sigma^2 \cong 6002500$$

$$D = \frac{B^2}{4} = \frac{400^2}{4} = 40000$$

$$n = \frac{N\sigma^2}{\left((N-1)D\right) + \sigma^2} = 139,65 \approx 140$$

$$D = \frac{B^2}{4} = \frac{0,05^2}{4} = 0,000625$$

$$p = q = 0,5 \qquad n = \frac{Npq}{(N-1)D + pq} = 333,47 \approx 334$$

Para conseguir estimar los dos parámetros con los niveles de error especificados necesitamos un tamaño muestral igual al máximo de 140 y 334. $n = 334$.

5. Se desea estimar el salario medio de los empleados de una empresa y la proporción de empleados que apoyan a la actual directiva. La empresa tiene 110 empleados y se sabe que el salario está comprendido entre los 1500 y 1800 euros mensuales. ¿Cuál debe ser el tamaño muestral para que al estimar el salario medio la cota de error se sitúe en 10 euros y al estimar la proporción de los que apoyan a la actual directiva el error máximo cometido sea del 2%?

SOLUCIÓN:

$$N = 110 \quad R = 1800 - 1500 = 300 \quad \Rightarrow \quad \sigma \cong \frac{R}{4} = 75 \quad \sigma^2 \cong 5625$$

$$D = \frac{B^2}{4} = \frac{10^2}{4} = 25$$

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = 74{,}1 \approx 75$$

$$D = \frac{B^2}{4} = \frac{0{,}02^2}{4} = 0{,}0001$$

$$p = q = 0{,}5 \qquad n = \frac{Npq}{(N-1)D + pq} = 105{,}4 \approx 106$$

6. Una empresa de trabajo temporal quiere investigar las necesidades de empleo de las empresas de un pueblo. Para ello decide seleccionar una muestra de 5 de las 25 inscritas en el registro mercantil. El número de bajas en el último año, el número de empleados y la respuesta de cada empresa sobre si utilizaría los servicios de la empresa de trabajo temporal fueron los siguientes:

Empresa	Bajas	Empleados	Respuesta
1	1	7	Si
2	2	15	No
3	9	85	Si
4	0	3	No
5	2	12	No

a) Estime el número de bajas en el último año en las empresas del pueblo. Calcule el límite para el error de estimación.

b) Estime el número de empresas que usarían los servicios ofertados. Calcule el límite para el error de estimación.

SOLUCIÓN:

a) $N = 25 \qquad n = 5$

$$\bar{y} = \frac{14}{5} = 2,8 \Rightarrow \hat{\tau} = N\bar{y} = 70$$

$$\hat{V}(\hat{\tau}) = N(N-n)\frac{S^2}{n} = 25 \times 20 \frac{12,7}{5} = 1270$$

$$B = 2\sqrt{\hat{V}(\hat{\tau})} = 71,2741$$

Nota: este apartado podrá resolverse de otra forma cuando estudiemos el muestreo por conglomerados. Véase ejercicio resuelto 4 del tema 5.

b)

$$\hat{p} = \frac{2}{5} = 0,4 \Rightarrow \hat{\tau} = N\hat{p} = 10$$

$$\hat{V}(\hat{\tau}) = N(N-n)\frac{\hat{p}\hat{q}}{n-1} = 25 \times 20 \frac{0,24}{4} = 30$$

$$B = 2\sqrt{\hat{V}(\hat{\tau})} = 10,9545$$

7. El consumo medio de combustible de los taxis de una ciudad es 5,6 litros cada 100 Km. Puesto que se considera que el consumo es demasiado elevado, en 600 taxis se monta un dispositivo para disminuirlo. Pasado cierto tiempo se toma una muestra aleatoria de 20 taxis, elegidos entre los 600 que colocaron el dispositivo. El consumo en litros de combustible por cada 100 Km. se recoge en la siguiente tabla

Taxi nº	Consumo	Taxi nº	Consumo	Taxi nº	Consumo	Taxi nº	Consumo
1	5,4	6	6,3	11	3,6	16	5,4
2	5,5	7	5,4	12	6,7	17	4,8
3	6,9	8	5	13	5,2	18	4,7
4	3,9	9	4,5	14	5,1	19	5,8
5	4,5	10	4,4	15	5,4	20	6,2

a) Estímese mediante un intervalo de confianza la proporción de taxis con un consumo inferior a 5,6 litros/100 Km.

b) ¿Cuántos taxis deben observarse para estimar la anterior proporción con un error menor o igual que un 10%?

SOLUCIÓN:

a) 15 de los 20 taxis no superan el consumo de 5,6 litros/100 Km, por tanto

$$\hat{p} = \frac{15}{20} = 0,75 \qquad \hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N} = \frac{0,75 \times 0,25}{19}\frac{580}{600} = 0,00954$$

$$2\sqrt{\hat{V}(\hat{p})} = 0,1953$$

$$(0,75 - 0,1953 \quad , \quad 0,75 + 0,1953) = (0,5547 \quad , \quad 0,9453) \qquad \boxed{(55,47\% \quad , \quad 94,53\%)}$$

b) $B = 0{,}10 \qquad D = \dfrac{(0{,}10)^2}{4} = 0{,}0025$

$$n = \frac{Npq}{(N-1)D + pq} = \frac{600 \times 0{,}75 \times 0{,}25}{(599 \times 0{,}0025) + (0{,}75 \times 0{,}25)} = 66{,}77 \approx 67$$

8. Una muestra aleatoria simple de 6 deudas de clientes de una farmacia es seleccionada para estimar la cantidad total de deuda de las 100 cuentas abiertas. Los valores de la muestra para estas seis cuentas son los siguientes:

Dinero adeudado (€)
35,50
32,00
43,00
41,00
44,00
42,50

Estime el total del dinero adeudado y establezca un límite para el error de estimación.

SOLUCIÓN:

y_i	y_i^2
35,50	1260,25
32,00	1024,00
43,00	1849,00
41,00	1681,00
44,00	1936,00
42,50	1806,25
$\sum_{i=1}^{n} y_i = 238{,}00$	$\sum_{i=1}^{n} y_i^2 = 9556{,}50$

$$\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{100}{6}238 = 3966,\widehat{6}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{1}{5}\left(9556{,}50 - \frac{238^2}{6}\right) = 23{,}1667$$

$$2\sqrt{\hat{V}(\hat{\tau})} = 2\sqrt{N(N-n)\frac{S^2}{n}} = 2\sqrt{100(100-6)\frac{23{,}1667}{6}} = 381{,}02$$

Los anteriores cálculos que se han realizado a mano o con ayuda de una calculadora básica se simplifican notablemente si utilizamos una calculadora científica de uso común. Estas

calculadoras nos proporcionan los valores de un grupo de funciones estadísticas $\sum x^2,\ \ \sum x,\ \ \bar{x},\ \ \sigma_n\ \ y\ \ \sigma_{n-1}$ de forma inmediata.

$\sigma_n = s_x = desviación\ \ típica \qquad \sigma_{n-1} = S_x = cuasidesviación\ \ típica$

9. En un estudio sociológico, realizado en una pequeña ciudad, se hicieron llamadas telefónicas para estimar la proporción de hogares donde habita por lo menos una persona mayor de 65 años de edad. La ciudad tiene 5000 hogares, según la guía de teléfonos más reciente. Una muestra aleatoria simple de 300 hogares fue seleccionada de la guía. Al terminar la investigación de campo, de los 300 hogares muestreados, en 51 habita al menos una persona mayor de 65 años. Contraste la hipótesis de que en el 25% de los hogares de esa ciudad habita al menos una persona mayor de 65 años.

SOLUCIÓN: N=5000, n=300

$$\hat{p} = \frac{51}{300} = 0{,}17 \quad \hat{q} = 1 - \hat{p} = 0{,}83 \quad \hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N} = 0{,}00044359197 \quad 2\sqrt{\hat{V}(\hat{p})} = 0{,}0421$$

$$25\% \notin (17\% \mp 4{,}21\%) = (12{,}79\%,\ 21{,}21\%)$$

luego se rechaza la hipótesis de que en el 25% de los hogares de esa ciudad habita al menos una persona mayor de 65 años.

10. El gerente de un taller de maquinaria desea estimar el tiempo medio que necesita un operador para terminar una tarea sencilla. El taller tiene 45 operadores. Se seleccionaron aleatoriamente 5 operadores y se les tomó el tiempo. Los resultados obtenidos son los siguientes:

Tiempo(minutos)	4,2	5,1	7,9	3,8	5,3

¿Se puede aceptar la hipótesis de que el tiempo medio que necesitan los operarios del taller para terminar dicha tarea es inferior a 6 minutos?

SOLUCIÓN: (*con las funciones del modo SD de la calculadora*)

N=45, n=5 $\quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 5{,}26 \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = 2{,}563$

$$\hat{V}(\bar{y}) = \frac{S^2}{n}\frac{N-n}{N} = 0{,}4556 \qquad 2\sqrt{\hat{V}(\bar{y})} = 1{,}35 \quad Intervalo\ de\ confianza: (3{,}91\ \text{min.}, 6{,}61\ \text{min.})$$

Valores mayores e igual a 6 minutos pertenecen al intervalo de confianza, por tanto no podemos aceptar esa hipótesis.

11. Con objetivos benéficos, una asociación filantrópica ha solicitado firmas para una petición en 700 hojas. Cada hoja tiene espacio suficiente para 40 firmas pero en muchas de las hojas se ha obtenido un número menor. Contando el número de firmas por hoja en una muestra aleatoria de 50 hojas se han observado los siguientes resultados:

$$\sum_{i=1}^{50} Y_i = 1450;\ \sum_{i=1}^{50} Y_i^2 = 54496$$

¿Cuál sería la previsión más optimista y más pesimista en cuanto al número total de firmas recogidas para la petición?

SOLUCIÓN: N=700, n=50

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1450}{50} = 29 \qquad S^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = 254$$

$$\hat{\tau} = N\bar{y} = 20300 \qquad \hat{V}(\hat{\tau}) = N(N-n)\frac{S^2}{n} = 2311400 \qquad B = 2\sqrt{\hat{V}(\hat{\tau})} = 3040,66$$

$$\left(20300 \mp 3040,66\right) = \left(17259,34\ ,\ 23340,66\right)$$

Previsión más optimista: 23340

Previsión más pesimista: 17259

2. Muestreo aleatorio estratificado.

2.1 Selección de una muestra aleatoria estratificada.

Una muestra aleatoria estratificada se obtiene mediante la separación de los elementos de la población en conjuntos que no presenten intersección, llamados estratos, y la selección posterior de una muestra aleatoria simple en cada estrato.

Los estratos deben formarse de manera que los elementos de cada estrato sean lo más homogéneos que se pueda entre sí (más homogéneos que el conjunto de la población) y las diferencias entre un estrato y otro sean las mayores posibles. Esta forma de construir los estratos conduce a muestras con poca variabilidad entre las mediciones que producirán pequeñas varianzas de los estimadores y por tanto menores límites para los errores de estimación que con otros diseños de la muestra.

Otras ventajas adicionales que presenta este tipo de muestreo son las siguientes:

- A veces los estratos se corresponden con zonas compactas bien definidas con lo que se reduce el coste de la muestra.
- Además de las estimaciones para toda la población, este muestreo permite hacer estimaciones de los parámetros poblacionales para los estratos.

Antes de continuar fijemos la notación que va a utilizarse:

$L =$ número de estratos

$N =$ tamaño de la población

$n =$ tamaño de la muestra

$N_i =$ tamaño del estrato

$n_i =$ tamaño de la muestra del estrato i

$N = \sum_{i=1}^{L} N_i$

$n = \sum_{i=1}^{L} n_i$

$\mu_i =$ media poblacional del estrato i

$\bar{y}_i =$ media muestral del estrato i

$\tau_i =$ total poblacional del estrato i

$\sigma_i^2 =$ varianza poblacional del estrato i

$S_i^2 =$ cuasivarianza muestral del estrato i

p_i = proporción poblacional del estrato i $\quad\quad$ $\hat{p}_i$ = proporción muestral del estrato i

c_i = coste de una observación del estrato i

2.2 Estimación de la media, proporción y total poblacionales.

En cada estrato se ha realizado un muestreo aleatorio simple, para **variables numéricas** sabemos que en cada estrato $N_i \bar{y}_i$ es un estimador insesgado del total τ_i, estimaremos $\tau = \sum_{i=1}^{L} \tau_i$ por $\boxed{\hat{\tau}_{st} = \sum_{i=1}^{L} N_i \bar{y}_i}$ y la media poblacional $\mu = \frac{\tau}{N}$ mediante $\boxed{\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_i}$

En poblaciones infinitas no se conoce N_i, pero sí conocemos $\frac{N_i}{N} \Rightarrow \bar{y}_{st} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$.

En poblaciones infinitas no tiene sentido la estimación del total.

NOTA: $\bar{y}_{st} \neq \bar{y}$ en general ($\bar{y}$ = media muestral de las *n* observaciones)

$\hat{\tau}_{st} \neq \hat{\tau}$ en general ($\hat{\tau} = N\bar{y}$ = estimador del total según un *m.a. simple.*)

Varianza estimada de $\bar{y}_{st}$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \, \hat{V}(\bar{y}_i) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i} \frac{N_i - n_i}{N_i}$$

Se obtiene aplicando que la varianza de la suma de variables independientes es la suma de sus varianzas y que $V(kX) = k^2 V(X)$.

En poblaciones infinitas, $\frac{N_i - n_i}{N_i} = 1$. Además no se conoce N_i, pero sí conocemos $\frac{N_i}{N}$.

Simplificándose la anterior expresión, $\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} \frac{N_i^2}{N^2} \frac{S_i^2}{n_i} = \sum_{i=1}^{L} \left(\frac{N_i}{N} \right)^2 \frac{S_i^2}{n_i}$

Varianza estimada de $\hat{\tau}_{st}$

$$\hat{V}(\hat{\tau}_{st}) = N^2 \hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i} \frac{N_i - n_i}{N_i}$$

En el caso de **variables dicotómicas** los estimadores de la proporción y total poblacionales así como sus varianzas toman valores similares a los anteriores, salvo las diferencias ya comentadas en la lección anterior.

Estimador de la proporción poblacional p

$$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i$$

En poblaciones infinitas no se conoce N_i, pero sí conocemos $\frac{N_i}{N} \quad \Rightarrow \quad \hat{p}_{st} = \sum_{i=1}^{L} \frac{N_i}{N} \hat{p}_i$

Varianza estimada de $\hat{p}_{st}$

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \, \hat{V}(\hat{p}_i) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i \hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i}$$

En poblaciones infinitas, $\frac{N_i - n_i}{N_i} = 1$. Además no se conoce N_i, pero sí conocemos $\frac{N_i}{N}$.

Simplificándose la anterior expresión, $\hat{V}(\hat{p}_{st}) = \sum_{i=1}^{L} \frac{N_i^2}{N^2}\frac{\hat{p}_i \hat{q}_i}{n_i - 1} = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \frac{\hat{p}_i \hat{q}_i}{n_i - 1}$

Estimador del total poblacional τ

$$\hat{\tau}_{st} = N\hat{p}_{st} = \sum_{i=1}^{L} N_i \hat{p}_i$$

Varianza estimada de $\hat{\tau}_{st}$

$$\hat{V}(\hat{\tau}_{st}) = N^2 \hat{V}(\hat{p}_{st}) = \sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i \hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i}$$

Ejemplo 2.1. Se está interesado en determinar la audiencia de la publicidad televisiva en una cadena local de un municipio, se decide realizar una encuesta por muestreo para estimar el número de horas por semana que se ve la televisión en las viviendas del municipio. Éste está formado por tres barrios con diferentes perfiles socio-culturales que afectan a la audiencia televisiva. Hay 210 hogares en el barrio A, 84 en el barrio B y 126 en el barrio C. La empresa publicitaria tiene tiempo y dinero suficientes para entrevistar 30 hogares y decide seleccionar muestras aleatorias de tamaños: 15 del barrio A, 6 del barrio B, y 9 del barrio C.

Se seleccionan las muestras aleatorias simples y se realizan las entrevistas. Los resultados, con mediciones del tiempo que se ve la televisión en horas por semana, se muestran en la siguiente tabla:

BARRIO A			BARRIO B		BARRIO C	
36	34	26	20	25	14	22
39	38	32	30		15	17
38	37	29	14		21	11
28	41	35	41		20	14
29	37	41	39		24	

Estime el tiempo medio que se ve la televisión y el límite para el error de estimación, en horas por semana, para:

a) Los hogares del barrio A.
b) Los hogares del barrio B.
c) Los hogares del barrio C.
d) Todos los hogares

Solución: en primer lugar se calculan las medias y cuasivarianzas muestrales en cada estrato

$$\bar{y}_1 = 34,67 \; horas/semana \quad \bar{y}_2 = 28,17 \; h/s \quad \bar{y}_3 = 17,56 \; h/s \quad S_1^2 = 23,24 \quad S_2^2 = 112,57 \quad S_3^2 = 19,28$$
$$\bar{y} = 28,23 \qquad S^2 = 92,74$$

A partir de estos valores calculamos las varianzas de los estimadores de la media en cada estrato y los límites para los errores de dichas estimaciones

$$N_1 = 210 \quad N_2 = 84 \quad N_3 = 126 \quad N = N_1 + N_2 + N_3 = 420$$

$$n_1 = 15 \quad n_2 = 6 \quad n_3 = 9 \quad n = n_1 + n_2 + n_3 = 30$$

$$\hat{V}(\bar{y}_1) = \frac{S_1^2}{n_1}\frac{N_1 - n_1}{N_1} = 1,44 \qquad \hat{V}(\bar{y}_2) = \frac{S_2^2}{n_2}\frac{N_2 - n_2}{N_2} = 17,42 \qquad \hat{V}(\bar{y}_3) = \frac{S_3^2}{n_3}\frac{N_3 - n_3}{N_3} = 1,99$$
$$2\sqrt{\hat{V}(\bar{y}_1)} = 2,40 \; h/s \qquad 2\sqrt{\hat{V}(\bar{y}_2)} = 8,35 \; h/s \qquad 2\sqrt{\hat{V}(\bar{y}_3)} = 2,82 \; h/s$$

Para el conjunto de todos los hogares el estimador de la media es

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{3} N_i \bar{y}_i = 28,23 \; h/s$$

y la varianza de este estimador la podemos calcular basándonos en las varianzas de los estimadores de la media en cada estrato mediante

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2 \, \hat{V}(\bar{y}_i) = 1,24$$

o, si se prefiere, utilizando

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i}$$

el error para la estimación de la media para todos los hogares está dado por

$$2\sqrt{\hat{V}(\bar{y}_{st})} = 2,22 \; h/s$$ ■

Ejemplo 2.2. En el caso anterior, también se desea saber qué proporción de hogares ven un determinado programa, para decidir la conveniencia de insertar un anuncio en los intermedios del mismo. La respuesta a la pregunta de si ven dicho programa en los hogares de la muestra anterior se recoge a continuación:

BARRIO A			BARRIO B		BARRIO C	
SI	NO	SI	SI	SI	NO	SI
SI	SI	SI	NO		SI	SI
NO	NO	NO	SI		SI	SI
NO	SI	NO	SI		NO	NO
SI	NO	NO	SI		SI	

Estime con un intervalo de confianza la proporción de hogares del municipio donde se ve el programa.

Solución: en primer lugar se calculan las proporciones muestrales en cada estrato

$$\hat{p}_1 = \frac{7}{15} = 0,4667 \qquad \hat{p}_2 = \frac{5}{6} = 0,8333 \qquad \hat{p}_3 = \frac{6}{9} = 0,6667$$

La estimación puntual de la proporción de hogares del municipio donde se ve el programa es

$$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{3} N_i \hat{p}_i = 0,60$$

la varianza y error de estimación asociados son

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2 \frac{\hat{p}_i \hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i} = 0,00748 \qquad 2\sqrt{\hat{V}(\hat{p}_{st})} = 0,173$$

y el intervalo de confianza expresado en porcentajes es

$$(60\% \mp 17,3\%) = (42,7\%, 77,3\%)$$ ■

2.3 Determinación del tamaño muestral.

El tamaño muestral para conseguir un límite para el error de estimación de la media, *B*, viene dado por $2\sqrt{V(\bar{y}_{st})} = B$ donde $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{\sigma_i^2}{n_i}\frac{N_i - n_i}{N_i - 1}$.

No podemos despejar el valor de todos los n_i de una sola ecuación a menos que conozcamos la relación entre los n_i y n. Hay diversas formas de asignar el tamaño muestral *n* en los diferentes estratos (problema de la asignación de la muestra que estudiaremos a continuación), sustituyendo $n_i = n\omega_i$ en $V(\bar{y}_{st})$ se puede despejar *n* en función de los ω_i obteniendo el tamaño muestral *aproximado* que se requiere para estimar μ con un límite para el error de estimación *B*.

En **variables numéricas**

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 \sigma_i^2}{\omega_i}}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$$

$D = \frac{B^2}{4}$ y la misma expresión vale para el total tomando $D = \frac{B^2}{4N^2}$.

Al igual que en el m.a.s. para poder usar la anterior ecuación necesitamos conocer las varianzas poblacionales de los estratos o valores aproximados de ellas, para lo cual se pueden usar las cuasivarianzas muestrales de una muestra previa o basarnos en el rango de variación de las observaciones dentro de cada estrato.

En el caso de **variables dicotómicas** se obtiene una expresión similar, teniendo en cuenta que en este caso particular $\sigma_i^2 = p_i q_i$

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 p_i q_i}{\omega_i}}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i}$$

$D = \frac{B^2}{4}$ (para estimar p) y la misma expresión vale para el total tomando $D = \frac{B^2}{4N^2}$.

2.4 Asignación de la muestra.

Hay diversas formas de asignar el tamaño muestral *n* en los distintos estratos.

El objetivo del diseño de una encuesta por muestreo es proporcionar estimadores con varianza pequeña (por tanto, pequeño error de estimación) al menor coste posible.

El mejor esquema de asignación está influido por:

- El número total de elementos en cada estrato.
- La variabilidad de las observaciones en cada estrato.
- El coste de obtener una observación en cada estrato.

2.4.1 Asignación Óptima.

La asignación que minimiza el coste para un límite para el error de estimación fijado, *B*, se denomina asignación Óptima y está dada en **variables numéricas** por

$$\omega_j = \frac{\frac{N_j \sigma_j}{\sqrt{c_j}}}{\sum_{i=1}^{L} \frac{N_i \sigma_i}{\sqrt{c_i}}}$$

sustituyendo los ω_j en la expresión que obteníamos antes para *n* se tiene el tamaño total de la muestra según la asignación Óptima

$$n = \frac{\sum_{i=1}^{L} N_i \sigma_i \sqrt{c_i} \quad \sum_{i=1}^{L} \frac{N_i \sigma_i}{\sqrt{c_i}}}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$$

En el **caso dicotómico** las anteriores expresiones toman los valores

$$\omega_j = \frac{N_j \sqrt{\frac{p_j q_j}{c_j}}}{\sum_{i=1}^{L} N_i \sqrt{\frac{p_i q_i}{c_i}}}$$

$$n = \frac{\sum_{i=1}^{L} N_i \sqrt{p_i q_i c_i} \quad \sum_{i=1}^{L} N_i \sqrt{\frac{p_i q_i}{c_i}}}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i}$$

$D = \frac{B^2}{4}$ (para estimar la media o p), $D = \frac{B^2}{4N^2}$ (para estimar el total).

En algunas ocasiones interesa encontrar la asignación que minimiza el error de estimación para un coste total fijo de obtención de la muestra, C. En este caso la asignación Óptima también es la respuesta y el tamaño total de la muestra, n, viene dado para **variables numéricas** por:

$$n = \frac{C \sum_{i=1}^{L} \frac{N_i \sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} N_i \sigma_i \sqrt{c_i}}$$

Análogamente para el **caso dicotómico** sustituyendo $\sigma_i = \sqrt{p_i q_i}$

$$n = \frac{C \sum_{i=1}^{L} N_i \sqrt{\frac{p_i q_i}{c_i}}}{\sum_{i=1}^{L} N_i \sqrt{p_i q_i c_i}}$$

En **poblaciones infinitas** no se conoce N_i, pero sí conocemos el tamaño relativo de cada estrato (peso del estrato) $\frac{N_i}{N} = W_i$, calculándose los anteriores valores así:

Variables numéricas

$$\omega_i = \frac{\frac{W_i\,\sigma_i}{\sqrt{c_i}}}{\sum_{j=1}^{L}\frac{W_j\,\sigma_j}{\sqrt{c_j}}} \qquad n = \frac{\sum_{i=1}^{L} W_i\,\sigma_i\sqrt{c_i}\ \ \sum_{i=1}^{L}\frac{W_i\,\sigma_i}{\sqrt{c_i}}}{D}$$

$D = \frac{B^2}{4}$ para estimar la media, en poblaciones infinitas no tiene sentido la estimación del total.

Para un coste total fijo de obtención de la muestra, C

$$n = \frac{C\sum_{i=1}^{L}\frac{W_i\,\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i\,\sigma_i\sqrt{c_i}}$$

Variables dicotómicas

$$\omega_i = \frac{W_i\sqrt{\frac{p_i q_i}{c_i}}}{\sum_{j=1}^{L} W_j\sqrt{\frac{p_j q_j}{c_j}}} \qquad n = \frac{\sum_{i=1}^{L} W_i\ \sqrt{p_i q_i c_i}\ \ \sum_{i=1}^{L} W_i\ \sqrt{\frac{p_i q_i}{c_i}}}{D}$$

$D = \frac{B^2}{4}$ para estimar la proporción, en poblaciones infinitas no tiene sentido la estimación del total.

Para un coste total fijo de obtención de la muestra, C

$$n = \frac{C\sum_{i=1}^{L} W_i\ \sqrt{\frac{p_i q_i}{c_i}}}{\sum_{i=1}^{L} W_i\ \sqrt{p_i q_i c_i}}$$

2.4.2 Asignación de Neyman.

A veces los costes de observación no se conocen. Si solo se consideran los tamaños de los estratos y su varianza, la mejor asignación se conoce como asignación de Neyman. Cuando los costes de observación de cada estrato son los mismos, la asignación Óptima y de Neyman coinciden.

Variables numéricas

$$\omega_j = \frac{N_j \sigma_j}{\sum_{i=1}^{L} N_i \sigma_i} \qquad n = \frac{\left(\sum_{i=1}^{L} N_i \sigma_i\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$$

Variables dicotómicas

$$\omega_j = \frac{N_j \sqrt{p_j q_j}}{\sum_{i=1}^{L} N_i \sqrt{p_i q_i}} \qquad n = \frac{\left(\sum_{i=1}^{L} N_i \sqrt{p_i q_i}\right)^2}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i}$$

En **poblaciones infinitas** no se conoce N_i, pero sí conocemos el tamaño relativo de cada estrato (peso del estrato) $\frac{N_i}{N} = W_i$, calculándose los anteriores valores así:

Variables numéricas

$$\omega_i = \frac{W_i \sigma_i}{\sum_{j=1}^{L} W_j \sigma_j} \qquad n = \frac{\left(\sum_{i=1}^{L} W_i \sigma_i\right)^2}{D}$$

Variables dicotómicas

$$\omega_i = \frac{W_i \sqrt{p_i q_i}}{\sum_{j=1}^{L} W_j \sqrt{p_j q_j}} \qquad n = \frac{\left(\sum_{i=1}^{L} W_i \sqrt{p_i q_i}\right)^2}{D}$$

$D = \frac{B^2}{4}$ para estimar la media o la proporción, en poblaciones infinitas no tiene sentido la estimación del total.

2.4.3 Asignación Proporcional.

Si solo se consideran los tamaños de los estratos, la mejor asignación se conoce como asignación Proporcional. Cuando los costes de observación de cada estrato son los mismos y también sus varianzas la asignación Óptima, de Neyman y Proporcional coinciden.

Variables numéricas

$$\omega_j = \frac{N_j}{N} \qquad n = \frac{\sum_{i=1}^{L} N_i \sigma_i^2}{ND + \frac{1}{N}\sum_{i=1}^{L} N_i \sigma_i^2}$$

Variables dicotómicas

$$\omega_j = \frac{N_j}{N} \qquad n = \frac{\sum_{i=1}^{L} N_i p_i q_i}{ND + \frac{1}{N}\sum_{i=1}^{L} N_i p_i q_i}$$

En **poblaciones infinitas** no se conoce N_i, pero sí conocemos el tamaño relativo de cada estrato (peso del estrato) $\frac{N_i}{N} = W_i$, calculándose los anteriores valores así:

Variables numéricas

$$\omega_i = W_i \qquad n = \frac{\sum_{i=1}^{L} W_i \ \sigma_i^2}{D}$$

Variables dicotómicas

$$\omega_i = W_i \qquad n = \frac{\sum_{i=1}^{L} W_i \ p_i q_i}{D}$$

$D = \frac{B^2}{4}$ para estimar la media o la proporción, en poblaciones infinitas no tiene sentido la estimación del total.

La asignación Proporcional puede y suele utilizarse cuando las varianzas y costes de observación no son iguales para cada estrato, por la simplicidad de los cálculos y por las ventajas que presenta frente a los anteriores tipos de asignaciones:

Cuando se utiliza la asignación Proporcional el estimador $\overline{y}_{st}$ coincide con la media muestral de la muestra que reúne a todas las muestras de cada estrato, $\overline{y}_{st} = \overline{y}$ (análogamente para $\widehat{p}_{st}$ y el total).

Cuando se observa más de una variable en cada unidad muestral para estimar más de un parámetro poblacional aparecen complicaciones en la asignación y determinación del tamaño muestral. Con la asignación Proporcional y tomando como n el máximo de los valores encontrados para cada estimación se resuelve el problema como puede verse en el siguiente ejemplo:

En la asignación Óptima y en la de Neyman los ω_i dependen de las varianzas y pueden ser distintos de una variable a otra

1ª estimación: $n = 100 \quad \omega_1 = 0,10 \Rightarrow n_1 = 10 \quad \omega_2 = 0,90 \Rightarrow n_2 = 90$

2ª estimación: $n = 40 \quad \omega_1 = 0,50 \Rightarrow n_1 = 20 \quad \omega_2 = 0,50 \Rightarrow n_2 = 20$

Aún tomando el mayor de los tamaños muestrales (100) y pasando la encuesta a 10 individuos del estrato 1 y 90 del estrato 2 no tenemos garantizado que se satisfaga el error de estimación fijado para la segunda estimación que necesita al menos 20 individuos de cada estrato.

En la asignación Proporcional no ocurre lo anterior pues los $\omega_j = \frac{N_j}{N}$ son iguales para todas las variables al no depender de sus varianzas, así si en dos estimaciones para los niveles de error requeridos tenemos lo siguiente

1ª estimación: $n = 100 \quad \omega_1 = 0,30 \Rightarrow n_1 = 30 \quad \omega_2 = 0,70 \Rightarrow n_2 = 70$

2ª estimación: $n = 40 \quad \omega_1 = 0,30 \Rightarrow n_1 = 12 \quad \omega_2 = 0,70 \Rightarrow n_2 = 28$

tomando como n el máximo de los dos (y en general para k variables), se tiene garantizado que se cumple con los límites para el error fijados para todas las estimaciones.

Ejemplo 2.3 Continuando con el ejemplo 2.1

a) ¿Qué tipo de asignación se ha utilizado?

Debido a los traslados necesarios no cuesta lo mismo obtener una observación en un barrio que en otro. Se estima que el coste de una observación del barrio A es de 1€, 9€ para el barrio B y 4€ para el barrio C.

b) Cuántos hogares deberían entrevistarse para estimar el número medio de horas a la semana que se ve la televisión en los hogares del municipio con un error inferior a 1 hora. (Tómese los anteriores datos como una muestra previa para estimar los parámetros necesarios).
c) Supóngase que se tiene sólo 600€ para gastar en el estudio, determine el tamaño de la muestra y la asignación que minimizan el error de estimación. (Como en el apartado anterior, tómese los datos de la tabla como una muestra previa para estimar las varianzas de los estratos).

Solución:

a) Podemos comprobar que se cumple $n_i = \frac{N_i}{N} n \quad \left(\omega_i = \frac{N_i}{N} \right) \quad \forall i$ o equivalentemente que $\frac{n_i}{n} = \frac{N_i}{N} \quad \forall i$

$$\frac{15}{30} = \frac{210}{420} = 0,5 \qquad \frac{6}{30} = \frac{84}{420} = 0,2 \qquad \frac{9}{30} = \frac{126}{420} = 0,3$$

luego la asignación utilizada ha sido la Proporcional.

b) Según los datos anteriores estimamos las varianzas de cada estrato por

$$\widehat{\sigma}_1^2 = S_1^2 = 23,24 \qquad \widehat{\sigma}_2^2 = S_2^2 = 112,56 \qquad \widehat{\sigma}_3^2 = S_3^2 = 19,28$$

N_i	σ_i	$\sqrt{c_i}$	$N_i\sigma_i\sqrt{c_i}$	$\frac{N_i\sigma_i}{\sqrt{c_i}}$	$N_i\sigma_i^2$
210	4,8208	1	1012,368	1012,368	4880,4
84	10,6094	3	2673,5688	297,0632	9455,04
126	4,3909	2	1106,5068	276,6267	2429,28
420			4792,4436	1586,0579	16764,72

$$D = \frac{B^2}{4} = \frac{1}{4} = 0,25$$

$$n = \frac{\sum_{i=1}^{3} N_i\,\sigma_i\,\sqrt{c_i} \quad \sum_{i=1}^{3} \frac{N_i\,\sigma_i}{\sqrt{c_i}}}{N^2 D + \sum_{i=1}^{3} N_i\sigma_i^2} = \frac{4792,4436 \times 1586,0579}{\left(420^2 \times 0,25\right) + 16764,72} = 124,89$$

$$\omega_1 = \frac{\frac{N_1\sigma_1}{\sqrt{c_1}}}{\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}} = 0,6383 \qquad \omega_2 = 0,1873 \qquad \omega_3 = 0,1744$$

$$n_1 = n\omega_1 = 79,71 \approx 80 \qquad n_2 = n\omega_2 = 23,39 \approx 24 \qquad n_3 = n\omega_3 = 21,78 \approx 22$$

$$n = 80 + 24 + 22 = 126$$

c) En el supuesto de que se disponga sólo de 600€ para realizar el estudio

$$n = \frac{600 \sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} N_i\sigma_i\sqrt{c_i}} = \frac{600 \times 1586,0679}{4792,4436} = 198,57$$

y la asignación de la muestra en cada estrato está dada por la asignación Óptima

$$n_1 = 0,6383n = 126,75 \approx 126 \qquad n_2 = 0,1873n = 37,19 \approx 37 \qquad n_3 = 0,1744n = 34,63 \approx 34$$
$$n = 126 + 37 + 34 = 197$$

o bien resolviendo la ecuación

$$c_1 n_1 + c_2 n_2 + c_3 n_3 = 600$$

donde $n_i = \omega_i n$

$$c_1\omega_1 n + c_2\omega_2 n + c_3\omega_3 n = 600$$

$$n = \frac{600}{c_1\omega_1 + c_2\omega_2 + c_3\omega_3} = \frac{600}{3,0216} = 198,57$$

A partir de *n* se obtienen los $n_i = \omega_i n$ según la asignación Óptima. ■

Ejemplo 2.4 Continuando con el ejemplo 2.2

a) Cuántos hogares deberían entrevistarse si se quisiera hacer dicha estimación con un error inferior al 5%. (Supóngase que se realiza la entrevista por teléfono y el coste de las observaciones es el mismo para todos los casos al no ser necesarios los traslados. Tómese los anteriores datos como una muestra previa para estimar los parámetros necesarios)

b) Respóndase a la pregunta anterior pero suponiendo que no se tiene ninguna información previa sobre la proporción de hogares donde se ve el programa.

Solución: a)

N_i	p_i	q_i	$N_i p_i q_i$	$N_i\sqrt{p_i q_i}$
210	0,4667	0,5333	52,2671	104,7669
84	0,8333	0,1667	11,6685	31,3075
126	0,6667	0,3333	27,9986	59,3955
420			91,9342	195,4699

$$D=\frac{B^2}{4}=\frac{0,05^2}{4}=0,000625$$

$$n=\frac{\left(\sum_{i=1}^{3}N_i\sqrt{p_i q_i}\right)^2}{N^2D+\sum_{i=1}^{3}N_i p_i q_i}=\frac{195,4699^2}{\left(420^2\times 0,000625\right)+91,9342}=188,98$$

$$n_1=n\omega_1=n\frac{N_1\sqrt{p_1 q_1}}{\sum_{i=1}^{3}N_i\sqrt{p_i q_i}}=188,98\frac{104,7669}{195,4699}=101,29\approx 102$$

análogamente $n_2=30,27\approx 31$ $\quad n_3=57,42\approx 58$ $\quad\Rightarrow\quad n=102+31+58=191$

b)

N_i	p_i	q_i	$N_i p_i q_i$
210	0,5	0,5	52,5
84	0,5	0,5	21
126	0,5	0,5	31,5
420			105

$$n=\frac{\sum_{i=1}^{L}N_i p_i q_i}{ND+\frac{1}{N}\sum_{i=1}^{L}N_i p_i q_i}=\frac{105}{\left(420\times 0,000625\right)+\frac{105}{420}}=204,878$$

$n_1=204,878\frac{210}{420}=102,439\approx 103$ análogamente $n_2=40,98\approx 41$ $\quad n_3=61,46\approx 62$

$$n=103+41+62=206$$ ■

El muestreo estratificado no siempre conduce a un estimador con menor error de estimación, esto suele ocurrir cuando los estratos no están formados por elementos suficientemente homogéneos. Muchas veces es debido a que predomina el deseo de obtener estimaciones en cada estrato (por ejemplo, en un estudio regional también se quieren obtener estimaciones a nivel provincial) frente al objetivo de minimizar los errores de los estimadores. Este problema queda bien ilustrado con el siguiente ejemplo.

Ejemplo 2.5 Un distribuidor de productos de limpieza desea conocer el consumo por hogar durante un año de un determinado producto en una comarca formada por cuatro municipios. Para estimar de paso también el consumo en cada municipio decide usar muestreo estratificado tomando cada municipio como un estrato. Se sabe que el 20% de la población de la comarca vive en el municipio 1, el 30% en el municipio 2, el 25% en el municipio 3 y el 25% restante en el municipio 4. El distribuidor tiene medios suficientes para controlar y obtener datos sobre el consumo anual de 20 hogares.

Dado que no tiene información previa respecto a las varianzas de los estratos y porque el coste del muestreo es el mismo en cada municipio, decide aplicar asignación Proporcional, la cual conduce a

$$n_1 = n\frac{N_1}{N} = 20\times 0,20 = 4 \quad \text{de forma similar} \quad n_2 = 6 \quad n_3 = 5 \quad n_4 = 5\,.$$

Obteniendo los resultados de la tabla siguiente (consumo expresado en euros).

Estrato 1	Estrato 2	Estrato 3	Estrato 4
470	490	540	450
510	500	480	560
500	470	500	460
550	520	470	440
	550	470	580
	500		
$\bar{y}_1 = 507,5 \quad S_1^2 = 1091,67$	$\bar{y}_2 = 505 \quad S_2^2 = 750$	$\bar{y}_3 = 492 \quad S_3^2 = 870$	$\bar{y}_4 = 498 \quad S_4^2 = 4420$

Estime el consumo anual medio por hogar y fije un límite para el error de estimación.

Solución: $\frac{N_1}{N} = 0,20 \qquad \frac{N_2}{N} = 0,30 \qquad \frac{N_3}{N} = 0,25 \qquad \frac{N_4}{N} = 0,25$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{4} N_i \bar{y}_i = \sum_{i=1}^{4} \frac{N_i}{N}\bar{y}_i = (0,20\times 507,5) + (0,30\times 505) + (0,25\times 492) + (0,25\times 498) = 500,5€$$

Obsérvese que cuando se utiliza la asignación Proporcional $\bar{y}_{st} = \bar{y}$, efectivamente

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{20} y_i = \frac{10010}{20} = 500,5€$$

En la siguiente expresión consideramos los coeficientes correctores para poblaciones finitas en cada estrato iguales a la unidad

$$\widehat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{4} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = \sum_{i=1}^{4} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i} = \sum_{i=1}^{2}\left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i} =$$

$$= \left(0,20^2 \frac{1091,67}{4}\right) + \left(0,30^2 \frac{750}{6}\right) + \left(0,25^2 \frac{870}{5}\right) + \left(0,25^2 \frac{4420}{5}\right) = 88,29$$

$$2\sqrt{\widehat{V}(\bar{y}_{st})} = 18,79\ €$$

Supongamos que el distribuidor hubiera decidido tomar una muestra aleatoria simple de 20 hogares, los mismos 20 de la tabla anterior, entonces el estimador de la media hubiera sido

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{20} y_i = 500,5\ €$$

que coincide con el estimador del muestreo estratificado por las razones mencionadas anteriormente, pero la varianza estimada y error de estimación asociados tomarían los valores

$$S_{n-1}^2 = 1520,79$$

$$\widehat{V}(\bar{y}) = \frac{S_{n-1}^2}{n}\frac{N-n}{N} = \frac{1520,79}{20} = 76,04 \qquad \text{, se supone } \frac{N-n}{N} \cong 1$$

$$2\sqrt{\widehat{V}(\bar{y})} = 17,44\ €$$

Se observa que el error de estimación es menor en el caso del muestreo aleatorio simple, esto es debido a que el distribuidor no tuvo en cuenta que el consumo varía mucho dentro del cuarto municipio. Pudo haber obtenido un error menor si hubiera estratificado en base al tamaño de las familias u hogares, esto es, colocando los hogares pequeños en un estrato, los medianos en otro, ..., es decir, formando los estratos con hogares que tengan un consumo similar. ■

2.5 Estratificación después de seleccionar la muestra.

A veces no se sabe a qué estrato pertenece un dato hasta que no se observa (por ejemplo, estratos según sexo y entrevista telefónica).

Supóngase una muestra aleatoria simple de n personas para una encuesta. La muestra puede ser dividida en n_1 masculinos y n_2 femeninos después de que ha sido realizada. Entonces en lugar de usar $\bar{y}$ para estimar μ, podemos usar $\bar{y}_{st}$ siempre que $\frac{N_i}{N}$ sea conocido para todo *i*.

Obsérvese que en esta situación los n_i son aleatorios, ya que varían de una muestra a otra aunque n sea fijo. Luego esto no es una muestra aleatoria estratificada en pleno sentido, pero

si $\frac{N_i}{N}$ es conocido y $n_i \geq 20 \quad \forall i$, entonces este método de estratificar después de la selección es casi tan exacto como el muestreo aleatorio estratificado con asignación Proporcional. Este método no debe usarse si $\frac{N_i}{N}$ o una buena aproximación de su valor se desconocen.

Ejemplo 2.6 En una ciudad se sabe que el 30% de los hogares tienen calefacción eléctrica. Al realizar una encuesta sobre el consumo de energía (valor en euros de la factura bimensual) se obtuvieron los siguientes resultados:

Tipo Calefacción	Nº casas	Valor total de las facturas	cuasidesviación típica muestral
Eléctrica	60	5730	200
No eléctrica	40	2080	90

Obtenga una estimación del valor medio de la factura de electricidad en la ciudad y el límite para el error de estimación.

Solución: Ya que la proporción observada de facturas de hogares con calefacción eléctrica (0,60=60/100) está muy alejada de la proporción verdadera (0,30), es conveniente la estratificación después de que se ha seleccionado la muestra aleatoria simple. Además el procedimiento se justifica pues tanto n_1 como n_2 superan 20.

$$\bar{y}_1 = \frac{5730}{60} = 95,5€ \qquad \bar{y}_2 = \frac{2080}{40} = 52€$$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{2} N_i \bar{y}_i = \sum_{i=1}^{2} \frac{N_i}{N} \bar{y}_i = (0,30 \times 95,5) + (0,70 \times 52) = 65,05€$$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{2} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = \sum_{i=1}^{2} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i}$$

omitiendo el coeficiente corrector por poblaciones finitas se tiene

$$\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{2} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i} = \sum_{i=1}^{2}\left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i} = \left(0,30^2 \frac{200^2}{60}\right) + \left(0,70^2 \frac{90^2}{40}\right) = 159,225$$

$$2\sqrt{\hat{V}(\bar{y}_{st})} = 25,24€$$ ■

A veces este método de estimación se utiliza para ajustar por no respuesta. Por ejemplo, si muchos de quienes no respondieron a una muestra aleatoria simple son varones, entonces la proporción de varones en la muestra va a ser pequeña, y se podría conseguir un estimador ajustado mediante la estratificación después del muestreo.

Así, en este ejemplo la baja representación en la muestra de facturas sin calefacción eléctrica y la alta de facturas con calefacción eléctrica conducen a una sobreestimación del valor medio de las facturas si se utiliza muestreo aleatorio simple y no se ajusta la estimación de la media con la estratificación después de seleccionar la muestra:

$$\bar{y} = \frac{5730 + 2080}{60 + 40} = \frac{7810}{100} = 78{,}10€$$

Con el muestreo aleatorio simple sobrevaloraríamos el consumo medio de electricidad por hogar (78,10>>65,05).

APÉNDICE: Estudio empírico de todas las posibles muestras en un muestreo aleatorio estratificado sobre una población finita.

Habitualmente el tamaño de la población es grande y no es posible comprobar empíricamente las propiedades del muestreo aleatorio estratificado que hemos estudiado en este tema. A continuación vamos a suponer una población con sólo 6 elementos, la misma que hemos utilizado en el anterior tema:

①②③ ⓫⓬⓭	Población finita	Media poblacional, μ	Varianza poblacional, σ^2
	N=6	$\mu = \frac{1+2+3+11+12+13}{6} = 7$	25,6667 **(#)**

$$\frac{(1-7)^2 + (2-7)^2 + (3-7)^2 + (11-7)^2 + (12-7)^2 + (13-7)^2}{6} = 25,6667 \quad \textbf{(\#)}$$

Según lo visto en el tema anterior podría pensarse que la única forma de reducir la variabilidad de los estimadores muestrales y por tanto el error de estimación es aumentando el tamaño de la muestra (como vimos al pasar de muestras de tamaño 4 a tamaño 5). Con el siguiente ejemplo mostramos como con un diseño de la muestra diferente, usando muestreo estratificado, con muestras también de tamaño 4 se consigue mucha menos variabilidad del estimador media muestral.

En el muestreo aleatorio simple al pasar de muestras de tamaño 4 a tamaño 5 la varianza de la media muestral se reduce algo más de la mitad (2,5667→1,0267). Con el muestreo aleatorio estratificado, manteniendo el tamaño muestral 4, la reducción ha sido de más de 30 veces (2,5667→0,0833). Veámoslo:

$$\frac{(1-7)^2+(2-7)^2+(3-7)^2+(11-7)^2+(12-7)^2+(13-7)^2}{6}=25,6667\ (+)$$

$$\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3}=0,6667\quad (\#)$$

①②③ ⓫⓬⓭	*Población finita*	*Media poblacional,* μ	*Varianza poblacional,* σ^2
	N=6	$\mu=\frac{1+2+3+11+12+13}{6}=7$	25,6667 (+)
①②③	*Estrato 1*	*Media poblacional,* μ_1	*Varianza poblacional,* σ_1^2
	N_1=3	$\mu_1=\frac{1+2+3}{3}=2$	0,6667 (#)
⓫⓬⓭	*Estrato 2*	*Media poblacional,* μ_2	*Varianza poblacional,* σ_2^2
	N_2=3	$\mu_2=\frac{11+12+13}{3}=12$	0,6667

MUESTRAS ***n*=4** **(n_1=2, n_2=2)**		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_{st}$
②③	⓬⓭	$\frac{2+3}{2}=2,5$	$\frac{12+13}{2}=12,5$	$\frac{(3\times 2,5)+(3\times 12,5)}{6}=7,5$
②③	⓫⓭	2,5	12	7,25
②③	⓫⓬	2,5	11,5	7
①③	⓬⓭	2	12,5	7,25
①③	⓫⓭	2	12	7
①③	⓫⓬	2	11,5	6,75
①②	⓬⓭	1,5	12,5	7
①②	⓫⓭	1,5	12	6,75
①②	⓫⓬	1,5	11,5	6,5
TOTAL:		18	108	63
MEDIA:		$E\left[\bar{y}_1\right]=\frac{18}{9}=2$	$E\left[\bar{y}_2\right]=\frac{108}{9}=12$	$E\left[\bar{y}_{st}\right]=\frac{63}{9}=7$
VARIANZA:		$V\left[\bar{y}_1\right]=0,1667$	$V\left[\bar{y}_2\right]=0,1667$	$V\left[\bar{y}_{st}\right]=0,0833$

Observemos cómo el muestreo aleatorio estratificado no conduce a mejores estimadores si los estratos no son más homogéneos que el conjunto de la población y/o no hay claras diferencias entre los elementos de uno y otro estrato. Consideramos la siguiente población:

①②③ ❶❷❸	Población finita	Media poblacional, μ	Varianza poblacional, σ^2
	N=6	$\mu=\frac{1+2+3+1+2+3}{6}=2$	0,6667

Tomamos todas las muestras aleatorias simples de tamaño 4 para estimar la media poblacional, usando la media muestral.

	MUESTRAS (*n*=4)	Media muestral, $\bar{y}$
1	③❶❷❸	$\frac{3+1+2+3}{4}=2,25$
2	②❶❷❸	2
3	②③❷❸	2,5
4	②③❶❸	2,25
5	②③❶❷	2
6	①❶❷❸	1,75
7	①③❷❸	2,25
8	①③❶❸	2
9	①③❶❷	1,75
10	①②❷❸	2
11	①②❶❸	1,75
12	①②❶❷	1,5
13	①②③❸	2,25
14	①②③❷	2
15	①②③❶	1,75
	TOTAL:	30

MUESTRAS (*n*=4)	Media muestral, $\bar{y}$
MEDIA:	$E\left[\bar{y}\right]=\frac{30}{15}=2$
VARIANZA:	$V\left[\bar{y}\right]=0,0667$

Ahora separamos la población en dos estratos, blancas y negras. Como puede observarse no hay diferencias entre los dos estratos y no presentan más homogeneidad que los datos de la población completa.

①②③ ❶❷❸	*Población finita*	*Media poblacional,* μ	*Varianza poblacional,* σ^2
	N=6	$\mu=\frac{1+2+3+1+2+3}{6}=2$	0,6667
①②③	*Estrato 1*	*Media poblacional,* μ_1	*Varianza poblacional,* σ_1^2
	N_1=3	$\mu_1=\frac{1+2+3}{3}=2$	0,6667
❶❷❸	*Estrato 2*	*Media poblacional,* μ_2	*Varianza poblacional,* σ_2^2
	N_2=3	$\mu_2=\frac{1+2+3}{3}=2$	0,6667

Tomamos todas las muestras aleatorias estratificadas de tamaño 4, asignando dos observaciones a cada estrato y estimamos la media poblacional usando el estimador $\bar{y}_{st}$ del muestreo aleatorio estratificado.

MUESTRAS n=4 (n_1=2, n_2=2)		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_{st}$
②③	❷❸	$\frac{2+3}{2}=2,5$	$\frac{2+3}{2}=2,5$	$\frac{(3\times 2,5)+(3\times 2,5)}{6}=2,5$
②③	❶❸	2,5	2	2,25
②③	❶❷	2,5	1,5	2,00
①③	❷❸	2	2,5	2,25
①③	❶❸	2	2	2,00
①③	❶❷	2	1,5	1,75
①②	❷❸	1,5	2,5	2,00
①②	❶❸	1,5	2	1,75
①②	❶❷	1,5	1,5	1,50
TOTAL:		18	18	18
MEDIA:		$E\left[\bar{y}_1\right]=\frac{18}{9}=2$	$E\left[\bar{y}_2\right]=\frac{18}{9}=2$	$E\left[\bar{y}_{st}\right]=\frac{18}{9}=2$
VARIANZA:		$V\left[\bar{y}_1\right]=0,1667$	$V\left[\bar{y}_2\right]=0,1667$	$V\left[\bar{y}_{st}\right]=0,0833$

En este caso el muestreo aleatorio simple conduce a un estimador con menor varianza que el muestreo aleatorio estratificado $\left(V\left[\bar{y}\right]=0,0667\right)<\left(V\left[\bar{y}_{st}\right]=0,0833\right)$.

EJERCICIOS RESUELTOS

1. Un analista de la opinión pública tiene un presupuesto de 20000 euros para realizar una encuesta sobre el número medio de coches por hogar. Se sabe que de los 10000 hogares de la ciudad, 9000 tienen teléfono. Las entrevistas por teléfono cuestan 10 euros por hogar llamado y las entrevistas personales cuestan 30 euros por hogar visitado. Suponga que las varianzas en los estratos con y sin teléfono son iguales. Con el objetivo de minimizar el límite de error de estimación ¿Cuántos hogares deben ser entrevistados en cada estrato si los hogares que cuentan con servicio telefónico son entrevistados por teléfono y los hogares sin teléfono son entrevistados personalmente?

   SOLUCIÓN:

$$n=\frac{C\sum_{i=1}^{L}\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L}N_i\sigma_i\sqrt{c_i}}=\frac{C\sum_{i=1}^{L}\frac{N_i\sigma}{\sqrt{c_i}}}{\sum_{i=1}^{L}N_i\sigma\sqrt{c_i}}=\frac{C\sigma\sum_{i=1}^{L}\frac{N_i}{\sqrt{c_i}}}{\sigma\sum_{i=1}^{L}N_i\sqrt{c_i}}=\frac{C\sum_{i=1}^{L}\frac{N_i}{\sqrt{c_i}}}{\sum_{i=1}^{L}N_i\sqrt{c_i}}=\frac{20000\times 3028,624}{33937,726}=1784,81$$

N_i	$\sqrt{c_i}$	$\frac{N_i}{\sqrt{c_i}}$	$N_i\sqrt{c_i}$	ω_i
9000	$\sqrt{10}$	2846,05	28460,5	2846,05/3028,624=0,9397
1000	$\sqrt{30}$	182,574	5477,226	182,574/3028,624=0,0603
10000		3028,624	33937,726	1,0000

$$n_1 = n\omega_1 = 1784,81 \times 0,9397 = 1677,2 \approx 1677$$
$$n_2 = n\omega_2 = 1784,81 \times 0,0603 = 107,59 \approx 107$$
$$n = n_1 + n_2 = 1784$$

O bien

$$c_1 n_1 + c_2 n_2 = 20000$$
$$c_1 \omega_1 n + c_2 \omega_2 n = 20000$$
$$9,397n + 1,809n = 11,206n = 20000$$
$$\frac{20000}{11,206} = 1784,8 = n$$

Y a partir de n se obtienen n_1 y n_2 como antes.

2. Se desea conocer el número de fines de semana que las familias de una gran ciudad salen fuera de ella. Se sabe que el 42'5% de las familias tienen de 0 a 2 hijos, el 30% tienen de 3 a 5 hijos y el 27'5% tienen más de 5 hijos. Se realizó un muestreo según el número de hijos y se preguntó a las familias sobre los fines de semana que pasan fuera, obteniéndose los siguientes datos:

Número de hijos	n_i	$\sum_{i=1}^{n} y_i$	S_i^2
0-2	25	239	60'76
3-5	19	174	63'01
Más de 5	16	78	78'24

Estimar el número medio de fines de semana que las familias pasan fuera de la ciudad y dar el límite de error de estimación. Omitir el corrector por población finita.

SOLUCIÓN:

$$\bar{y}_1 = \frac{239}{25} = 9,56 \quad \bar{y}_2 = \frac{174}{19} = 9,16 \quad \bar{y}_3 = \frac{78}{16} = 4,87$$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i = (0,425 \times 9,56) + (0,30 \times 9,16) + (0,275 \times 4,87) = 8,15$$

$$Si \quad \frac{N_i - n_i}{N_i} = 1 \quad \Rightarrow \quad \hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i} = \sum_{i=1}^{L} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i} = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i} =$$

$$= \left(0,425^2 \frac{60,77}{25}\right) + \left(0,30^2 \frac{63,01}{19}\right) + \left(0,275^2 \frac{78,24}{16}\right) = 1,107$$

$$2\sqrt{1,107} = 2,1$$

3. Una compañía de autobuses está planeando una nueva ruta para dar servicio a cuatro barrios. Se tomaron muestras aleatorias de hogares en cada barrio y se solicitó a los miembros de la muestra que valorasen en una escala de 1 (totalmente opuesto) a 5 (totalmente a favor) su opinión sobre el servicio propuesto. Los resultados se resumen en la tabla adjunta:

	Barrio			
	1	2	3	4
N_i	240	190	350	220
n_i	25	25	25	25
$\bar{y}_i$	3,5	3,6	3,9	3,8
S_i	0,8	0,9	1,2	0,7

a) Halle un intervalo de confianza para la opinión media de los hogares que dispondrán del nuevo servicio.

b) Si se asigna la muestra de 100 hogares de la mejor forma, determine cuántos pertenecerían al barrio 3. (Suponga iguales los costes de observación)

SOLUCIÓN:

a)

$$N = \sum_{i=1}^{L} N_i = 1000 \qquad \bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = 3,725 \qquad \hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = 0,00973$$

$$B = 2\sqrt{\hat{V}(\bar{y}_{st})} = 0,1973 \qquad \boxed{\mu \in (3,5277\ ,\ 3,9223)}$$

b)

$$n_3 = n\omega_3 = 100\frac{N_3\sigma_3}{\sum_{i=1}^{4} N_i\sigma_i} = 100\frac{350 \times 1,2}{(240 \times 0,8) + (190 \times 0,9) + (350 \times 1,2) + (220 \times 0,7)} =$$

$$= 100 \times 0,4482 = 44,82 \approx \boxed{45}$$

4. Una empresa especializada en seguros está pensando en ofrecer sus servicios a las empresas de los polígonos industriales de una ciudad. Para ajustar sus tarifas desea estimar el gasto en pequeñas reparaciones de mantenimiento (objeto del seguro) de dichas empresas. Se clasifican las empresas en función de su tamaño. El número de empresas de cada tipo, el coste de obtención de esta información en cada empresa así como los valores

mínimos, medios y máximos de un estudio similar hecho hace dos años se expresan en la siguiente tabla (los costes y gastos están expresados en euros)

Tipo de empresa	Número de empresas	Costes de observación	Gastos de reparación		
			Mínimo	Media	Máximo
A	100	16	400	500	600
B	500	9	240	300	360
C	700	4	70	100	130

Si la empresa de seguros dispone de hasta 600 € para llevar a cabo la estimación, ¿cuántas empresas de cada tipo tiene que observar para conseguir que sea mínimo el error de estimación asociado?

SOLUCIÓN:

La asignación que minimiza la cota del error de estimación para un coste fijo es la asignación Óptima.

Usamos que $R \approx 4\sigma$ y por tanto estimamos que $\sigma \approx \frac{R}{4}$.

N_i	c_i	$\sqrt{c_i}$	R_i	σ_i	$\frac{N_i\sigma_i}{\sqrt{c_i}}$	ω_i
100	16	4	600-400	50	1250	0'1087
500	9	3	360-240	30	5000	0'4348
700	4	2	130-70	15	5250	0'4565
					11500	1

$600 = 16n_1 + 9n_2 + 4n_3 \quad (n_i = \omega_i n) \quad 600 = 1'7392n + 3'9132n + 1'826n = 7'4784n$

$$n = 600/7'4784 = 80'231$$

$n_1 = \omega_1 n = 8'72 \approx 8$	$n_2 = \omega_2 n = 34'88 \approx 34$	$n_3 = \omega_3 n = 36'63 \approx 36$

$$C = (16\times 8) + (9\times 34) + (4\times 36) = 578 < 600$$

5. En una población compuesta por aproximadamente igual número de hombres que de mujeres se desea estimar el gasto medio mensual por habitante en ocio. Se lleva a cabo la encuesta por teléfono mediante una muestra aleatoria simple de 500 números de teléfono del citado municipio. Después de obtenidos los datos se observa que sólo 100 de los encuestados fueron hombres y el resto mujeres. Por ello se decide llevar a cabo una estratificación después de seleccionar la muestra obteniéndose los siguientes datos

	HOMBRES	MUJERES
N_i	2500	2700
n_i	100	400
$\bar{y}_i$	120	250
S_i^2	9000	16000

Estime la media poblacional de gasto mensual en ocio y su cota de error, mediante muestreo aleatorio estratificado después de seleccionar la muestra.

SOLUCIÓN:

N_i	n_i	$\bar{y}_i$	S_i^2	$N_i \bar{y}_i$	$\frac{N_i - n_i}{N_i}$	$N_i^2 \frac{S_i^2}{n_i} \frac{N_i - n_i}{N_i}$
2500	100	120	9000	300000	0,96	540000000
2700	400	250	16000	675000	0,85185	248399460
5200	500			975000		788399460

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \frac{975000}{5200} = 187,5$$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = \frac{788399460}{5200^2} = 29,16 \qquad 2\sqrt{29,16} = 10,8$$

6. En una población compuesta por aproximadamente igual número de hombres que de mujeres se desea estimar la proporción de individuos que ven un determinado programa de televisión. Se lleva a cabo la encuesta por teléfono mediante una muestra aleatoria simple de 300 números de teléfono. Después de obtenidos los datos se observa que sólo 50 de los encuestados fueron hombres y el resto mujeres. Por ello se decide llevar a cabo una estratificación después de seleccionar la muestra obteniéndose los siguientes datos

	HOMBRES	MUJERES
Encuestados	50	250
Ven el programa	12	130

Estime la proporción de la población que ven el programa de televisión y su cota de error, mediante muestreo aleatorio estratificado después de seleccionar la muestra.

SOLUCIÓN:

$$\hat{p}_1 = \frac{12}{50} = 0,24 \quad \hat{p}_2 = \frac{130}{250} = 0,52 \quad \hat{q}_i = 1 - \hat{p}_i$$

$$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i = \sum_{i=1}^{L} \frac{N_i}{N}\hat{p}_i = (0,50 \times 0,24) + (0,50 \times 0,52) = 0,38 \Rightarrow \hat{p}_{st} = 38\%$$

$$Si \quad \frac{N_i - n_i}{N_i} = 1 \quad \Rightarrow$$

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1} = \sum_{i=1}^{L} \frac{N_i^2}{N^2}\frac{\hat{p}_i\hat{q}_i}{n_i - 1} = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1} =$$

$$= \left(0,50^2 \frac{0,24 \times 0,76}{49}\right) + \left(0,50^2 \frac{0,52 \times 0,48}{249}\right) = 0,0011812146$$

$$2\sqrt{\hat{V}(\hat{p}_{st})} = 0,0687 \quad \Rightarrow \quad 6,87\%$$

7. Se quiere hacer un estudio sobre gasto en ropa en una comarca donde el 41% de los habitantes son mujeres.

   a) Se decide tomar una muestra aleatoria estratificada de 300 observaciones con asignación proporcional. ¿Cuántos hombres y mujeres deben entrevistarse?

   Se toma la anterior muestra, obteniéndose los siguientes valores:

	media muestral (en euros)	cuasivarianza muestral
HOMBRES	120	4000
MUJERES	170	9000

   b) Estime el gasto medio en ropa para toda la comarca y el límite del error de estimación asociado.

   c) Quiere repetirse el estudio sólo en la población de mujeres para estimar el gasto medio en ropa de ellas con un error inferior a 10 euros. ¿A cuántas mujeres habría que preguntarle?

   SOLUCIÓN:

   a) $\frac{N_1}{N} = 0,59 \quad \frac{N_2}{N} = 0,41 \quad n = 300 \quad \Rightarrow \quad n_1 = 0,59n = 177 \quad n_2 = 0,41n = 123$

   b) $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i = (0,59 \times 120) + (0,41 \times 170) = 140,5$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = \sum_{i=1}^{L} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i} = \left(0,59^2 \times \frac{4000}{177}\right) + \left(0,41^2 \times \frac{9000}{123}\right) = 20,17$$

$$2\sqrt{\hat{V}(\bar{y}_{st})} = 8,98146$$

   c) $B = 10 \quad D = \frac{B^2}{4} = 25 \quad \Rightarrow \quad n = \frac{\sigma^2}{D} = \frac{9000}{25} = 360$

8. Una corporación desea estimar el número total de horas perdidas debido a accidentes de sus empleados, en un determinado mes. Ya que los obreros, técnicos y administrativos tienen diferentes tasas de accidentes, la corporación decide usar muestreo estratificado, formando con cada grupo un estrato. Datos de años previos sugieren las cuasivarianzas mostradas en la siguiente tabla para el número de horas perdidas por empleado en los tres grupos, y de datos actuales se obtienen los tamaños de los estratos. No habiendo diferencia entre los costes de observación de cada grupo, determine la mejor asignación para una muestra de 40 empleados.

	Obreros	**Técnicos**	**Administrativos**
S_i^2	36	25	9
N_i	132	92	27

SOLUCIÓN:

N_i	$\sigma_i \approx S_i$	$N_i\sigma_i$	ω_i
132	6	792	$792/1333 = 0,5941$
92	5	460	$460/1333 = 0,3451$
27	3	81	$81/1333 = 0,0608$
		1333	1

Donde se ha aplicado la asignación de Neyman al ser los costes de observación iguales:

$$\omega_j = \frac{N_j\sigma_j}{\sum_{i=1}^{L} N_i\sigma_i}$$

$$n_1 = 40 \times 0,5941 = 23,8 \approx 24$$
$$n_2 = 40 \times 0,3451 = 13,8 \approx 14$$
$$n_3 = 40 \times 0,0608 = 2,4 \approx 2 \qquad n = 40$$

9. Se dispone de la siguiente información sobre tamaños poblacionales de los estratos, costes de observación y estimaciones de las proporciones

	Tamaño del estrato	Coste de observación	Proporciones en %
ESTRATO 1	5000	9	90
ESTRATO 2	2000	25	55
ESTRATO 3	3000	16	70

Determine la mejor asignación para una muestra de 200 observaciones.

SOLUCIÓN:

N_i	$\sqrt{c_i}$	$\hat{p}_i$	$\hat{q}_i$	$\sqrt{\hat{p}_i\hat{q}_i}$	$N_i\sqrt{\hat{p}_i\hat{q}_i}/\sqrt{c_i}$	ω_i
5000	3	0,90	0,10	0,3	500	$500/1042,695 = 0,4795$
2000	5	0,55	0,45	0,4975	199	$199/1042,695 = 0,1909$
3000	4	0,70	0,30	0,45826	343,695	$343,695/1042,695 = 0,3296$
					1042,695	1

Donde se ha aplicado la asignación Óptima:

$$\omega_j = \frac{N_j\sqrt{\frac{p_jq_j}{c_j}}}{\sum_{i=1}^{L} N_i\sqrt{\frac{p_iq_i}{c_i}}}$$

$$n_1 = 200 \times 0,4795 = 95,9 \approx 96$$
$$n_2 = 200 \times 0,1909 = 38,2 \approx 38$$
$$n_3 = 200 \times 0,3296 = 65,9 \approx 66$$
$$n = 200$$

10. La producción de piezas de una factoría se realiza en dos máquinas. El 40% de las piezas las produce la máquina A y el 60% restante la máquina B. Se les pasó control de calidad a 200 piezas; 67 producidas por la máquina A y dos de ellas resultaron defectuosas; las 133 restantes procedían de la máquina B, siendo 6 de ellas defectuosas. Estimar la proporción de piezas defectuosas de la factoría y dar el límite de error de estimación. Omita el coeficiente corrector por población finita.

SOLUCIÓN:

Estrato	N_i	n_i	$\hat{p}_i$	$\frac{\hat{p}_i\hat{q}_i}{n_i-1}$
A	$0,40\times N$	67	2/67=0,030	0,000441
B	$0,60\times N$	133	6/133=0,045	0,000326
	N	200		

$$\hat{p}=\frac{1}{N}\left(\left(0,40\times N\times 0,030\right)+\left(0,60\times N\times 0,045\right)\right)=\left(\left(0,40\times 0,030\right)+\left(0,60\times 0,045\right)\right)=0,039\quad(3,9\%)$$

$$\hat{V}\left(\hat{p}\right)=\frac{1}{N^2}\left(\left(0,40^2\times N^2\times 0,000441\right)+\left(0,60^2\times N^2\times 0,000326\right)\right)=$$
$$=\left(\left(0,40^2\times 0,000441\right)+\left(0,60^2\times 0,000326\right)\right)=0,000188$$
$$B=2\sqrt{0,000188}=0,0274\quad(2,74\%)$$

11. Para la comercialización de un producto se le clasifica, atendiendo al calibre, en tres categorías: pequeña, mediana y grande. Un establecimiento dispone de 300 piezas pequeñas, 500 medianas y 200 piezas grandes. Para estimar el peso total de producto almacenado se decide tomar una muestra aleatoria que contenga piezas de todas las categorías, resultando

Categoría	Nº de piezas	Peso en gramos
Pequeña	5	12, 14, 12, 15, 12
Mediana	6	16, 22, 24, 20, 20, 18
Grande	4	30, 33, 31, 34

Considerando los anteriores datos como una muestra previa, obtenga el número de unidades que cada categoría debe aportar a la muestra para que el error en la estimación del peso total no supere el medio kilo.

SOLUCIÓN:

Peso en gramos	(*con las funciones del modo SD de la calculadora*)
12, 14, 12, 15, 12	$S_1=1,4142\qquad S_1^2=2$
16, 22, 24, 20, 20, 18	$S_2=2,8284\qquad S_2^2=8$
30, 33, 31, 34	$S_3=1,8257\qquad S_3^2=3,3333$

N_i	σ_i	σ_i^2	$N_i\sigma_i$	$N_i\sigma_i^2$	$\omega_j=\frac{N_j\sigma_j}{\sum_{i=1}^{L}N_i\sigma_i}$	$n_i=71,66\omega_i$
300	1,4142	2	424,26	600	0,1925	$13,79\approx 14$
500	2,8284	8	1414,2	4000	0,6418	$45,99\approx 46$
200	1,8257	3,3333	365,14	666,66	0,1657	$11,87\approx 12$
$N=1000$			2203,6	5266,66	1	$n=72$

$$D = \frac{B^2}{4N^2} = \frac{250000}{4000000} = 0,0625 \qquad n = \frac{\left(\sum_{i=1}^{L} N_i \sigma_i\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} = 71,66$$

12. Una inspectora de control de calidad debe estimar la proporción de circuitos integrados de ordenador defectuosos que provienen de dos diferentes operaciones de ensamble. Ella sabe que de entre los circuitos integrados que van a ser inspeccionados, 60% procede de la operación de ensamble A y 40% de la operación de ensamble B. En una muestra aleatoria de 100 circuitos integrados resulta que 20 provienen de la operación A y 80 de la operación B. De entre los circuitos integrados muestreados de la operación A, 2 son defectuosos. De entre las piezas muestreadas de la operación B, 16 son defectuosas.
    a) Considerando únicamente la muestra aleatoria simple de 100 circuitos integrados, estime la proporción de los defectuosos en el lote, y establezca un límite para el error de estimación.
    b) Estratifique la muestra, después de la selección, en circuitos integrados provenientes de la operación A y B, estime la proporción de los defectuosos en la población, y fije un límite para el error de estimación.
    c) ¿Qué respuesta encuentra más aceptable? ¿Por qué?

    SOLUCIÓN:

    a. $\hat{p} = \frac{18}{100} = 0,18 \quad (18\%) \quad \hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1} = 0,001491 \quad 2\sqrt{\hat{V}(\hat{p})} = 0,0772 \quad (7,72\%)$

    b. $\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i = \sum_{i=1}^{L} \frac{N_i}{N}\hat{p}_i = \left(0,60\frac{2}{20}\right) + \left(0,40\frac{16}{80}\right) = 0,14 \quad (14\%)$

    $$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i} = \sum_{i=1}^{L} \frac{N_i^2}{N^2}\frac{\hat{p}_i\hat{q}_i}{n_i - 1} = \sum_{i=1}^{L}\left(\frac{N_i}{N}\right)^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1} =$$

    $$= (0,60)^2 \frac{0,10 \times 0,90}{19} + (0,40)^2 \frac{0,20 \times 0,80}{79} = 0,00203$$

    $$2\sqrt{\hat{V}(\hat{p}_{st})} = 0,0901 \quad (9,01\%)$$

    c. Aunque en el conjunto de la población hay más elementos que proceden de A (60%) que de B (40%), la muestra global no representa adecuadamente este hecho, predominando los elementos de B (80) frente a los de A (20), esto ocasiona que en el apartado a. la estimación esté sesgada hacia el valor de B ($\hat{p}_2 = 0,20$) frente al de A

($\hat{p}_1 = 0{,}10$). En el apartado b. este hecho se corrige dando a $\hat{p}_1$ y $\hat{p}_2$ las ponderaciones 0,60 y 0,40 respectivamente para estimar *p*.

13. Una cadena de restaurantes tiene 100 establecimientos en Madrid, 70 en Barcelona y 30 en Sevilla. La dirección está considerando añadir un nuevo producto en el menú. Para contrastar la posible demanda de este producto, lo introdujo en el menú de muestras aleatorias de 10 restaurantes de Madrid, 5 de Barcelona y 5 de Sevilla. Usando los índice 1, 2 y 3 para designar Madrid, Barcelona y Sevilla, respectivamente, las medias y las desviaciones típicas muestrales del número de pedidos de este producto recibidos por restaurante en las tres ciudades durante una semana fueron:

$$\bar{y}_1 = 21{,}2 \quad S_1 = 12$$
$$\bar{y}_2 = 13{,}3 \quad S_2 = 11$$
$$\bar{y}_3 = 26{,}1 \quad S_3 = 9$$

a. Estimar el número medio de pedidos semanales por restaurante para los restaurantes de la cadena. Dar un límite del error de estimación.
b. Determinar el tamaño muestral y la asignación para repetir el estudio anterior cometiendo un error inferior a 3 pedidos.

SOLUCIÓN:

a. $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \frac{3834}{200} = 19{,}17 \ \textit{pedidos / semana}$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = 6{,}2965 \qquad 2\sqrt{\hat{V}(\bar{y}_{st})} = 5{,}02 \ \textit{pedidos / semana}$$

b.

N_i	σ_i	σ_i^2	$N_i\sigma_i$	$N_i\sigma_i^2$	$\omega_j = \frac{N_j\sigma_j}{\sum_{i=1}^{L} N_i\sigma_i}$	$n_i = 43{,}52\omega_i$
100	12	144	1200	14400	0,5357	$23{,}31 \approx 24$
70	11	121	770	8470	0,3438	$14{,}96 \approx 15$
30	9	81	270	2430	0,1205	$5{,}24 \approx 6$
$N = 200$			2240	25300	1	$n = 45$

$$D = \frac{B^2}{4} = \frac{9}{4} = 2{,}25 \qquad n = \frac{\left(\sum_{i=1}^{L} N_i\,\sigma_i\right)^2}{N^2 D + \sum_{i=1}^{L} N_i\sigma_i^2} = 43{,}52$$

14. De una ciudad con 350 casas, se sabe que 164 de ellas tienen calefacción eléctrica. Al realizar una encuesta sobre el consumo de energía (en kilovatios-hora) se obtuvieron los siguientes resultados:

Tipo Calefacción	Nº casas	Media muestral	Cuasivarianza muestral
Eléctrica	24	972	202,396
No eléctrica	36	463	96,721

a. Obtenga una estimación del número medio de kilovatios-hora utilizado en la ciudad. Dé un límite para el error de estimación.

b. Obtenga una estimación del número medio de kilovatios-hora utilizado por las casas que no tienen calefacción eléctrica. Dé un límite para el error de estimación.

SOLUCIÓN:

a.

N_i	n_i	$\bar{y}_i$	S_i^2	$N_i \bar{y}_i$	$\frac{N_i - n_i}{N_i}$	$N_i^2 \frac{S_i^2}{n_i} \frac{N_i - n_i}{N_i}$
164	24	972	202,396	159408	0,854	193699,13
186	36	463	96,721	86118	0,806	74925,32
350				245526		268624,45

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \frac{245526}{350} = 701,50$$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} = \frac{268624,45}{350^2} = 2,19$$

$$2\sqrt{2,19} = 2,96$$

b.

$$\bar{y}_2 = 463$$

$$\hat{V}(\bar{y}_2) = \frac{S_2^2}{n_2}\frac{N_2 - n_2}{N_2} = \frac{96,721}{36}\,\frac{186-36}{186} = 2,17$$

$$2\sqrt{2,17} = 2,94$$

3. Muestreo con información auxiliar.

3.1 Introducción.

Si entre dos variables existe una fuerte relación es posible utilizar la **información auxiliar** que tengamos de una variable, como puede ser la media o el total poblacional, para estimar la media o el total de la otra variable.

Notaremos por

$Y \rightarrow$ Variable bajo estudio

$X \rightarrow$ Variable que proporciona la información auxiliar

De las que tomaremos una muestra constituida por n pares de datos:

$$(x_1, y_1),...,(x_n, y_n)$$

A partir de los datos muestrales se puede estimar la **relación existente entre ambas variables**.

Pueden utilizarse distintos diseños de muestreo en la **estimación con información auxiliar**. Aquí suponemos que se emplea el **muestreo aleatorio simple.**

La estimación con información auxiliar es importante cuando se pretende estimar el total sin conocer el número de elementos de la población pero sí el valor total de la variable que proporciona la información auxiliar.

Por ejemplo, debido a que existe una fuerte relación entre renta y ahorro, se puede estimar el valor total de los ahorros de los individuos de una población si se conoce el valor total de las rentas de dichos individuos. Así, si se sabe que por término medio el 10% de la renta se dedica al ahorro y se conoce la renta total, el ahorro total se estima igual al 10% de la renta total. Observemos que la estimación del total de ahorro se ha llevado a cabo sin necesidad de conocer el número de individuos de la población, *N*.

Dependiendo de la **relación entre las variables** X e Y utilizaremos:

- Estimadores de razón $\left(y = bx \quad o\ con\ otra\ notación \quad y = rx\right)$
- Estimadores de regresión $\left(y = a + bx\right)$
- Estimadores de diferencia $\left(y = a + x \quad o\ con\ otra\ notación \quad y = d + x\right)$

Estos estimadores sólo se deben utilizar si entre las dos variables existe una fuerte relación lineal positiva, $r_{xy} > \frac{1}{2}$.

3.2 Estimación de razón

Dada una población de tamaño N en la que se consideran las variables X e Y, se define la **razón** poblacional como el cociente:

$$R = \frac{\tau_y}{\tau_x}$$

Es decir, la proporción del total poblacional de Y respecto del total poblacional de X.

Puesto que $\tau_y = N\mu_y$ y $\tau_x = N\mu_x$, obtenemos

$$R = \frac{\tau_Y}{\tau_X} = \frac{N\mu_Y}{N\mu_X} = \frac{\mu_Y}{\mu_X}$$

De esta definición se deduce que

$$\tau_y = R\tau_x \qquad \mu_y = R\mu_x$$

Por tanto, si se conocen los valores de la media poblacional y el total poblacional de la variable X, para estimar la media poblacional y el total poblacional de Y sólo hay que estimar el valor de R (que notaremos como $\hat{R} = r$) y sustituirlo en las anteriores igualdades:

$$\hat{\tau}_y = r\tau_x \qquad \hat{\mu}_y = r\mu_x$$

Puesto que la razón R es el cociente entre las medias poblacionales, $R = \frac{\mu_Y}{\mu_X}$, tomando una muestra aleatoria simple: $\left(y_1, x_1\right),...,\left(y_n, x_n\right)$, podemos estimar R tomando el cociente entre las medias muestrales:

- ESTIMADOR DE LA RAZÓN: $\hat{R} = r = \frac{\bar{y}}{\bar{x}} = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$
- VARIANZA ESTIMADA DE r: $\hat{V}(r) = \frac{1}{\mu_x^2}\frac{S_r^2}{n}\left(\frac{N-n}{N}\right)$, $\quad S_r{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - rx_i\right)^2$

En el caso de no conocer μ_x, esta se estima con $\bar{x}$ en $\hat{V}(r)$.

En poblaciones infinitas, $\frac{N-n}{N}=1$ y $\hat{V}(r)=\frac{1}{\mu_x^2}\frac{S_r^2}{n}$.

3.2.1 Estimación de la media y el total poblacionales

Hemos de suponer que entre X e Y existe una alta correlación lineal positiva y que el modelo lineal, donde X es la variable explicativa e Y la explicada, pasa por el origen, $\left(y=bx,\quad \textit{en este contexto se nota } b=r \textit{ dado su significado},\ y=rx\right)$

- ESTIMADOR DE LA MEDIA: $\hat{\mu}_y = r\mu_x$

Para estimar $\hat{\mu}_y$ necesitamos conocer el verdadero valor de μ_x. No vale la estimación $\mu_x \cong \bar{x}$

- VARIANZA ESTIMADA DE $\hat{\mu}_y$: $\hat{V}\left(\hat{\mu}_y\right)=\mu_x^2\hat{V}(r)=\frac{S_r^2}{n}\left(\frac{N-n}{N}\right)$

  En poblaciones infinitas, $\frac{N-n}{N}=1$.

- ESTIMADOR DEL TOTAL: $\hat{\tau}_y = r\tau_x$

- VARIANZA ESTIMADA DE $\hat{\tau}_y$: $\hat{V}(\hat{\tau}_y)=\tau_x^2\hat{V}(r)=\frac{\tau_x^2}{\mu_x^2}\frac{S_r^2}{n}\left(\frac{N-n}{N}\right)=N\left(N-n\right)\frac{S_r^2}{n}$

En poblaciones infinitas, $\frac{N-n}{N}=1$. Además, si no conocemos μ_x, esta se estima con $\bar{x}$ en $\hat{V}(\hat{\tau}_y) \Rightarrow \hat{V}(\hat{\tau}_y)\cong\frac{\tau_x^2}{\bar{x}^2}\frac{S_r^2}{n}$.

Comentarios sobre estos estimadores:

- Son estimadores **sesgados** (Véase última tabla del apéndice de esta lección)**.**
- Cuando N es desconocido y si estimamos que $n \le 5\%N$, es decir que $\frac{N-n}{N}\ge 0{,}95$, entonces $\frac{N-n}{N}\cong 1$. (Véase ejercicio resuelto 4)

- De la relación $\mu_x = \frac{\tau_x}{N}$, se sigue que conociendo dos de estos elementos se puede calcular el tercero. (Véase ejemplo 3.1: $\mu_x = \frac{\tau_x}{N} = \frac{3840}{750} = 5{,}12\, ha/socio$)
- De su expresión, $\hat{\tau}_y = r\tau_x$, se sigue que no es necesario conocer el tamaño de la población, *N*, para estimar el total.
- A la hora de estimar el total, aunque conozcamos el tamaño de la población, cuando existe una fuerte correlación entre las variables, se comporta mejor el muestreo con información auxiliar $\left(\hat{\tau}_y = r\tau_x\right)$ que el muestreo aleatorio simple $\left(\hat{\tau} = N\bar{y}\right)$. (Véase ejemplo 3.5)

Ejemplo 3.1

Mediante una tasación previa se desea estimar la producción media y la producción total de los 750 socios de una cooperativa agrícola. Se sabe que el total de superficie plantada es de 3840 hectáreas. Se realizó un sorteo entre los socios para elegir a 20 de ellos a los que se les preguntó por la superficie plantada y se les tasó su producción. Los resultados fueron:

Superficie	Producción	Superficie	Producción
3,7	12	3	8
4,3	14	7	20
4,1	11	5,4	16
5	15	4,4	14
5,5	16	5,5	18
3,8	12	5	15
8	24	5,9	18
5,1	15	5,6	17
5,7	18	5	15
6	20	7,2	22

Estime la producción media y total mediante los estimadores de razón y muestreo aleatorio simple. Calcule sus respectivos límites para el error de estimación y compárelos.

Solución

$Y =$ producción (toneladas, tm) $X =$ *S*uperficie plantada (hectáreas, ha)

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	3,7	12	13,69	144	44,4
	4,3	14	18,49	196	60,2
	4,1	11	16,81	121	45,1
	5	15	25	225	75
	5,5	16	30,25	256	88
	3,8	12	14,44	144	45,6
	8	24	64	576	192
	5,1	15	26,01	225	76,5
	5,7	18	32,49	324	102,6
	6	20	36	400	120
	3	8	9	64	24
	7	20	49	400	140
	5,4	16	29,16	256	86,4
	4,4	14	19,36	196	61,6
	5,5	18	30,25	324	99
	5	15	25	225	75
	5,9	18	34,81	324	106,2
	5,6	17	31,36	289	95,2
	5	15	25	225	75
	7,2	22	51,84	484	158,4
TOTALES	**105,2**	**320**	**581,96**	**5398**	**1770,2**

Del enunciado y de la tabla anterior obtenemos:

$n = 20 \qquad N = 750$ socios $\qquad \tau_x = 3840$ ha

$$\sum_{i=1}^{n} x_i = 105,2 \qquad \sum_{i=1}^{n} y_i = 320 \qquad \sum_{i=1}^{n} x_i^2 = 581,96 \qquad \sum_{i=1}^{n} y_i^2 = 5398 \qquad \sum_{i=1}^{n} x_i y_i = 1770,2$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{105,2}{20} = 5,26 \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{581,96}{20} - 5,26^2 = 1,4304$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{320}{20} = 16 \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{5398}{20} - 16^2 = 13,9$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} = \frac{1770,2}{20} - \left(5,26 \times 16\right) = 4,35$$

Si queremos calcular las cuasivarianzas a partir de las varianzas:

$$S_x^2 = \frac{n}{n-1} s_x^2 = \frac{20}{19} 1,4304 = 1,5057 \qquad S_y^2 = \frac{n}{n-1} s_y^2 = \frac{20}{19} 13,9 = 14,6316$$

y hallando las raíces cuadradas obtenemos las desviaciones típicas $\left(s_x, s_y\right)$ y las cuasidesviaciones típicas $\left(S_x, S_y\right)$.

Los anteriores *cálculos* que se han realizado a mano o con ayuda de una calculadora básica se simplifican notablemente si utilizamos una calculadora científica de uso común. Estas calculadoras nos proporcionan los valores de un grupo de funciones estadísticas de forma inmediata:

$\sum x^2$ $\sum x$ $\bar{x}$ $\sigma_n = s_x = desviación\ \ típica$ $\sigma_{n-1} = S_x = cuasidesviación\ \ típica$

La relación entre las variables es alta, $r_{xy} = \dfrac{s_{xy}}{s_x\, s_y} = \dfrac{4,35}{1,196 \times 3,728} = 0,9756$. Esto, junto con la información auxiliar que disponemos de la variable *X,* justifica el uso de estimadores de razón. Por otra parte, dado el contexto, es lógico que la relación pase por el origen (a 0 ha de superficie le corresponde una producción de 0 tm).

$$r = \frac{\sum_{i=1}^{20} y_i}{\sum_{i=1}^{20} x_i} = \frac{320}{105,2} = 3,042 \text{ tm/ha}$$

$$\hat{\tau}_y = r\tau_x = 3,042 \times 3840 = 11680,6 \text{ tm}$$

$$\mu_x = \frac{\tau_x}{N} = \frac{3840}{750} = 5,12\, ha/socio$$

$$\hat{\mu}_y = r\mu_x = 3,042 \times 5,12 = 15,57\ tm/socio$$

$$S_r^{\,2} = \frac{1}{n-1}\sum_{i=1}^{20}\left(y_i - rx_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{20} y_i^2 + r^2\sum_{i=1}^{20} x_i^2 - 2r\sum_{i=1}^{20} x_i y_i\right) = 0,706$$

$$\hat{V}\left(\hat{\mu}_y\right) = \frac{S_r^2}{n}\left(\frac{N-n}{N}\right) = 0,0344 \quad \Rightarrow \quad B_\mu = 2\sqrt{\hat{V}\left(\hat{\mu}_y\right)} = 0,37 \text{ tm/socio}$$

$$\hat{V}(\hat{\tau}_y) = \frac{\tau_x^2}{\mu_x^2}\frac{S_r^2}{n}\left(\frac{N-n}{N}\right) = N^2\frac{S_r^2}{n}\left(\frac{N-n}{N}\right) = 19326,75 \quad \Rightarrow \quad B_\tau = 2\sqrt{\hat{V}(\hat{\tau}_y)} = 278,04 \text{ tm}$$

o $B_\tau = 750 \times B_\mu = 750 \times 0,37 = 277,5\, tm$ *(los dos procedimientos no coinciden por simples errores de redondeo en el valor de* B_μ*).*

A continuación lo estimaremos utilizando muestro aleatorio simple:

$$\bar{y} = \frac{320}{20} = 16\, tm/socio \qquad \hat{V}\left(\bar{y}\right) = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = \frac{14,63}{20}\left(\frac{750-20}{750}\right) = 0,712$$

$$B_\mu = 2\sqrt{0,712} = 1,69\, tm/socio$$

$$\hat{\tau} = N\bar{y} = 750\frac{320}{20} = 12000\, tm$$

$$\hat{V}(\hat{\tau}) = N^2 \frac{S^2}{n}\left(\frac{N-n}{N}\right) = 750^2 \frac{14,63}{20}\left(\frac{750-20}{750}\right) = 400539,47$$

$$B_\tau = 2\sqrt{400539,47} = 1265,76\ tm \quad \text{o} \quad B_\tau = 750 \times B_\mu$$

Observemos que el límite del error de estimación, tanto para la media como para el total, es mucho mayor que el cometido utilizando estimadores de razón. ■

3.2.2 Determinación del tamaño muestral

Tamaño muestral mínimo para que la estimación de la razón, la media y el total no supere una cota de error *B*

$$n = \frac{N\sigma_r^2}{\sigma_r^2 + ND}$$

donde para estimar:

- la razón: $D = \frac{B^2 \mu_x^2}{4}$
- la media: $D = \frac{B^2}{4}$
- el total: $D = \frac{B^2}{4N^2}$

Comentarios:

- En el caso de poblaciones infinitas $n = \frac{\sigma_r^2}{D}$
- σ_r^2 se estima utilizando una muestra previa: $\hat{\sigma}_r^2 = S_r^2$.
- Si μ_x es desconocido, $\hat{\mu}_x^2 = \bar{x}^2$.

Ejemplo 3.2 (continuación del ejemplo 3.1)

Supongamos que queremos reducir el límite para el error de estimación de la media a 0,25 tm/socio y el del total no debe superar las 200 tm ¿a cuántos socios se les debe tasar su producción antes de realizar una nueva estimación?

Solución

MEDIA: $n = \frac{N\sigma_r^2}{\sigma_r^2 + N\frac{B^2}{4}} = \frac{750 \times 0,706}{0,706 + \left(750 \times \frac{0,25^2}{4}\right)} = 42,6 \cong 43$ socios

TOTAL: $n = \dfrac{N\sigma_r^2}{\sigma_r^2 + N\dfrac{B^2}{4N^2}} = \dfrac{N\sigma_r^2}{\sigma_r^2 + \dfrac{B^2}{4N}} = \dfrac{750 \times 0,706}{0,706 + \left(\dfrac{200^2}{4 \times 750}\right)} = 37,7 \cong 38$ socios

Necesitamos al menos 43 socios para cumplir con ambos niveles de error. ■

3.3 Estimación de regresión

El uso del estimador de razón es más efectivo cuando la relación entre las variables X e Y es lineal y pasa por el origen de coordenadas. En caso de relación lineal que no pase por el origen de coordenadas es preferible utilizar estimadores de regresión.

En el modelo lineal simple $Y = a + bX$, el método de mínimos cuadrados permite estimar a y b de la siguiente forma:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

donde

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\,\bar{y}$$

3.3.1 Estimación de la media y el total poblacionales

- ESTIMADOR DE LA MEDIA: $\hat{\mu}_{yL} = a + b\mu_x = \bar{y} - b\bar{x} + b\mu_x = \bar{y} + b(\mu_x - \bar{x})$

  Para estimar $\hat{\mu}_{yL}$ necesitamos conocer el verdadero valor de μ_x. No vale la estimación $\mu_x \cong \bar{x}$.

- VARIANZA ESTIMADA DE $\hat{\mu}_{yL}$: $\hat{V}(\hat{\mu}_{yL}) = \dfrac{S_L^2}{n}\left(\dfrac{N-n}{N}\right)$

  siendo S_L^2 la varianza residual en el modelo lineal simple:

  $$S_L^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \left(\bar{y} + b(x_i - \bar{x})\right)\right)^2 = \frac{n}{n-2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = \frac{n}{n-2}s_y^2\left(1 - r_{xy}^2\right)$$

  En poblaciones infinitas, $\dfrac{N-n}{N} = 1$.

- ESTIMADOR DEL TOTAL: $\hat{\tau}_{yL} = N\hat{\mu}_{yL}$

En este caso para estimar el total es necesario conocer el tamaño de la población N. No se puede estimar como $\hat{\tau}_{yL} = a + b\tau_x$ *ya que la recta de regresión no pasa por el punto* (τ_x, τ_y).

- VARIANZA ESTIMADA DE $\hat{\tau}_{yL}$: $\hat{V}(\hat{\tau}_{yL}) = N^2\hat{V}(\hat{\mu}_{yL})$

Ejemplo 3.3

Para un grupo de 1000 pequeños establecimientos se desea realizar un estudio sobre las ventas diarias. Se tiene información de que, por término medio, el gasto en publicidad es de 5 euros. Se elige al azar una muestra de 18 establecimientos y se toman datos de su gasto en publicidad y ventas diarios. Los resultados son:

Gastos	Ventas
3,7	120
4,3	140
4,1	135
5	150
5,5	160
3,8	120
8	160
5,1	150
5,7	125
6	130
0	80
7	150
5,4	150
4,4	120
5,5	140
5	150
5,9	150
6,6	170

Estime el total de ventas diarias y la media utilizando estimadores de regresión. Obtenga el límite para el error de estimación.

Solución

$Y =$ ventas diaria (euros) $\quad X =$ gastos diarios en publicidad (euros)

$n = 18$ establecimientos $\quad N = 1000$ establecimientos $\quad \mu_x = 5€$

Tal y como se explicó en la resolución del ejemplo 3.1 obtenemos:

$\bar{x} = 5,0556€ \quad \bar{y} = 138,889€$

$s_x = 1,6375 \Rightarrow s_x^2 = 2,6814$

$s_y = 20,314 \quad \Rightarrow \quad s_y^2 = 412,654$

$s_{xy} = 27,7284$

La relación entre las variables es fuerte: $r_{xy} = 0,8336$.

$$b = \frac{s_{xy}}{s_x^2} = \frac{27,7284}{2,6814} = 10,341 \qquad \hat{\mu}_{yL} = \bar{y} + b\left(\mu_x - \bar{x}\right) = 138,889 + 10,341\left(5 - 5,0556\right) = 138,314€$$

$$\hat{\tau}_{yL} = N\hat{\mu}_{yL} = 138314€$$

$$S_L^2 = \frac{n}{n-2} s_y^2 \left(1 - r_{xy}^2\right) = 141,6 \qquad \hat{V}\left(\hat{\mu}_{yL}\right) = \frac{S_L^2}{n}\left(\frac{N-n}{N}\right) = 7,73$$

$$B_\mu = 2\sqrt{\hat{V}\left(\hat{\mu}_{yL}\right)} = 5,56 \qquad B_\tau = N \times B_\mu = 1000 \times 5,56 = 5560€ \qquad \blacksquare$$

3.3.2 Determinación del tamaño muestral

Tamaño muestral mínimo necesario para que al estimar la media y el total poblacionales la cota de error no supere el valor *B*

$$n = \frac{N\sigma_L^2}{\sigma_L^2 + ND}$$

donde para estimar:

- la media: $D = \frac{B^2}{4}$
- el total: $D = \frac{B^2}{4N^2}$

En el caso de poblaciones infinitas $n = \frac{\sigma_L^2}{D}$.

σ_L^2 se estima utilizando una muestra previa: $\hat{\sigma}_L^2 = S_L^2$

Ejemplo 3.4 (continuación del ejemplo 3.3)

Se quiere repetir el estudio anterior de forma que el error para la estimación del total no supere los 1000 euros ¿cuál debe ser el tamaño muestral?

Solución

$$n = \frac{N\sigma_L^2}{\sigma_L^2 + N\frac{B^2}{4N^2}} = \frac{1000 \times 141,6}{141,6 + \left(1000\frac{1000^2}{4 \times 1000^2}\right)} = 361,6 \cong 362 \text{ establecimientos.}$$

$\blacksquare$

3.4 Estimación de diferencia

El uso del estimador de diferencia tiene un buen comportamiento (cota de error más baja) cuando la relación entre las variables es lineal y la pendiente del modelo es uno. $\left(y = a + x \quad ó \quad y = \bar{y} + (x - \bar{x}) \quad a = \bar{y} - \bar{x} = \bar{d}\right)$

Comúnmente se emplea en procedimientos de auditoría.

3.4.1 Estimación de la media y el total poblacionales

- ESTIMADOR DE LA MEDIA: $\hat{\mu}_{yD} = \bar{y} + (\mu_x - \bar{x}) = \mu_x + \bar{d}$

Donde $\bar{d} = \bar{y} - \bar{x} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) = \frac{1}{n}\sum_{i=1}^{n} d_i$. *Para estimar* $\hat{\mu}_{yD}$ *necesitamos conocer el verdadero valor de* μ_x *. No vale la estimación* $\mu_x \cong \bar{x}$

- VARIANZA ESTIMADA DE $\hat{\mu}_{yD}$: $\hat{V}(\hat{\mu}_{yD}) = \frac{S_D^2}{n}\left(\frac{N-n}{N}\right)$

$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \left(x_i + \bar{d}\right)\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2$, *donde* $d_i = y_i - x_i$, *por tanto* S_D^2 *es la cuasivarianza de los* d_i .

- ESTIMADOR DEL TOTAL: $\hat{\tau}_{yD} = N\hat{\mu}_{yD}$

En este caso para estimar el total es necesario conocer el tamaño de la población N. No se puede estimar como $\hat{\tau}_{yL} = \bar{y} + (\tau_x - \bar{x}) = \tau_x + \bar{d}$ *por análogas razones a las expuestas en el estimador de regresión.*

- VARIANZA ESTIMADA DE $\hat{\tau}_{YD}$: $\hat{V}(\hat{\tau}_{yD}) = N^2\hat{V}(\hat{\mu}_{yD})$

Ejemplo 3.5

Para un grupo de 200 establecimientos se desea realizar un estudio sobre el gasto diario. Se tiene información de que los ingresos medios diarios son de 500 euros. Se elige al azar una muestra de 10 establecimientos y se toman datos de ingresos y gastos, obteniéndose:

X=Ingresos	Y=Gastos
470	405
650	585
710	650
300	240
475	410
505	435
610	550
380	320
540	480
520	460

Estime el gasto medio y el gasto total diario para los 200 establecimientos utilizando muestreo aleatorio simple, estimadores de razón, regresión y diferencia. Obtenga el límite para el error de estimación en cada caso.

Solución

$Y =$ gasto diario (euros) $\quad X =$ ingresos diarios (euros)

$n = 10$ establecimientos $\quad N = 200$ establecimientos $\quad \mu_x = 500€$

Tal y como se explicó en la resolución del ejemplo 3.1 obtenemos:

$\overline{x} = 516€ \quad \overline{y} = 453,5€$

$s_x = 115,797 \quad \Rightarrow \quad s_x^2 = 13409$

$s_y = 115,738 \quad \Rightarrow \quad s_y^2 = 13395,3$

$S_y^2 = 14883,7 \quad s_{xy} = 13396,5$

La relación entre las variables es muy fuerte: $r_{xy} = 0,99958 \quad r_{xy}^2 = 0,99916$.

MUESTREO ALEATORIO SIMPLE

$\hat{\mu} = \overline{y} = 453,5€ \quad \hat{\tau} = N\overline{y} = 90700€$

$\hat{\mathrm{V}}(\hat{\mu}) = \dfrac{\mathrm{S}_y^2}{n}\left(\dfrac{N-n}{N}\right) = 1413,94 \quad \mathrm{B}_\mu = 2\sqrt{\hat{\mathrm{V}}(\hat{\mu})} = 75,20€ \quad \mathrm{B}_\tau = 200 \times \mathrm{B}_\mu = 15040,97€$

ESTIMADORES DE RAZÓN

$$r = \frac{\overline{y}}{\overline{x}} = 0,879 \quad \tau_x = 200\mu_x = 100000 \quad \hat{\tau}_y = r\tau_x = 87900€ \quad \hat{\mu}_y = r\mu_x = 439,5€$$

$$S_r{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - rx_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + r^2\sum_{i=1}^{n} x_i^2 - 2r\sum_{i=1}^{n} x_i y_i\right) = 227,717$$

$$\hat{V}\left(\hat{\mu}_y\right) = \frac{S_r^2}{n}\left(\frac{N-n}{N}\right) = 21,63 \quad \Rightarrow \quad B_\mu = 9,3€$$

$$B_\tau = N \times B_\mu = 1860€$$

ESTIMADORES DE REGRESIÓN

$\hat{b} = \frac{s_{xy}}{s_x^2} = \frac{13396,5}{13409} = 0,99907$ $\quad \hat{\mu}_{yL} = \overline{y} + \hat{b}\left(\mu_x - \overline{x}\right) = 437,515€$ $\quad \hat{\tau}_{yL} = N\hat{\mu}_{yL} = 87503€$

$S_L^2 = \frac{n}{n-2} s_y^2 \left(1 - r_{xy}^2\right) = 14,05$ $\quad \hat{V}\left(\hat{\mu}_{yL}\right) = \frac{S_L^2}{n}\left(\frac{N-n}{N}\right) = 1,33$

$B_\mu = 2,3104€$ $\quad B_\tau = NB_\mu = 462,09€$

ESTIMADORES DE DIFERENCIA

$\overline{d} = -62,5$ $\quad \hat{\mu}_{yD} = \mu_x + \overline{d} = 437,5€$ $\quad \hat{\tau}_{yD} = N\hat{\mu}_{yD} = 87500€$

(con la calculadora hallamos σ_{n-1} sobre las diferencias d_i y lo elevamos al cuadrado)

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2 = 12,5$$

$\hat{V}\left(\hat{\mu}_{yD}\right) = \frac{S_D^2}{n}\left(\frac{N-n}{N}\right) = 1,1875$ $\quad B_\mu = 2\sqrt{\hat{V}\left(\hat{\mu}_{yD}\right)} = 2,179$ $\quad B_\tau = NB_\mu = 435,8899$ ■

3.4.2 Determinación del tamaño muestral

Tamaño muestral mínimo necesario para que la estimación no supere un cota de error B al estimar la media y el total poblacionales

$$n = \frac{N\sigma_D^2}{\sigma_D^2 + ND}$$

donde para estimar:

- la media: $D = \frac{B^2}{4}$
- el total: $D = \frac{B^2}{4N^2}$

En el caso de poblaciones infinitas $n = \frac{\sigma_D^2}{D}$.

σ_D^2 se estima utilizando una muestra previa: $\hat{\sigma}_D^2 = S_D^2$

Ejemplo 3.6 (continuación ejemplo 3.5)

Se quiere repetir el estudio anterior utilizando un estimador de diferencia y cometiendo un error como máximo de 300 euros al estimar el total ¿cuál debe ser el tamaño muestral?

Solución

$$n = \frac{N\sigma_D^2}{\sigma_D^2 + N\dfrac{B^2}{4N^2}} = \frac{200 \times 12,5}{12,5 + \dfrac{300^2}{4 \times 200}} = 20 \text{ establecimientos}$$ ■

APÉNDICE: Estudio empírico de todas las posibles muestras en un muestreo con información auxiliar sobre una población finita.

Habitualmente el tamaño de la población es grande y no es posible comprobar empíricamente las propiedades del muestreo con información auxiliar que hemos estudiado en este tema. A continuación vamos a suponer una población con sólo 6 elementos, la misma que hemos utilizado en los temas anteriores:

①②③ ⓫⓬⓭	Población finita	Media poblacional, μ	Varianza poblacional, σ^2
	N=6	$\mu = \dfrac{1+2+3+11+12+13}{6} = 7$	25,6667 **(#)**

$$\frac{(1-7)^2 + (2-7)^2 + (3-7)^2 + (11-7)^2 + (12-7)^2 + (13-7)^2}{6} = 25,6667 \quad \textbf{(\#)}$$

Supongamos que disponemos de la información auxiliar X.

(aparece como subíndice de las bolas, ①$_{var.aux.\ X}$)

①$_4$ ②$_8$ ③$_{14}$ ⓫$_{56}$ ⓬$_{62}$ ⓭$_{66}$	Población finita	Total poblacional, τ_Y	Media poblacional, μ_Y	Varianza poblacional, σ_Y^2
	N=6	$\tau_Y = 1+2+3+11+12+13 = 42$	$\mu_Y = \dfrac{42}{6} = 7$	25,6667
		Total poblacional, τ_X	Media poblacional, μ_X	Varianza poblacional, σ_X^2
		$\tau_X = 4+8+14+56+62+66 = 210$	$\mu_X = \dfrac{210}{6} = 35$	710,3333

$$R = \frac{\tau_Y}{\tau_X} = \frac{\mu_Y}{\mu_X} = \frac{42}{210} = \frac{7}{35} = 0,2$$

Tomamos todas las muestras aleatorias simples de tamaño 4 para estimar la media poblacional de Y usando muestreo aleatorio simple y muestreo con información auxiliar.

$$\mu_X = \frac{210}{6} = 35$$

	MUESTRAS (*n*=4)	Media muestral, $\bar{x}$	Media muestral, $\bar{y}$	$r = \frac{\bar{y}}{\bar{x}}$	$\hat{\mu}_Y = r\mu_X$
1	③$_{14}$ ⓫$_{56}$ ⓬$_{62}$ ⓭$_{66}$	$\frac{14+56+62+66}{4} = 49,5$	$\frac{3+11+12+13}{4} = 9,75$	$\frac{9,75}{49,5} = 0,197$	6,8939
2	②$_{8}$ ⓫$_{56}$ ⓬$_{62}$ ⓭$_{66}$	48	9,5	0,1979	6,9271
3	②$_{8}$ ③$_{14}$ ⓬$_{62}$ ⓭$_{66}$	37,5	7,5	0,2000	7,0000
4	②$_{8}$ ③$_{14}$ ⓫$_{56}$ ⓭$_{66}$	36	7,25	0,2014	7,0486
5	②$_{8}$ ③$_{14}$ ⓫$_{56}$ ⓬$_{62}$	35	7	0,2000	7,0000
6	①$_{4}$ ⓫$_{56}$ ⓬$_{62}$ ⓭$_{66}$	47	9,25	0,1968	6,8883
7	①$_{4}$ ③$_{14}$ ⓬$_{62}$ ⓭$_{66}$	36,5	7,25	0,1986	6,9521
8	①$_{4}$ ③$_{14}$ ⓫$_{56}$ ⓭$_{66}$	35	7	0,2000	7,0000
9	①$_{4}$ ③$_{14}$ ⓫$_{56}$ ⓬$_{62}$	34	6,75	0,1985	6,9485
10	①$_{4}$ ②$_{8}$ ⓬$_{62}$ ⓭$_{66}$	35	7	0,2000	7,0000
11	①$_{4}$ ②$_{8}$ ⓫$_{56}$ ⓭$_{66}$	33,5	6,75	0,2015	7,0522
12	①$_{4}$ ②$_{8}$ ⓫$_{56}$ ⓬$_{62}$	32,5	6,5	0,2000	7,0000
13	①$_{4}$ ②$_{8}$ ③$_{14}$ ⓭$_{66}$	23	4,75	0,2065	7,2283
14	①$_{4}$ ②$_{8}$ ③$_{14}$ ⓬$_{62}$	22	4,5	0,2045	7,1591
15	①$_{4}$ ②$_{8}$ ③$_{14}$ ⓫$_{56}$	20,5	4,25	0,2073	7,2561
	TOTAL:	525	105	3,01012	105,3542

MUESTRAS (*n*=4)	Media muestral, $\bar{x}$	Media muestral, $\bar{y}$	$r = \frac{\bar{y}}{\bar{x}}$	$\hat{\mu}_Y = r\mu_X$
MEDIA:	$E\left[\bar{x}\right] = \frac{525}{15} = 35$	$E\left[\bar{y}\right] = \frac{105}{15} = 7$	$E\left[r\right] = 0,20067$	$E\left[\hat{\mu}_Y\right] = 7,0236$
VARIANZA:	$V\left[\bar{x}\right] = 71,0333$	$V\left[\bar{y}\right] = 2,5667$	$V\left[r\right] = 0,00001$	$V\left[\hat{\mu}_Y\right] = 0,0116$

Los estimadores *r* y $\hat{\mu}_Y$ son estimadores sesgados, a pesar de ello se utilizan debido a su pequeña varianza. $V\left[\hat{\mu}_Y\right] = 0,0116$ es más de 200 veces menor que $V\left[\bar{y}\right] = 2,5667$ en el muestreo aleatorio simple. Y también es menor que la varianza del estimador de la media en el muestreo aleatoria estratificado.

EJERCICIOS RESUELTOS

1. En una población de 500 hogares, para la que es conocido que el gasto total general durante un año es de 15000000 €, se quiere estimar el gasto total en alimentación durante un año, para lo que se obtiene una muestra aleatoria simple de 4 hogares que proporciona los siguientes valores anuales en €:

Gasto en alimentación	12500	15000	10000	17500
Gasto general	24000	31000	20000	36000

Estime con un estimador de razón el total de gasto en alimentación mediante un intervalo de confianza.

SOLUCIÓN (trabajaremos en cientos de euros)

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
240	125	57600	15625	30000
310	150	96100	22500	46500
200	100	40000	10000	20000
360	175	129600	30625	63000
1110	550	323300	78750	159500

$$N = 500 \quad n = 4 \quad r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{550}{1110} = 0,4955 \quad \hat{\tau}_y = r\tau_x = 0,4955 \times 150000 = 74325 \text{ cientos de } €$$

$$\hat{\tau}_y = 7432500 \text{ €}$$

$$S_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - rx_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + r^2\sum_{i=1}^{n} x_i^2 - 2r\sum_{i=1}^{n} x_i y_i\right) = \frac{62,2}{3} = 20,73$$

$$\widehat{V}(\hat{\tau}_y) = N(N-n)\frac{S_r^2}{n} = 1285260 \qquad 2\sqrt{\widehat{V}(\hat{\tau}_y)} = 2267,39$$

$\tau_y \in (72057,61\,;\,76592,39)$ *en cientos de* € **.** *Para expresarlo en € se multiplica por cien.*

2. Un trabajador social quiere estimar la ratio personas/habitación en un determinado barrio. El trabajador social selecciona una muestra aleatoria simple de 25 viviendas de las 275 del barrio. Sea x el número de personas en cada vivienda e y el número de habitaciones por vivienda. A partir de los datos siguientes:

$$\overline{x} = 9,1; \quad \overline{y} = 2,6; \quad \sum_{i=1}^{25} x_i^2 = 2240; \quad \sum_{i=1}^{25} y_i^2 = 169; \quad \sum_{i=1}^{25} x_i y_i = 522$$

Estime la razón personas/habitación en el barrio y establezca el límite para el error de estimación con una confianza del 95%.

SOLUCIÓN (los papeles de las variables x e y deben permutarse en las expresiones del formulario)

$$N = 275 \quad n = 25 \quad r = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} = \frac{\bar{x}}{\bar{y}} = 3,5 \; pers./hab. \quad \mu_y^2 \cong \bar{y}^2 = 2,6^2 = 6,76$$

$$S_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - r y_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + r^2 \sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} x_i y_i\right) = 27,34375$$

$$\hat{V}(r) = \frac{1}{\mu_y^2}\frac{(N-n)}{N}\frac{S_r^2}{n} = 0,1471 \qquad 2\sqrt{\hat{V}(r)} = 0,767$$

3. Se desea estimar el agua utilizada en la presente campaña por una comunidad de riego constituida por 250 parcelas. Se seleccionan al azar 10 parcelas cuyo tamaño y metros cúbicos utilizados en riego aparecen en la siguiente tabla

m^3	600	1800	750	900	1100	1400	950	700	1000	720
Hectáreas	50	150	60	70	100	120	80	60	90	60

Estime la media de m^3/hectárea que utiliza la comunidad de regantes y la cota del error de dicha estimación.

SOLUCIÓN:

Y = consumo de m^3 *de agua, X = tamaño de la parcela en hectáreas*

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
50	600	2500	360000	30000
150	1800	22500	3240000	270000
60	750	3600	562500	45000
70	900	4900	810000	63000
100	1100	10000	1210000	110000
120	1400	14400	1960000	168000
80	950	6400	902500	76000
60	700	3600	490000	42000
90	1000	8100	1000000	90000
60	720	3600	518400	43200
840	9920	79600	11053400	937200

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{9920}{840} = 11,81 \, m^3 / hectarea$$

$$S_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - r x_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + r^2\sum_{i=1}^{n} x_i^2 - 2r\sum_{i=1}^{n} x_i\, y_i\right) =$$

$$= \frac{1}{9}\left(11053400 + 11102297,56 - 22136664\right) = 2114,84$$

$$\widehat{\mu}_x = \bar{x} = \frac{840}{10} = 84 \qquad \widehat{V}(r) = \frac{1}{\mu_x^2}\frac{N-n}{N}\frac{S_r^2}{n} = \frac{1}{84^2}\frac{240}{250}\frac{2114,84}{10} = 0,02877$$

$$2\sqrt{\widehat{V}(r)} = 0,3392$$

4. Se desea estimar el consumo mensual de una ciudad. Se sabe que los ingresos en dicha ciudad, vía declaración de la renta, ascienden a 1502530 euros mensuales. Se realiza una encuesta entre 12 hogares elegidos al azar y los resultados de renta y consumo se recogen en esta tabla.

Renta	Consumo
1702,44	1204
1339,56	1000
981,06	800
2537,04	1800
1519,85	1200
3080,19	2600
1502,53	1080
1702,87	1240
1402,36	1000
1803,04	1400
2053,46	1484
3005,06	2000

Estime el consumo total mensual para todos los hogares de la ciudad mediante el estimador de razón. Obtenga el límite para el error de estimación.

SOLUCIÓN:

Denotemos por $Y = consumo\ mensual$ $X = ingresos\ mensuales$

De la información muestral obtenemos

$$n = 12 \qquad \sum_{i=1}^{12} y_i = 16808 \text{ euros} \qquad \sum_{i=1}^{12} x_i = 22629,46 \text{ euros}$$

y como información auxiliar sabemos que $\tau_x = 1502530$ euros.

Podemos comprobar que el coeficiente de correlación lineal es alto, $r_{xy} = \frac{s_{xy}}{s_x s_y} = 0,9677$.

Esto junto con la información auxiliar nos permite utilizar muestreo con información auxiliar, en concreto utilizaremos estimadores de razón.

$$r = \frac{\sum_{i=1}^{12} y_i}{\sum_{i=1}^{12} x_i} = 0,7427$$

$$\hat{\tau}_y = r\tau_x = 1116002,07€$$

Para calcular $\hat{V}(\hat{\tau}_Y) = \frac{\tau_x^2}{\mu_x^2}\frac{S_r{}^2}{n}\left(\frac{N-n}{N}\right)$ tenemos en cuenta que:

No conocemos N, pero en la ciudad hay muchos hogares, observando $\sum_{i=1}^{12} x_i < (5\%\ \tau_x)$

estimamos que $n < (5\% N) \quad \Rightarrow \quad \frac{N-n}{N} \cong 1$

$\widehat{\mu}_x = \overline{x} = 1885,79€$

$S_r{}^2 = \frac{1}{n-1}\sum_{i=1}^{12}(y_i - rx_i)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{12} y_i^2 + r^2\sum_{i=1}^{12} x_i^2 - 2r\sum_{i=1}^{12} x_i y_i\right) = 16479,7$

$\hat{V}(\hat{\tau}_Y) = 871825002,67 \qquad \Rightarrow \qquad B = 2\sqrt{\hat{V}(\hat{\tau}_Y)} = 59053,37€$

5. Las diferencias entre ingresos y gastos, en 5 de las 250 oficinas que tiene abiertas una agencia de seguros, en el presente mes, han sido (en euros)

570	721	650	650	569

Este mes el gasto medio para el conjunto de todas las oficinas ha sido 12764 euros, estime el total de ingresos y el límite para el error de estimación.

SOLUCIÓN:

N=250, n=5, $\mu_x = 12764$, *X=gastos, Y=ingresos*

(*con las funciones del modo SD de la calculadora*): $\overline{d} = 632 \qquad S_D^2 = 4095,5$

$\widehat{\mu}_{yD} = \mu_x + \overline{d} = 13396\ € \qquad \hat{\tau}_{yD} = N\widehat{\mu}_{yD} = 3349000\ €$

$\widehat{V}(\hat{\tau}_{yD}) = N^2\frac{N-n}{N}\frac{S_D^2}{n} = N(N-n)\frac{S_D^2}{n} = 50169875\ €^2 \qquad 2\sqrt{\widehat{V}(\hat{\tau}_{yD})} = 14166,14\ €$

6. Una cadena de electrodomésticos está interesada en estimar el total de ganancias por las ventas de televisores al final de un periodo de tres meses. Se tienen cifras del total de ganancias de todas las tiendas de la cadena para ese mismo periodo de tres meses correspondiente al año anterior, ese total es de 128200 €. Una muestra aleatoria simple de 5 tiendas es seleccionada de las 123 tiendas de la cadena resultando los datos de la siguiente tabla:

Oficinas	*Datos de 3 meses del año anterior*	*Datos de 3 meses del año actual*
1	550	610
2	720	780
3	1500	1600
4	1020	1030
5	620	600

Usando un estimador de razón, estime el total de ganancias con un intervalo de confianza.

SOLUCIÓN:

N=123, *n*=5, $\tau_x = 128200\,€$, *X=ganancias del año anterior, Y=ganancias del año actual* (*con las funciones del modo SD de la calculadora*):

$\bar{x} = 882 \qquad \sum_{i=1}^{5} x_i = 4410 \qquad \sum_{i=1}^{5} x_i^2 = 4495700$

$\bar{y} = 924 \qquad \sum_{i=1}^{5} y_i = 4620 \qquad \sum_{i=1}^{5} y_i^2 = 4961400$

$x_i y_i$
335500
561600
2400000
1050600
372000
$\sum_{i=1}^{5} x_i y_i = 4719700$

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} = 1,047619 \qquad \hat{\tau}_y = r\tau_x = 134304,76\,€$$

$$S_r^{\,2} = \frac{1}{n-1}\sum_{i=1}^{5}\left(y_i - rx_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{5} y_i^2 + r^2\sum_{i=1}^{5} x_i^2 - 2r\sum_{i=1}^{5} x_i y_i\right) = 1640,25$$

$$\hat{V}(\hat{\tau}_y) = N\left(N-n\right)\frac{S_r^2}{n} = 4761314,071 \qquad 2\sqrt{\hat{V}(\hat{\tau}_y)} = 4364,09$$

$$\tau_y \in \left(129940,67\,,138668,85\right)$$

7. Una agencia de publicidad está interesada en el efecto de una nueva campaña de promoción regional sobre las ventas totales de un producto en particular. Una muestra aleatoria simple de 5 tiendas es seleccionada de 452 tiendas regionales en las cuales se vende el producto. Los datos de las ventas trimestrales son obtenidos para el periodo actual de tres meses y para el periodo de tres meses previo a la nueva campaña.

Tienda	*Ventas antes de la campaña*	*Ventas actuales*
1	208	239
2	400	428
3	440	472
4	259	276
5	351	363

Usando los anteriores datos para estimar los parámetros necesarios, determine el tamaño de la muestra para estimar $\hat{\tau}_y$ con un límite para el error de estimación de 2000€, cuando se utiliza el estimador de razón.

SOLUCIÓN:

N=452, *n'*=5, *X=ventas antes de la campaña, Y=ventas actuales*

(*con las funciones del modo SD de la calculadora*):

$\bar{x} = 331,6 \qquad \sum_{i=1}^{5} x_i = 1658 \qquad \sum_{i=1}^{5} x_i^2 = 587146$

$\bar{y} = 355,6 \qquad \sum_{i=1}^{5} y_i = 1778 \qquad \sum_{i=1}^{5} y_i^2 = 671034$

$x_i y_i$
49712
171200
207680
71484
127413
$\sum_{i=1}^{5} x_i y_i = 627489$

$$r = \frac{\sum_{i=1}^{5} y_i}{\sum_{i=1}^{5} x_i} = \frac{\bar{y}}{\bar{x}} = 1,072376$$

$$S_r^{\,2} = \frac{1}{n'-1}\sum_{i=1}^{5}\left(y_i - r x_i\right)^2 = \frac{1}{n'-1}\left(\sum_{i=1}^{5} y_i^2 + r^2\sum_{i=1}^{5} x_i^2 - 2r\sum_{i=1}^{5} x_i y_i\right) = 109,4775$$

$$D = \frac{B^2}{4N^2} = 4,8947 \qquad \hat{\sigma}_r^2 = S_r^2 = 109,4775$$

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} = 21,3 \approx 22$$

4. Muestreo sistemático.

4.1 Selección de una muestra sistemática. Usos. Ventajas.
4.2 Estimación de la media, proporción y total poblacionales.
4.3 Comparación con el muestreo aleatorio simple: Poblaciones ordenadas, aleatorias y periódicas.
4.4 Determinación del tamaño muestral.

4.1 Selección de una muestra sistemática. Usos. Ventajas.

En el muestreo sistemático los elementos de la población se enumeran, o se ordenan. Una *muestra sistemática de "1 en k"* es la que se extrae de la siguiente forma:

1. Se selecciona aleatoriamente un elemento (*llamado punto de inicio o pivote*) de los primeros k elementos de la población.
2. Después se seleccionan cada k-ésimo elemento hasta conseguir una muestra de tamaño n.

k se toma como el número entero menor o igual que el cociente $\frac{N}{n}$: $\quad k \le \frac{N}{n}$.

Nos podemos encontrar con las siguientes situaciones:

1. $k = \frac{N}{n}$ entero. Entonces se obtienen exactamente n observaciones.
2. $\frac{N}{n}$ no es entero. Por ejemplo si $N = 103$ y $n = 5$, entonces $\frac{N}{n} = 20{,}6$ y tomamos $k = 20$. Según el punto inicial nos podemos encontrar con:
   - **a.** Si elegimos, por ejemplo, el 2º como punto inicial, obtendríamos:

     2º, 22º, 42º, 62º, 82º, 102º

     Al dividir la población en 5 intervalos de 20 elementos, sobran 3. Podríamos elegir también el 102º y la muestra sería de tamaño 6.
   - **b.** Si se elige, por ejemplo, la observación 18º como la inicial, obtendríamos una muestra de tamaño 5:

     18º, 38º, 58º, 78º, 98º
3. N es desconocido. En este caso, la decisión sobre el valor de k se tomará de forma que se asegure el número mínimo deseado de elementos de la muestra. N se estima por defecto, así k será menor de lo necesario y, por tanto, el tamaño muestral será mayor o igual de lo requerido.

Ventajas del muestreo sistemático frente al muestreo aleatorio simple:

- En la práctica **el muestreo sistemático es más fácil de llevar a cabo** y está expuesto a menos errores del encuestador.

  En el muestreo aleatorio simple podría ser un problema si dos números aleatorios fueran consecutivos o muy próximos. Por ejemplo, sería difícil escoger una muestra aleatoria simple de personas entre las que entran a un supermercado. Al seleccionar las personas al azar podríamos encontrarnos que no hemos acabado de hacer la encuesta a un cliente cuando el siguiente a encuestar ya ha pasado. Pero sí sería fácil coger 1 de cada 20 personas que pasen hasta completar la muestra.
- Frecuentemente, con igual tamaño de muestra, **el muestreo sistemático proporciona mejor información** que el muestreo aleatorio simple. Esto se debe a que la muestra sistemática se extiende uniformemente a lo largo de toda la población, mientras que en el muestreo aleatorio simple puede ocurrir que un gran número de observaciones se concentre en una parte de la población y descuide otras.

  Por ejemplo, supongamos que en una fábrica los primeros 3000 motores se fabrican correctamente y los últimos 3000 son defectuosos por un desajuste en la línea de montaje. Una muestra aleatoria simple podría seleccionar un gran número o incluso todos del mismo grupo, dando una mala estimación de la proporción de defectuosos. El muestreo sistemático, en cambio, selecciona el mismo número de motores de ambos grupos, dando una estimación mejor. En este caso, donde en cierta medida **hay un orden en la población**, el muestreo sistemático es mejor que el muestreo aleatorio simple.

***Usos*:**

Este tipo de muestreo es muy utilizado por los planes de muestreo para el **control de calidad** dentro del proceso de fabricación, los auditores cuando se enfrentan a **largas listas** de apuntes para comprobar y los investigadores de mercados cuando se enfrentan a **personas en movimiento**.

4.2 Estimación de la media, proporción y el total poblacionales

- ESTIMADOR DE LA MEDIA POBLACIONAL: $\hat{\mu} = \overline{y}_{sy} = \frac{1}{n}\sum_{j=1}^{n} y_{i+(j-1)k}$

  $1 \le i \le k$ *i=punto de inicio o pivote*
- VARIANZA ESTIMADA DE $\overline{y}_{sy}$: $\hat{V}\left(\overline{y}_{sy}\right) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$

Comentarios.

- Si se desconoce el tamaño poblacional por su gran magnitud, entonces $\frac{N-n}{N} \cong 1$.
- Cuando N no es múltiplo exacto de n, el estimador es sesgado.

Como puede observarse, la varianza del estimador de la media se estima igual que en el muestreo aleatorio simple (véase *4.3 Comparación con el muestreo aleatorio simple*). Aunque las varianzas de los estimadores no son realmente iguales, éstas son:

$$V(\bar{y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \qquad \text{y} \qquad V(\bar{y}_{sy}) = \frac{\sigma^2}{n}\left[1+(n-1)\rho\right]$$

donde ρ = *coeficiente de correlación entre los elementos de una muestra sistemática*.

El tamaño poblacional se desconoce en muchas situaciones prácticas en las que se usa el muestreo sistemático. Cuando N es conocido también se puede estimar el total poblacional.

- ESTIMADOR DEL TOTAL POBLACIONAL: $\hat{\tau} = N\bar{y}_{sy}$
- VARIANZA ESTIMADA DE $\hat{\tau}$: $\hat{V}(\hat{\tau}) = N^2\hat{V}(\bar{y}_{sy}) = N^2\frac{S^2}{n}\left(\frac{N-n}{N}\right)$

Ejemplo 4.1

Los funcionarios de un museo están interesados en el número total de personas que visitaron el lugar durante un periodo de 180 días cuando una costosa colección de antigüedades estuvo en exhibición. Puesto que el control de visitantes en el museo cada día es muy costoso, los funcionarios decidieron obtener estos datos cada diez días. La información de esta muestra sistemática de 1 en 10 se resume en esta tabla

Día	*Nº personas que visitan el museo*
3	160
13	350
23	225
⋮	⋮
173	290
	$\sum_{i=1}^{18} y_i = 4868 \qquad \sum_{i=1}^{18} y_i^2 = 1321450$

Use estos datos para estimar el número total de personas que visitaron el museo durante el periodo especificado y el límite para el error de estimación.

Solución

$N = 180 \qquad \hat{\tau} = N\bar{y}_{sy} = 180\dfrac{4868}{18} = 48680$ visitantes

$$S^2 = \frac{\left(1321450 - \dfrac{(4868)^2}{n}\right)}{n-1} = 289,79$$

$$\hat{V}(\hat{\tau}) = N^2 \frac{S^2}{n}\left(\frac{N-n}{N}\right) = 469461,18 \qquad B_{\tau} = 1370,34 \qquad \blacksquare$$

Como en el muestreo aleatorio simple, las propiedades del estimador de la proporción son análogas a las propiedades de la media muestral:

- ESTIMADOR DE LA PROPORCIÓN POBLACIONAL: $\hat{p}_{sy} = \dfrac{1}{n}\sum_{j=1}^{n} y_{i+(j-1)k}$, $y_i = 0,1$

  $1 \le i \le k$ *i=punto de inicio o pivote*

- VARIANZA ESTIMADA DE $\hat{p}_{sy}$: $\hat{V}(\hat{p}_{sy}) = \dfrac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}\left(\dfrac{N-n}{N}\right)$

Notemos, de nuevo, que las varianzas estimadas son iguales a las del muestreo aleatorio simple.

Ejemplo 4.2

La Guardia Civil de Tráfico está interesada en la proporción de automovilistas que llevan el permiso de conducir. Se instala un puesto de control en una carretera nacional y se detiene un conductor de cada siete. Use los datos de la tabla adjunta para estimar la proporción de conductores que portan su licencia. Establezca un límite para el error de estimación. Suponga que 2800 autos pasan por el puesto de verificación durante el periodo de muestreo.

Automóvil	*Respuesta*
1	1
8	1
15	0
⋮	⋮
2794	1
	$\sum_{i=1}^{400} y_i = 324$

Solución

$$\hat{p}_{sy} = \bar{y}_{sy} = \frac{324}{400} = 0,81$$

$$\hat{V}(\hat{p}_{sy}) = \frac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}\left(\frac{N-n}{N}\right) = \frac{0,81(1-0,81)}{400-1}\left(\frac{2800-400}{2800}\right) = 0,000330612 \Rightarrow \qquad B = 0,0364 \quad \blacksquare$$

Si la estratificación de la población fuese ventajosa, el muestreo sistemático puede utilizarse dentro de cada estrato en lugar del muestreo aleatorio simple, aplicándose las fórmulas del muestreo aleatorio estratificado, análogamente a como se han utilizado las del muestreo aleatorio simple para aproximar el comportamiento del muestreo sistemático.

4.3 Comparación con el muestreo aleatorio simple: Poblaciones ordenadas, aleatorias y periódicas

Veamos bajo qué condiciones la varianza estimada de los estimadores en el muestreo sistemático se puede suponer igual a la del muestreo aleatorio simple.

Según las expresiones

$$V(\bar{y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \qquad \text{y} \qquad V(\bar{y}_{sy}) = \frac{\sigma^2}{n}\left[1+(n-1)\rho\right]$$

éstas serán similares cuando $\frac{N-n}{N-1} \cong 1$ y $\rho \cong 0$, pero en otros casos no.

Distinguimos los siguientes casos:

A. ***Población ordenada*** $(\rho \leq 0)$

Una población es ordenada cuando los elementos que la constituyen están ordenados de acuerdo a los valores, crecientes o decrecientes, de una determinada característica. En este caso es preferible el uso del muestreo sistemático, ya que la muestra se extiende uniformemente a lo largo de toda la población:

$$\rho \leq 0 \quad \Rightarrow \quad V(\bar{y}_{sy}) \leq V(\bar{y})$$

Por ejemplo, en una lista de cuentas por cobrar que estén ordenadas de mayor a menor cantidad, las estimaciones de una muestra sistemática tendrían en general una varianza menor que las de una muestra aleatoria simple (es posible que ésta última contenga solo cantidades grandes o cantidades pequeñas).

Al utilizar las varianzas estimadas de los estimadores del muestreo aleatorio simple en el muestreo sistemático conseguimos una estimación conservadora del error (mayor que el error real que cometemos en el muestreo sistemático).

B. ***Población aleatoria*** $(\rho \cong 0)$

Se dice que una población es aleatoria cuando sus elementos están ordenados al azar. En este caso es indiferente el uso del muestreo aleatorio simple y el muestreo sistemático ya que

$$\rho \cong 0 \quad \Rightarrow \quad V(\bar{y}_{sy}) \cong V(\bar{y}).$$

Por ejemplo, en una lista de estudiantes por orden alfabético, la estimación de sus calificaciones sería similar con ambos muestreos ya que las calificaciones no dependen del apellido del estudiante.

C. *Población periódica* $(\rho \geq 0)$

Una población es periódica cuando los valores de la variable objeto de estudio tienen una variación cíclica. En este caso es preferible el muestreo aleatorio simple dado que

$$\rho \geq 0 \quad \Rightarrow \quad V(\bar{y}_{sy}) > V(\bar{y}).$$

Por ejemplo:

a. Supongamos que tenemos una lista en la que los nombres de mujeres y hombres se alternan. Una muestra sistemática con k par proporcionaría solo una lista de mujeres o de hombres.
b. Ventas diarias de un supermercado a partir de una muestra sistemática con $k = 7$.

Para evitar este problema, el investigador puede cambiar varias veces el punto de inicio aleatorio. Esto tiene el efecto de mezclar los elementos de la población y comportarse como una población aleatoria, en cuyo caso el uso de las expresiones del muestreo aleatorio simple en el muestreo sistemático estaría justificado.

4.4 Determinación del tamaño muestral

El tamaño muestral requerido para estimar la media poblacional con un límite *B* para el error de estimación se obtiene de las expresiones del muestreo aleatorio simple. Lo que conduce a obtener muestras más grandes de las necesarias en poblaciones ordenadas y muestras más pequeñas en poblaciones periódicas (*si no se mezclaran los elementos cambiando el punto de inicio*). En poblaciones aleatorias no tendremos problemas.

Tamaño muestral requerido para estimar μ *y* τ *con un límite B para el error de estimación*

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \qquad \text{con} \quad D = \begin{cases} \dfrac{B^2}{4} & \text{para estimar la media} \\ \dfrac{B^2}{4N^2} & \text{para estimar el total} \end{cases}$$

Tamaño muestral requerido para estimar p *y* τ *con un límite B para el error de estimación*

$$n = \frac{Npq}{(N-1)D + pq} \qquad \text{con} \quad D = \begin{cases} \dfrac{B^2}{4} & \text{para estimar p} \\ \dfrac{B^2}{4N^2} & \text{para estimar el total} \end{cases}$$

Ejemplo 4.3 (continuación del ejemplo 4.2)

En un nuevo control, la Guardia Civil de Tráfico espera que pasen unos 5000 automóviles por el puesto de verificación. Determine el tamaño de muestra y k para estimar p con un error inferior al 2%. Repita esta ejercicio pero suponiendo que no se dispone de información previa.

Solución

UTILIZANDO LA INFORMACIÓN PREVIA

$\hat{p} = 0,81 \qquad \hat{q} = 1 - \hat{p} = 0,19$

$$n = \frac{Npq}{(N-1)\frac{B^2}{4} + pq} = \frac{5000 \times 0,81 \times (1-0,81)}{\left((5000-1)\frac{0,02^2}{4}\right) + \left(0,81 \times (1-0,81)\right)} = 1176,97 \cong 1177 \text{ } automóviles$$

$$k \leq \frac{N}{n} = 4,25$$

Si tomáramos $k=5 \Rightarrow n = \frac{5000}{5} = 1000$. Tomando $k=4 \Rightarrow n = \frac{5000}{4} = 1250 \geq 1177$.

SIN UTILIZAR INFORMACIÓN PREVIA

$\hat{p} = 0,5 \qquad \hat{q} = 0,5$

$$n = \frac{Npq}{(N-1)\frac{B^2}{4} + pq} = \frac{5000 \times 0,5 \times 0,5}{\left((5000-1)\frac{0,02^2}{4}\right) + \left(0,5 \times 0,5\right)} = 1666,889 \cong 1667 \text{ } automóviles$$

$$k \leq \frac{N}{n} = 2,9994$$

Con $k=2$ debemos obtener un tamaño muestral superior a 1667 automóviles. En efecto, se obtiene:

$$n = \frac{5000}{2} = 2500$$

Con $k=3$ se debería obtener un tamaño inferior a 1667, tal como se ha explicado con anterioridad, pero en esta ocasión el cociente $\frac{N}{n} = 2,9994$ está muy próximo a 3 y sucede lo siguiente:

$$n = \frac{5000}{3} = 1666,667$$

En definitiva, si la muestra de tamaño 2500 resulta muy elevada se puede optar por utilizar el valor $k=3$, y detener el muestreo sistemático cuando hayan pasado por el control de la Guardia Civil 5001 automóviles, puesto que en este caso la muestra será del tamaño deseado:

$$n = \frac{5001}{3} = 1667$$

■

APÉNDICE: Estudio empírico de todas las posibles muestras en un muestreo sistemático sobre distintos tipos de poblaciones.

Consideramos las siguientes tres poblaciones: En la primera los datos están ordenados en sentido creciente desde 1 a 60, en la segunda los mismos 60 elementos anteriores están desordenados aleatoriamente (por eso ambas poblaciones comparten la misma media y varianza) y en la tercera población se repite un patrón periódicamente, 6-24-60 … 6-24-60.

POBLACIÓN ORDENADA	POBLACIÓN ALEATORIA	POBLACIÓN PERIÓDICA	POBLACIÓN ORDENADA	POBLACIÓN ALEATORIA	POBLACIÓN PERIÓDICA
1	32	6	31	43	6
2	49	24	32	4	24
3	35	60	33	31	60
4	18	6	34	19	6
5	13	24	35	10	24
6	6	60	36	24	60
7	22	6	37	2	6
8	55	24	38	53	24
9	57	60	39	44	60
10	1	6	40	16	6
11	59	24	41	52	24
12	50	60	42	27	60
13	8	6	43	25	6
14	12	24	44	7	24
15	46	60	45	41	60
16	20	6	46	21	6
17	26	24	47	39	24
18	14	60	48	30	60
19	11	6	49	58	6
20	3	24	50	51	24
21	23	60	51	40	60
22	34	6	52	37	6
23	15	24	53	42	24
24	47	60	54	29	60
25	36	6	55	33	6
26	54	24	56	45	24
27	56	60	57	5	60
28	60	6	58	28	6
29	48	24	59	38	24
30	9	60	60	17	60
		N	**60**	**60**	**60**
		MEDIA (μ)	**30,5**	**30,5**	**30**
		VARIANZA (σ^2)	**299,92**	**299,92**	**504**

Veremos las ventajas e inconvenientes de utilizar un muestreo sistemático en cada tipo de población, confirmando así las propiedades teóricas estudiadas en este tema.

Hacemos un muestreo sistemático de 1 en 6 sobre la población ordenada. Todas las posibles muestras sistemáticas se recogen en la siguiente tabla:

***k*=6**	**POBLACIÓN ORDENADA**					
***n*=10**	**muestra 1**	**muestra 2**	**muestra 3**	**muestra 4**	**muestra 5**	**muestra 6**
1	1	2	3	4	5	6
2	7	8	9	10	11	12
3	13	14	15	16	17	18
4	19	20	21	22	23	24
5	25	26	27	28	29	30
6	31	32	33	34	35	36
7	37	38	39	40	41	42
8	43	44	45	46	47	48
9	49	50	51	52	53	54
10	55	56	57	58	59	60
$\bar{y}_{sy}=$	**28**	**29**	**30**	**31**	**32**	**33**

En el m.a.s. $V\left(\bar{y}\right)=25,42$

$E\left(\bar{y}_{sy}\right)=30,5$

$V\left(\bar{y}_{sy}\right)=2,92$

Como puede verse, en este caso, el muestreo sistemático conduce a un estimador de la media con mucha menos varianza que el muestreo aleatorio simple, $\left(V\left(\bar{y}_{sy}\right)=2,92\right)<\left(V\left(\bar{y}\right)=25,42\right)$, y por tanto con menor error de estimación.

Repetimos el proceso anterior, ahora sobre la población desordenada aleatoriamente:

***k*=6**	**POBLACIÓN ALEATORIA**					
***n*=10**	**muestra 1**	**muestra 2**	**muestra 3**	**muestra 4**	**muestra 5**	**muestra 6**
1	32	49	35	18	13	6
2	22	55	57	1	59	50
3	8	12	46	20	26	14
4	11	3	23	34	15	47
5	36	54	56	60	48	9
6	43	4	31	19	10	24
7	2	53	44	16	52	27
8	25	7	41	21	39	30
9	58	51	40	37	42	29
10	33	45	5	28	38	17
$\bar{y}_{sy}=$	**27**	**33,3**	**37,8**	**25,4**	**34,2**	**25,3**

En el m.a.s. $V\left(\bar{y}\right)=25,42$

$E\left(\bar{y}_{sy}\right)=30,5$

$V\left(\bar{y}_{sy}\right)=23,35$

En este caso, el muestreo sistemático conduce a un estimador de la media con una varianza parecida a la del muestreo aleatorio simple, $\left(V\left(\bar{y}_{sy}\right)=23,35\right)\cong\left(V\left(\bar{y}\right)=25,42\right)$, y por tanto con similar error de estimación.

Repetimos ahora el proceso anterior sobre la población periódica:

***k*=6**	**POBLACIÓN PERIÓDICA**						
***n*=10**	**muestra 1**	**muestra 2**	**muestra 3**	**muestra 4**	**muestra 5**	**muestra 6**	
1	6	24	60	6	24	60	
2	6	24	60	6	24	60	
3	6	24	60	6	24	60	
4	6	24	60	6	24	60	
5	6	24	60	6	24	60	
6	6	24	60	6	24	60	
7	6	24	60	6	24	60	
8	6	24	60	6	24	60	
9	6	24	60	6	24	60	En el m.a.s. $V\left(\bar{y}\right)=42,71$
10	6	24	60	6	24	60	
$\bar{y}_{sy}=$	**6**	**24**	**60**	**6**	**24**	**60**	$E\left(\bar{y}_{sy}\right)=30$ $V\left(\bar{y}_{sy}\right)=504$

En poblaciones periódicas el muestreo sistemático conduce a un estimador de la media con mucha mayor varianza que el muestreo aleatorio simple, $\left(V\left(\bar{y}_{sy}\right)=504\right)>\left(V\left(\bar{y}\right)=25,42\right)$, y por tanto con mayor error de estimación. Esto es así porque *k*=6 es múltiplo de 3, el tamaño del ciclo 6-24-60 que se repite periódicamente en esta población.

¿Qué ocurriría si *k* fuese distinto de 3 o de cualquier otro múltiplo de 3?

Vamos a comprobarlo tomando sobre las mismas poblaciones todas las muestras sistemáticas de 1 en 5.

k=5	POBLACIÓN ORDENADA				
n=12	**muestra 1**	**muestra 2**	**muestra 3**	**muestra 4**	**muestra 5**
1	1	2	3	4	5
2	6	7	8	9	10
3	11	12	13	14	15
4	16	17	18	19	20
5	21	22	23	24	25
6	26	27	28	29	30
7	31	32	33	34	35
8	36	37	38	39	40
9	41	42	43	44	45
10	46	47	48	49	50
11	51	52	53	54	55
12	56	57	58	59	60
$\bar{y}_{sy}$ =	**28,5**	**29,5**	**30,5**	**31,5**	**32,5**

En el m.a.s. $V\left(\bar{y}\right)=20,33$

$E\left(\bar{y}_{sy}\right)=30,5$

$V\left(\bar{y}_{sy}\right)=2$

En la población ordenada seguimos teniendo un estimador con menor varianza, $\left(V\left(\bar{y}_{sy}\right)=2\right)<\left(V\left(\bar{y}\right)=20,33\right)$, y por tanto con menor error de estimación.

k=5	POBLACIÓN ALEATORIA				
n=12	**muestra 1**	**muestra 2**	**muestra 3**	**muestra 4**	**muestra 5**
1	32	49	35	18	13
2	6	22	55	57	1
3	59	50	8	12	46
4	20	26	14	11	3
5	23	34	15	47	36
6	54	56	60	48	9
7	43	4	31	19	10
8	24	2	53	44	16
9	52	27	25	7	41
10	21	39	30	58	51
11	40	37	42	29	33
12	45	5	28	38	17
$\bar{y}_{sy}$ =	**34,92**	**29,25**	**33**	**32,33**	**23**

En el m.a.s. $V\left(\bar{y}\right)=20,33$

$E\left(\bar{y}_{sy}\right)=30,5$

$V\left(\bar{y}_{sy}\right)=17,39$

En la población con un orden aleatorio seguimos teniendo un estimador con una varianza no muy diferente de la obtenida en el muestreo aleatorio simple, $\left(V\left(\bar{y}_{sy}\right)=17,39\right)\approx\left(V\left(\bar{y}\right)=20,33\right)$, y por tanto con parecido error de estimación.

k=5	POBLACIÓN PERIÓDICA				
n=12	**muestra 1**	**muestra 2**	**muestra 3**	**muestra 4**	**muestra 5**
1	6	24	60	6	24
2	60	6	24	60	6
3	24	60	6	24	60
4	6	24	60	6	24
5	60	6	24	60	6
6	24	60	6	24	60
7	6	24	60	6	24
8	60	6	24	60	6
9	24	60	6	24	60
10	6	24	60	6	24
11	60	6	24	60	6
12	24	60	6	24	60
$\bar{y}_{sy}$ =	**30**	**30**	**30**	**30**	**30**

En el m.a.s. $V\left(\bar{y}\right)=34{,}17$

$E\left(\bar{y}_{sy}\right)=30$

$V\left(\bar{y}_{sy}\right)=0$

Y curiosamente, en la población con un orden periódico, hemos obtenido un estimador exacto, con una varianza nula. Por tanto, el hecho de que la población sea periódica no es ningún inconveniente para aplicar muestreo sistemático siempre y cuando el valor de k no sea igual o múltiplo del tamaño del ciclo que periódicamente se repite, en nuestro caso $k \neq 3,\ k \neq 6,\ k \neq 9$ ...

EJERCICIOS RESUELTOS

1. La gerencia de una compañía privada con 2000 empleados está interesada en estimar la proporción de empleados que están a favor de una nueva política de inversión. Una muestra sistemática de 1 en 10 es obtenida de los empleados que salen del edificio al final de un día de trabajo (*las respuestas a favor se han representado como 1*)

Empleado muestreado	*Respuesta*
3	1
13	0
23	1
⋮	⋮
1993	1

$$\sum_{i=1}^{200} y_i = 110$$

Se quiere volver a repetir el anterior estudio pero con un error de estimación inferior al 5% (*considerando la muestra anterior como una muestra previa para estimar los parámetros necesarios*). ¿Qué tipo de muestra sistemática deberá obtenerse? (indique *n* y *k*).

SOLUCIÓN

$$N = 2000 \quad \hat{p} = \frac{110}{200} = 0,55 \quad \hat{q} = 1 - \hat{p} = 0,45 \quad D = \frac{0,05^2}{4} = 0,000625$$

$$n = \frac{Npq}{(N-1)D + pq} = 330,7 \approx 331 \qquad k \leq \frac{N}{n} = 6,04 \Rightarrow k = 6$$

2. Un auditor se enfrenta a una larga lista de 1000 cuentas por cobrar de una empresa. El valor de cada una de estas cuentas no suele superar los 21000€. El auditor quiere estimar el valor total de las deudas por cobrar con un error inferior a 1000000€ y con una confianza del 95%. Para ello decide tomar una muestra sistemática de 1 en k. Determine el valor de *k*.

SOLUCIÓN

$$N = 1000 \quad R = 21000 \quad \sigma^2 \cong \frac{21000^2}{4^2} = 27562500 \quad D = \frac{1000000^2}{4 \times 1000^2} = 250000$$

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = 99,39 \approx 100 \quad k = \frac{N}{n} = 10$$

3. La tabla anexa muestra el número de nacimientos y la tasa de natalidad por cada 1000 individuos para Estados Unidos durante seis años seleccionados sistemáticamente.

Año	Nac.Masculinos	Nac.Femeninos	Total de Nac.	Natalidad
1955	2073719	1973576	4047295	26,0
1960	2179708	2078142	4257850	23,7
1965	1927054	1833304	3760358	19,4
1970	1915378	1816008	3731386	18,4
1975	1613135	1531063	3144198	14,6
1980	1852616	1759642	3612258	15,9

Estime el número medio de varones nacidos por año para el periodo 1955-1980, y establezca un límite para el error de estimación.

SOLUCIÓN

Desde 1955 hasta 1980, ambos inclusive, hay 26 años. $N = 26$.

$$\hat{\mu} = \bar{y}_{sy} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}11561610 = 1926935$$

$S^2 = 37913412871,20$ (*con las funciones estadísticas en el modo SD de la calculadora*)

$$\hat{V}\left(\bar{y}_{sy}\right) = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = 4860693957,85 \qquad \text{B} = 139437,35$$

4. La sección de control de calidad de una empresa usa el muestreo sistemático para estimar la cantidad media de llenado en latas de 33*cl* que salen de una línea de producción. Los datos de la tabla adjunta representan una muestra sistemática 1 en 300 de una producción diaria de 1800 latas.

Cantidad de llenado en cl					
33	32,5	33,5	33	32	31

Determine el tamaño de la muestra y *k* para estimar el contenido medio de las latas con un error de estimación inferior a 0,42*cl*, considerando la muestra anterior como una muestra previa para estimar los parámetros necesarios.

SOLUCIÓN: N=1800 n'=6

(*con las funciones estadísticas del modo SD de la calculadora*): $S^2_{n'-1} = 0,8$ $\widehat{\sigma^2} = S^2_{n'-1}$

$$D = \frac{B^2}{4} = 0,0441 \qquad n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = 17,97 \approx 18 \qquad k = \frac{1800}{18} = 100$$

5. Los funcionarios de cierta sociedad profesional desean determinar la proporción de miembros que apoyan varias enmiendas propuestas en las prácticas de arbitraje. Los funcionarios tomaron una muestra sistemática de 1 en 10, a partir de una lista en orden alfabético de los 650 miembros registrados, obteniendo que 47 estaban a favor de los cambios propuestos. Se quiere repetir el estudio anterior con un error de estimación inferior al 5%. Considerando la muestra anterior como una muestra previa para estimar los parámetros necesarios, ¿qué tipo de muestra sistemática deberá obtenerse? (indique *n* y *k*).

SOLUCIÓN:

N=650 n'=65 $\hat{p} = \frac{47}{65} = 0,7231$ $\hat{q} = 1 - 0,7231 = 0,2769$

$$B = 0,05 \qquad D = \frac{B^2}{4} = 0,000625$$

$$n = \frac{Npq}{(N-1)D + pq} = 214,8 \approx 215 \qquad k \le \frac{650}{215} = 3,02 \qquad k = 3$$

5. Muestreo por conglomerados.

5.1 Necesidad y ventajas del muestreo por conglomerados.

Una muestra por conglomerados es una muestra aleatoria en la cual cada unidad de muestreo es una colección (o conglomerado) de elementos.

El muestreo por conglomerados es útil para obtener información en las siguientes situaciones:

- Es complicado disponer de una lista de los elementos de la población, mientras que es fácil lograr un marco que liste los conglomerados. (Alumnos que asisten a clase = elemento, aulas = conglomerados)
- El coste de obtención de las observaciones es menor debido al agrupamiento de los elementos.

5.2 Formación de los conglomerados. Conglomerados y estratos.

Los elementos de un conglomerado deben ser diferentes entre sí, así una muestra con pocos conglomerados recogería gran cantidad de información sobre el parámetro poblacional. Si los elementos dentro de un conglomerado presentan características similares, tomar varias observaciones dentro de un conglomerado no aporta más información.

Recordemos que los estratos debían ser tan homogéneos como fuera posible y diferir tanto como se pudiera uno de otro con respecto a la característica que está siendo estudiada. Los conglomerados, sin embargo, deben ser tan heterogéneos dentro de ellos como sea posible y muy similar uno a otro para que el muestreo por conglomerados esté indicado y proporcione buenos resultados.

Una vez especificados los conglomerados, se selecciona una muestra aleatoria simple de conglomerados.

5.3 Estimación de la media, proporción y total poblacionales.

Vamos a utilizar la siguiente notación:

$N =$ conglomerados en la población. $n =$ conglomerados en la muestra.

$m_i =$ elementos en el conglomerado i

$y_i =$ suma de las observaciones en el conglomerado i

$M = \sum_{i=1}^{N} m_i =$ elementos en la población (*con frecuencia es desconocido*)

$m = \sum_{i=1}^{n} m_i =$ elementos en la muestra

$\overline{M} = \frac{1}{N}\sum_{i=1}^{N} m_i =$ tamaño medio de los conglomerados de la población (*con frecuencia es desconocido*)

$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i =$ tamaño medio de los conglomerados de la muestra (*se utiliza para estimar* $\overline{M}$).

(A) Estimación de la media.

El estimador de la media poblacional μ es la media $\bar{y}$,

$$\hat{\mu} = \bar{y} = \frac{1}{m}\sum_{i=1}^{n} y_i = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

Es un estimador sesgado. Tiene la forma de un estimador de razón, por lo que la varianza estimada de $\bar{y}$ toma la misma forma que la varianza de un estimador de razón.

$$\hat{V}(\bar{y}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} \qquad \text{donde} \quad S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2$$

($\overline{M}$ *es estimado por* $\overline{m}$*, si se desconoce*)

La varianza estimada es también sesgada y sería un buen estimador de $V(\bar{y})$ si *n* es grande ($n \geq 20$). El sesgo, tanto para $\hat{\mu}$ como para $\hat{V}(\bar{y})$, desaparece cuando los tamaños de los conglomerados son iguales ($m_1 = m_2 = ... = m_N$), por lo que este tipo de muestreo no es adecuado cuando hay una gran disparidad entre los tamaños de los conglomerados.

Notas:

- La expresión de $\hat{V}(\bar{y}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n}$ no se suele simplificar como $\hat{V}(\bar{y}) = \frac{N(N-n)}{M^2}\frac{S_c^2}{n}$ para poder calcularla fácilmente cuando *N* y *M* sean desconocidos.

  A veces *N* no se conoce debido a su gran tamaño y $\frac{N-n}{N}$ se aproxima por 1. Si M es desconocido $\overline{M}$ debe ser estimada por $\overline{m}$.
- Si la variable que estamos estudiando es dicotómica, hablaremos de la proporción poblacional p y de la proporción muestral $\hat{p}$. En este caso al número total de elementos

en el conglomerado i que poseen la característica de interés se nota como a_i en lugar de y_i como es habitual en variables numéricas. Así tendremos que

$$\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}$$

Salvo esta diferencia en la notación, todo lo anteriormente expuesto para variables numéricas es válido para variables dicotómicas.

(B) Estimación del total.

De la relación entre la media y el total poblacional $\mu = \frac{\tau}{M}$ se sigue que $\tau = M\mu$, siendo el estimador del total poblacional τ

$$\hat{\tau} = M\bar{y}$$

y la varianza estimada del mismo

$$\hat{V}(\hat{\tau}) = M^2\hat{V}(\bar{y}) = N(N-n)\frac{S_c^2}{n} \cong \frac{M^2}{\bar{m}^2}\frac{S_c^2}{n}$$

(*sea cual sea el valor de M, éste no afecta a la varianza ni al error del estimador, aunque sí al valor del estimador del total*)

Como en la estimación del total con un estimador de razón, cuando $\bar{M}$ es desconocida y se estima por $\bar{m}$ y $\frac{N-n}{N}$ se aproxima por 1 debido al gran tamaño de *N*, la varianza del estimador del total se aproxima por $\frac{M^2}{\bar{m}^2}\frac{S_c^2}{n}$.

(C) Estimación del total cuando se desconoce el tamaño de la población.

Frecuentemente el número de elementos en la población no es conocido en problemas donde se aplica el muestreo por conglomerados. En ese caso no podemos utilizar el estimador del total $\hat{\tau} = M\bar{y}$, debemos construir un estimador del total que no dependa de M. La cantidad $\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i$, es el promedio de los totales de los conglomerados de la muestra y un estimador insesgado del promedio de los *N* totales de los conglomerados de la población. Por el mismo razonamiento empleado en el muestreo aleatorio simple, $N\bar{y}_t$ es un estimador insesgado de la suma de los totales de todos los conglomerados, o equivalentemente del total poblacional τ.

En resumen

$$\hat{\tau}_t = N\bar{y}_t$$

$$\hat{V}(\hat{\tau}_t) = N^2\hat{V}(\bar{y}_t) = N(N-n)\frac{S_t^2}{n}$$

donde $\hat{V}(\bar{y}_t) = \frac{N-n}{N}\frac{S_t^2}{n}$ $\quad S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2$ $\quad \bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i$

Si existe una gran variación entre los tamaños de los conglomerados y además los tamaños están altamente correlacionados con los totales de los conglomerados, la varianza de $\hat{\tau}_t = N\bar{y}_t$ es generalmente mayor que la varianza de $\hat{\tau} = M\bar{y}$. Esto es debido a que el estimador $\hat{\tau}_t = N\bar{y}_t$ no usa la información proporcionada por los tamaños de los conglomerados, m_i, y por ello puede ser menos preciso.

Cuando los tamaños de los conglomerados son iguales, los dos estimadores del total coinciden, además el estimador de la media, $\bar{y}$, es un estimador insesgado de la media poblacional y también es insesgado el estimador de su varianza, $\hat{V}(\bar{y})$ (*lo mismo vale para el total*).

Ejemplo 5.1 En una ciudad se quiere estimar la proporción de hogares interesados en contratar el sistema de televisión digital, para lo cual se considera la ciudad dividida en 200 manzanas de viviendas. Se extrae una muestra piloto de 5 manzanas y se interroga a cada familia acerca de si estaría interesada en contratar la televisión digital. Los datos de la encuesta se encuentran en la tabla:

Manzana	Nº hogares en la manzana	Nº hogares interesados
1	8	2
2	7	2
3	9	3
4	6	3
5	5	3

a) Estime la proporción de hogares interesados en contratar el sistema de televisión digital. Calcule el límite para el error de estimación.
b) Con un intervalo de confianza estime el número de hogares interesados en contratar dicho sistema.
c) Responda al apartado b) suponiendo que el número de hogares en la ciudad es 1500.

SOLUCIÓN

Aunque en un caso de variables dicotómicas como éste se suele usar en los textos la notación a_i en lugar de y_i , utilizaremos esta última para unificar la notación a emplear en el muestreo por conglomerados, tanto para variables numéricas como dicotómicas.

m_i	y_i	m_i^2	y_i^2	$m_i y_i$
8	2	64	4	16
7	2	49	4	14
9	3	81	9	27
6	3	36	9	18
5	3	25	9	15
35	13	255	35	90

a) N=200 n=5

$$\boxed{\hat{p} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{13}{35} = 0,3714 \qquad \hat{p} = 37,14\%}$$

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2\bar{y}\sum_{i=1}^{n} y_i m_i + \bar{y}^2\sum_{i=1}^{n} m_i^2\right) = \frac{3,3222}{4} = 0,8306$$

Ya que M es desconocido, $\overline{M}$ debe ser estimada por $\bar{m}$

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i = \frac{35}{5} = 7 \; hogares/manzana$$

$$\hat{V}(\bar{y}) = \frac{1}{\bar{m}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 0,003305 \qquad \boxed{2\sqrt{\hat{V}(\bar{y})} = 0,115 \qquad 11,5\%}$$

b) $\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{13}{5} = 2,6 \qquad \hat{\tau}_t = N\bar{y}_t = 520$

$$S_t^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}{n-1} = 0,3 \qquad \hat{V}(\hat{\tau}_t) = \frac{N(N-n)S_t^2}{n} = 2340$$

$$2\sqrt{\hat{V}(\hat{\tau}_t)} = 96,75 \qquad \boxed{(423,25\,;\,616,75)}$$

c)

$$\hat{\tau} = M\bar{y} = 557,14 \qquad \overline{M} = \frac{1500}{200} = 7,5 \qquad \hat{V}(\bar{y}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 0,0028795$$

Observe que al conocer $\overline{M}$, la estimación de $V(\bar{y})$ es diferente de la obtenida en **a)**.

$$\hat{V}(\hat{\tau}) = M^2\hat{V}(\bar{y}) = 6478,8 \qquad 2\sqrt{\hat{V}(\hat{\tau})} = 160,98 \qquad \boxed{(396,16\,;\,718,12)}$$

El límite para el error de estimación es más pequeño en b) que en c), debido a que los tamaños de los conglomerados no están correlacionados con los totales de los conglomerados ($r_{my}^2 = 0,08$). En otras palabras, los tamaños de los conglomerados proporcionan poca información sobre los totales de los conglomerados. ■

5.4 Determinación del tamaño muestral.

Supongamos que los conglomerados ya están formados y vamos a seleccionar el número de conglomerados *n* para conseguir un determinado límite para el error de estimación *B*

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2}$$

donde σ_c^2 se estima mediante $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}m_i\right)^2$ de una muestra previa, siendo

$D = \frac{B^2\overline{M}^2}{4}$ para la estimación de la media y $D = \frac{B^2}{4N^2}$ para la estimación del total.

Habitualmente el tamaño promedio de los conglomerados de la población $\overline{M}$ no se conoce y tiene que estimarse por el tamaño medio $\overline{m}$ de los conglomerados de una muestra previa.

Cuando se utiliza $N\overline{y}_t$ para estimar el total, el número de conglomerados en la muestra para obtener un determinado límite para el error de estimación *B* viene dado por

$$n = \frac{N\sigma_t^2}{ND + \sigma_t^2}$$

$D = \frac{B^2}{4N^2}$ y σ_t^2 se estima mediante $S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}_t\right)^2$ de una muestra previa.

Ejemplo 5.2 Suponiendo que los datos del ejemplo 5.1 representan una muestra previa, cómo debe tomarse una nueva muestra para estimar la proporción poblacional del apartado a) con un límite para el error de estimación del 1%.

SOLUCIÓN

$$S_c^2 = 0,8306 \qquad \overline{M} \cong \overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i = \frac{35}{5} = 7 \qquad D = \frac{B^2\overline{M}^2}{4} = \frac{0,01^2 \times 7^2}{4} = 0,001225$$

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = 154,4 \approx 155$$ ■

APÉNDICE: Estudio empírico de todas las posibles muestras en un muestreo por conglomerados sobre una población finita.

Consideramos para nuestro ejemplo la siguiente población con 36 elementos:

101	201	301	401	501	601
102	202	302	402	502	602
103	203	303	403	503	603
104	204	304	404	504	604
105	205	305	405	505	605
106	206	306	406	506	606

Media (μ)*:* 353,5

Varianza (σ^2)*:* 29169,58

N (*número de elementos*)*:* 36

Si estimamos la media de la anterior población usando muestreo aleatorio simple con un tamaño muestral de n=24, la varianza del estimador media muestral es $V(\bar{y}) = 416,71$.

Consideramos la anterior población dividida en los siguientes 6 conglomerados:

Conglomerado						
1	101	201	301	401	501	601
2	102	202	302	402	502	602
3	103	203	303	403	503	603
4	104	204	304	404	504	604
5	105	205	305	405	505	605
6	106	206	306	406	506	606

Media (μ)*:* 353,5

Varianza (σ^2)*:* 29169,58

M (*número de elementos*)*:* 36

N (*número de conglomerados*)*:* 6

Como podemos observar los conglomerados son similares unos a otros, en cada uno hay un valor próximo a 100, otro valor próximo a 200, … y otro valor próximo a 600. Hay homogeneidad entre los conglomerados.

Por otra parte, dentro de cada conglomerado hay bastante heterogeneidad pues en todos ellos hay valores que van desde aproximadamente 100 hasta aproximadamente 600. Es decir hay heterogeneidad dentro de cada uno de los conglomerados.

Tomamos todas las posibles muestras de n=4 conglomerados, en total 4x6=24 elementos en cada muestra.

Muestra	Conglomerados en la muestra, n=4				m_i				y_i				$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$
1	3	4	5	6	6	6	6	6	2118	2124	2130	2136	354,5
2	2	4	5	6	6	6	6	6	2112	2124	2130	2136	354,25
3	2	3	5	6	6	6	6	6	2112	2118	2130	2136	354
4	2	3	4	6	6	6	6	6	2112	2118	2124	2136	353,75
5	2	3	4	5	6	6	6	6	2112	2118	2124	2130	353,5
6	1	4	5	6	6	6	6	6	2106	2124	2130	2136	354
7	1	3	5	6	6	6	6	6	2106	2118	2130	2136	353,75
8	1	3	4	6	6	6	6	6	2106	2118	2124	2136	353,5
9	1	3	4	5	6	6	6	6	2106	2118	2124	2130	353,25
10	1	2	5	6	6	6	6	6	2106	2112	2130	2136	353,5
11	1	2	4	6	6	6	6	6	2106	2112	2124	2136	353,25
12	1	2	4	5	6	6	6	6	2106	2112	2124	2130	353
13	1	2	3	6	6	6	6	6	2106	2112	2118	2136	353
14	1	2	3	5	6	6	6	6	2106	2112	2118	2130	352,75
15	1	2	3	4	6	6	6	6	2106	2112	2118	2124	352,5

MEDIA DEL ESTIMADOR, $E\left(\bar{y}\right) =$ 353,5

VARIANZA DEL ESTIMADOR, $V\left(\bar{y}\right) =$ 0,29

Como podemos observar, el estimador es insesgado $E\left(\bar{y}\right) = 353,5 = \mu$ y tiene una varianza muy pequeña, mucho menor que la que obteníamos en el muestreo aleatorio simple, $\left(V\left(\bar{y}\right) = 0,29\right) << \left(V\left(\bar{y}\right) = 416,71\right)$.

Vamos a compararlo con la siguiente población: la misma población anterior pero se han formado los conglomerados con distinto criterio, de distinta forma. Hay una mayor homogeneidad dentro de sus conglomerados y mayores diferencias de un conglomerado a otro:

Conglomerados					
1	2	3	4	5	6
101	201	301	401	501	601
102	202	302	402	502	602
103	203	303	403	503	603
104	204	304	404	504	604
105	205	305	405	505	605
106	206	306	406	506	606

Media (μ)*:* 353,5

Varianza (σ^2)*:* 29169,58

M (*número de elementos*)*:* 36

N (*número de conglomerados*)*:* 6

Muestra	Conglomerados en la muestra, n=4				m_i				y_i				$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$
1	3	4	5	6	6	6	6	6	1821	2421	3021	3621	453,5
2	2	4	5	6	6	6	6	6	1221	2421	3021	3621	428,5
3	2	3	5	6	6	6	6	6	1221	1821	3021	3621	403,5
4	2	3	4	6	6	6	6	6	1221	1821	2421	3621	378,5
5	2	3	4	5	6	6	6	6	1221	1821	2421	3021	353,5
6	1	4	5	6	6	6	6	6	621	2421	3021	3621	403,5
7	1	3	5	6	6	6	6	6	621	1821	3021	3621	378,5
8	1	3	4	6	6	6	6	6	621	1821	2421	3621	353,5
9	1	3	4	5	6	6	6	6	621	1821	2421	3021	328,5
10	1	2	5	6	6	6	6	6	621	1221	3021	3621	353,5
11	1	2	4	6	6	6	6	6	621	1221	2421	3621	328,5
12	1	2	4	5	6	6	6	6	621	1221	2421	3021	303,5
13	1	2	3	6	6	6	6	6	621	1221	1821	3621	303,5
14	1	2	3	5	6	6	6	6	621	1221	1821	3021	278,5
15	1	2	3	4	6	6	6	6	621	1221	1821	2421	253,5

MEDIA DEL ESTIMADOR, $E(\bar{y}) =$ 353,5

VARIANZA DEL ESTIMADOR, $V(\bar{y}) =$ 2916,67

Como podemos observar, el estimador es insesgado $E(\bar{y}) = 353,5 = \mu$ y tiene una varianza mucho mayor que la que obteníamos en el muestreo aleatorio simple, $\left(V(\bar{y}) = 2916,67\right) > \left(V(\bar{y}) = 416,71\right)$.

En conclusión, el muestreo por conglomerados es adecuado (nos conduce a una menor varianza del estimador y menor error de estimación) cuando los conglomerados son similares

unos de otros y tienen la máxima variabilidad dentro de ellos (máxima representación de los elementos de toda la población)

Por último, vamos a ver qué ocurriría si tuviéramos conglomerados de diferentes tamaños, m_i.

Tomamos como ejemplo la siguiente población:

m_i	Conglomerado						
6	1	101	201	301	401	501	601
5	2	102	202		402	502	602
6	3	103	203	303	403	503	603
4	4	104		304		504	604
6	5	105	205	305	405	505	605
5	6	106	206		406	506	606

Media (μ)*:*	359,69
Varianza (σ^2)*:*	31845,96
M (*número de elementos*)*:*	32
N (*número de conglomerados*)*:*	6

Si estimamos la media de la anterior población usando muestreo aleatorio simple con un tamaño muestral de n=21, la varianza del estimador media muestral es $V(\bar{y}) = 538{,}1$.

Muestra	Conglomerados en la muestra, n=4				m_i				y_i				$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$
1	3	4	5	6	6	4	6	5	2118	1516	2130	1830	361,62
2	2	4	5	6	5	4	6	5	1810	1516	2130	1830	364,30
3	2	3	5	6	5	6	6	5	1810	2118	2130	1830	358,55
4	2	3	4	6	5	6	4	5	1810	2118	1516	1830	363,70
5	2	3	4	5	5	6	4	6	1810	2118	1516	2130	360,67
6	1	4	5	6	6	4	6	5	2106	1516	2130	1830	361,05
7	1	3	5	6	6	6	6	5	2106	2118	2130	1830	355,83
8	1	3	4	6	6	6	4	5	2106	2118	1516	1830	360,48
9	1	3	4	5	6	6	4	6	2106	2118	1516	2130	357,73
10	1	2	5	6	6	5	6	5	2106	1810	2130	1830	358,00
11	1	2	4	6	6	5	4	5	2106	1810	1516	1830	363,10
12	1	2	4	5	6	5	4	6	2106	1810	1516	2130	360,10
13	1	2	3	6	6	5	6	5	2106	1810	2118	1830	357,45
14	1	2	3	5	6	5	6	6	2106	1810	2118	2130	354,96
15	1	2	3	4	6	5	6	4	2106	1810	2118	1516	359,52

MEDIA DEL ESTIMADOR, $E(\bar{y}) =$	359,80
VARIANZA DEL ESTIMADOR, $V(\bar{y}) =$	7,10

Si usamos muestreo por conglomerados el estimador de la media no es insesgado, $\left(E\left(\bar{y}\right)=359{,}80\right)\neq\left(\mu=359{,}69\right)$, aunque el sesgo es despreciable. Sin embargo tiene una varianza mucho menor que la asociada al muestreo aleatorio simple, $\left(V\left(\bar{y}\right)=7{,}10\right)<<\left(V\left(\bar{y}\right)=538{,}1\right)$. Por tanto, el uso de conglomerados es adecuado cuando estos son similares unos de otros y tienen la máxima variabilidad dentro de ellos (máxima representación de los elementos de toda la población) aunque no sean todos los conglomerados de igual tamaño, lo que conlleva un sesgo, en muchos casos despreciable.

EJERCICIOS RESUELTOS

1. Con motivo del cuarto centenario del Quijote, el Ministerio de Cultura desea estimar el número de libros comprados cada mes en una localidad. Se selecciona una localidad con 6200 hogares agrupados en 700 manzanas de viviendas. Se tiene una encuesta piloto en la cual se seleccionó una muestra de 4 manzanas y se entrevistaron a todas las familias, obteniéndose los siguientes resultados:

manzana	*libros comprados cada mes por familia*										
1	1	2	1	0	3	2	1	0	1	2	
2	1	0	2	2	0	0	1	3			
3	2	1	1	1	1	0	2	1	2	2	2
4	1	1	0	2	1	0	3				

Determine, usando los datos de la encuesta piloto, cuántas manzanas debe tener una nueva muestra si se quiere estimar los libros comprados cada mes con un error de estimación inferior a 140 unidades.

SOLUCIÓN

m_i	y_i	m_i^2	y_i^2	$m_i y_i$
10	13	100	169	130
8	9	64	81	72
11	15	121	225	165
7	8	49	64	56
36	45	334	539	423

$$M=6200 \quad N=700 \quad \bar{y}=\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}=1{,}25 \quad D=\frac{B^2}{4N^2}=0{,}01$$

$$\sigma_c^2 \cong S_c^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}m_i\right)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2+\bar{y}^2\sum_{i=1}^{n} m_i^2-2\bar{y}\sum_{i=1}^{n} m_i y_i\right)=1{,}125$$

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = 96,92 \approx 97$$

2. Una industria está considerando la revisión de su política de jubilación y quiere estimar la proporción de empleados que apoyan la nueva política. La industria consta de 57 plantas. Se selecciona una muestra aleatoria simple de 5 plantas y se obtienen las opiniones de los empleados en estas plantas a través de un cuestionario. Los resultados se presentan en esta tabla:

Planta	*Nº empleados*	*Nº empleados que apoyan la nueva política*
1	51	42
2	62	53
3	49	40
4	73	45
5	101	63

a. Estime la proporción de empleados en la industria que apoyan la nueva política de jubilación y establezca un límite para el error de estimación.

b. La industria modificó su política de jubilación después de obtener los resultados de la encuesta. Ahora se quiere estimar la proporción de empleados a favor de la política modificada ¿Cuántas plantas deben ser muestreadas para tener un límite del 5% para el error de estimación? Use los datos anteriores para aproximar los resultados de la nueva encuesta.

SOLUCIÓN:

a) $N = 57 \qquad n = 5$

m_i	y_i	m_i^2	y_i^2	$m_i y_i$
51	42	2601	1764	2142
62	53	3844	2809	3286
49	40	2401	1600	1960
73	45	5329	2025	3285
101	63	10201	3969	6363
336	243	24376	12167	17036

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{243}{336} = 0,7232 \Rightarrow \hat{p} = 72,32\%$$

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{p}m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2\hat{p}\sum_{i=1}^{n} y_i m_i + \hat{p}^2\sum_{i=1}^{n} m_i^2\right) = 68,7$$

$$\overline{M}^2 \approx \overline{m}^2 = \left(\frac{336}{5}\right)^2 = 4515,84$$

$$\hat{V}(\hat{p}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 0,00278 \qquad 2\sqrt{\hat{V}(\hat{p})} = 0,1054 \Rightarrow 10,54\%$$

b)

$$D = \frac{B^2 \overline{M}^2}{4} = \frac{0,05^2 \times 4515,84}{4} = 2,8224 \qquad \sigma_c^2 \approx S_c^2 \qquad n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = 17,06 \approx 18$$

3. Un sociólogo quiere estimar el ingreso medio por persona en cierta ciudad pequeña donde no existe una lista disponible de adultos residentes. Por esta razón para el diseño de la encuesta utiliza muestreo por conglomerados. Se divide la ciudad en bloques rectangulares y el sociólogo decide que cada bloque rectangular va a ser considerado como un conglomerado. Los conglomerados son numerados del 1 al 415. El investigador tiene tiempo y dinero suficientes para hacer un muestreo de 25 conglomerados y entrevistar a cada hogar dentro de cada uno. Se seleccionan aleatoriamente 25 conglomerados y se realizan las entrevistas, obteniéndose estos datos:

Conglomerado (i)	*Nº de residentes (m_i)*	*Ingreso total por conglomerado en € (y_i)*
1	8	96000
2	12	121000
3	4	42000
4	5	65000
5	6	52000
6	6	40000
7	7	75000
8	5	65000
9	8	45000
10	3	50000
11	2	85000
12	6	43000
13	5	54000
14	10	49000
15	9	53000
16	3	50000
17	6	32000
18	5	22000
19	5	45000
20	4	37000
21	6	51000
22	8	30000
23	7	39000
24	3	47000
25	8	41000
	151 residentes	1329000 €

a) Estime el ingreso medio por persona en la ciudad y establezca un límite para el error de estimación.

b) Estime el ingreso total de todos los residentes de la ciudad y el límite para el error de estimación, suponiendo que M es desconocido.

c) Suponiendo que existen 2500 residentes en la ciudad, estime el ingreso total de todos los residentes de la ciudad mediante un intervalo de confianza.

NOTA: *Repetir este ejemplo con todos los* m_i *iguales (por ejemplo,* $m_i = 6 \ \forall i$*, supongamos conocido* $M = 6 \times 415 = 2490$*) y estime el total por los dos métodos estudiados* $\left(\hat{\tau} = M\bar{y} \quad \hat{\tau}_t = N\bar{y}_t\right)$*. Observe como coinciden las dos estimaciones así como la varianza del estimador y el límite para el error de estimación.*

d) Tomando los anteriores datos como una muestra previa, cómo debe tomarse la muestra en una encuesta futura para estimar el ingreso promedio por persona con un límite para el error de estimación de 500€.

SOLUCIÓN:

a) (*este ejemplo no se puede resolver con una calculadora de 10 dígitos de forma exacta por la dificultad de trabajar con cantidades muy grandes*)

n=25 N=415

$$\boxed{\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{1329000}{151} = 8801{,}32 \text{ €} / \textit{residente}}$$

$$\sum_{i=1}^{n} y_i^2 = 96000^2 + \ldots = 82039000000 \qquad \sum_{i=1}^{n} m_i^2 = 8^2 + \ldots = 1047$$

$$\sum_{i=1}^{n} y_i m_i = \left(96000 \times 8\right) + \ldots = 8403000$$

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2\bar{y}\sum_{i=1}^{n} y_i m_i + \bar{y}^2\sum_{i=1}^{n} m_i^2\right) = \frac{15227502247}{24} = 634501213{,}40$$

Ya que M es desconocido, $\overline{M}$ debe ser estimada por $\bar{m}$,

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i = \frac{151}{25} = 6{,}04 \ \textit{residente} / \textit{bloque}$$

$$\hat{V}(\bar{y}) = \frac{1}{\bar{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 653785{,}19 \qquad \boxed{2\sqrt{\hat{V}(\bar{y})} = 1617{,}14€}$$

b)

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1329000}{25} = 53160 \text{ €} / \textit{bloque} \qquad \boxed{\hat{\tau}_t = N\bar{y}_t = 22061400 \text{ €}}$$

$$S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}_t\right)^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}{n-1} = \frac{82039000000 - \frac{1}{25}(1329000)^2}{24} =$$

$$= \frac{11389360000}{24} = 474556666,6$$

$$\widehat{V}(\hat{\tau}_t) = N(N-n)\frac{S_t^2}{n} = 3072279860000 \qquad \boxed{2\sqrt{\widehat{V}(\hat{\tau}_t)} = 3505584,04\ €}$$

c) $N = 415 \qquad n = 25 \qquad \overline{M} = \frac{2500}{415} = 6,0241 \qquad \hat{\tau} = M\overline{y} = 22003311,26€$

$$S_c^2 = 634501213,40 \qquad \widehat{V}(\overline{y}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 657240,9482$$

$$\widehat{V}(\hat{\tau}) = M^2\widehat{V}(\overline{y}) = 4107755926250 \qquad 2\sqrt{\widehat{V}(\hat{\tau})} = 4053519,92$$

$$\boxed{(17949791,34€\ ;\ 26056831,18€)}$$

Como puede observarse el límite para el error de estimación es más pequeño en b) que en c) debido a que los tamaños de los conglomerados no están altamente correlacionados con los totales de los conglomerados en este ejemplo ($r_{my}^2 = 0,0919$). En otras palabras, los tamaños de los conglomerados proporcionan poca información referente a los totales de los conglomerados.

d) $S_c^2 = 634501213,40 \qquad D = \frac{B^2\overline{M}^2}{4} = \frac{500^2 \times 6,04^2}{4} = 2280100$

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = 166,58 \approx 167$$

4. Una empresa de trabajo temporal quiere investigar las necesidades de empleo de las empresas de un pueblo. Para ello decide seleccionar una muestra de 10 de las 85 inscritas en el registro mercantil. El número de bajas en el último año, el número de empleados y la respuesta de cada empresa sobre si utilizaría los servicios de la empresa de trabajo temporal fueron los siguientes:

Empresa	Bajas	Empleados	Respuesta
1	1	7	Si
2	2	15	No
3	9	85	Si
4	0	3	No
5	2	12	No
6	0	8	No
7	1	21	Si
8	0	4	No
9	4	35	No
10	6	92	Si

(a) Estime el número de bajas en el último año en las empresas del pueblo y el límite del error de estimación.

(b) Estime la proporción de empresas que usarían los servicios ofertados y el límite del error de estimación.

SOLUCIÓN:

a) Se trata de un muestreo por conglomerados (*cada empresa es un conglomerado*) donde no se conoce el número total de empleados para toda la población, por tanto para estimar el total consideraremos un muestreo aleatorio simple tomando como elementos muestrales las empresas.

y_i	$\left(y_i - \bar{y}_t\right)^2$
1	2,25
2	0,25
9	42,25
0	6,25
2	0,25
0	6,25
1	2,25
0	6,25
4	2,25
6	12,25
25	80,5

$$\bar{y}_t = \frac{25}{10} = 2,5 \; bajas/empresa \qquad \boxed{\hat{\tau}_t = 85 \times 2,5 = 212,5 \; bajas}$$

$$S_t^2 = \frac{80,5}{9} = 8,9\hat{4} \quad \Rightarrow \quad \hat{V}(\bar{y}_t) = \left(\frac{85-10}{85}\right)\frac{8,9\hat{4}}{10} = 0,7892157 \quad \Rightarrow \quad \hat{V}(\hat{\tau}_t) = 85^2 \, \hat{V}(\bar{y}_t) = 5702,08$$

$$\boxed{B_\tau = 2\sqrt{5702.08} = 151,02 \; bajas}$$

b)

$$\boxed{\hat{p} = \frac{4}{10} = 0,40 \quad (40\%)}$$

$$\hat{V}(\hat{p}) = \frac{85-10}{85}\frac{0,4 \times 0,6}{10-1} = 0,02353 \qquad \boxed{B = 2\sqrt{0,02353} = 0,3068 \quad (30,68\%)}$$

5. Se diseña una encuesta económica para estimar la cantidad media gastada en servicios por hogar de una ciudad formada por 3600 hogares. Se selecciona una muestra aleatoria de 3 barrios de la ciudad de un total de 60. Los entrevistadores obtienen el gasto en servicios de cada hogar en los barrios seleccionados; los gastos totales se muestran en esta tabla:

Barrio	*Nº hogares*	*Cantidad total gastada en servicios (€)*
1	55	2210
2	60	2390
3	63	2430

Estime la cantidad media de gastos en servicios por hogar en la ciudad y el límite para el error de estimación.

SOLUCIÓN:

$m_i y_i$
121550 143400 153090
$\sum_{i=1}^{n} m_i y_i = 418040$

$$N = 60 \qquad n = 3 \qquad \bar{M} = \frac{3600}{60} = 60$$

$$\sum_{i=1}^{n} m_i = 178 \qquad \sum_{i=1}^{n} m_i^2 = 10594 \qquad \sum_{i=1}^{n} y_i = 7030 \qquad \sum_{i=1}^{n} y_i^2 = 16501100$$

$$\bar{y} = \hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = 39,49 \text{ €}$$

$$S_c^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \bar{y}^2\sum_{i=1}^{n} m_i^2 - 2\bar{y}\sum_{i=1}^{n} m_i y_i\right) = 2612,04$$

$$\hat{V}(\bar{y}) = \frac{1}{\bar{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 0,23 \qquad 2\sqrt{\hat{V}(\bar{y})} = 0,96 \text{ €}$$

6. En un proceso de control del volumen envasado por una fábrica de bebidas se eligen 3 de los 40 paquetes envasados en una hora, cada uno de los cuales contiene 4 envases, y se mide el volumen que cada envase contiene. Las observaciones se presentan en la tabla adjunta:

Paquete nº	*Volumen envasado en cl*			
1	33,5	32,5	31	34
2	32,5	32	33	32,5
3	30,5	33	33	33,5

Estime el volumen medio de los envases y la cota del error de estimación.

SOLUCIÓN:

N=40, n=3,

m_i	y_i	$m_i y_i$
4	131	524
4	130	520
4	130	520
		$\sum_{i=1}^{3} m_i y_i = 1564$

(*con las funciones del modo SD de la calculadora*):

$$\overline{M} = \overline{m} = 4 \qquad \sum_{i=1}^{3} m_i = 12 \qquad \sum_{i=1}^{3} m_i^2 = 48$$

$$\overline{y}_t = 130{,}33 \qquad \sum_{i=1}^{3} y_i = 391 \qquad \sum_{i=1}^{3} y_i^2 = 50961$$

$$\overline{y} = \hat{\mu} = \frac{\sum_{i=1}^{3} y_i}{\sum_{i=1}^{3} m_i} = \frac{\overline{y}_t}{\overline{m}} = 32{,}5833\ cl$$

$$S_c^{\,2} = \frac{1}{n-1}\sum_{i=1}^{3}\left(y_i - \overline{y} m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{3} y_i^2 + \overline{y}^2 \sum_{i=1}^{3} m_i^2 - 2\overline{y}\sum_{i=1}^{3} m_i y_i\right) = 0{,}3333$$

$$\hat{V}(\overline{y}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 0{,}006423 \qquad 2\sqrt{\hat{V}(\overline{y})} = 0{,}1603\ cl$$

7. Un fabricante de sierras quiere estimar el coste medio de reparación mensual para las sierras que ha vendido a ciertas industrias. El fabricante no puede obtener un coste de reparación para cada sierra, pero puede obtener la cantidad total gastada en reparación y el número de sierras que tiene cada industria. Entonces decide usar muestreo por conglomerados, con cada industria como un conglomerado. El fabricante selecciona una muestra aleatoria simple de 5 de 100 industrias a las que da servicio. Los datos sobre coste total de reparaciones por industria y el número de sierras son:

Industria	*Nº sierras*	*Costo total de reparación para el mes pasado (€)*
1	3	50
2	7	110
3	11	230
4	9	140
5	2	60

Estime el coste medio de reparación por sierra para el mes pasado y el límite para el error de estimación.

SOLUCIÓN: N=100 n=5

$m_i y_i$
150
770
2530
1260
120
$\sum_{i=1}^{n} m_i y_i = 4830$

(*con las funciones del modo SD de la calculadora*):

$$\widehat{M} = \overline{m} = 6,4 \qquad \sum_{i=1}^{n} m_i = 32 \qquad \sum_{i=1}^{n} m_i^2 = 264$$

$$\overline{y}_t = 118 \qquad \sum_{i=1}^{n} y_i = 590 \qquad \sum_{i=1}^{n} y_i^2 = 90700$$

$$\overline{y} = \widehat{\mu} = \frac{\sum_{i=1}^{5} y_i}{\sum_{i=1}^{5} m_i} = \frac{\overline{y}_t}{\overline{m}} = 18,4375\ €$$

$$S_c^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \overline{y}^2\sum_{i=1}^{n} m_i^2 - 2\overline{y}\sum_{i=1}^{n} m_i y_i\right) = 584,57$$

$$\widehat{V}(\overline{y}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 2,7116 \qquad 2\sqrt{\widehat{V}(\overline{y})} = 3,2934\ €$$

8. Un periódico quiere estimar la proporción de votantes que apoyan a cierto candidato A. Ya que la selección y entrevista de una muestra aleatoria simple de votantes registrados es muy costosa, se utiliza muestreo por conglomerados, con distritos como conglomerados. Se selecciona una muestra aleatoria de 5 distritos de un total de 495. El periódico quiere hacer la estimación el día de la elección, pero antes de que se haya hecho el recuento final de los votos. Los reporteros son enviados a los lugares de votación de cada distrito en la muestra, para obtener la información pertinente directamente de los votantes. Los resultados se muestran en la tabla:

Nº votantes	*Nº votantes que apoyan A*
1290	680
1170	631
840	475
1620	935
1381	472

Estime la proporción de votantes que apoyan al candidato A y el límite para el error de estimación.

SOLUCIÓN:

N=495 n=5

$m_i y_i$
877200
738270
399000
1514700
651832
$\sum_{i=1}^{n} m_i y_i = 4181002$

(*con las funciones del modo SD de la calculadora*):

$\widehat{\overline{M}} = \overline{m} = 1260,2$ $\quad \sum_{i=1}^{n} m_i = 6301$ $\quad \sum_{i=1}^{n} m_i^2 = 8270161$

$\overline{y}_t = 638,6$ $\quad \sum_{i=1}^{n} y_i = 3193$ $\quad \sum_{i=1}^{n} y_i^2 = 2183195$

$$\hat{p} = \hat{\mu} = \frac{\sum_{i=1}^{5} y_i}{\sum_{i=1}^{5} m_i} = \frac{\overline{y}_t}{\overline{m}} = 0,506745 \qquad (50,67\%)$$

$$S_c^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{p} m_i\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \hat{p}^{\,2}\sum_{i=1}^{n} m_i^2 - 2\hat{p}\sum_{i=1}^{n} m_i y_i\right) = 17372,505$$

$$\hat{V}(\hat{p}) = \frac{1}{\overline{M}^2}\frac{N-n}{N}\frac{S_c^2}{n} = 0,00216573 \qquad 2\sqrt{\hat{V}(\overline{y})} = 0,0930748 \quad (9,31\%)$$

6. Estimación del tamaño de la población.

6.1 Estimación del tamaño de la población usando muestreo directo

En el muestreo directo se realizan los siguientes pasos:

1. Se selecciona una muestra aleatoria de tamaño t, se marcan y se devuelven a la población.
2. Posteriormente se selecciona una muestra aleatoria de tamaño n (tamaño fijado de antemano) de la misma población y se observa cuántos de ellos están marcados (s =*número de elementos marcados en esta 2ª muestra*)

Sea p = *proporción de elementos marcados en la población*, $p = \frac{t}{N}$, $N = \frac{t}{p}$, pero p es desconocido. Entonces estimamos p mediante la proporción muestral:

$$\hat{p} = \frac{s}{n} = \textit{proporción de elementos marcados en la } 2^a \textit{ muestra}$$

Por tanto,

- ESTIMADOR DE N : $\hat{N} = \frac{t}{\hat{p}} = \frac{t}{s/n} = \frac{nt}{s}$ $\begin{pmatrix} n, t = \text{constantes} \\ s = \text{aleatoria} \end{pmatrix}$
- VARIANZA ESTIMADA DE $\hat{N}$: $\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^3}$

Comentarios

- s = *número de elementos marcados en la 2ª muestra* ha de ser mayor que 0 para que las fórmulas estén bien definidas. Si en la segunda muestra no aparece ningún elemento marcado, se aumenta el tamaño muestral.
- $\hat{N}$ no es un estimador insesgado de N :

$$E\left[\hat{N}\right] = N + N\frac{(N-t)}{nt} \neq N$$

  Cuanto mayor sean n y t menor será el sesgo $N\frac{(N-t)}{nt}$.

- $\hat{N}$ tiende a sobreestimar el valor real de N .

Ejemplo 6.1

Un club deportivo se interesa por el número de truchas de río en un arroyo. Durante un periodo de varios días se atrapan 100 truchas, se marcan y se devuelven al arroyo. Obsérvese que la muestra representa 100 peces diferentes, ya que cualquier pez atrapado que ya hubiera sido marcado se devolvía inmediatamente. Varias semanas después se atrapó una muestra de 120 peces y se observó el número de peces marcados. Supongamos que este número fue de 27 en la segunda muestra. Estime el tamaño total de la población de truchas y dé un límite de error de estimación.

Solución

$$\hat{N} = \frac{nt}{s} = \frac{120 \times 100}{27} = 444,4$$

$$\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^3} = \frac{100^2 \times 120(120-27)}{27^3} = 5669,87$$

$$B = 2\sqrt{\hat{V}\left(\hat{N}\right)} = 150,60$$ ■

6.2 Estimación del tamaño de la población usando muestreo inverso

La diferencia con el muestreo directo es que aquí el tamaño de la segunda muestra no está fijado (es aleatorio), lo que se fija es $s =$ número de elementos marcados en la segunda muestra.

Los pasos para realizar este método son:

1. Se selecciona una muestra inicial de t elementos, se marcan y se devuelven a la población.
2. Se selecciona una segunda muestra aleatoria hasta que se obtienen s elementos marcados (sea n el tamaño final de dicha muestra).

- ESTIMADOR DE N: $\hat{N} = \frac{t}{\hat{p}} = \frac{t}{s/n} = \frac{nt}{s}$ $\quad \begin{pmatrix} t, s = \text{ constantes}, & s \neq 0 \\ n = \text{ aleatoria} \end{pmatrix}$
- VARIANZA ESTIMADA DE $\hat{N}$: $\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^2(s+1)}$

Comentario. $\hat{N}$ es un estimador insesgado de N, por ello, si se pueden aplicar ambos tipos de muestreo se prefiere el inverso.

Ejemplo 6.2

Una zoóloga desea estimar el tamaño de la población de tortugas en determinada área geográfica. Ella cree que el tamaño de la población está entre 500 y 1000; por lo que una muestra inicial de 100 parece ser suficiente. Las 100 tortugas son capturadas, marcadas y liberadas. Toma una segunda muestra un mes después y decide continuar muestreando hasta que se recapturen 15 tortugas marcadas. Atrapa 160 tortugas para obtener las 15 marcadas. Estime el tamaño total de la población de tortugas y establezca un límite de error de estimación.

Solución

$$\hat{N} = \frac{nt}{s} = \frac{160 \times 100}{15} = 1066,67$$

$$\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^2(s+1)} = \frac{100^2 \times 160(160-15)}{15^2(15+1)} = 64444,44$$

$$B = 2\sqrt{\hat{V}\left(\hat{N}\right)} = 507,72$$ ■

6.3.1 Estimación de la densidad y del tamaño de la población usando muestreo por cuadros

Con este método se estudia el tamaño de la población contenida en un área delimitada A conocida. Los pasos a seguir son:

1. Dividir a la población en N cuadros de igual área a. Sea

   m_i = número de elementos en el cuadro i-ésimo

2. Tomar una muestra de n cuadros entre los N existentes. Se observa el número total de elementos que contiene la muestra:

   $$m = \sum_{i=1}^{n} m_i$$

3. Calcular la densidad de elementos en la muestra (densidad muestral):

   $$\hat{\lambda} = \frac{\text{nº elementos en la muestra}}{\text{área de la muestra}} = \frac{m}{na}$$

4. La densidad poblacional es

   $$\lambda = \frac{\text{nº elementos en la población}}{\text{área de la población}} = \frac{M}{Na} = \frac{M}{A}$$

   entonces $M = A\lambda$. Por tanto:

- ESTIMADOR DE LA DENSIDAD: $\hat{\lambda} = \frac{m}{na}$
- VARIANZA ESTIMADA DE $\hat{\lambda}$: $\hat{V}(\hat{\lambda}) = \frac{m}{a^2 n^2} = \hat{\lambda}\frac{1}{na}$
- ESTIMADOR DEL TAMAÑO POBLACIONAL: $\hat{M} = A\hat{\lambda} = A\frac{m}{na}$
- VARIANZA ESTIMADA DE $\hat{M}$: $\hat{V}(\hat{M}) = A^2\hat{V}(\hat{\lambda}) = \frac{A^2 m}{a^2 n^2}$

Ejemplo 6.3

La policía de Madrid está interesada en conocer el número de aficionados que se reunieron en torno a la fuente de Neptuno para celebrar el triunfo de su equipo. Con este dato se puede conocer la cuantía de medios materiales y humanos (policía, protección civil, personal sanitario, etc.) necesaria para atender futuras concentraciones. Para estimar el número de aficionados se toma una fotografía aérea de la zona ocupada por éstos, tras lo cual se traza sobre ella una cuadrícula que divide el área total en 300 cuadros de 10 metros de lado cada uno. Posteriormente se numeran y se extrae una muestra aleatoria de 20 de estos cuadros; por último se cuenta el número de aficionados que hay en cada uno de los cuadros seleccionados, obteniéndose los resultados de la tabla:

Nº del cuadro	Número de aficionados en el cuadro	Nº del cuadro	Número de aficionados en el cuadro
1	193	11	160
2	216	12	220
3	250	13	163
4	163	14	306
5	209	15	319
6	195	16	289
7	232	17	205
8	174	18	210
9	215	19	209
10	198	20	198

a) Estime la densidad de aficionados por metro cuadrado y obtenga su intervalo de confianza.

b) Estime el número total de aficionados concentrados en la plaza de Neptuno y obtenga su intervalo de confianza.

Solución:

a) $a = 10 \times 10 = 100 \qquad m = 193 + 216 + ... + 198 = 4324 \qquad \hat{\lambda} = \dfrac{m}{na} = \dfrac{4324}{20 \times 100} = 2{,}162$

$\hat{V}(\hat{\lambda}) = \dfrac{\hat{\lambda}}{na} = \dfrac{2{,}162}{2000} = 0{,}001081 \quad \Rightarrow \quad B = 2\sqrt{0{,}001081} = 0{,}065757$

$\hat{\lambda} = 2{,}162 \; aficionados / m^2 \qquad (2{,}09624 \;\; , \;\; 2{,}22776)$

b) $A = 300 \times 100 = 30000 \; m^2$

$\hat{M} = A\hat{\lambda} = 30000 \times 2{,}162 = 64860 \; aficionados \qquad B_M = AB_\lambda = 30000 \times 0{,}065757 = 1972{,}71$

$(62887{,}29 \;\; , \;\; 66832{,}71)$ ■

6.3.2 Muestreo en el espacio temporal

En determinadas ocasiones podemos tomar los cuadros como intervalos temporales. Veámoslo con un ejemplo.

Ejemplo 6.4

Se desea estimar el número total de personas que diariamente solicitan información en una oficina turística. Se observa que 114 personas solicitan información, durante 12 intervalos de 5 minutos cada uno, repartidos aleatoriamente entre las 8 horas que permanece abierta la oficina. Estime el total de personas que visitan la oficina diariamente y calcule la cota del error de estimación.

Solución

$A = 8$ horas= 480 minutos $\qquad n$=12 intervalos $\qquad a = 5$ minutos $\qquad m$=114 personas

$$\hat{\lambda} = \frac{114}{5 \times 12} = 1{,}9 \; personas / minuto \qquad \hat{M} = A\frac{m}{na} = 912 \; personas$$

$\hat{V}(\hat{M}) = \dfrac{A^2 m}{a^2 n^2} = 7296 \Rightarrow B = 170{,}8$ ■

6.3.3 Determinación del tamaño muestral.

Supongamos que para el habitual nivel de confianza (95%, z_c=2) queremos estimar la densidad poblacional λ con un límite para el error de estimación B

$$B = 2\sqrt{\frac{\lambda}{na}}$$

¿Cuál debe ser el número de cuadros en la muestra n? Es fácil despejar n en la anterior igualdad

$$B^2 = 4\frac{\lambda}{na} \quad \Leftrightarrow \quad n = \frac{4\lambda}{aB^2}$$

Donde el valor λ debe estimarse con una muestra o estudios previos.

Análogamente si se desea estimar el tamaño de la población M con un límite para el error de estimación *B*.

$$B = 2\sqrt{A^2 \frac{\lambda}{na}} = 2A\sqrt{\frac{\lambda}{na}}$$

¿Cuál debe ser el número de cuadros en la muestra *n*?

$$B^2 = 4A^2 \frac{\lambda}{na} \quad \Leftrightarrow \quad n = \frac{4A^2\lambda}{aB^2}$$

Donde, como en el caso anterior, el valor λ debe estimarse con una muestra o estudios previos.

Resumiendo:

$$n = \frac{\lambda}{aD}$$

siendo $D = \frac{B^2}{4}$ cuando estimamos λ y $D = \frac{B^2}{4A^2}$ cuando estimamos *M*.

Ejemplo 6.5

Se quiere conocer el número de peatones que pasan por la Carrera del Darro a lo largo de un mes. Para ello, en un mes de 30 días, se seleccionaron al azar 6 días, observándose en esas 144 horas que 31032 peatones pasaron por la calle.

a) Estime con un intervalo de confianza el ***nº de peatones*** que pasaron por dicha calle ese mes.

b) Si se quisiera repetir en otro mes de 30 días el mismo estudio, garantizando para la anterior estimación un error inferior a 1000 *peatones*, ¿cuantos días deberíamos observar?

Solución

a)

$n = 6\ días \quad a = 1\ día = 24\ horas \quad A = 30 \times 24 = 720\ horas \quad m = \sum_{i=1}^{n} m_i = 31032 \quad \overline{m} = \frac{m}{n} = 5172$

$\hat{\lambda} = \frac{m}{na} = \frac{\overline{m}}{a} = \frac{5172}{24} = 215{,}5\ peatones/hora \qquad \widehat{M} = \hat{\lambda}A = 215{,}5 \times 720 = 155160\ peatones\ al\ mes$

$\hat{V}\left(\widehat{M}\right) = \frac{A^2\hat{\lambda}}{an} = \frac{720^2 \times 215{,}5}{24 \times 6} = 775800 \qquad 2\sqrt{\hat{V}\left(\widehat{M}\right)} = 1761{,}59$

$\left(155160 \mp 1761{,}59\right) = \left(153398{,}41 \quad , \quad 156921{,}59\right)$

b)

$$2\sqrt{\hat{V}\left(\widehat{M}\right)} = 1000 \quad \Rightarrow \quad \hat{V}\left(\widehat{M}\right) = \frac{1000000}{4} = 250000 \quad \Rightarrow \quad \frac{A^2\hat{\lambda}}{an} = \frac{720^2 \times 215{,}5}{24 \times n} = 250000$$

$$\Rightarrow \quad \frac{720^2 \times 215{,}5}{24 \times 250000} = n = 18{,}62 \; días$$

6.3.4 Cuadros cargados

En este tipo de muestreo también se divide a la población en cuadros, pero el método se utiliza cuando después de hecha la división son muchos los cuadros que no contienen elementos y otros contienen pocos, es decir, la densidad de elementos por unidad de superficie es muy pequeña.

Este tipo de muestreo se basa en la identificación de la presencia o ausencia de elementos en cada uno de los cuadros de la muestra. Un cuadro se dice *cargado* cuando contiene al menos un elemento objeto de estudio.

Los pasos a seguir son:

1. Se divide a la población en N cuadros de igual área a.
2. Se toma una muestra de n cuadros entre los N existentes. Se observa el número total de cuadros no cargados de la muestra, a este número de cuadros sin presencia de elementos se le designa por y. Es importante tener en cuenta que y no puede ser cero ni n ($0 < y < n$). Si una vez observada la muestra $y = 0$ ó $y = n$, ampliaremos el tamaño muestral
3. La **densidad poblacional** se **estima** como

$$\hat{\lambda} = -\frac{1}{\mathrm{a}} \ln\left(\frac{y}{n}\right)$$

y su **varianza** como

$$\hat{V}(\hat{\lambda}) = \frac{1}{a^2}\frac{n-y}{ny}$$

Dado que $M = A\lambda$ obtenemos

- ESTIMADOR DEL TAMAÑO POBLACIONAL: $\hat{M} = A\hat{\lambda} = -\frac{A}{a}\ln\left(\frac{y}{n}\right)$
- VARIANZA ESTIMADA DE $\hat{M}$: $\hat{V}(\hat{M}) = \frac{A^2}{a^2}\frac{n-y}{ny}$

Ejemplo 6.6

Se desea estimar el número total de autobuses que, entre las 6 y las 24 horas del domingo, circulan por un determinado punto kilométrico de una carretera. La observación se realiza mediante 40 intervalos, de 10 minutos cada uno, repartidos a lo largo del periodo en estudio. En 18 ocasiones, de las cuarenta que se estableció el control, no circuló por el punto en cuestión ningún autobús. Estimar el número total de autobuses que circularon entre las 6 y las 24 horas. Dar un límite de error de estimación.

Solución

$A = 24\text{-}6 = 18$ horas$=1080$ minutos $\quad n = 40$ intervalos $\quad a = 10$ minutos

$y = 18$ intervalos sin autobuses $\quad \hat{M} = -\frac{A}{a}\ln\left(\frac{y}{n}\right) = -\frac{1080}{10}\ln\left(\frac{18}{40}\right) = 86,24$

$$\hat{V}(\widehat{M}) = \frac{A^2}{a^2}\frac{n-y}{ny} = \frac{1080^2}{10^2}\frac{40-18}{40\times 18} = 356,4 \quad \Rightarrow \quad B = 37,8 \qquad \blacksquare$$

EJERCICIOS RESUELTOS

1. En una plantación de pinos de 200 acres, se va a estimar la densidad de árboles que presentan hongos parásitos. Se toma una muestra de 10 cuadros de 0,5 acres cada uno. Las diez parcelas muestreadas tuvieron una media de 2,8 árboles infectados por cuadro.
    a) Estime la densidad de árboles infectados y establezca un límite de error de estimación.
    b) Estime el total de árboles infectados en los 200 acres de la plantación y establezca un límite de error de estimación.

    SOLUCIÓN:

    a) $\hat{\lambda} = \frac{m}{na} = \frac{2,8\times 10}{10\times 0,5} = 5,6 \; arb.infectados/acre$;

    $$\hat{V}(\hat{\lambda}) = \hat{\lambda}\frac{1}{na} = 5,6\frac{1}{10\times 0,5} = 1,12 \Rightarrow B = 2,1$$

    b) $\hat{M} = A\hat{\lambda} = 200\times 5,6 = 1120; \quad B = AB_{\lambda} = 200\times 2,1 = 423,32$

2. Se desea estimar el número de vehículos de un modelo determinado que el mes próximo utilizarán el aparcamiento de Puerta Real. Durante las 720 horas del mes se van a

establecer 5 controles aleatorios de 1 hora de duración cada uno. Transcurrido el mes, se ha observado en los 5 controles los siguientes resultados:

Control	Número de vehículos de ese modelo que usan el aparcamiento
1	0
2	1
3	2
4	0
5	3

Estime el número total de vehículos del modelo en estudio que utilizaron el aparcamiento. Dé el límite del error de estimación.

SOLUCIÓN:

$A = 720\,h \quad a = 1\,h \quad n = 5\,contr. \quad m = 0+1+2+0+3 = 6\,veh. \quad \overline{m} = \frac{6}{5} = 1,2 \quad \hat{\lambda} = \frac{\overline{m}}{a} = 1,2\,veh./h$

$$\boxed{\widehat{M} = \hat{\lambda} A = 1,2 \times 720 = 864\,veh.}$$

$$\hat{V}\left(\widehat{M}\right) = \frac{A^2 \hat{\lambda}}{an} = 124416 \qquad \boxed{B = 2\sqrt{124416} = 705,45\,veh.}$$

3. El hermano de un alumno de T.A.M. está pensando en abrir una farmacia de 24 horas. Para saber si los ingresos compensarían los gastos de esta inversión deciden observar un establecimiento similar. Este asiduo alumno de T.A.M. conoce perfectamente que es una pérdida de tiempo innecesaria observar el flujo de clientes las 24 horas del día por lo que decide observar la afluencia de clientes en distintos periodos de igual duración, obteniendo los datos de la siguiente tabla

	clientes
10:00-10:30	15
14:00-14:30	13
18:00-18:30	18
22:00-22:30	8
02:00-02:30	2
06:00-06:30	4

Estime el número de clientes diarios de la farmacia observada y el correspondiente límite para el error de estimación.

SOLUCIÓN:

$A = 24h \quad a = 0,5h \quad N = 48 \quad n = 6 \quad m = 60 \quad \overline{m} = 10$

$\widehat{M} = \hat{\lambda} A = \frac{\overline{m}}{a} A = 480\,clientes \quad \hat{V}\left(\widehat{M}\right) = \frac{A^2 \hat{\lambda}}{an} = \frac{A^2 \overline{m}}{a^2 n} = 3840 \qquad 2\sqrt{\hat{V}\left(\widehat{M}\right)} = 123,94\,clientes$

4. El ayuntamiento de Barcelona está interesado en conocer el número de aficionados que acudieron al aeropuerto para vitorear al equipo campeón. Para ello, dividieron la sala de espera, de dimensiones 100 metros de largo por 40 metros de ancho, en 100 cuadros de igual tamaño y seleccionaron 20, observando que el número de personas era 1.100.
   Estime el número total de asistentes y el límite para el error de estimación.
   SOLUCIÓN:

$$A = 4000 \quad a = 40 \quad N = 100 \quad n = 20 \quad m = 1100 \quad \overline{m} = 55$$

$$\widehat{M} = \hat{\lambda} A = \frac{\overline{m}}{a} A = 5500 \qquad \widehat{V}\left(\widehat{M}\right) = \frac{A^2 \hat{\lambda}}{an} = \frac{A^2 \overline{m}}{a^2 n} = 27500 \qquad 2\sqrt{\widehat{V}\left(\widehat{M}\right)} = 331,66$$

5. Un alumno de A.T.C. desea estimar el número de alumnos que una determinada mañana han ido a la Facultad. Para ello se basa en que dicho día una conocida marca comercial ha repartido a primeras horas de la mañana en la entrada de la Facultad 500 carpetas. En un intercambio de clase, sentado en un banco del pasillo, decide contar los alumnos que pasan hasta observar a 100 que portan la carpeta, para lo que fue necesario contar hasta 382 alumnos.
   Estime con un intervalo de confianza el número de alumnos que asistieron esa mañana a la Facultad.
   SOLUCIÓN: muestreo inverso

$$t = 500 \quad n = 382 \quad s = 100$$

$$\widehat{N} = \frac{t}{\widehat{p}} = \frac{nt}{s} = 1910 \; alumnos$$

$$\widehat{V}\left(\widehat{N}\right) = \frac{t^2 n(n-s)}{s^2(s+1)} = 26664,35643 \qquad 2\sqrt{\widehat{V}\left(\widehat{N}\right)} = 326,58 \; alumnos$$

$$\left(1910 \mp 326,58\right)$$

6. Se quiere conocer el número de enfermos que utilizan el Servicio de Urgencias de un hospital. Para ello, de un mes (30 días) se seleccionaron al azar 3 días, observándose en esas 72 horas que 4320 personas usaron el servicio.
   a) Estime con un intervalo de confianza el ***nº de enfermos/hora*** que acudieron al servicio de urgencias.
   b) Si se quisiera repetir en el próximo mes el mismo estudio, garantizando para la anterior estimación un error inferior a un *enfermo/hora*, ¿cuantos días deberíamos observar?

SOLUCIÓN:

$n = 3 \qquad a = 1\ dia = 24\ horas \qquad m = \sum_{i=1}^{n} m_i = 4320 \qquad \overline{m} = \frac{m}{n} = 1440$

a) $\hat{\lambda} = \frac{m}{na} = \frac{\overline{m}}{a} = \frac{1440}{24} = 60\ enfermos / hora$

$\hat{V}\left(\hat{\lambda}\right) = \frac{\hat{\lambda}}{an} = \frac{60}{24 \times 3} = 0,8333 \qquad 2\sqrt{\hat{V}\left(\hat{\lambda}\right)} = 1,8257$

$$\left(60 \mp 1,8257\right) = \left(58,1743 \ , \ 61,8257\right)$$

b) $2\sqrt{\hat{V}\left(\hat{\lambda}\right)} = 1 \ \Rightarrow \ 4\hat{V}\left(\hat{\lambda}\right) = 1 \ \Rightarrow \ 4\frac{\hat{\lambda}}{an} = 4\frac{60}{24 \times n} = 1 \ \Rightarrow \ n = \frac{4 \times 60}{24} = 10$

7. Muestreo con probabilidades desiguales.

7.1.1 Probabilidades de inclusión.

En el capítulo 1.3 se ha abordado el problema de seleccionar una muestra aleatoria simple (sin reemplazamiento) en una población finita. Al método utilizado para la selección de una muestra aleatoria de individuos se le conoce como diseño muestral. El muestreo aleatorio simple se caracteriza por el hecho de que todas las muestras posibles con el mismo tamaño n tienen la misma probabilidad de ser seleccionadas. A la probabilidad de seleccionar una determinada muestra s la denotaremos por $p(s)$.

Adicionalmente, cabe destacar que también los N individuos de la población finita tienen la misma probabilidad de ser seleccionados en una muestra de tamaño n cuando se aplica un muestreo aleatorio simple sin reemplazamiento y, en particular, esta probabilidad es n/N. Sin embargo, en la práctica es muy frecuente que los organismos oficiales de estadística, empresas, instituciones y otros centros que realizan encuestas por muestreo hagan uso de diseños muestrales donde las probabilidades de selección de muestras y/o individuos sean desiguales. **En esta situación, los estimadores estudiados en el capítulo 1 no son válidos y podrían producir importantes errores**. **Los estimadores de parámetros** como totales, medias o proporciones, e incluso los estimadores de sus correspondientes varianzas **deben incluir en sus expresiones las probabilidades que tienen los individuos de ser seleccionados en la muestra,** para ser más precisos.

La probabilidad de que el *i*-ésimo individuo esté incluido en una muestra se denotará por π_i, y su definición es la siguiente:

$$\pi_i = \sum_{s \ni i} p(s)$$

A estas probabilidades también se les denominan probabilidades de inclusión de primer orden. Por su parte, también pueden definirse las probabilidades de inclusión de segundo orden, o

bien la probabilidad de que ambos individuos *i* y *j* estén incluidos simultáneamente en la muestra. Esta probabilidad viene dada por:

$$\pi_{ij} = \sum_{s \ni i \& j} p(s)$$

En un muestreo aleatorio simple sin reemplazamiento (capítulo 1.3), las probabilidades de inclusión de primer y segundo orden vienen dadas, respectivamente, por:

$$\pi_i = \frac{n}{N}$$

$$\pi_{ij} = \frac{n}{N}\frac{n-1}{N-1}$$

Por tanto, dado un muestreo aleatorio simple sin reemplazamiento es posible utilizar tanto las expresiones que estudiaremos a continuación como las expresiones ya estudiadas del capítulo 1.3, y obtendremos en ambos casos los mismos resultados.

7.1.2 Pesos del diseño muestral (o factores de elevación).

En general, los organismos oficiales de estadística, tal como EUROSTAT, suelen llevar a cabo para sus estudios muestras basadas en diseños muestrales con probabilidades desiguales, y sus correspondientes bases de datos públicas incorporan, para cada individuo de la muestra, los llamados pesos muestrales, pesos del diseño muestral o factores del diseño, en lugar de las probabilidades de inclusión ya explicadas.

Los pesos muestrales se definen simplemente como la inversa de las probabilidades de inclusión de primer orden, es decir, el peso muestral del *i*-ésimo individuo se define como

$$d_i = \frac{1}{\pi_i}$$

Por tanto, las expresiones de los estimadores en diseños desiguales se pueden expresar tanto en términos de las probabilidades de inclusión como en función de los pesos muestrales. Si se desean utilizar las probabilidades de inclusión de primer orden pero sólo se conocen los pesos muestrales, es bastante fácil comprobar a partir de la expresión anterior que tales probabilidades de inclusión de primer orden se pueden obtener como la inversa de los pesos muestrales:

$$\pi_i = \frac{1}{d_i}$$

Destacamos que es bastante común por las empresas y agencias estadísticas utilizar diseños muestrales donde los individuos tienen distinta probabilidad de ser seleccionados, y este hecho se debe a varias razones. En primer lugar, las muestras con un gran tamaño utilizan diseños muestrales complejos donde combinan, por ejemplo, muestreo por conglomerados en varias etapas con muestreo estratificado. En estas situaciones resulta muy conveniente y simple proporcionar los pesos muestrales asociados a cada individuo, de forma que cualquier investigador o usuario interesado en los datos pueda realizar estimaciones correctas y de una forma muy simple basadas en los pesos muestrales.

Por otra parte, también existen situaciones en las que interesa asignar a ciertos individuos una mayor probabilidad de ser seleccionados en comparación con otros individuos. Por ejemplo, los conocidos diseños muestrales basados en Probabilidades Proporcionales al Tamaño (PPT) y que estudiaremos a continuación hacen uso de esta metodología.

Encuestas destinadas a empresas suelen utilizar con bastante frecuencia muestras extraídas mediante PPT debido a la importancia de incluir en la muestra a las empresas más grandes, las cuales aportan una mayor cantidad de producción de bienes o servicios.

7.1.3 Algunos métodos con probabilidades desiguales.

En el muestreo aleatorio simple y en el muestro sistemático los individuos tienen la misma probabilidad de ser seleccionados, es decir, las probabilidades de inclusión de primer orden son iguales para todos los individuos. A continuación se describen algunos métodos de muestreo donde los individuos tienen distintas probabilidades de ser seleccionados, y se proporcionan las correspondientes probabilidades de inclusión.

Métodos con probabilidades proporcionales al tamaño.

En primer lugar destacamos los diseños muestrales con probabilidades proporcionales al tamaño (PPT), donde los individuos de la población tienen probabilidades de ser seleccionados proporcionales a los valores $\{x_1,\ldots,x_N\}$ de una variable auxiliar *X*, y a la cual también se le suele denominar "tamaño". En estos diseños se asume que los valores $\{x_1,\ldots,x_N\}$ son conocidos y positivos. Las probabilidades de inclusión de primer orden de un diseño muestral PPT vienen dados por:

$$\pi_i = n\frac{x_i}{\tau_x} \quad , \quad \text{con} \quad \tau_x = \sum_{i=1}^{N} x_i$$

Algunos métodos muy conocidos basados en PPT son: método de Lahiri, método de Brewer, método de Midzuno, método de Madow, método de Sampford,…

Muestreo aleatorio estratificado

En el muestreo aleatorio estratificado, salvo que se aplique asignación proporcional, los individuos de un estrato tendrán una probabilidad de ser seleccionados distinta de la probabilidad de ser seleccionados los individuos del resto de estratos de la población, y por tanto, el muestreo aleatorio estratificado también puede considerarse como un diseño muestral con probabilidades desiguales. En un muestreo aleatorio estratificado, las probabilidades de inclusión vienen dadas por:

$\pi_i = \dfrac{n_h}{N_h}$ si el individuo *i* pertenece al estrato *h*.

$$\pi_{ij} = \begin{cases} \dfrac{n_h}{N_h}\dfrac{n_h - 1}{N_h - 1} & \text{si ambos individuos } i \text{ y } j \text{ pertenecen al estrato } h. \\ \dfrac{n_h}{N_h}\dfrac{n_k}{N_k} & \text{si el individuo } i \text{ pertenece al estrato } h, \text{ y el individuo } j \text{ al estrato } k \end{cases}$$

Por tanto, en un muestreo estratificado se pueden aplicar tanto las expresiones estudiadas en el capítulo 2 como las expresiones que se estudian a continuación basadas en las probabilidades de inclusión.

7.2 Estimación de la media, proporción y total poblacionales.

Estimación de la media poblacional.

Dada una muestra seleccionada con probabilidades desiguales, en la práctica se pueden utilizar dos estimadores diferentes para la media poblacional (μ_y) de una variable de interés *y*: el estimador de tipo Horvitz-Thompson y el estimador de tipo Hájek.

El ***estimador de tipo Horvitz-Thompson*** para la media poblacional viene dado por la expresión:

$$\overline{y}_{HT} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i}$$

Este estimador se puede expresar fácilmente en función de los pesos muestrales:

$$\overline{y}_{HT} = \frac{1}{N}\sum_{i=1}^{n} d_i y_i$$

El estimador de tipo Horvitz-Thompson tiene la propiedad de ser insesgado.

Para estimar la varianza de este estimador podemos utilizar las siguientes expresiones (complicadas por apoyarse en las probabilidades de inclusión de segundo orden):

1. Estimador de la varianza de tipo Horvitz-Thompson:

$$\hat{V}_{HT}\left(\bar{y}_{HT}\right)=\frac{1}{N^2}\sum_{i=1}^{n}\left(1-\pi_i\right)\frac{y_i^2}{\pi_i^2}+\frac{2}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

2. Estimador de la varianza de tipo Sen-Yates-Grundy:

$$\hat{V}_{SYG}\left(\bar{y}_{HT}\right)=\frac{1}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

El ***estimador de tipo Hájek*** para la media poblacional viene dado por la expresión:

$$\bar{y}_H=\frac{1}{\hat{N}}\sum_{i=1}^{n}\frac{y_i}{\pi_i}=\frac{1}{\hat{N}}\sum_{i=1}^{n}d_i y_i \qquad \text{donde} \quad \hat{N}=\sum_{i=1}^{n}\frac{1}{\pi_i}=\sum_{i=1}^{n}d_i$$

Si comparamos ambos estimadores de la media poblacional podemos destacar que el estimador de tipo Hájek no es insesgado, a diferencia del estimador de tipo Horvitz-Thompson que si lo es. No obstante, el estimador de tipo Hájek es aproximadamente insesgado, lo cual quiere decir que el valor esperado de este estimador se aproxima a la media poblacional para tamaños muestrales elevados. Por otra parte, el estimador de tipo Hájek suele tener una varianza menor que la varianza del estimador de tipo Horvitz-Thompson. Además, podemos observar que si el tamaño de la población N es conocido se podrían obtener ambos estimadores, mientras que si N es desconocido tan sólo sería posible obtener el estimador de tipo Hájek.

Existen varias técnicas para estimar la varianza del estimador de tipo Hájek. Por ejemplo, el estimador de la varianza Quenouille-Tukey (utilizando el método *Jackknife*) viene dado por la siguiente expresión:

$$\hat{V}_J\left(\bar{y}_H\right)=\frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{y}_{H(i)}-\bar{y}_H\right)^2$$

donde $\bar{y}_{H(i)}$ denota el estimador de tipo Hájek pero después de eliminar de la muestra el i-ésimo individuo, es decir,

$$\bar{y}_{H(i)}=\frac{1}{\hat{N}_{(i)}}\sum_{j\in s,j\neq i}\frac{y_j}{\pi_j}=\frac{1}{\hat{N}_{(i)}}\sum_{j\in s,j\neq i}d_j y_j \qquad \text{con} \quad \hat{N}_{(i)}=\sum_{j\in s,j\neq i}\frac{1}{\pi_j}=\sum_{j\in s,j\neq i}d_j$$

Para muestras con tamaños muestrales elevados ($n>30$) también se suele asumir que los estimadores de la media anteriores siguen una distribución Normal, y por tanto el límite del

error de estimación y los intervalos de confianza se definen de la misma forma que en capítulos anteriores.

Estimación de la proporción poblacional.

En el caso de variables dicotómicas, los estimadores de la proporción poblacional p se obtienen siguiendo las mismas expresiones que las proporcionadas para la estimación de la media poblacional. De este modo, los estimadores de tipo Horvitz-Thompson y de tipo Hájek de la proporción poblacional son, respectivamente:

$$\hat{p}_{HT} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{n} d_i y_i \qquad \text{y} \qquad \hat{p}_H = \frac{1}{\hat{N}}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{1}{\hat{N}}\sum_{i=1}^{n} d_i y_i$$

En el caso de que la proporción poblacional p tome valores próximos a 1, se recomienda utilizar el estimador de tipo Hájek, debido a que este estimador toma el valor 1 cuando todas las observaciones y_i de la muestra son unos. Esta propiedad no se cumple para el estimador de tipo Horvitz-Thompson, e incluso se podrían obtener estimaciones mayores de 1 en estas situaciones donde la proporción poblacional se aproxima a la unidad.

Por otra parte, las expresiones de las varianzas estimadas de los estimadores de la media poblacional también pueden utilizarse para las varianzas estimadas de los estimadores de una proporción, y por tanto se tiene:

Estimador de tipo Horvitz-Thompson de la proporción:

1. Estimador de la varianza de tipo Horvitz-Thompson:

$$\hat{V}_{HT}\left(\hat{p}_{HT}\right) = \frac{1}{N^2}\sum_{i=1}^{n}\left(1-\pi_i\right)\frac{y_i^2}{\pi_i^2} + \frac{2}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

2. Estimador de la varianza de tipo Sen-Yates-Grundy:

$$\hat{V}_{SYG}\left(\hat{p}_{HT}\right) = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

Estimador de tipo Hájek de la proporción:

Estimador de la varianza usando el método Jackknife:

$$\hat{V}_J\left(\hat{p}_H\right) = \frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{p}_{H(i)}-\hat{p}_H\right)^2$$

$$\hat{p}_{H(i)} = \frac{1}{\hat{N}_{(i)}}\sum_{j\in s, j\neq i}\frac{y_j}{\pi_j} = \frac{1}{\hat{N}_{(i)}}\sum_{j\in s, j\neq i} d_j y_j \qquad \hat{N}_{(i)} = \sum_{j\in s, j\neq i}\frac{1}{\pi_j} = \sum_{j\in s, j\neq i} d_j$$

Ejemplo 7.1. Según la agencia de estadística EUROSTAT, una persona se considera pobre si sus ingresos por unidad de consumo son inferiores al umbral o línea de pobreza fijado en el país de dicha persona. Supongamos que en una población con 1000 habitantes se selecciona una muestra con probabilidades desiguales de tamaño 10, obteniéndose los siguientes resultados

i	Ingresos	π_i
1	1500	0,020
2	600	0,010
3	2000	0,005
4	450	0,010
5	700	0,020

i	Ingresos	π_i
6	1100	0,010
7	1300	0,020
8	1350	0,005
9	960	0,010
10	1700	0,010

Donde suponemos los π_i conocidos (la agencia EUROSTAT proporciona los pesos del diseño muestral).

Si el umbral de pobreza en esta población lo ha fijado EUROSTAT en 750 euros, obtenga un intervalo de confianza para la proporción de pobres de esta población con un nivel de confianza del 95% y basado en el estimador de tipo Hájek.

Solución: En primer lugar hay que definir la variable dicotómica *y*, la cual tomará el valor 1 si la persona se considera pobre y 0 en caso contrario. En otras palabras, la variable *y* tomará el valor 1 si su ingreso por unidad de consumo es inferior al umbral de pobreza (750 euros) y la variable tomará el valor 0 en caso contrario.

I	y_i	π_i	$\frac{1}{\pi_i}$	$\frac{y_i}{\pi_i}$	$\hat{N}_{(i)} = \hat{N} - \frac{1}{\pi_i}$	$\sum_{j \in s, j \neq i} \frac{y_j}{\pi_j}$	$\hat{p}_{H(i)}$	$\left(\hat{p}_{H(i)} - \hat{p}_H\right)^2$
1	0	0,020	50	0	1000	250	0,2500	0,00014
2	1	0,010	100	100	950	150	0,1579	0,00643
3	0	0,005	200	0	850	250	0,2941	0,00314
4	1	0,010	100	100	950	150	0,1579	0,00643
5	1	0,020	50	50	1000	200	0,2000	0,00145
6	0	0,010	100	0	950	250	0,2632	0,00063
7	0	0,020	50	0	1000	250	0,2500	0,00014
8	0	0,005	200	0	850	250	0,2941	0,00314
9	0	0,010	100	0	950	250	0,2632	0,00063
10	0	0,010	100	0	950	250	0,2632	0,00063
			$\hat{N} = 1050$	**250**				**0,02276**

El estimador de tipo Hájek de la proporción de pobres es:

$$\hat{p}_H = \frac{1}{\hat{N}} \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \frac{250}{1050} = 0,2381 \qquad \text{donde } \hat{N} = \sum_{i=1}^{n} \frac{1}{\pi_i}$$

Se estima que el 23,81% de esta población es considerada como pobre según EUROSTAT.

Para obtener el estimador de la varianza de tipo Hájek necesitaremos incluir en la tabla anterior los resultados de las expresiones:

$$\hat{p}_{H(i)} = \frac{1}{\hat{N}_{(i)}} \sum_{j \in s, j \neq i} \frac{y_j}{\pi_j} \quad ; \quad \hat{N}_{(i)} = \sum_{j \in s, j \neq i} \frac{1}{\pi_j} = \left(\sum_{j=1}^{n} \frac{1}{\pi_j} \right) - \frac{1}{\pi_i} = \hat{N} - \frac{1}{\pi_i},$$

El estimador de la varianza que resulta es:

$$\hat{V}_J\left(\hat{p}_H\right) = \frac{N-n}{N} \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{p}_{H(i)} - \hat{p}_H \right)^2 = \frac{1000-10}{1000} \frac{10-1}{10} 0,02276 = 0,0203$$

El intervalo de confianza viene dado por:

$$\hat{p}_H \mp 2\sqrt{\hat{V}_J\left(\hat{p}_H\right)} = \left(0,2381 - 2 \times 0,1425 \ ; \ 0,2381 + 2 \times 0,1425 \right) = (-0,05 \ ; 0,52) = [0 \ ; 0,52) \ \blacksquare$$

Estimación del total poblacional.

Para estimar el total poblacional, τ, dado que $\mu = \frac{\tau}{N} \Leftrightarrow \tau = N\mu$, el procedimiento es similar al estudiado en capítulos anteriores, es decir, para estimar el total sólo hay que multiplicar por N el estimador de la media. Además, dado que se cumple la propiedad $V(kX) = k^2 V(X)$, para obtener la varianza del estimador del total habría que multiplicar por N^2 el estimador de la varianzas estudiadas en el problema de la estimación de la media. En particular, las expresiones asociadas a la estimación del total poblacional son:

Estimador de tipo Horvitz-Thompson para el total poblacional:

$$\widehat{\tau}_{HT} = N \, \overline{y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i$$

Estimadores de la varianza del estimador de tipo Horvitz-Thompson:

1. Estimador de la varianza de tipo Horvitz-Thompson:

$$\hat{V}_{HT}\left(\hat{\tau}_{HT}\right)=N^2\hat{V}_{HT}\left(\overline{y}_{HT}\right)=\sum_{i=1}^{n}\left(1-\pi_i\right)\frac{y_i^2}{\pi_i^2}+2\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

2. Estimador de la varianza de tipo Sen-Yates-Grundy:

$$\hat{V}_{SYG}\left(\hat{\tau}_{HT}\right)=N^2\hat{V}_{SYG}\left(\overline{y}_{HT}\right)=\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

Estimador de tipo Hájek para el total poblacional:

$$\hat{\tau}_H=N\,\overline{y}_H=\frac{N}{\hat{N}}\sum_{i=1}^{n}\frac{y_i}{\pi_i}=\frac{N}{\hat{N}}\sum_{i=1}^{n}d_i y_i \qquad \text{donde} \quad \hat{N}=\sum_{i=1}^{n}\frac{1}{\pi_i}=\sum_{i=1}^{n}d_i$$

Si N es conocido se pueden obtener ambos estimadores. En el caso de que N sea desconocido sólo es posible obtener el estimador de tipo Horvitz-Thompson.

Estimadores de la varianza del estimador de tipo Hájek (método Jackknife):

$$\hat{V}_J\left(\hat{\tau}_H\right)=N^2\hat{V}_J\left(\overline{y}_H\right)=N^2\,\frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\overline{y}_{H(i)}-\overline{y}_H\right)^2=\frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\tau}_{H(i)}-\hat{\tau}_H\right)^2$$

$$\hat{\tau}_{H(i)}=\frac{N}{\hat{N}_{(i)}}\sum_{j\in s,j\neq i}\frac{y_j}{\pi_j}=\frac{N}{\hat{N}_{(i)}}\sum_{j\in s,j\neq i}d_j y_j \qquad\qquad \hat{N}_{(i)}=\sum_{j\in s,j\neq i}\frac{1}{\pi_j}=\sum_{j\in s,j\neq i}d_j$$

Ejemplo 7.2. Para los datos del Ejemplo 7.1, obtenga la estimación del número total de pobres en la población en estudio así como el límite del error de estimación.

Solución:

Para estimar el total de pobres hay que multiplicar el estimador de la proporción por el tamaño de la población:

$\hat{\tau}_H=N\;\hat{\mathrm{p}}_H=1000\times 0,2381=238,1$

Es decir, se estima que en la población hay 238 personas consideradas como pobres.

El límite del error de estimación viene dado por:

$$B=2\sqrt{\hat{V}\left(\hat{\tau}_H\right)}=2\sqrt{\hat{V}\left(N\hat{p}_H\right)}=2\sqrt{N^2\hat{V}\left(\hat{p}_H\right)}=2N\sqrt{\hat{V}\left(\hat{p}_H\right)}=2\times 1000\times\sqrt{0,0203}=284,96 \quad \blacksquare$$

Ejemplo 7.3. Siguiendo con el ***Ejemplo 2.1***, se está interesado de nuevo en determinar la audiencia de la publicidad televisiva en una cadena local de un municipio, y se decide realizar una encuesta por muestreo para estimar el número de horas por semana que se ve la televisión en las viviendas del municipio. Éste está formado por tres barrios con diferentes perfiles socio-culturales que afectan a la audiencia televisiva. Hay 210 hogares en el barrio A, 84 en el barrio B y 126 en el barrio C. ***En esta ocasión***, la empresa publicitaria tiene tiempo y dinero suficientes para entrevistar 23 hogares y decide seleccionar muestras aleatorias de tamaños: 11 del barrio A, 6 del barrio B, y 6 del barrio C.

Se seleccionan las muestras aleatorias simples y se realizan las entrevistas. Los resultados, con mediciones del tiempo que se ve la televisión en horas por semana, se muestran en la siguiente tabla:

BARRIO A			BARRIO B		BARRIO C	
39	34	24	22	40	16	21
27	42		28		17	
37	38		17		23	
41	37		24		18	
30	36		37		25	

Estimar el tiempo medio que se ve la televisión por semana en todos los hogares del municipio usando:

a) Las expresiones del muestreo aleatorio estratificado.
b) Las expresiones para diseños muestrales con probabilidades desiguales.

Solución:

a) Muestreo aleatorio estratificado

Barrio	N_i	$\overline{y}_i$	$N_i\overline{y}_i$
A	210	35	7350
B	84	28	2352
C	126	20	2520
	420		**12222**

Usando muestreo aleatorio estratificado, el número medio de horas que se ve la televisión a la semana en todo el municipio es:

$$\overline{y}_{st} = \frac{1}{N}\sum_{i=1}^{3} N_i\overline{y}_i = 29{,}1 \quad \text{horas/semana}$$

b) Muestreo con probabilidades desiguales

Las probabilidades de inclusión de los individuos según al barrio al que pertenecen son:

$\pi_1 = \frac{n_A}{N_A} = \frac{11}{210} = 0{,}05238$ si el individuo i pertenece al barrio A $\quad \frac{1}{\pi_1} = \frac{N_A}{n_A} = \frac{210}{11} = 19{,}090909$

$\pi_2 = \frac{n_B}{N_B} = \frac{6}{84} = 0{,}07143$ si el individuo i pertenece al barrio B $\quad \frac{1}{\pi_2} = \frac{N_B}{n_B} = \frac{84}{6} = 14$

$\pi_3 = \frac{n_C}{N_C} = \frac{6}{126} = 0{,}04762$ si el individuo i pertenece al barrio C $\quad \frac{1}{\pi_3} = \frac{N_C}{n_C} = \frac{126}{6} = 21$

A partir de las probabilidades de inclusión podemos observar que los individuos del barrio B tienen una mayor probabilidad de pertenecer a la muestra, mientras que los individuos del barrio C son los que tienen la menor probabilidad.

Estimador de tipo Horvitz-Thompson:

$$\bar{y}_{HT} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{12222{,}04}{420} = 29{,}1 \text{ horas/semana}$$

Estimador de tipo Hájek:

$$\bar{y}_{H} = \frac{1}{\hat{N}}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{12222{,}04}{420} = 29{,}1 \text{ horas/semana}$$

Como ocurre en este ejercicio y en general, el estimador de la media tipo Horvitz-Thompson y de tipo Hájek coinciden con el estimador de la media basado en muestras aleatorias simples y en muestras aleatorias estratificadas.

Barrio	y_i	π_i	$\frac{y_i}{\pi_i}$	$\frac{1}{\pi_i}$
A	39	0,05238	744,5590	19,0909
	27	0,05238	515,4639	19,0909
	37	0,05238	706,3765	19,0909
	41	0,05238	782,7415	19,0909
	30	0,05238	572,7377	19,0909
	34	0,05238	649,1027	19,0909
	42	0,05238	801,8328	19,0909
	38	0,05238	725,4677	19,0909
	37	0,05238	706,3765	19,0909
	36	0,05238	687,2852	19,0909
	24	0,05238	458,1901	19,0909
B	22	0,07143	307,9938	14
	28	0,07143	391,9922	14
	17	0,07143	237,9952	14
	24	0,07143	335,9933	14
	37	0,07143	517,9896	14
	40	0,07143	559,9888	14
C	16	0,04762	335,9933	21
	17	0,04762	356,9929	21
	23	0,04762	482,9903	21
	18	0,04762	377,9924	21
	25	0,04762	524,9895	21
	21	0,04762	440,9912	21
			12222,04	$\hat{N} = 420$

■

7.3 El problema de la estimación de la varianza de estimadores: métodos de remuestreo.

En un muestreo con probabilidades desiguales es bastante común que las probabilidades de inclusión de segundo orden sean desconocidas. Este hecho supone un problema para la estimación de la varianza de estimadores puesto que tales expresiones dependen de estas probabilidades de inclusión de segundo orden, tal como se ha podido observar a lo largo de

este capítulo. Este problema se puede resolver mediante diferentes técnicas. Por ejemplo, una opción podría ser aproximar las probabilidades de inclusión de segundo orden a partir de las probabilidades de inclusión de primer orden. Sin embargo, el procedimiento más común que suele utilizarse en la práctica es recurrir a un método de estimación de la varianza por "remuestreo". En general, este procedimiento consiste en obtener distintas estimaciones del parámetro en estudio a partir de los datos muestrales, y aproximar la varianza del estimador a través de la variabilidad de estas estimaciones adicionales. Los métodos de remuestreo más comunes son: el método Jackknife y el método Bootstrap.

Para explicar estos métodos de remuestreo consideremos que el objetivo es estimar un determinado parámetro θ (por ejemplo, la media, la proporción o el total poblacional) a partir de una muestra de tamaño *n*, y donde $\hat{\theta}$ representa el correspondiente estimador de θ.

1. ***Método Jackknife***. Denotando por $\hat{\theta}_{(i)}$ al estimador $\hat{\theta}$ después de eliminar de la muestra la observación *i*-ésima, el estimador de la varianza de $\hat{\theta}$ por el método Jackknife tiene la siguiente expresión:

$$\hat{V}_J\left(\hat{\theta}\right) = \frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

2. ***Método Bootstrap*** Consiste en replicar d_i veces las observaciones y_i de la muestra, de forma que se obtiene una nueva población llamada población Bootstrap, y la cual tiene un tamaño aproximado de *N* valores (igual que la población original). A partir de esta población se extraen *B* muestras (llamadas muestras Bootstrap). En general, habría que considerar un valor de *B* mayor que 100 para obtener una buena aproximación de la varianza (algunos autores sugieren hasta *B*>1000). Denotaremos por $\hat{\theta}_{(b)}$ al estimador $\hat{\theta}$ obtenido en la *b*-ésima muestra Bootstrap. El estimador de la varianza de $\hat{\theta}$ por el método Bootstrap tiene la siguiente expresión:

$$\hat{V}_B\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{(b)} - \overline{\theta}_B\right)^2 \qquad ; \qquad \overline{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{(b)}$$

En el caso del muestreo aleatorio simple $\left(d_i = \frac{N}{n}\right)$ el anterior estimador de la varianza no es insesgado. Pero lo es si las muestras Bootstrap se toman de tamaño *n*-1.

Los métodos de remuestreo, tal como el Jackknife y el Bootstrap, tienen una carga computacional muy elevada y suelen utilizarse programas específicos para su cálculo.

Ejemplo 7.4. Un banco tiene repartidos 100 cajeros automáticos en la provincia de Granada. El responsable de este banco desea estimar la cantidad media de efectivo que se retiran en todos sus cajeros automáticos en un determinado día, por lo que selecciona una muestra aleatoria simple sin reemplazamiento de 5 cajeros y obtiene las siguientes cantidades (en miles de euros):

Cajero	1	2	3	4	5
Efectivo	15	20	10	25	30

Obtenga intervalos de confianza para el importe medio que se retira en los cajeros utilizando la fórmula de la varianza del muestreo aleatorio simple y las varianzas obtenidas por los métodos Jackknife y Bootstrap. Para el método Bootstrap considere que se obtuvieron 10 muestras Bootstrap, y éstas fueron las siguientes:

b	Muestra Bootstrap
1	$\{1,1,2,4,5\}$
2	$\{1,2,2,2,5\}$
3	$\{1,2,3,4,5\}$
4	$\{3,3,4,5,5\}$
5	$\{1,1,3,4,4\}$

b	Muestra Bootstrap
6	$\{1,2,3,3,4\}$
7	$\{1,2,2,3,5\}$
8	$\{3,3,4,5,5\}$
9	$\{3,3,3,4,4\}$
10	$\{2,3,4,5,5\}$

Solución:

El estimador de la media en un muestreo aleatorio simple es:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{100}{5} = 20$$

Se estima que en los cajeros se retiran de media 20000 euros en el día del estudio.

Intervalo de confianza usando la expresión del muestreo aleatorio simple:

$$\hat{V}(\bar{y}) = \frac{S^2}{n}\frac{N-n}{N} = \frac{62,5}{5}\frac{100-5}{100} = 11,875$$

$$\left(\bar{y} - 2\sqrt{\hat{V}(\bar{y})}\,,\, \bar{y} + 2\sqrt{\hat{V}(\bar{y})}\right) = (13,11,\ 26,89)$$

Intervalo de confianza usando el método Jackknife:

i	$\bar{y}_{(i)}$	$\left(\bar{y}_{(i)} - \bar{y}\right)^2$
1	21,25	1,5625
2	20,00	0,0000
3	22.50	6,2500
4	18,75	1,5625
5	17,50	6,2500
		15,625

$$\hat{V}_J\left(\bar{y}\right) = \frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{y}_{(i)} - \bar{y}\right)^2 = \frac{100-5}{100}\frac{5-1}{5}15,625 = 11,875$$

$$\left(\bar{y} - 2\sqrt{\hat{V}_J(\bar{y})}\,,\, \bar{y} + 2\sqrt{\hat{V}_J(\bar{y})}\right) = (13,11\,,\ 26,89)$$

Intervalo de confianza usando el método Bootstrap:

$$\bar{y}_{(1)} = \frac{1}{n}\sum_{i=1}^{n} y_{1i} = \frac{1}{5}(15+15+20+25+30) = 21$$

$$\bar{y}_{(2)} = \frac{1}{n}\sum_{i=1}^{n} y_{2i} = \frac{1}{5}(15+20+20+20+30) = 21$$

$$\vdots$$

$$\bar{y}_{(10)} = \frac{1}{n}\sum_{i=1}^{n} y_{2i} = \frac{1}{5}(20+10+25+30+30) = 23$$

b	Muestra Bootstrap	$\bar{y}_{(b)}$	$\left(\bar{y}_{(b)}-20,5\right)^2$	
1	$\{1,1,2,4,5\}$	21	0,25	
2	$\{1,2,2,2,5\}$	21	0,25	
3	$\{1,2,3,4,5\}$	20	0,25	
4	$\{3,3,4,5,5\}$	25	20,25	
5	$\{1,1,3,4,4\}$	20	0,25	
6	$\{1,2,3,3,4\}$	16	20,25	
7	$\{1,2,2,3,5\}$	19	2,25	
8	$\{3,3,4,5,5\}$	25	20,25	
9	$\{3,3,3,4,4\}$	16	20,25	
10	$\{2,3,4,5,5\}$	23	6,25	
		205	**90,5**	

$$\bar{y}_B = \frac{1}{B}\sum_{b=1}^{B}\bar{y}_{(b)} = \frac{205}{10} = 20,5$$

$$\hat{V}_B\left(\bar{y}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{y}_{(b)}-\bar{y}_B\right)^2 = \frac{1}{10-1}90,5 = 10,0556$$

$$\left(\bar{y}-2\sqrt{\hat{V}_B(\bar{y})}\,,\,\bar{y}+2\sqrt{\hat{V}_B(\bar{y})}\right) = (13,66;\ 26,34)$$ ■

7.4 Aplicaciones en encuestas oficiales

Oficina Europea de Estadística – Eurostat

Encuesta de la Unión Europea sobre ingresos y condiciones de vida (European Union Survey on Income and Living Condition, EU-SILC)**.**

El objetivo principal de esta encuesta es la recogida de microdatos multidimensionales transversales y longitudinales sobre ingresos de individuos, pobreza, exclusión social y condiciones de vida. El diseño muestral puede variar entre los distintos países, aunque el más

común es el ***muestreo polietápico estratificado*** (stratified multi-stage sampling), tal como puede verse en la siguiente tabla extraída directamente de Eurostat:

Sampling design	Country
Without stratification	
Simple random sampling	DK, MT, IS, NO
Systematic sampling	SE
With stratification	
Stratified sampling according to different design by rotational group	HU
Stratified simple random sampling	DE*, CY, LT, LU, AT, SK, CH
Stratified and systematic sampling	EE
Stratified two-stage sampling	IT, HR, LV, NL, PT, SI
Stratified multi-stage sampling	BE, BG, CZ, IE, EL, ES, FR, IT, PL, RO, UK
Stratified two-phase sampling	FI

* from former participants of micro census

Figura 7.1. *Diseños muestrales utilizados por Eurostat en distintos países de la Unión Europa. Figura tomada de la web oficial de Eurostat.*

Los microdatos de esta encuesta están disponibles para su explotación por parte de investigadores, empresas, etc. Bajo estos diseños muestrales, los individuos de una misma población tienen distintas probabilidades de ser seleccionados, y por esta razón, los microdatos incluyen los pesos muestrales asociados a cada individuo, y los cuales deben tenerse en cuenta en la etapa de estimación de parámetros.

A modo de ejemplo se presenta en la figura siguiente resultados realizados por Lelkes et al. (2009) y basados en datos de la encuesta EU-SILC del año 2006. El objetivo es estimar la proporción de pobres de cada país. Las estimaciones han tenido en cuenta los pesos del diseño muestral proporcionados junto con los valores observados de las variables en estudio.

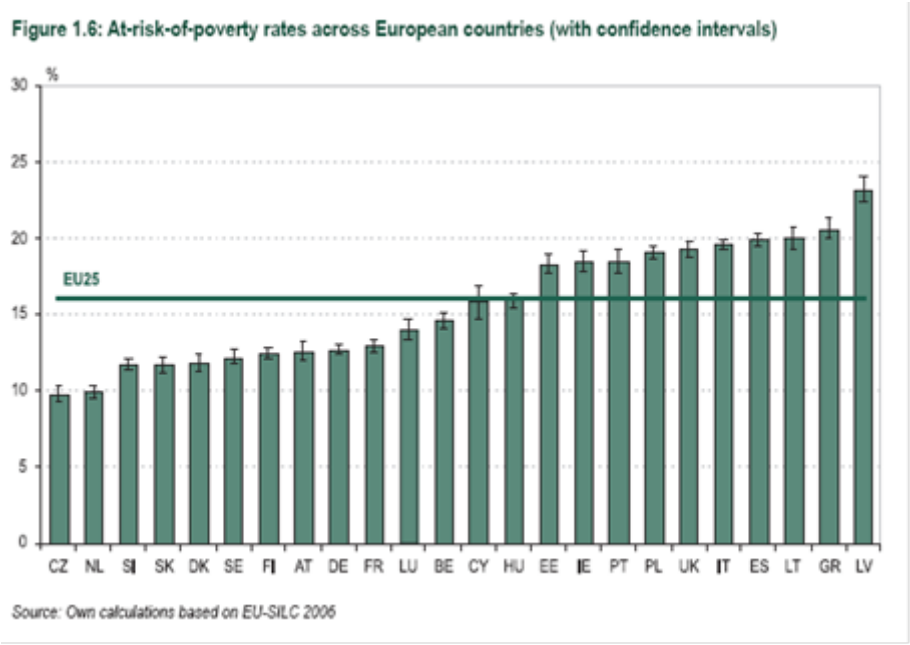

Figura 7.2. *Estimaciones puntuales y por intervalos de confianza usando Bootstrap de la proporción de pobres en distintos países de la Unión Europa. Los datos se han tomado de la encuesta EU-SILC del año 2006. Estudio realizado por Lelkes et al. (2009).*

Referencia: Lelkes, O., Medgyesi, M., Tóth, I. G. and Ward, T. (2009) 'Income Distribution and the Risk of Poverty' In T. Ward, O. Lelkes, H. Sutherland, and I. G. Tóth (eds) European Inequalities. Social Inclusion and Income Distribution in the European Union (Budapest: Tárki), pp. 17–44.

8. Decisión en ambiente de incertidumbre.

8.1 Elementos de un problema de decisión.

En esta parte de la asignatura (***Teoría de la Decisión***) abordaremos la toma racional (coherente) de decisiones. Estudiaremos cómo deberían tomarse las decisiones en contraste con la forma incoherente en que muchas veces se suelen tomar en la práctica.

El **primer elemento** a tener en cuenta en un problema de toma de decisiones es la persona (empresa,…) que tiene que decidir, al que se denomina ***decisor***.

El decisor tiene unas características, como son sus preferencias, sus creencias y la información de que dispone, que deben tenerse muy en cuenta al estudiar el problema.

Nos ocupamos exclusivamente del problema de decisión unipersonal, en el que hay un único decisor. Como caso particular de toma de decisiones están aquellas situaciones en que dos individuos han de decidir en conflicto de intereses, de cuyo estudio se encarga la ***Teoría de Juegos***.

El **segundo elemento** de un problema de decisión es el ***conjunto de posibles acciones o alternativas*** de que dispone el decisor.

En la construcción de un modelo para la toma de decisiones, la **primera tarea** del decisor, en la que habrá de poner mucha atención, es la enumeración exhaustiva de ese conjunto de alternativas.

La Teoría de la Decisión no entra en la elaboración del conjunto de alternativas, es algo que ya viene dado. Nosotros suponemos además que ese conjunto es finito: $a_1, a_2,.....,a_m$, y entre ellas hay que elegir sólo una.

La elección de alternativa se hará en función de sus consecuencias. Consecuencias que dependerán de cierta cantidad de factores externos que están fuera del control del decisor.

Se denomina ***estado de la naturaleza*** o simplemente estado a una descripción completa de cómo son esos factores externos en una determinada situación. Si el decisor conociese el estado de la naturaleza imperante, es decir, si conociese perfectamente cómo van a

manifestarse los factores externos, podría predecir con exactitud las consecuencias de sus acciones.

La **segunda tarea** en la construcción del modelo para la toma de decisiones consiste en el análisis de esos factores externos, juzgando cuáles son relevantes y cuáles no, y en la enumeración exhaustiva de los estados de la naturaleza.

El ***conjunto de todos los posibles estados de la naturaleza*** constituye el **tercer elemento** de un problema de decisión. Suponemos que este conjunto es *finito*, tendremos una lista *exhaustiva* de estados de la naturaleza $e_1, e_2, ..., e_n$ mutuamente *excluyentes*.

La conjunción de una decisión con un estado de la naturaleza determina perfectamente la consecuencia resultante, notaremos por x_{ij} a la valoración del **resultado** de adoptar la alternativa a_i, cuando el estado de la naturaleza es e_j.

8.2 Tablas de decisión

Muchos procesos de toma de decisiones son tratados por medio de las ***tablas de decisión***, también llamadas ***matriz de pagos***, en las que se representan los elementos característicos de estos problemas:

- Los diferentes ***estados*** que puede presentar la naturaleza: $\boldsymbol{e_1, e_2, ..., e_n}$.
- Las ***acciones*** o ***alternativas*** entre las que seleccionará el decisor: $\boldsymbol{a_1, a_2,..., a_m}$.
- Las ***consecuencias*** o ***resultados*** $\boldsymbol{x_{ij}}$ de la elección de la alternativa $\boldsymbol{a_i}$ cuando la naturaleza presenta el estado $\boldsymbol{e_j}$.

Se supone un *número finito* de estados y alternativas. El ***formato general*** de una ***tabla de decisión*** es el siguiente:

Estados de la Naturaleza

Alternativas	e_1	. . .	e_j	. . .	e_n
a_1	x_{11}	. . .	x_{1j}	. . .	x_{1n}
. . .	. . .	. . .	. . .	. . .	. . .
a_i	x_{i1}	. . .	x_{ij}	. . .	x_{in}
. . .	. . .	. . .	. . .	. . .	. . .
a_m	x_{m1}	. . .	x_{mj}	. . .	x_{mn}

8.3 Valoración de los resultados.

Aunque las consecuencias o resultados x_{ij} no son necesariamente números, supondremos que el decisor puede valorarlos numéricamente. En el contexto económico esta valoración suele ser monetaria. En lo que sigue identificaremos cada resultado con su valoración numérica. Así, x_{ij} hará referencia tanto al propio resultado como al valor asignado por el decisor.

Ejemplo 8.1. En cierta ciudad se va a construir un Centro de Congresos en una de dos posibles localizaciones *X* e *Y*, que será decidido el próximo año. Una cadena de restaurantes está interesada en abrir un restaurante cerca del nuevo Centro de Congresos, para lo cual tiene que decidir qué inmueble comprar. La siguiente tabla muestra el precio de los inmuebles, el beneficio estimado que obtendrá el restaurante en cada posible localización, si el Centro de Congresos se localiza allí, y el valor de venta de cada inmueble si finalmente el Centro de Congresos no se construye en ese lugar. ¿Cuál es la decisión más adecuada?

	en *X*	en *Y*
Precio del inmueble	126	84
Beneficio estimado del restaurante	217	161
Valor de venta del inmueble	42	28

Las ***alternativas*** posibles de que dispone el decisor son las siguientes:

- Comprar el inmueble en *X*.
- Comprar el inmueble en *Y*.
- Comprar ambos inmuebles.
- No comprar ninguno.

Por otra parte, los posibles ***estados de la naturaleza*** son:

- El Centro de Congresos se construye en *X*.
- El Centro de Congresos se construye en *Y*.

Así, si la cadena de restaurantes compra el inmueble en *X* y el Centro de Congresos se construye allí finalmente, obtendrá como rendimiento final el correspondiente a la explotación del restaurante, 217, menos la inversión realizada en la compra del inmueble, 126, es decir, 217-126 = 91. Por el contrario, si el Centro de Congresos se construye en *Y*, el inmueble adquirido en *X* deberá ser vendido, por lo que se obtendrá un beneficio de 42, al que habrá que restar la inversión inicial en la compra, 126. Esto proporciona un rendimiento final de 42-126 = -84.

De manera análoga se determinan los resultados de las restantes alternativas ante cada uno de los posibles estados de la naturaleza, dando lugar a la siguiente tabla de decisión:

Alternativas Inmueble comprado en:	Estados de la Naturaleza	
	Centro de Congresos en *X*	Centro de Congresos en *Y*
X	91	- 84
Y	- 56	77
X e Y	35	- 7
Ninguno	0	0

8.4 Clasificación de los problemas de decisión.

Se clasifican en función del grado de información que sobre los estados de la naturaleza tenga el decisor:

- **Problemas de decisión en ambiente de certeza o certidumbre:** El decisor conoce el estado de la naturaleza que ocurrirá con absoluta certeza. Solamente hay una consecuencia para cada alternativa y puede conocer con certeza el resultado de sus acciones. La tabla de decisión sólo tiene una columna.
- **Problemas de decisión en ambiente de riesgo:** Ocurre cuando hay dos o más estados de la naturaleza y se conoce la probabilidad de que se presente cada uno de ellos. Es decir, se conoce una distribución de probabilidad sobre la ocurrencia de los estados. Cada alternativa tiene diferentes consecuencias (resultados) con una probabilidad conocida.
- **Problemas de decisión en ambiente de incertidumbre:** En estos problemas también hay múltiples consecuencias (resultados) para cada alternativa, pero no se conoce la probabilidad de ocurrencia de cada una de ellas. El decisor solo conoce cuales son los estados posibles y las consecuencias de sus alternativas según el estado que se presente. La decisión final se basa en criterios subjetivos del decisor, jugando un papel importante su actitud.

8.5 Toma de decisiones en ambiente de incertidumbre.

En los procesos de decisión bajo ***incertidumbre***, el decisor conoce cuáles son los posibles estados de la naturaleza, aunque no dispone de información alguna sobre cuál de ellos ocurrirá. No sólo es incapaz de predecir el estado real que se presentará, sino que además no puede cuantificar de ninguna forma esta incertidumbre.

A continuación se describen las diferentes **reglas de decisión** en ambiente de incertidumbre, que serán sucesivamente aplicadas al ejemplo de construcción del Centro de Congresos:

- Criterio de Laplace
- Criterio de Wald

- Criterio Maximax
- Criterio de Hurwicz
- Criterio de Savage

8.5.1 Criterio de Laplace.

Este criterio, propuesto por Laplace en 1825, está basado en el **principio de razón insuficiente**. Como a priori no existe ninguna razón para suponer que un estado se puede presentar más fácilmente que los demás, podemos considerar que **todos los estados tienen la misma probabilidad de ocurrencia**, es decir, la ausencia de conocimiento sobre el estado de la naturaleza equivale a afirmar que todos los estados son equiprobables. Así, para un problema de decisión con ***n*** posibles estados de la naturaleza, asignaríamos **probabilidad *1/n*** a cada uno de ellos.

La regla de Laplace selecciona como alternativa óptima aquella que proporciona el mayor resultado esperado:

$$\max_{a_i}\left\{\frac{1}{n}\sum_{j=1}^{n} x_{ij}\right\}$$

Ejemplo 8.2. Partiendo del ejemplo de construcción del Centro de Congresos, la siguiente tabla muestra los resultados esperados para cada una de las alternativas:

Alternativas Inmueble comprado en:	**Estados de la Naturaleza**		**Resultado esperado**
	Centro de Congresos en *X*	**Centro de Congresos en *Y***	
X	91	- 84	3,5
Y	- 56	77	10,5
X e Y	35	- 7	**14**
Ninguno	0	0	0

En este caso, cada estado de la naturaleza tendría probabilidad de ocurrencia igual a 1/2.

$$3,5 = \frac{1}{2}91 - \frac{1}{2}84 \quad \ldots \quad 14 = \frac{1}{2}35 - \frac{1}{2}7$$

El resultado esperado máximo se obtiene para la tercera alternativa, por lo que la decisión óptima según el criterio de Laplace sería comprar ambos inmuebles.

La **objeción** que se suele hacer al criterio de Laplace es la siguiente: **ante una misma realidad, pueden tenerse distintas probabilidades, según los estados de la naturaleza que se consideren**. Por ejemplo, los estados de la naturaleza podrían ser: *"Centro de Congresos en X", "Centro de Congresos en Y" y "Centro de Congresos en otra localización diferente a las anteriores",* con lo que la probabilidad de cada estado de la naturaleza sería 1/3.

Desde un punto de vista práctico, la dificultad de aplicación de este criterio reside en la necesidad de elaboración de una **lista exhaustiva y mutuamente excluyente de todos los posibles estados de la naturaleza**.

Por otra parte, al ser un criterio basado en el **concepto de valor esperado**, su resultado es correcto tras muchas repeticiones del proceso de toma de decisiones. Sin embargo, en aquellos casos en que la elección sólo va a realizarse una vez, puede conducir a decisiones poco acertadas si la distribución de resultados presenta una gran dispersión, como se muestra en la siguiente tabla:

	Estados de la Naturaleza		
Alternativas	e_1	e_2	**Resultado esperado**
a_1	30	-10	**10**
a_2	10	8	9

Este criterio seleccionaría la alternativa a_1, que puede ser poco conveniente si la toma de decisiones se realiza una única vez, ya que podría conducirnos a una pérdida (-10).

8.5.2 Criterio de Wald (maximin).

Este es el criterio más conservador, que está basado en lograr lo mejor en las peores condiciones posibles. Si x_{ij} representa ganancias para el decisor, para a_i la peor ganancia, independientemente de lo que e_j pueda ser, es $\min_{e_j}\{x_{ij}\}$. Este resultado recibe el nombre de *nivel de seguridad* (al elegir a_i se garantiza al menos un beneficio de $\min_{e_j}\{x_{ij}\}$ unidades).

Wald sugirió que el decisor debe adoptar aquella alternativa que tenga el mayor nivel de seguridad, es decir, elegir a_i asociada a

$$\max_{a_i} \min_{e_j}\{x_{ij}\}$$

Este criterio recibe el nombre de ***criterio maximin***, y corresponde a un **pensamiento pesimista**, pues se basa en lo peor que le puede ocurrir al decisor cuando elige una alternativa.

Ejemplo 8.3. Partiendo del ejemplo de construcción del Centro de Congresos, la siguiente tabla muestra los resultados obtenidos junto con los niveles de seguridad de las diferentes alternativas:

Alternativas Inmueble comprado en:	**Estados de la Naturaleza**		$\min_{e_j}\{x_{ij}\}$
	Centro de Congresos en *X*	**Centro de Congresos en *Y***	
X	91	- 84	-84
Y	- 56	77	-56
X e Y	35	- 7	-7
Ninguno	0	0	**0**

La alternativa óptima según el criterio de Wald sería no comprar ningún inmueble, pues proporciona el mayor de los niveles de seguridad.

En ocasiones, el criterio de Wald puede conducir a decisiones poco adecuadas. Por ejemplo, consideremos la siguiente tabla de decisión, en la que se muestran los niveles de seguridad de las diferentes alternativas.

Alternativas	**Estados de la Naturaleza**		$\min_{e_j}\{x_{ij}\}$
	e_1	e_2	
a_1	3000	199	199
a_2	200	200	**200**

El criterio de Wald seleccionaría la alternativa a_2, aunque lo más razonable parece ser elegir la alternativa a_1, ya que en el caso más favorable proporciona una recompensa mucho mayor, mientras que en el caso más desfavorable la recompensa es similar.

Este criterio sería adecuado si la naturaleza fuese un contrincante que intentase hacernos perder (*Teoría de Juegos*), pero no es así: la naturaleza presenta un estado u otro independientemente de los resultados.

8.5.3 Criterio de Hurwicz.

Este criterio representa un rango de actitudes desde la más optimista hasta la más pesimista. En las condiciones más optimistas se elegiría la acción que proporcione el $\max_{a_i}\max_{e_j}\{x_{ij}\}$ (*criterio maximax*). Se supone que x_{ij} representa ganancia o beneficio. A $\max_{e_j}\{x_{ij}\}$ se le denomina *nivel optimista* de la acción a_i.

Igual de racional es adoptar el criterio pesimista (Wald) que el optimista. A la vista de que muy poca gente es tan optimista o tan pesimista como estos criterios obligan a ser, Hurwicz (1951) propuso un criterio intermedio: el decisor ordenará sus alternativas según una media ponderada de los niveles optimista y de seguridad por los pesos respectivos α y (1- α), donde $0 \le \alpha \le 1$.

Si x_{ij} representa beneficio, se selecciona la acción que proporcione:

$$\max_{a_i}\left\{\alpha \max_{e_j}\{x_{ij}\} + (1-\alpha)\min_{e_j}\{x_{ij}\}\right\}$$

El parámetro α se conoce como ***índice de optimismo*:** cuando $\alpha=1$, el criterio es el más optimista; cuando $\alpha=0$, es el más pesimista. Un valor de α entre cero y uno puede ser seleccionado dependiendo de si el decisor tiende hacia el pesimismo o al optimismo. En

ausencia de una sensación fuerte de una circunstancia u otra, un valor de α=1/2 parece ser una selección razonable[1].

Ejemplo 8.4. Partiendo del ejemplo de construcción del Centro de Congresos, la siguiente tabla muestra los resultados junto con la media ponderada de los niveles optimista y de seguridad de las diferentes alternativas para un índice de optimismo $\alpha = 0{,}4$:

Alternativas Inmueble comprado en:	**Estados de la Naturaleza**		$\min_{e_j}\{x_{ij}\}$	$\max_{e_j}\{x_{ij}\}$	Media ponderada (α=0,4)
	Centro de Congresos en *X*	**Centro de Congresos en *Y***			
X	91	- 84	-84	91	-14
Y	- 56	77	-56	77	-2,8
X e Y	35	- 7	-7	35	**9,8**
Ninguno	0	0	0	0	0

$$-14 = 0{,}4\times 91 - 0{,}6\times 84 \quad \ldots \quad 9{,}8 = 0{,}4\times 35 - 0{,}6\times 7$$

La alternativa óptima según el criterio de Hurwicz sería comprar los inmuebles *X* e *Y*, pues proporciona la mayor de las medias ponderadas para el valor de α seleccionado.

8.5.4 Criterio de Savage (minimax).

En 1951 Savage argumenta que al utilizar los valores x_{ij} para realizar la elección, el decisor compara el resultado de una alternativa bajo un estado de la naturaleza con todos los demás resultados, independientemente del estado de la naturaleza bajo el que ocurran. Sin embargo, el resultado de una alternativa sólo debería ser comparado con los resultados de las demás alternativas bajo el mismo estado de la naturaleza. Una consecuencia puede ser muy pobre en el contexto de la tabla completa y sin embargo ser la mejor que puede ocurrir bajo un determinado estado.

Con este propósito Savage define el concepto de **pérdida relativa** o **pérdida de oportunidad** p_{ij} asociada a un resultado x_{ij} como la diferencia entre el resultado de la mejor alternativa dado que e_j es el verdadero estado de la naturaleza y el resultado de la alternativa a_i bajo ese mismo estado e_j:

[1] Para determinar α , se le ofrece al decisor el siguiente problema:

	e_1	e_2
a_1	1	0
a_2	v	v

Haciendo variar v hasta que ambas alternativas le sean indiferentes.
Para la primera alternativa el nivel de seguridad es cero y el nivel optimista es 1, entonces según el índice de optimismo del decisor, su media ponderada es: $\alpha\times 1 + (1-\alpha)\times 0 = \alpha$. Para la segunda alternativa su media ponderada es $\alpha v + (1-\alpha)v = v$. Por lo que en caso de indiferencia se tendrá $\alpha = v$.

$$p_{ij} = \max_{a_k} \{x_{kj}\} - x_{ij}$$

Así, si el verdadero estado en que se presenta la naturaleza es e_j y el decisor elige la alternativa a_k que proporciona el máximo resultado x_{kj}, entonces no ha dejado de ganar nada, pero si elige otra alternativa cualquiera a_i , entonces obtendría como ganancia x_{ij} y dejaría de ganar $x_{kj} - x_{ij}$.

Savage opina que en lugar de trabajar con la tabla de beneficios, debe trabajarse con la tabla de pérdidas de oportunidad y en ella debe actuarse con una filosofía pesimista: propone seleccionar la alternativa que proporcione la menor de las mayores pérdidas relativas, es decir,

$$\min_{a_i} \max_{e_j} \{p_{ij}\}$$

Conviene destacar que, como paso previo a la aplicación de este criterio, se debe calcular la matriz de pérdidas relativas, formada por los elementos p_{ij}. Cada columna de esta matriz se obtiene calculando la diferencia entre el valor máximo de esa columna y cada uno de los valores que aparecen en ella.

Ejemplo 8.5. Partiendo del ejemplo de construcción del Centro de Congresos, la siguiente tabla muestra la matriz de pérdidas relativas:

Alternativas Inmueble comprado en:	**Estados de la Naturaleza**		$\max_{e_j} \{p_{ij}\}$
	Centro de Congresos en *X*	**Centro de Congresos en *Y***	
X	91-91=0	77-(-84)=161	161
Y	91-(-56)=147	77-77=0	147
X e Y	91-35=56	77-(-7)=84	**84**
Ninguno	91-0=91	77-0=77	91

La decisión óptima según el criterio de Savage sería comprar ambas parcelas.

El criterio de Savage puede dar lugar en ocasiones a decisiones poco razonables. Para comprobarlo, consideremos la siguiente tabla de resultados:

Alternativas	**Estados de la Naturaleza**	
	e_1	e_2
a_1	45	10
a_2	20	30

La **tabla de pérdidas relativas** correspondiente a esta tabla de resultados es la siguiente:

Alternativas	**Estados de la Naturaleza**		$\max_{e_j} \{p_{ij}\}$
	e_1	e_2	
a_1	0	20	**20**
a_2	25	0	25

La alternativa óptima es a_1. Supongamos ahora que se añade una alternativa, dando lugar a la siguiente tabla de resultados:

Alternativas	Estados de la Naturaleza	
	e_1	e_2
a_1	45	10
a_2	20	30
a_3	15	45

La nueva tabla de pérdidas relativas sería:

Alternativas	Estados de la Naturaleza		$\max_{e_j}\{p_{ij}\}$
	e_1	e_2	
a_1	0	35	35
a_2	25	15	**25**
a_3	30	0	30

El criterio de Savage selecciona ahora como alternativa óptima a_2, cuando antes seleccionó a_1. Este cambio de alternativa resulta un poco paradójico: supongamos que a una persona se le da a elegir entre carne o verduras, y prefiere carne. Si posteriormente se la da a elegir entre carne, verduras o pescado, ¡ahora prefiere verduras!

Ejemplo 8.6. En este ejemplo la matriz de decisión recoge gastos en lugar de beneficios por lo que los anteriores criterios de decisión varían en su forma de aplicación.

Una cafetería debe planificar su nivel de abastecimiento para satisfacer la demanda de sus clientes en un día de fiesta. El número exacto de clientes no se conoce, pero se espera que esté en una de cuatro categorías: 300, 400, 600 y 650 clientes. Se plantean, por lo tanto, cuatro niveles de abastecimiento. La desviación respecto del número de clientes esperado resulta en costes adicionales, ya sea por un abastecimiento excesivo sin necesidad o porque la demanda no puede satisfacerse. La tabla que sigue recoge estos costes en cientos de euros:

		$e_1(300)$	$e_2(400)$	$e_3(600)$	$e_4(650)$
Nivel de abastecimiento	$a_1(300)$	15	30	54	75
	$a_2(400)$	24	21	24	69
	$a_3(600)$	63	54	36	63
	$a_4(650)$	90	66	57	45

Determine cuál es el nivel de aprovisionamiento óptimo según los diferentes criterios.

Solución: Como los valores de la tabla representan costes, todos los criterios de decisión estudiados anteriormente donde buscábamos la alternativa con mayor beneficio se transforman en elegir la alternativa con menor coste. Otra manera sencilla de resolver los problemas de decisión con costes es cambiar el signo a los resultados y actuar como en los problemas de decisión con beneficios.

Criterio de Laplace:

El principio de Laplace establece que e_1, e_2, e_3, e_4 tienen la misma probabilidad de suceder. Por consiguiente las probabilidades asociadas son $P(e_j)=1/4$ y los costes esperados para las distintas alternativas son:

E(a_1)=(15+30+54+75)/4=43,5

E(a_2)=(24+21+24+69)/4=**34,5**

E(a_3)=(63+54+36+63)/4=54

E(a_4)=(90+66+57+45)/4=64,5

		e_1(300)	*e_2(400)*	*e_3(600)*	*e_4(650)*	*Valor esperado*
Nivel de abastecimiento	*a_1(300)*	15	30	54	75	43,5
	a_2(400)	24	21	24	69	**34,5**
	a_3(600)	63	54	36	63	54
	a_4(650)	90	66	57	45	64,5

Por lo tanto, el mejor nivel de abastecimiento es el asociado al **menor coste esperado** (34,5) especificado por a_2.

Criterio de Wald:

		e_1(300)	*e_2(400)*	*e_3(600)*	*e_4(650)*	$\max_{e_j}\{x_{ij}\}$
Nivel de abastecimiento	*a_1(300)*	15	30	54	75	75
	a_2(400)	24	21	24	69	69
	a_3(600)	63	54	36	63	**63**
	a_4(650)	90	66	57	45	90

Este criterio pesimista basado en suponer que se dan las peores condiciones posibles, supondrá que para cada alternativa nos enfrentamos al mayor coste posible (nivel de seguridad), $\max_{e_j}\{x_{ij}\}$, y elegirá aquella alternativa (a_3) para la que este valor es el menor de todos, $\min_{a_i}\max_{e_j}\{x_{ij}\}=63$.

Criterio de Hurwicz:

Supongamos α=1/2, valor que suele tomarse cuando no se especifica un determinado índice de optimismo.

		$e_1(300)$	$e_2(400)$	$e_3(600)$	$e_4(650)$	Nivel de seguridad: $\max_{e_j}\{x_{ij}\}$	Nivel optimista: $\min_{e_j}\{x_{ij}\}$	*Media ponderada*
Nivel de abastecimiento	$a_1(300)$	15	30	54	75	75	15	**45**
	$a_2(400)$	24	21	24	69	69	21	**45**
	$a_3(600)$	63	54	36	63	63	36	49,5
	$a_4(650)$	90	66	57	45	90	45	67,5

Al tratarse de costes, elegiremos la alternativa asociada a la **menor media ponderada**. La solución óptima está dada por a_1 ó a_2.

Si el índice de optimismo es α=0,7, se multiplicaría por este valor el nivel optimista de cada alternativa y por 0,3=1-α su nivel de seguridad, resultando:

		$e_1(300)$	$e_2(400)$	$e_3(600)$	$e_4(650)$	Nivel de seguridad: $\max_{e_j}\{x_{ij}\}$	Nivel optimista: $\min_{e_j}\{x_{ij}\}$	*Media ponderada*
Nivel de abastecimiento	$a_1(300)$	15	30	54	75	75	15	**33**
	$a_2(400)$	24	21	24	69	69	21	35,4
	$a_3(600)$	63	54	36	63	63	36	44,1
	$a_4(650)$	90	66	57	45	90	45	58,5

En este caso la solución óptima es a_1.

Criterio de Savage:

La matriz de pérdidas de oportunidad, al tratarse de costes, se calcularía como:

$$p_{ij} = x_{ij} - \min_{a_k}\{x_{kj}\}$$

		$e_1(300)$	$e_2(400)$	$e_3(600)$	$e_4(650)$	$\max_{e_j}\{p_{ij}\}$
Nivel de abastecimiento	$a_1(300)$	0=15-15	9=30-21	30=54-24	30=75-45	30
	$a_2(400)$	9=24-15	0=21-21	0=24-24	24=69-45	**24**
	$a_3(600)$	48=63-15	33=54-21	12=36-24	18=63-45	48
	$a_4(650)$	75=90-15	45=66-21	33=57-24	0=45-45	75

La solución óptima está dada por a_2.

EJERCICIOS RESUELTOS

1. Una editorial tiene que decidir sobre el número de ejemplares que imprimirá de un nuevo libro de apuntes de una asignatura del Grado en Marketing e Investigación de Mercados. Simplificando el problema, se barajan tres posibles alternativas: 100, 200 y 300 ejemplares, esperándose una demanda de 100, 200 o 300 libros. La edición del libro tiene un coste general de 1000€ que se pagan al autor, independientemente del número de libros que se vendan, más un coste de 10€ por cada libro (papel, impresión, encuadernación,…). El libro se venderá en 25€/ejemplar.

a) Construya la tabla de decisión.

b) Si la editorial tiene un índice de optimismo del 80%, ¿cuántos libros debería imprimir? ¿cuál sería el beneficio ponderado para dicha decisión?

c) Basándose en la pérdida de oportunidad, ¿qué decisión debería tomar? ¿A qué perdida de oportunidad, como máximo, se arriesgaría?

SOLUCIÓN:

a)

Beneficios de la editorial: ↘	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros
Impresión de 100 libros	500	500	500
Impresión de 200 libros	-500	2000	2000
Impresión de 300 libros	-1500	1000	3500

Por ejemplo: 2000=-1000-(10x200)+(25x200)
-1500=-1000-(10x300)+(25x100)
1000=-1000-(10x300)+(25x200) …

b)

Beneficios de la editorial(€):	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros	$\min_{e_j}\{x_{ij}\}$	$\max_{e_j}\{x_{ij}\}$	Media ponderada (α=0,80)
Impresión de 100 libros	500	500	500	500	500	500
Impresión de 200 libros	-500	2000	2000	-500	2000	1500
Impresión de 300 libros	-1500	1000	3500	-1500	3500	2500

Decisión de imprimir 300 libros con un beneficio ponderado de 2500€.

c)

Pérdida de oportunidad(€):	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros	$\max_{e_j}\{p_{ij}\}$
Impresión de 100 libros	0	1500	3000	3000
Impresión de 200 libros	1000	0	1500	**1500**
Impresión de 300 libros	2000	1000	0	2000

Decisión de imprimir 200 libros con una pérdida de oportunidad máxima de 1500€.

2. Los dueños de un club deportivo está planteándose realizar obras en su estadio de cara a la próxima temporada, y han considerado la posibilidad de ampliar la capacidad del estadio, mejorar la calidad de los asientos, ampliar los accesos al estadio o no hacer obras. La evolución de la asistencia al estadio de cara a la próxima temporada podría ser alta, media o baja. En función de la evolución de la asistencia, las ganancias (en miles de €) que piensan que puede obtener en la próxima temporada con cada una de las alternativas descritas son las siguientes:

	Ampliar estadio	Reformar asientos	Ampliar accesos	No hacer obras
Asistencia alta	4200	3000	3300	2400
Asistencia media	1700	3100	2500	1800
Asistencia baja	700	1400	2100	1000

a) ¿Qué decisión tomarían los dueños utilizando el criterio de Laplace?

b) Si la tienda tiene un índice de optimismo del 70%, ¿qué tipo de obras debería acometer? ¿cuál sería el beneficio ponderado para dicha decisión?

c) Basándose en la pérdida de oportunidad, ¿qué decisión deberían tomar? ¿A qué perdida de oportunidad, como máximo, se arriesgaría?

SOLUCIÓN:

Beneficios próxima temporada (miles de €):	Asistencia alta	Asistencia media	Asistencia baja	Resultado esperado	$\min_{e_j}\{x_{ij}\}$	$\max_{e_j}\{x_{ij}\}$	Media ponderada (α=0,70)
Ampliar estadio	4200	1700	700	2200	700	4200	**3150**
Reformar asientos	3000	3100	1400	2500	1400	3100	2590
Ampliar accesos	3300	2500	2100	**2633.3333**	2100	3300	2940
No hacer obras	2400	1800	1000	1733.3333	1000	2400	1980

a) Decisión de solicitar ampliar el accesos con un beneficio esperado de 2633,3333 miles de euros.

b) Decisión de solicitar ampliar el estadio con un beneficio esperado de 3150 miles de €.

c)

Pérdida de oportunidad (miles de €):	Asistencia alta	Asistencia media	Asistencia baja	$\max_{e_j}\{p_{ij}\}$
Ampliar estadio	0	1400	1400	1400
Reformar asientos	1200	0	700	1200
Ampliar accesos	900	600	0	**900**
No hacer obras	1800	1300	1100	1800

Decisión de ampliar accesos con una pérdida de oportunidad máxima de 900 miles €.

9. Decisión en ambiente de riesgo.

9.1 El criterio del valor monetario esperado (VME).

Cada estado de la naturaleza puede considerarse un suceso. Como ya se dijo, los estados de la naturaleza son sucesos mutuamente excluyentes, es imposible que se presenten dos estados simultáneamente, y además constituyen una enumeración exhaustiva de todos los factores externos que pueden presentarse. Cuando un conjunto de sucesos cumplen con las anteriores propiedades, se dice que son una partición de un espacio muestral y sobre ellos se puede definir una probabilidad.

En el ambiente de riesgo cada alternativa lleva asociado un conjunto de consecuencias que dependerán del estado de la naturaleza y por tanto cada una de ellas tendrá asociada una probabilidad conocida.

Ejemplo 9.1. Suponga que tiene un negocio de ventas de pinos para Navidad y debe decidir cuántos pinos pedir para la próxima Navidad. Se debe pagar 5€ por cada árbol, se pueden pedir solo lotes de 100 y se planea venderlos a 9€ cada uno. Si no se venden, no tienen valor de recuperación. Se revisan las ventas pasadas, llegando a las siguientes estimaciones para la próxima Navidad:

Venta de pinos (e_j)	*Probabilidad*
200	0,4
300	0,4
400	0,2

Con estos datos pueden calcularse los beneficios para cada combinación de cantidad pedida y vendida. Por ejemplo, si se ordenan 300 pinos y se venden sólo 200, el beneficio será de 4€ por cada árbol vendido menos una pérdida de 5€ por cada árbol no vendido, es decir:

$$200x(9€-5€)-100x5€=800€-500€=300€$$

Haciendo esto para cada posibilidad, se obtienen los resultados mostrados en la tabla de decisión o matriz de pagos siguiente:

Probabilidad $p(e_j)$		0,4	0,4	0,2
		Demanda de árboles		
		$e_1(200)$	$e_2(300)$	$e_3(400)$
Nivel de abastecimiento	$a_1(200)$	800	800	800
	$a_2(300)$	300	1200	1200
	$a_3(400)$	-200	700	1600

El problema de decisión lo tenemos tabulado, y las diversas consecuencias valoradas numéricamente. Al desarrollarse en ambiente de riesgo, el decisor no conoce cuál es el verdadero estado de la naturaleza, pero si conoce la probabilidad de que lo sea cada uno de los estados considerados, es decir, conoce $p(e_1)$ $p(e_2)$ .....$p(e_n)$. Al ser los estados considerados exhaustivos y excluyentes ha de verificarse que $\sum_{j=1}^{n} p\left(e_j\right)=1$. En general, la tabla de decisión en ambiente de riesgo adopta la siguiente forma:

Probabilidad $p(e_j)$		$p(e_1)$	$p(e_2)$	. . .	$p(e_n)$
		Estados de la naturaleza			
		e_1	e_2	. . .	e_n
Alternativas	a_1	x_{11}	x_{12}	. . .	x_{1n}
	a_2	x_{21}	x_{22}	. . .	x_{2n}
	. . .	. . .	. . .	. . .	. . .
	a_m	x_{m1}	x_{m2}	. . .	x_{mn}

Como puede observarse la tabla de decisión es análoga a la de decisión en ambiente de incertidumbre, con la información añadida sobre las probabilidades de que se presente cada estado de la naturaleza.

En esta situación es evidente que adoptada la decisión a_i, el decisor puede obtener cualquiera de las recompensas que figuran en la i-ésima fila dependiendo de cuál sea el verdadero estado de la naturaleza: el decisor obtendrá la recompensa x_{i1} con probabilidad $p(e_1)$, la recompensa x_{i2} con probabilidad $p(e_2)$ y así sucesivamente.

Cuando tenemos un conjunto de cantidades numéricas, cada una de ellas con una probabilidad asociada, tomamos como representante de ese conjunto al valor medio o valor esperado. A cada alternativa le asociamos la esperanza matemática de los diversos valores monetarios a los que conduce (*Valor Monetario Esperado*):

$$VME(a_i) = x_{i1}p(e_1) + x_{i2}p(e_2) + \cdots + x_{in}p(e_n)$$

Según este criterio de decisión, la alternativa óptima será aquella que conduzca al **máximo valor monetario esperado.**

Ejemplo 9.2. Resolvamos mediante este criterio el ejemplo de los árboles de Navidad

Probabilidad p(ej)		0,4	0,4	0,2	
		Demanda de árboles			
		e1(200)	*e2(300)*	*e3(400)*	$VME(a_i)$
Nivel de abastecimiento	*a1(200)*	800	800	800	800
	a2(300)	300	1200	1200	**840**
	a3(400)	-200	700	1600	520

$$VME(a_1) = 0,4 \times 800 + 0,4 \times 800 + 0,2 \times 800 = 800$$
$$VME(a_2) = 0,4 \times 300 + 0,4 \times 1200 + 0,2 \times 1200 = 840$$
$$VME(a_3) = 0,4 \times (-200) + 0,4 \times 700 + 0,2 \times 1600 = 520$$

La alternativa óptima es la segunda: *pedir 300 pinos*.

9.1.1 Inconvenientes del criterio del valor monetario esperado.

Analicemos la situación sobre nuestro ejemplo. El dueño del negocio de árboles de Navidad adoptaba la decisión de pedir 300 árboles por ser la de mayor VME (840€), pero al tomar esa decisión pueden obtenerse bien 300€ o bien 1200€, según cuál sea el estado de la naturaleza, pero nunca 840€. El VME no representa el beneficio de una única decisión, es un promedio. Si adoptase esa decisión varias veces, cada una con una demanda posiblemente distinta, el vendedor de árboles obtendría un promedio de 840€ por cada vez. Desde esta perspectiva de **decisiones repetitivas** el **criterio del VME es adecuado**.

Pero si las decisiones no se toman de forma repetitiva, ¿sigue siendo válido el criterio de VME? Eso lo responderá el decisor, para lo cual habrá de contemplar otros aspectos de la distribución de probabilidad además de su media, como es la dispersión.

9.2 El criterio de la pérdida de oportunidad esperada (POE).

Cuando el problema de decisión se desarrolla en ambiente de riesgo o en ambiente de incertidumbre, en lugar de decidir en base a una matriz de recompensas o beneficios puede hacerse con una matriz de pérdidas de oportunidad, como se vio en el tema anterior.

La pérdida de oportunidad es lo que podría haberse ganado de más si se hubiera conocido el estado de la naturaleza. Fijado un estado de la naturaleza, la pérdida de oportunidad asociada a cada alternativa es el beneficio que ha dejado de obtenerse por no ser esa alternativa la mejor frente al estado fijado. En otras palabras, en qué medida podría haber sido mejor el resultado de mi alternativa. Por tanto, como ya definimos en el tema anterior, el concepto de **pérdida relativa** o **pérdida de oportunidad** p_{ij} asociada a un resultado x_{ij} se define como la diferencia entre el resultado de la mejor alternativa dado que e_j es el verdadero estado de la naturaleza y el resultado de la alternativa a_i bajo ese mismo estado e_j:

$$p_{ij} = \max_{a_k} \{x_{kj}\} - x_{ij}$$

Así, si el verdadero estado en que se presenta la naturaleza es e_j y el decisor elige la alternativa a_k que proporciona el máximo resultado x_{kj}, entonces no ha dejado de ganar nada, pero si elige otra alternativa cualquiera a_i , entonces obtendría como ganancia x_{ij} y dejaría de ganar $x_{kj} - x_{ij}$.

Ejemplo 9.3. En el ejemplo de los pinos de Navidad, su matriz de pérdidas de oportunidad sería:

Probabilidad p(ej)		0,4	0,4	0,2
		Demanda de árboles		
		e1(200)	*e2(300)*	*e3(400)*
Nivel de abastecimiento	*a1(200)*	800-800=0	1200-800=400	1600-800=800
	a2(300)	800-300=500	1200-1200=0	1600-1200=400
	a3(400)	800-(-200)=1000	1200-700=500	1600-1600=0

Salvo por la interpretación de las cantidades contenidas en la tabla, antes eran beneficios y ahora son pérdidas de oportunidad, la situación es la misma: a cada alternativa le corresponden varias cantidades, cada una con una probabilidad asociada. De nuevo, como representante de todas ellas consideramos su media, que ahora representa la pérdida de oportunidad esperada. **La alternativa óptima será la que conduzca a la menor pérdida de oportunidad esperada**.

Probabilidad $p(e_j)$		0,4	0,4	0,2	
		Demanda de árboles			
		$e_1(200)$	$e_2(300)$	$e_3(400)$	$POE(a_i)$
Nivel de abastecimiento	$a_1(200)$	0	400	800	320
	$a_2(300)$	500	0	400	**280**
	$a_3(400)$	1000	500	0	600

$$POE(a_1) = 0,4 \times 0 + 0,4 \times 400 + 0,2 \times 800 = 320$$
$$POE(a_2) = 0,4 \times 500 + 0,4 \times 0 + 0,2 \times 400 = 280$$
$$POE(a_3) = 0,4 \times 1000 + 0,4 \times 500 + 0,2 \times 0 = 600$$

La acción óptima según este criterio es a_2: *pedir 300 pinos*.

Obsérvese que la decisión óptima por el criterio de VME y la decisión óptima por el criterio de POE coinciden. Esta coincidencia se dará siempre: maximizar el VME es equivalente a minimizar la POE.

9.3 Valor monetario esperado con información perfecta (VMEIP).

Como hemos visto, el decisor incurre en pérdidas de oportunidad por tener que adoptar una alternativa desconociendo el verdadero estado de la naturaleza. Si pudiese disponer de esa información ¿qué beneficios adicionales le proporcionaría?, o lo que es lo mismo, ¿cuánto estaría dispuesto a pagar por ella?

Concretando sobre el problema de los árboles de navidad: si el decisor supiese que el verdadero estado es e_1 (200 árboles) elegiría a_1 (200 árboles) y obtendría una beneficio de 800€; si el verdadero estado fuese e_2 (300 árboles) elegiría a_2 (300 árboles) y obtendría 1200€ y si fuese e_3 (400 árboles) adoptaría a_3 (400 árboles) y ganaría 1600€. A priori, a la hora de pagar por la información, no se sabe cuál es el verdadero estado de la naturaleza, pero si conoce la probabilidad que cada uno tiene de serlo, por tanto el decisor puede calcular el valor monetario esperado disponiendo de información perfecta (será la media ponderada de los máximos de las columnas):

$$VMEIP = 0,4 \times 800 + 0,4 \times 1200 + 0,2 \times 1600 = 1120$$

9.3.1 El valor de la información perfecta (VIP).

No disponiendo de información perfecta el máximo valor monetario esperado que puede lograr el decisor es 840 como se vio en el ejemplo 9.2.

Por tanto, el beneficio adicional que proporciona la información perfecta es:

$$VIP = VMEIP - VME(máximo) = 1120 - 840 = 280$$

Evidentemente:

a) Esta sería la máxima cantidad que un decisor racional pagaría por disponer de información perfecta.

b) Cualquier otra información no perfecta valdrá menos (como veremos en el próximo tema).

El valor de la información perfecta (VIP) también conocido como valor esperado de la información perfecta es un valor esperado, es decir, una media de lo que vale esa información perfecta bajo cada posible estado de la naturaleza.

VIP también podría haberse calculado a partir de la pérdida de oportunidad. En efecto, con información perfecta nunca se incurre en pérdidas de oportunidad, por lo que la pérdida de oportunidad esperada con información perfecta (POEIP) es nula.

Sin información perfecta la mínima POE en nuestro ejemplo de los pinos es 280 (ejemplo 9.3), por tanto: $VIP = POE(mínima) - POEIP = 280 - 0 = 280$

Ejemplo 9.4. El Sr. Ramírez ha heredado 100000€ y ha decidido invertir su dinero. Un asesor de inversiones le ha sugerido cinco inversiones posibles: oro, bonos, negocio en desarrollo, depósitos y acciones. El heredero debe decidir en qué opción invertir. La siguiente tabla representa los beneficios que obtendría para cada posible comportamiento del mercado. Calcule el valor máximo que debería pagar por un estudio donde le aseguren la ocurrencia de un determinado estado de la naturaleza.

Probabilidad		0,25	0,25	0,2	0,2	0,1	
		Estados de la naturaleza (comportamiento del mercado)					
		Gran alza	*Pequeña alza*	*Sin cambios*	*Pequeña baja*	*Gran baja*	$VME(a_i)$
Alternativas	Oro	-2000	2000	*4000*	*6000*	0	2000
	Bonos	5000	4000	3000	-2000	-3000	**2150**
	Negocio	*10000*	*5000*	2000	-4000	-12000	**2150**
	Depósitos	1200	1200	1200	1200	*1200*	1200
	Acciones	4000	3000	3000	-4000	-3000	1250

Primero se calcula el Valor Monetario Esperado para cada alternativa de decisión:

$$VME(Oro) = -2000 \times 0,25 + 2000 \times 0,25 + 4000 \times 0,2 + 6000 \times 0,2 + 0 \times 0,1 = 2000$$

$$VME(Bonos) = 5000 \times 0,25 + 4000 \times 0,25 + 3000 \times 0,2 - 2000 \times 0,2 - 3000 \times 0,1 = 2150$$

...

El Valor Monetario Esperado Máximo es 2150€.

Para calcular el Valor Monetario Esperado disponiendo de información perfecta, se eligen los máximos de cada columna y se calcula su media ponderada:

$$VMEIP = 10000 \times 0,25 + 5000 \times 0,25 + 4000 \times 0,2 + 6000 \times 0,2 + 1200 \times 0,1 = 5870$$

Por tanto el beneficio adicional que proporciona la información perfecta es

$$VIP = VMEIP - VME(máximo) = 5870 - 2150 = 3720$$

Si puede obtener información que le asegure la ocurrencia de un determinado estado de la naturaleza, podría pagar por ella hasta un *máximo* de 3720€.

EJERCICIOS RESUELTOS

1. Una editorial tiene que decidir sobre el número de ejemplares que imprimirá de un nuevo libro de apuntes de una asignatura del Grado en Marketing e Investigación de Mercados. Simplificando el problema, se barajan tres posibles alternativas: 100, 200 y 300 ejemplares, esperándose una demanda de 100, 200 o 300 libros. La edición del libro tiene un coste general de 1000€ que se pagan al autor, independientemente del número de libros que se vendan, más un coste de 10€ por cada libro (papel, impresión, encuadernación,…). El libro se venderá en 25€/ejemplar. Siendo la tabla de decisión (véanse ejercicios resueltos del tema 8):

Beneficios de la editorial: ↘	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros
Impresión de 100 libros	500	500	500
Impresión de 200 libros	-500	2000	2000
Impresión de 300 libros	-1500	1000	3500

En los últimos diez años la demanda de libros de apuntes en dicho Grado ha sido:

Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros
1	6	3

a) Teniendo en cuenta dicha información, ¿Qué decisión debería tomar la editorial y cuál sería el beneficio esperado para tal decisión?

b) Valor esperado de la información perfecta.

SOLUCIÓN:

a)

$p(e_j)$	*0,1*	*0,6*	*0,3*	
Beneficios de la editorial (€):	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros	$VME(a_i)$
Impresión de 100 libros	500	500	500	500
Impresión de 200 libros	-500	2000	2000	**1750**
Impresión de 300 libros	-1500	1000	3500	1500

Decisión de imprimir 200 libros con un beneficio esperado de 1750€.

b)

$VMEIP = (0{,}1 \times 500) + (0{,}6 \times 2000) + (0{,}3 \times 3500) = 2300$

$VIP = VMEIP - VME(máximo) = 2300 - 1750 = 550$

2. Los dueños de un club deportivo está planteándose realizar obras en su estadio de cara a la próxima temporada, y han considerado la posibilidad de ampliar la capacidad del estadio, mejorar la calidad de los asientos, ampliar los accesos al estadio o no hacer obras. La evolución de la asistencia al estadio de cara a la próxima temporada podría ser alta, media o baja. En función de la evolución de la asistencia, las ganancias (en miles de €) que piensan que puede obtener en la próxima temporada con cada una de las alternativas descritas son las siguientes:

	Ampliar estadio	Reformar asientos	Ampliar accesos	No hacer obras
Asistencia alta	4200	3000	3300	2400
Asistencia media	1700	3100	2500	1800
Asistencia baja	700	1400	2100	1000

La probabilidad estimada de que la asistencia sea alta, media o baja es, respectivamente, 0,2; 0,5 y 0,3.

a) Teniendo en cuenta dicha información, ¿Cuál es la mejor decisión que podrían adoptar los dueños del club? ¿Cuáles serían los beneficios esperados?

b) Valor esperado de la información perfecta.

SOLUCIÓN:

a)

$p(e_j)$	*0,2*	*0,5*	*0,3*	
Beneficios próxima temporada (miles de €):	Asistencia alta	Asistencia media	Asistencia baja	$VME(a_i)$
Ampliar estadio	4200	1700	700	1900
Reformar asientos	3000	3100	1400	**2570**
Ampliar accesos	3300	2500	2100	2540
No hacer obras	2400	1800	1000	1680

Decisión de reformar los asientos con un beneficio esperado de 2570€.

b)

$VMEIP = (0{,}2 \times 4200) + (0{,}5 \times 3100) + (0{,}3 \times 2100) = 3020$

$VIP = VMEIP - VME(máximo) = 3020 - 2570 = 450$

10. Decisión bayesiana.

10.1 Probabilidad condicionada. Probabilidad total. Teorema de Bayes.

Sean A y B dos sucesos y se sabe seguro que ha ocurrido A, ¿cómo afecta esto a la probabilidad de B?

Notamos por $P\left(B/A\right)$ a la probabilidad de B condicionada a A, es decir a la probabilidad de B una vez que sabemos seguro que ha ocurrido A.

Se define la **probabilidad de *B* condicionada a *A*** como:

$$P\left(B/A\right)=\frac{P\left(A\cap B\right)}{P\left(A\right)} \qquad \text{, siendo } P\left(A\right)>0\,.$$

Obsérvese que a partir de esta definición: $P\left(A\cap B\right)=P\left(A\right)P\left(B/A\right)$

Puede ocurrir que la probabilidad de B no se vea alterada por la ocurrencia de A, es decir, que $P\left(B/A\right)=P\left(B\right)$. En este caso se dice que los sucesos A y B son independientes, y se tiene que: $P\left(A\cap B\right)=P\left(A\right)P\left(B\right)$

Teorema de la Probabilidad Total.

Sea $A_1,\ A_2, \ldots A_n$ una partición del espacio muestral $\left(A_i\cap A_j=\varnothing,\ \ \bigcup_{i=1}^{n}A_i=\Omega\right)$, entonces para cualquier suceso B se tiene que:

$$P\left(B\right)=P\left(B/A_1\right)P\left(A_1\right)+P\left(B/A_2\right)P\left(A_2\right)+\ldots+P\left(B/A_n\right)P\left(A_n\right)=\sum_{i=1}^{n}P\left(B/A_i\right)P\left(A_i\right)$$

Teorema de Bayes

Sea $A_1,\ A_2, \ldots A_n$ una partición del espacio muestral y sea B un suceso cualquiera con probabilidad estrictamente mayor que cero. Entonces:

$$P\left(A_k/B\right)=\frac{P\left(B/A_k\right)P\left(A_k\right)}{\sum_{i=1}^{n}P\left(B/A_i\right)P\left(A_i\right)}=\frac{P\left(B/A_k\right)P\left(A_k\right)}{P(B)}$$

En nuestro caso, la partición estará constituida por el conjunto de estados de la naturaleza, $e_1, e_2, ..., e_n$ y B representará la información que tenemos acerca de ellos.

10.2 Interpretaciones del concepto de probabilidad.

Sobre el significado que tiene la probabilidad de un suceso hay una fuerte controversia.

El primer concepto de probabilidad surge en el contexto de los juegos de azar, donde se considera que todos los resultados posibles son equiprobables. Laplace definió la probabilidad de un suceso como el cociente entre el número de resultados favorables a dicho suceso, y el número total de resultados posibles.

Este ***concepto clásico de la probabilidad*** es escasamente aplicable, pues en la mayoría de las circunstancias inciertas es imposible clasificar los posibles resultados en clases igualmente probables. Es fácil aplicarlo al lanzamiento de un dado, pero ya no lo es tanto para determinar la probabilidad de que mañana suba el índice de la bolsa. Incluso no es aplicable cuando el dado está trucado.

Otra ***interpretación*** es la ***frecuentista*** según la cual la probabilidad sólo tiene sentido en el marco de un experimento infinitamente repetible. La probabilidad de un suceso será la frecuencia relativa de ocurrencia del suceso en infinidad de pruebas repetidas de ese experimento.

Según esta interpretación, obsérvese que la probabilidad sólo podrá definirse sobre sucesos que estén ligados a los resultados de un experimento repetible. Por ejemplo podrá aplicarse al lanzamiento de un dado sin necesidad de suponer, de antemano, que el dado no está cargado, puesto que ese dado puede lanzarse indefinidamente, pero no puede aplicarse a sucesos tales como el resultado de las próximas elecciones generales, pues éstas se desarrollarán en unas circunstancias irrepetibles.

La tercera ***interpretación*** es la ***subjetiva***, según la cual la probabilidad representa el grado de creencia del observador sobre el estado que adoptará el sistema. Según esta interpretación, en la Teoría de la Decisión, la probabilidad de un estado de la naturaleza representa el grado de creencia del decisor en que acontezca ese estado, lo que dependerá de la información que tenga. Cuanto mayor sea su creencia mayor será la probabilidad. Obsérvese que admitiendo esta interpretación:

1.- Puede definirse la probabilidad de prácticamente cualquier suceso, sin necesidad de que se produzca en el seno de un experimento repetible, ahora sí que puede asignarse una probabilidad a los resultados de las próximas elecciones generales pues el observador dispondrá de unas creencias sobre los mismos (creencias que estarán basadas en unas informaciones).

2.- Como cada uno tiene sus propias creencias, dos observadores distintos pueden asignar probabilidades distintas al mismo suceso. Por tanto la probabilidad es subjetiva (no objetiva), es más propia del sujeto que observa que del suceso (objeto) observado.

De inmediato surge una cuestión: imaginemos un suceso que se produce en el seno de un experimento repetible, y al que por tanto se le puede asignar una probabilidad frecuentista (objetiva) e imaginemos un observador racional ¿qué relación existe entre esas dos probabilidades? Inicialmente el decisor podría asignarle una probabilidad que nada tuviera que ver con la frecuencia relativa, pero a medida que va observando los resultados de las sucesivas repeticiones, el observador va revisando sus creencias, y su probabilidad subjetiva se irá aproximando a la objetiva del suceso.

10.3 Modificación de las creencias del decisor.

Las creencias del decisor no son fijas, sino que cambian según asimila nuevos hechos, informaciones, evidencias, ... ¿Cómo debe revisar sus creencias un decisor racional a la luz de una nueva información?

La información extra, aquella de la que no disponía y ahora se incorpora, se expresa mediante la ocurrencia de un suceso *B*. Antes el decisor tenía unas creencias y después de disponer de esta información tendrá otras, nuestro objetivo consiste en obtener estas nuevas creencias (probabilidades) a partir de aquellas.

La probabilidad condicionada representa perfectamente las nuevas creencias del decisor y se puede utilizar para revisar las probabilidades o creencias iniciales. Suele decirse que $P\left(e_j\right)$ es la ***probabilidad a priori*** (antes de saber que *B* ha ocurrido) y $P\left(e_j / B\right)$ es la ***probabilidad a posteriori*** (después de saber que ha ocurrido *B*).

El teorema de Bayes proporciona la herramienta para obtener las probabilidades a posteriori:

Suponga que el decisor está considerando una tabla de decisión en la que aparecen los estados e_1, e_2,...,e_n. Sus creencias a priori están reflejadas en las probabilidades $P(e_1)$, $P(e_2)$,...,$P(e_n)$, y si tuviese que tomar una decisión, con ellas calcularía los valores monetarios esperados que le indicarían la acción a tomar.

Pero si antes de tomar la decisión ha ocurrido un suceso *B* (se tiene nueva información adicional), entonces el decisor deberá construir sus probabilidades a posteriori $P(e_1/B)$,...., $P(e_n/B)$ para calcular con ellas los valores monetarios esperados. Estas probabilidades a posteriori vendrán dadas por el teorema de Bayes, según:

$$P\left(\frac{e_j}{B}\right)=\frac{P\left(\frac{B}{e_j}\right)P(e_j)}{P(B)}=\frac{P\left(\frac{B}{e_j}\right)P(e_j)}{\sum_{i=1}^{n}P\left(\frac{B}{e_i}\right)P(e_i)} \qquad j=1,...,n$$

Obsérvese que para aplicar este teorema hace falta calcular (o conocer) previamente, las probabilidades $P(B/e_j)$, $j= 1, 2, ..., n$, es decir la probabilidad de ocurrencia del suceso B, bajo cada uno de los estados posibles de la naturaleza. A estas probabilidades se les llama ***verosimilitudes***. Normalmente son conocidas o fáciles de obtener.

He aquí la importancia del teorema de Bayes: ***describe cómo debemos ir aprendiendo de la experiencia, cómo deben modificarse nuestras creencias, expresadas mediante probabilidades, al incorporar nueva información.***

10.4 Valor monetario esperado con información imperfecta. Valor de la información imperfecta.

La información adicional que se incorpora no siempre es perfecta, muchas veces los estudios adicionales que se encargan tienen cierto margen de error. La información adicional (imperfecta) mejora el conocimiento sobre la probabilidad de ocurrencia de los estados de la naturaleza (sin llegar a asegurarnos qué estado va a ocurrir), y ese mejor conocimiento ayuda a tomar mejores decisiones.

Ejemplo 10.1. Una comunidad de vecinos desea incorporar una red local de internet en su edificio. Recibe los presupuestos de dos empresas instaladoras: el de la empresa A por 10000€ y el de la empresa B por 7500€. No obstante, la existencia de antenas colectivas en la inmediaciones puede provocar interferencias, en el caso de que se produzcan sería necesario añadir un nuevo aparato que generaría unos costes adicionales de 4000€ de los cuales la empresa A sólo incluiría el 20% de los mismos en el presupuesto, mientras que la empresa B los incluiría en su totalidad.

Después de tratar de recabar información sobre las posibilidades de que haya interferencias, el presidente de la comunidad no logra obtenerla, por lo que considera que ambas situaciones son equiprobables. ¿Cuál es la mejor alternativa para la comunidad de vecinos?

La tabla de decisión (gastos) con los Valores Monetarios Esperados (VME) es:

Probabilidades	0,5	0,5	
ESTADOS → Alternativas ↓	NO HAY INTERFERENCIAS	SÍ HAY INTERFERENCIAS	VME
Empresa A	10000	10800	10400
Empresa B	7500	11500	9500

$10400 = (10000 \times 0,5) + (10800 \times 0,5)$ $9500 = (7500 \times 0,5) + (11500 \times 0,5)$

La empresa escogida sería la B y el gasto medio asociado sería de 9500 euros.

Supongamos ahora que el presidente de la comunidad puede contratar un experto, quien por 100 euros realizaría unas mediciones. La probabilidad de que el experto acierte cuando no hay interferencias es de 0,9 y de que no acierte cuando hay interferencias es de 0,2 (obsérvese que la información que se pagaría no es perfecta en cuanto a total seguridad en las afirmaciones del experto).

¿Cuál sería entonces la decisión a adoptar por la comunidad de vecinos?

La primera decisión que debe tomar el presidente es si contratar o no un experto. Ya sabemos que sin experto la decisión sería contratar a la empresa B con un gasto medio asociado de 9500 euros. Por tanto, debemos comprobar si la contratación de un experto conlleva más gastos o no.

Las **probabilidades de acierto y fallo del experto** son, según los estados de la naturaleza

ESTADOS → Informe del experto ↓	NO HAY INTERFERENCIAS	SÍ HAY INTERFERENCIAS
No hay interferencias	0,9	0,2
Sí hay interferencias	0,1	0,8

Con esta información podemos actualizar y modificar las probabilidades de los estados de la naturaleza (sin interferencias, con interferencias).

Notaremos: $e_1 = NO\ HAY\ INTERFERENCIAS$ y $e_2 = SÍ\ HAY\ INTERFERENCIAS$

$c_1 = Informa:\ no\ hay\ interferencias$ y $c_2 = Informa:\ sí\ hay\ interferencias$

La anterior tabla recoge las siguientes probabilidades condicionadas (verosimilitudes):

Verosimilitudes		
ESTADOS → Informe del experto ↓	e_1 NO HAY INTERFERENCIAS	e_2 SÍ HAY INTERFERENCIAS
c_1 No hay interferencias	$P\left(c_1/e_1\right)=0,9$	$P\left(c_1/e_2\right)=0,2$
c_2 Sí hay interferencias	$P\left(c_2/e_1\right)=0,1$	$P\left(c_2/e_2\right)=0,8$
SUMA:	1	1

A partir de las *verosimilitudes* y de las ***probabilidades a priori***, $P\left(e_1\right)=0,5$ y $P\left(e_2\right)=0,5$, con la ayuda del teorema de Bayes obtenemos las **probabilidades a posteriori**:

$$P\left(e_1/c_1\right)=\frac{P\left(c_1/e_1\right)P(e_1)}{\sum_{j=1}^{2}P\left(c_1/e_j\right)P(e_j)}=\frac{0,9\times0,5}{\left(0,9\times0,5\right)+\left(0,2\times0,5\right)}=0,81818$$

$$P\left(e_2/c_1\right)=\frac{P\left(c_1/e_2\right)P(e_2)}{\sum_{j=1}^{2}P\left(c_1/e_j\right)P(e_j)}=\frac{0,2\times0,5}{\left(0,9\times0,5\right)+\left(0,2\times0,5\right)}=0,18182$$

$$P\left(e_1/c_2\right)=\frac{P\left(c_2/e_1\right)P(e_1)}{\sum_{j=1}^{2}P\left(c_2/e_j\right)P(e_j)}=\frac{0,1\times0,5}{\left(0,1\times0,5\right)+\left(0,8\times0,5\right)}=0,11111$$

$$P\left(e_2/c_2\right)=\frac{P\left(c_2/e_2\right)P(e_2)}{\sum_{j=1}^{2}P\left(c_2/e_j\right)P(e_j)}=\frac{0,8\times0,5}{\left(0,1\times0,5\right)+\left(0,8\times0,5\right)}=0,88889$$

Una forma cómoda de ordenar los anteriores cálculos mediante una tabla es:

Informe del experto: No hay interferencias (c_1)			
	e_1 NO HAY INTERFERENCIAS	e_2 SÍ HAY INTERFERENCIAS	suma
$P\left(e_j\right)$	**0,5**	**0,5**	**1**
$P\left(c_1/e_j\right)$	0,9	0,2	
$P\left(c_1/e_j\right)P\left(e_j\right)$	0,45	0,10	0,55
$P\left(e_j/c_1\right)=\frac{P\left(c_1/e_j\right)P\left(e_j\right)}{P\left(c_1\right)}$	**0,81818**	**0,18182**	**1**

$$P(c_1)=\sum_{j=1}^{n}P\left(c_1/e_j\right)P\left(e_j\right)=0,55$$

Informe del experto: Sí hay interferencias (c_2)			
	e_1 NO HAY INTERFERENCIAS	e_2 SÍ HAY INTERFERENCIAS	suma
$P(e_j)$	**0,5**	**0,5**	**1**
$P(c_2/e_j)$	0,1	0,8	
$P(c_2/e_j)P(e_j)$	0,05	0,40	0,45
$P(e_j/c_2)=\frac{P(c_2/e_j)P(e_j)}{P(c_2)}$	**0,11111**	**0,88889**	**1**

$$P(c_2)=\sum_{j=1}^{n}P(c_2/e_j)P(e_j)=0,45$$

Con estas nuevas probabilidades se actualizan las probabilidades sobre los estados de la naturaleza y la decisión óptima:

Informe del experto: No hay interferencias (c_1)			
$P\left(\frac{e_j}{c_1}\right)\rightarrow$	**0,81818**	**0,18182**	
ESTADOS → Alternativas ↓	e_1 NO HAY INTERFERENCIAS	e_2 SÍ HAY INTERFERENCIAS	VME/ c_1
Empresa A	10000	10800	10145,456
Empresa B	7500	11500	**8227,28**

$$10145,456=(10000\times 0,81818)+(10800\times 0,18182)$$

$$8227,28=(7500\times 0,81818)+(11500\times 0,18182)$$

Si el informe del experto es que no hay interferencias, la empresa escogida sería la B y el gasto medio asociado sería de 8227,28 euros.

Informe del experto: Sí hay interferencias (c_2)			
$P\left(\frac{e_j}{c_2}\right)\rightarrow$	**0,11111**	**0,88889**	
ESTADOS → Alternativas ↓	e_1 NO HAY INTERFERENCIAS	e_2 SÍ HAY INTERFERENCIAS	VME/ c_2
Empresa A	10000	10800	**10711,11**
Empresa B	7500	11500	11055,56

$$10711{,}112 = (10000 \times 0{,}11111) + (10800 \times 0{,}88889)$$

$$11055{,}56 = (7500 \times 0{,}11111) + (11500 \times 0{,}88889)$$

Si el informe del experto es que sí hay interferencias, la empresa escogida sería la A y el gasto medio asociado sería de 10711,11 euros.

El gasto esperado con la información adicional del experto depende de cuál sea esta.

El experto dirá que no hay interferencias (c_1) con una probabilidad:

$$P(c_1) = \sum_{j=1}^{2} P\left(c_1/e_j\right)P\left(e_j\right) = (0{,}9 \times 0{,}5) + (0{,}2 \times 0{,}5) = 0{,}55$$

El experto dirá que si hay interferencias (c_2) con una probabilidad:

$$P(c_2) = \sum_{j=1}^{2} P\left(c_2/e_j\right)P\left(e_j\right) = (0{,}1 \times 0{,}5) + (0{,}8 \times 0{,}5) = 0{,}45$$

Evidentemente, $\sum_{k=1}^{2} P\left(c_k\right) = 1$.

El gasto esperado medio con la información adicional del experto es (en general llamado: ***valor monetario esperado con información imperfecta***):

$$\left(8227{,}28 \times P\left(c_1\right)\right) + \left(10711{,}11 \times P\left(c_2\right)\right) = \left(8227{,}28 \times 0{,}55\right) + \left(10711{,}11 \times 0{,}45\right) = 9345$$

Y sin ella es 9500, luego estaríamos dispuestos a pagar:

9500-9345=155 euros

Cantidad que denominamos ***valor de la información imperfecta, VII***.

Por tanto se decidiría contratar al experto, pues su información que nos costaría 100 euros conduciría a un ahorro mayor, de 155 euros.

Cuando, en lugar de costes, trabajemos con beneficios en la tabla de decisión, el valor de la información imperfecta se calculará restando al valor monetario esperado con información imperfecta (que será mayor) el valor monetario esperado sin dicha información (que será menor).

EJERCICIOS RESUELTOS

1. Una editorial tiene que decidir sobre el número de ejemplares que imprimirá de un nuevo libro de apuntes de una asignatura del Grado en Marketing e Investigación de Mercados. Simplificando el problema, se barajan tres posibles alternativas: 100, 200 y 300 ejemplares, esperándose una demanda de 100, 200 o 300 libros. La edición del libro tiene un coste general de 1000€ que se pagan al autor, independientemente del número de libros que se vendan, más un coste de 10€ por cada libro (papel, impresión, encuadernación,…). El libro se venderá en 25€/ejemplar. Siendo la tabla de decisión (véanse ejercicios resueltos del tema 8):

Beneficios de la editorial: ↘	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros
Impresión de 100 libros	500	500	500
Impresión de 200 libros	-500	2000	2000
Impresión de 300 libros	-1500	1000	3500

En los últimos diez años la demanda de libros de apuntes en dicho Grado ha sido:

Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros
1	6	3

Teniendo en cuenta dicha información, el beneficio esperado para la decisión óptima, *impresión de 200 libros*, es 1750€ (véanse ejercicios resueltos del tema 9).

Un grupo de alumnos de TC3 ofrecen a la editorial recabar información entre sus compañeros para saber si la demanda superará o no los 250 ejemplares.

Cuando la demanda real vaya a ser de 100 o de 200 libros, la probabilidad de que la información aportada por los alumnos de TC3 sea que no se superarán los 250 libros es del 80%, mientras que la probabilidad de que afirmen que la demanda superará los 250 libros cuando la demanda real vaya a ser de 300 libros es del 90%.

a) Obtenga bajo cada supuesto las probabilidades a posteriori para los estados de la naturaleza.
b) ¿Cuál es el máximo precio que podrían pedir los alumnos de TC3 a la editorial por facilitarle la información?

SOLUCIÓN:

a)

Informe de los alumnos: Demanda inferior a 250 (c_1)				
	e_1 Demanda de 100 libros	e_2 Demanda de 200 libros	e_3 Demanda de 300 libros	suma
$P(e_j)$	**0,1**	**0,6**	**0,3**	**1**
$P(c_1/e_j)$	0,8	0,8	0,1	
$P(c_1/e_j)P(e_j)$	0,08	0,48	0,03	$P(c_1)=0,59$
$P(e_j/c_1)=\frac{P(c_1/e_j)P(e_j)}{P(c_1)}$	**0,1356**	**0,8136**	**0,0508**	**1**

Informe de los alumnos: Demanda superior a 250 (c_2)				
	e_1 Demanda de 100 libros	e_2 Demanda de 200 libros	e_3 Demanda de 300 libros	suma
$P(e_j)$	**0,1**	**0,6**	**0,3**	**1**
$P(c_2/e_j)$	0,2	0,2	0,9	
$P(c_2/e_j)P(e_j)$	0,02	0,12	0,27	$P(c_2)=0,41$
$P(e_j/c_2)=\frac{P(c_2/e_j)P(e_j)}{P(c_2)}$	**0,0488**	**0,2927**	**0,6585**	**1**

b)

Informe de los alumnos: Demanda inferior a 250 (c_1)				
$P\left(\frac{e_j}{c_1}\right)$	*0,1356*	*0,8136*	*0,0508*	
Beneficios de la editorial (€):	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros	$VME(a_i/c_1)$
Impresión de 100 libros	500	500	500	500
Impresión de 200 libros	-500	2000	2000	**1661,02**
Impresión de 300 libros	-1500	1000	3500	788,14

Informe de los alumnos: Demanda superior a 250 (c_2)				
$P\left(e_j / c_2\right)$	*0,0488*	*0,2927*	*0,6585*	
Beneficios de la editorial (€):	Demanda de 100 libros	Demanda de 200 libros	Demanda de 300 libros	$VME\left(a_i / c_2\right)$
Impresión de 100 libros	500	500	500	500
Impresión de 200 libros	-500	2000	2000	1878,05
Impresión de 300 libros	-1500	1000	3500	**2524,39**

$$P(c_1)=\sum_{j=1}^{3} P\left(c_1 / e_j\right)P\left(e_j\right)=0,59 \qquad P(c_2)=\sum_{j=1}^{3} P\left(c_2 / e_j\right)P\left(e_j\right)=0,41$$

$$VMEII=\left(0,59\times 1661,02\right)+\left(0,41\times 2524,39\right)=2015 \qquad VII=2015-1750=265$$

2. Los dueños de un club deportivo está planteándose realizar obras en su estadio de cara a la próxima temporada, y han considerado la posibilidad de ampliar la capacidad del estadio, mejorar la calidad de los asientos, ampliar los accesos al estadio o no hacer obras. La evolución de la asistencia al estadio de cara a la próxima temporada podría ser alta, media o baja. En función de la evolución de la asistencia, las ganancias (en miles de €) que piensan que puede obtener en la próxima temporada con cada una de las alternativas descritas son las siguientes:

	Ampliar estadio	Reformar asientos	Ampliar accesos	No hacer obras
Asistencia alta	4200	3000	3300	2400
Asistencia media	1700	3100	2500	1800
Asistencia baja	700	1400	2100	1000

La probabilidad estimada de que la asistencia sea alta, media o baja es, respectivamente, 0,2; 0,5 y 0,3.

Supongamos que los dueños del club pueden recurrir a un estudio de mercado que le informe acerca de la evolución futura de la asistencia al estadio. La probabilidad de que si la asistencia va a ser alta o media, este hecho haya sido pronosticado por el estudio de mercado, es de 0,8, repartiéndose por igual la probabilidad de error en las otras dos posibilidades. Si la asistencia va a ser baja, es seguro que lo habrá pronosticado así.

a) Obtenga bajo cada supuesto las probabilidades a posteriori para los estados de la naturaleza.

b) ¿Cuál es el máximo precio que estarían dispuestos a pagar los dueños del club por la información contenida en el estudio de mercado?

SOLUCIÓN:

a)

Estudio de mercado: Asistencia alta (c_1)				
	e_1 Asistencia alta	e_2 Asistencia media	e_3 Asistencia baja	suma
$P(e_j)$	**0,2**	**0,5**	**0,3**	**1**
$P(c_1/e_j)$	0,8	0,1	0	
$P(c_1/e_j)P(e_j)$	0,16	0,05	0	$P(c_1)=0,21$
$P(e_j/c_1)=\frac{P(c_1/e_j)P(e_j)}{P(c_1)}$	**0,7619**	**0,2381**	**0**	**1**

Estudio de mercado: Asistencia media (c_2)				
	e_1 Asistencia alta	e_2 Asistencia media	e_3 Asistencia baja	suma
$P(e_j)$	**0,2**	**0,5**	**0,3**	**1**
$P(c_2/e_j)$	0,1	0,8	0	
$P(c_2/e_j)P(e_j)$	0,02	0,4	0	$P(c_2)=0,42$
$P(e_j/c_2)=\frac{P(c_2/e_j)P(e_j)}{P(c_2)}$	**0,0476**	**0,9524**	**0**	**1**

Estudio de mercado: Asistencia baja (c_3)				
	e_1 Asistencia alta	e_2 Asistencia media	e_3 Asistencia baja	suma
$P(e_j)$	**0,2**	**0,5**	**0,3**	**1**
$P(c_3/e_j)$	0,1	0,1	1	
$P(c_3/e_j)P(e_j)$	0,02	0,05	0,3	$P(c_3)=0,37$
$P(e_j/c_3)=\frac{P(c_3/e_j)P(e_j)}{P(c_3)}$	**0,0541**	**0,1351**	**0,8108**	**1**

b)

Estudio de mercado: Asistencia alta (c_1)				
$P\left(e_j / c_1\right)$	0,7619	0,2381	0	
Beneficios próxima temp. (miles de €):	Asistencia alta	Asistencia media	Asistencia baja	$VME\left(a_i / c_1\right)$
Ampliar estadio	4200	1700	700	**3604,76**
Reformar asientos	3000	3100	1400	3023,81
Ampliar accesos	3300	2500	2100	3109,52
No hacer obras	2400	1800	1000	2257,14

Estudio de mercado: Asistencia media (c_2)				
$P\left(e_j / c_2\right)$	0,0476	0,9524	0	
Beneficios próxima temp. (miles de €):	Asistencia alta	Asistencia media	Asistencia baja	$VME\left(a_i / c_2\right)$
Ampliar estadio	4200	1700	700	1819,05
Reformar asientos	3000	3100	1400	**3095,24**
Ampliar accesos	3300	2500	2100	2538,1
No hacer obras	2400	1800	1000	1828,57

Estudio de mercado: Asistencia baja (c_3)				
$P\left(e_j / c_3\right)$	0,0541	0,1351	0,8108	
Beneficios próxima temp. (miles de €):	Asistencia alta	Asistencia media	Asistencia baja	$VME\left(a_i / c_2\right)$
Ampliar estadio	4200	1700	700	1024,32
Reformar asientos	3000	3100	1400	1716,22
Ampliar accesos	3300	2500	2100	**2218,92**
No hacer obras	2400	1800	1000	1183,78

$$P(c_1)=\sum_{j=1}^{3} P\left(c_1 / e_j\right) P\left(e_j\right)=0,21 \qquad P(c_2)=\sum_{j=1}^{3} P\left(c_2 / e_j\right) P\left(e_j\right)=0,42$$

$$P(c_3)=\sum_{j=1}^{3} P\left(c_3 / e_j\right) P\left(e_j\right)=0,37$$

$$VMEII=\left(0,21\times 3604,76\right)+\left(0,42\times 3095,24\right)+\left(0,37\times 2218,92\right)=2878$$

$$VII=2878-2570=308$$

3. Un ahorrador ha de decidir en qué tipo de fondo de inversión va a depositar su dinero. Puede optar por un fondo de inversión muy agresivo, un fondo mixto que entraña un riesgo moderado o decidirse por títulos de renta fija. En función de la evolución de los mercados, las ganancias que puede obtener con cada uno de los fondos son:

	Evolución buena	Evolución mala
Fondo agresivo	140	-140
Fondo mixto	84	-28
Títulos renta fija	56	7

La probabilidad estimada de que la evolución de los mercados sea buena o mala es, respectivamente, de un 0,6 y 0,4.

a) ¿Cuánto estaría dispuesto a pagar, como máximo, por conocer la evolución de los mercados?

b) Suponga que el inversor puede recurrir a una consultora que pronostica bastante acertadamente cuál va a ser la marcha del mercado. Las probabilidades de si el mercado sigue una determinada evolución, este hecho haya sido pronosticado por la consultora, están recogidas en la siguiente tabla:

	La consultora pronostica una evolución	
	buena	*mala*
Evolución buena	0,7	0,3
Evolución mala	0,2	0,8

¿Cuál es el valor máximo de esta información?

SOLUCIÓN:

a)

Probabilidad a priori	**0,6**	**0,4**	
	Evolución buena	*Evolución mala*	*VME(a_i)*
Fondo agresivo	140	-140	*28*
Fondo mixto	84	-28	***39,2***
Títulos renta fija	56	7	*36,4*

Según el criterio del valor monetario esperado (con las probabilidades a priori) la decisión óptima sería **"Fondo mixto"** y las ganancias esperadas serían **39,2**.

El valor monetario esperado conociendo con certeza la evolución del mercado sería:

$$VMEIP = (0,6 \times 140) + (0,4 \times 7) = 86,8$$

Por tanto, el valor de la información perfecta (valor de conocer con certeza la evolución del mercado) es:

$$VIP = VMEIP - VME(máximo) = 86,8 - 39,2 = 47,6$$

b) $$P\left(\frac{e_j}{c_k}\right)=\frac{P\left(c_k/e_j\right)P(e_j)}{\sum_{i=1}^{2}P\left(c_k/e_i\right)P(e_i)}=\frac{P\left(c_k/e_j\right)P(e_j)}{P\left(c_k\right)}$$

Pronóstico de la consultora: Evolución buena (c_1)			
	e_1 *Evolución buena*	e_2 *Evolución mala*	suma
$P\left(e_j\right)$	**0,6**	**0,4**	**1**
$P\left(c_1/e_j\right)$	0,7	0,2	
$P\left(c_1/e_j\right)P\left(e_j\right)$	0,42	0,08	$P(c_1)=0{,}50$
$P\left(e_j/c_1\right)=\frac{P\left(c_1/e_j\right)P\left(e_j\right)}{P\left(c_1\right)}$	**0,84**	**0,16**	**1**

$P\left(\frac{e_j}{c_1}\right)$	**0,84**	**0,16**	
	Evolución buena	*Evolución mala*	*VME(a_i) / c_1*
Fondo agresivo	140	-140	***95,2***
Fondo mixto	84	-28	*66,08*
Títulos renta fija	56	7	*48,16*

Según el criterio del valor monetario esperado, si el pronóstico de la consultora es de una evolución buena, la decisión óptima sería **"Fondo agresivo"** y las ganancias esperadas serían **95,2**.

Pronóstico de la consultora: Evolución mala (c_2)			
	e_1 *Evolución buena*	e_2 *Evolución mala*	suma
$P\left(e_j\right)$	**0,6**	**0,4**	**1**
$P\left(c_2/e_j\right)$	0,3	0,8	
$P\left(c_2/e_j\right)P\left(e_j\right)$	0,18	0,32	$P(c_2)=0{,}50$
$P\left(e_j/c_2\right)=\frac{P\left(c_2/e_j\right)P\left(e_j\right)}{P\left(c_2\right)}$	**0,36**	**0,64**	**1**

$P\left(\frac{e_j}{c_2}\right)$	**0,36**	**0,64**	
	Evolución buena	*Evolución mala*	*VME(a_i) / c_2*
Fondo agresivo	140	-140	*-39,2*
Fondo mixto	84	-28	*12,32*
Títulos renta fija	56	7	***24,64***

Según el criterio del valor monetario esperado, si el pronóstico de la consultora es de una evolución mala, la decisión óptima sería **"Títulos renta fija"** y las ganancias esperadas serían **24,64**.

Las ganancias esperadas medias con el pronóstico de la consultora son (***valor monetario esperado con información imperfecta***):

$$VMEII = \left(95{,}2 \times P\left(c_1\right)\right) + \left(24{,}64 \times P\left(c_2\right)\right) = 59{,}92$$

Y sin el informe de la consultora serían 39,2, luego estaríamos dispuestos a pagar como máximo por dicha información (***valor de la información imperfecta***):

$$VII = 59{,}92 - 39{,}2 = 20{,}72$$

4. Un ahorrador ha decidido invertir su dinero comprando acciones de una sola empresa. Puede optar por comprar acciones de la empresa 1, acciones de la empresa 2 o acciones de la empresa 3. En función de la evolución de los mercados (buena, regular o mala) las ganancias que puede obtener por su venta después de un tiempo son:

	Evolución buena	Evolución regular	Evolución mala
Empresa 1	150	53	-150
Empresa 2	89	46	-23
Empresa 3	56	35	19

La probabilidad estimada de que la evolución de los mercados sea buena, regular o mala es, respectivamente, de un 0,45; 0,35 y 0,20.

a) ¿Cuánto estaría dispuesto a pagar, como máximo, por conocer la evolución de los mercados?

Suponga que el inversor puede recurrir a una consultora que pronostica bastante acertadamente cuál va a ser la marcha del mercado. Las probabilidades de si el mercado sigue una determinada evolución, este hecho haya sido pronosticado por la consultora, están recogidas en la siguiente tabla:

	La consultora pronostica una evolución		
	buena	*regular*	*mala*
Evolución buena	0,65	0,25	0,10
Evolución regular	0,2	0,8	0
Evolución mala	0,15	0,25	0,6

b) ¿En qué empresa invertiría el ahorrador si prevé una evolución de los mercados buena?

c) ¿Cuál es el valor máximo de la información que nos aporta la consultora?

SOLUCIÓN:

e_1 = Evolución buena del mercado.

e_2 = Evolución regular del mercado.

e_3 = Evolución mala del mercado.

c_1 = Pronóstico de evolución buena.

c_2 = Pronóstico de evolución regular.

c_3 = Pronóstico de evolución mala.

a)

$VMEIP = (0,45 \times 150) + (0,35 \times 53) + (0,2 \times 19) = 89.85$

$VIP = VMEIP - VME(máximo)$ = 89,85 –56,05 = 33,8

$p(e_i)$	0,45	0,35	0,20	
	Evolución buena	Evolución regular	Evolución mala	$VME(a_i)$
Empresa 1	150	53	-150	**56,05**
Empresa 2	89	46	-23	51,55
Empresa 3	56	35	19	41,25

b)

Informe consultora: Evolución buena (c_1)				
	e_1 Evolución buena	e_2 Evolución regular	e_3 Evolución mala	suma
$P(e_j)$	**0,45**	**0,35**	**0,20**	**1**
$P(c_1/e_j)$	0,65	0,25	0,10	
$P(c_1/e_j)P(e_j)$	0,2925	0,07	0,03	$P(c_1) = 0,3925$
$P(e_j/c_1) = \frac{P(c_1/e_j)P(e_j)}{P(c_1)}$	**0,74522293**	**0,17834395**	**0,07643312**	**1**

Informe consultora: Evolución buenas (c_1)				
$P\left(\frac{e_j}{c_1}\right)$	0,74522293	0,17834395	0,07643312	
	Evolución buena	Evolución regular	Evolución mala	$VME(a_i/c_1)$
Empresa 1	150	53	-150	**109,770701**
Empresa 2	89	46	-23	72,7707006
Empresa 3	56	35	19	49,4267516

El ahorrador invertirá en la empresa 1.

c)

Informe consultora: Evolución regular (c_2)				
	e_1 Evolución buena	e_2 Evolución regular	e_3 Evolución mala	suma
$P(e_j)$	**0,45**	**0,35**	**0,20**	**1**
$P(c_2/e_j)$	0,25	0,80	0,25	
$P(c_2/e_j)P(e_j)$	0,1125	0,28	0,05	$P(c_2)=0{,}4425$
$P(e_j/c_2)=\frac{P(c_2/e_j)P(e_j)}{P(c_2)}$	**0,25423729**	**0,63276836**	**0,11299435**	**1**

Informe consultora: Evolución regular (c_2)				
$P\left(\frac{e_j}{c_2}\right)$	0,25423729	0,63276836	0,11299435	
	Evolución buena	Evolución regular	Evolución mala	$VME(a_i/c_2)$
Empresa 1	150	53	-150	**54,7231638**
Empresa 2	89	46	-23	49,1355932
Empresa 3	56	35	19	38,5310734

Informe consultora: Evolución mala (c_3)				
	e_1 Evolución buena	e_2 Evolución regular	e_3 Evolución mala	suma
$P(e_j)$	**0,45**	**0,35**	**0,20**	**1**
$P(c_3/e_j)$	0,10	0	0,6	
$P(c_3/e_j)P(e_j)$	0,045	0	0,12	$P(c_3)=0{,}165$
$P(e_j/c_3)=\frac{P(c_3/e_j)P(e_j)}{P(c_3)}$	**0,27272727**	**0**	**0,72727273**	**1**

Informe consultora: Evolución mala (c_3)				
$P\left(\frac{e_j}{c_3}\right)$	0,27272727	0	0,72727273	
	Evolución buena	Evolución regular	Evolución mala	$VME(a_i/c_3)$
Empresa 1	150	53	-150	-68,1818182
Empresa 2	89	46	-23	7,54545455
Empresa 3	56	35	19	**29,0909091**

$$P(c_1) = \sum_{j=1}^{3} P\left(c_1 / e_j\right) P\left(e_j\right) = 0,3925$$

$$P(c_2) = \sum_{j=1}^{3} P\left(c_2 / e_j\right) P\left(e_j\right) = 0,4425$$

$$P(c_3) = \sum_{j=1}^{3} P\left(c_3 / e_j\right) P\left(e_j\right) = 0,165$$

$$VMEII = \left(0,3925 \times 109,7707\right) + \left(0,4425 \times 54.7232\right) + \left(0,165 \times 29.0909\right) = 72,1$$

$$VII = 72,1 - 56,05 = 16,05$$

RELACIÓN DE EJERCICIOS

1. Muestreo Aleatorio Simple

1. Un dentista está interesado en la efectividad de una nueva pasta dental. Un grupo de 1000 niños de escuela participó en el estudio. Los registros de un estudio anterior mostraron que había un promedio de 2,2 caries cada seis meses para el grupo. Después de un año de iniciado el estudio, el dentista muestreó 10 niños para determinar cuánto habían progresado con la nueva pasta dental. Usando los datos de la siguiente tabla:

Niño	Número de caries en seis meses
1	0
2	4
3	2
4	3
5	2
6	0
7	3
8	4
9	1
10	1

¿Se puede decir que la incidencia media de las caries ha disminuido?

Solución: $2{,}2 \in (1{,}06,\quad 2{,}94) \Rightarrow No$

2. Un psicólogo desea estimar el tiempo de reacción medio para un estímulo en 200 pacientes de un hospital especializado en trastornos nerviosos. Una muestra aleatoria simple de 20 pacientes fue seleccionada, y fueron medidos sus tiempos de reacción, con los resultados siguientes: $\bar{y} = 2{,}1$ segundos y $S = 0{,}4$ segundos. Estime la media poblacional y establezca un límite para el error de estimación.

Solución: $\hat{\mu} = 2{,}1;\quad B = 0{,}1697$

3. En un estudio sociológico, realizado en una pequeña ciudad, se hicieron llamadas telefónicas para estimar la proporción de hogares donde habita por lo menos una persona mayor de 65 años de edad. La ciudad tiene 621 hogares, según la guía de teléfonos más reciente. Una muestra aleatoria simple de 60 hogares fue seleccionada de la guía. Al terminar la investigación de campo, de los 60 hogares muestreados, en 11 habita al menos una persona mayor de 65 años. Estime la proporción poblacional y establezca un límite para el error de estimación.

Solución: $\hat{p} = 0{,}1833;\quad B = 0{,}0958$

4. Un investigador está interesado en estimar el número total de árboles mayores de un cierto tamaño específico en una plantación de 1500 acres. Esta información se utiliza para estimar el volumen total de madera en la plantación. Una muestra aleatoria simple de 100 parcelas de 1 acre fue seleccionada, y cada parcela fue examinada en relación con el número de árboles de tamaño grande. La media muestral para las 100 parcelas de 1 acre fue $\bar{y} = 25{,}2$ árboles, con una cuasivarianza muestral de $S^2 = 136$. Estime el número total de árboles de tamaño grande en la plantación. Establezca un límite para el error de estimación.

   Solución: $\hat{\tau} = 37.800; \quad B = 3.379{,}9408$

5. Usando los datos del ejercicio anterior, determine el tamaño de muestra requerido para estimar el número total de árboles grandes en la plantación, con un límite para el error de estimación de 1500 árboles.

   Solución: $n = 399{,}413 \cong 400$

6. Una muestra aleatoria de 30 familias fue extraída de una zona de cierta ciudad que contiene 14848 familias. El número de personas por familia en la muestra obtenida fue el siguiente:

5	6	3	3	2	3	3	3	4	4	3	2	7	4	3
5	4	4	3	3	4	3	3	1	2	4	3	4	2	4

   Estimar el número total de personas en la zona, construyendo un intervalo de confianza al 95%.

   Solución: $(44842{,}09, \quad 58104{,}04)$

2. Muestreo Aleatorio Estratificado

1. Una gran empresa sabe que el 40% de las facturas que emite son al por mayor y el 60% al por menor. Sin embargo, identificar las facturas individuales sin consultar un archivo es complicado. Un auditor desea muestrear 100 de sus facturas para estimar el valor medio de las facturas de la empresa (Nota para estimar el total necesitaríamos conocer N). Una muestra aleatoria simple presentó 70 facturas al por mayor y 30 al por menor. Los datos son separados en facturas al por mayor y al por menor después del muestreo, con los siguientes resultados en €:

Por mayor	Por menor
Valor total facturas=36400€	*Valor total facturas*=8400€
$n_1 = 70 \quad \bar{y}_1 = 520€ \quad S_1 = 210€$	$n_2 = 30 \quad \bar{y}_2 = 280€ \quad S_2 = 90€$

Estime el valor medio de las facturas de la empresa, y fije un límite para el error de estimación.

Solución: $\overline{y}_{st} = 376€;\ B = 28{,}14€$

2. De las 1395 universidades de Estados Unidos, 364 imparten estudios universitarios de dos años y 1031 estudios universitarios de cuatro años. Se recogieron de manera independiente, una muestra aleatoria simple de 40 universidades con estudios de dos años y otra de 60 con estudios de 4 años. Las medias muestrales y las desviaciones típicas del número de estudiantes matriculados el pasado año en asignaturas de estadística aparecen a continuación.

	Carreras de 2 años	Carreras de 4 años
Media	154,3	411,8
Desviación típica	87,3	219,9

   a) Estimar el número total de estudiantes matriculados en asignaturas de estadísticas. Dar un límite de error de estimación.

   b) En el estudio del ejercicio anterior, se investigó también en qué proporción de las universidades la asignatura de estadística para economistas era impartida por miembros del departamento de economía. En la muestra se halló que en 7 de las universidades con carreras de dos años y en 13 de las que tienen carreras de cuatro años sucedía esto. Estimar la proporción de universidades en las que esta asignatura es impartida por profesores del departamento de economía. Dar un límite de error de estimación.

   Solución: (a) $\hat{\tau}_{st} = 480731;\ B = 57594,84$ (b) $\hat{p}_{st} = 0{,}2058;\quad B = 0{,}0826$

3. Una universidad tiene 152 profesores ayudantes, 127 profesores asociados y 208 profesores titulares. Una reportera del periódico de los estudiantes quiere averiguar si los profesores están realmente en sus despachos durante las horas de tutorías. Decide investigar muestras de 40 profesores ayudantes, 40 asociados y 50 titulares. Algunos estudiantes voluntarios llamaron a la puerta de los profesores de la muestra durante sus horas de tutorías. Se halló que 31 de los profesores ayudantes, 29 de los asociados y 34 de los titulares se encontraban realmente en sus despachos. Hallar un intervalo de confianza para la proporción de profesores que permanecen en sus despachos durante las horas de tutorías.

   Solución: $\hat{p}_{st} = 0{,}7214;\quad B = 0{,}0685$

4. Un auditor quiere estimar el valor medio de las facturas por cobrar de una compañía. La población se divide en cuatro estratos que contienen 500, 400, 300 y 200 facturas, respectivamente. Basándose en una experiencia previa, se estima que las desviaciones

típicas en estos estratos son de 15, 20, 30 y 40 euros, respectivamente. Determinar el tamaño muestral y la asignación para estimar el valor medio de las facturas por cobrar cometiendo un error de como mucho 5 euros.

Solución: $n_1 = 18{,}59; \quad n_2 = 19{,}83; \quad n_3 = 22{,}31; \quad n_4 = 19{,}83; \quad n = 80{,}55$

5. Un ayuntamiento está interesado en ampliar las instalaciones de un centro de atención diurna para niños. Se va a realizar una encuesta para estimar la proporción de familias con niños que utilizarán las instalaciones ampliadas. Las familias están divididas en aquellas que en la actualidad usan las instalaciones y las que aún no la usan. Aproximadamente el 90% de los que usan las instalaciones y el 50% de los que no las usan van a utilizar las nuevas instalaciones. Los costos por efectuar la observación de un cliente actual es de 4€ y de 8€ para uno que no lo es. Registros existentes nos dan que existen 97 familias que en la actualidad utilizan las instalaciones y 145 que no lo hacen.
   a) Encuentre el tamaño muestral aproximado y la asignación necesaria para estimar la proporción poblacional con un límite de 0,05 para el error de estimación.
   b) Suponga que el costo total de muestreo se fija en 400 € . Elija el tamaño de la muestra y la asignación que minimiza la varianza del estimador para este costo fijo.

   Solución: (a) $n_1 = 47;\ n_2 = 83;\ n = 130$ (b) $n_1 = 22; \quad n_2 = 39; \quad n = 61$

6. En un centro escolar se quiere realizar una encuesta para conocer la proporción de padres que estarían dispuestos a participar en actividades. Se quiere estimar la proporción de padres tanto a nivel global como para cada grupo de edad de los alumnos por lo que se decide estratificar según la edad de los alumnos. A partir de la información proporcionada por la siguiente tabla, obtener el número óptimo de padres que, de cada estrato, hay que encuestar para que la proporción de participación de los padres con hijos de edades entre 6 y 8 años sea estimada con un error menor o igual al 10%. (Suponemos que cada padre tiene un solo hijo en el centro)

Años	Alumnos matriculados	Porcentaje de participación en años anteriores	Coste de encuestar a un elemento
4-6	150	40%	4
6-8	130	30%	9
8-12	120	25%	16
12-14	100	20%	25

Sol. $n = 200{,}3; \quad n_1 = 94{,}84 \cong 95; \quad n_2 = 51{,}27 \cong 52; \quad n_3 = 33{,}53 \cong 34; \quad n_4 = 20{,}65 \cong 21 \Rightarrow n = 202$

7. El coste de transportar mercancías en avión depende del peso. Un determinado embarque de una fábrica consistía en las máquinas producidas por la citada fábrica a lo largo de las dos últimas semanas. Se decide estratificar basándose en las semanas, con el fin de observar si existe variación semanal en la cantidad producida. Las muestras aleatorias

simples de los pesos (en kilos) de las máquinas transportadas en el embarque, para las dos semanas, mostraron las siguientes mediciones:

Semana A	Semana B
58,3	59,2
60,4	60,1
59,3	59,6
58,7	59,2
59,1	58,8
59,6	60,5

a. Estimar el peso total del embarque de maquinaria, sabiendo que el número total de máquinas producidas ha sido de 162 en la semana A y de 170 en la semana B.
b. Obtenga un intervalo de confianza para el peso total del embarque de maquinaria.
c. Determinar el tamaño de la muestra y su asignación, en el caso de que se quiera estimar el peso total del embarque, con un límite para el error de estimación de 50 kg. Las dispersiones en los pesos se suponen diferentes de una semana a otra. Considere las muestras anteriores como muestras previas para estimar los parámetros necesarios.

Solución: (a) $\hat{\tau} = 19722{,}13$ (b) $(19593{,}71;\ 19850{,}56)$

(c) $n = 65{,}67;\quad n_1 = 34{,}37 \cong 35;\quad n_2 = 31{,}30 \cong 32 \Rightarrow n = 67$

8. Una cadena de almacenes está interesada en estimar la proporción de cuentas no cobradas. La cadena está formada por 4 almacenes, siendo el coste de muestreo igual para todos. Se usa muestreo aleatorio estratificado, con cada tienda como un estrato.

	Estrato I	Estrato II	Estrato III	Estrato IV
Nº cuentas por cobrar	$N_1 = 65$	$N_2 = 42$	$N_3 = 93$	$N_4 = 25$
Tamaño muestra	$n_1 = 14$	$n_2 = 9$	$n_3 = 21$	$n_4 = 6$
Nº cuentas no cobradas	4	2	8	1

a. Estime la proporción de cuentas no cobradas para la cadena y fije un límite para el error de estimación.
b. Utilice los datos anteriores para determinar la asignación y el tamaño de la muestra necesarios para estimar la proporción de cuentas no cobradas, con un límite del error de estimación del 5%.

Solución: (a) $\hat{p} = 0{,}30;\quad B = 0{,}1173$

(b) $n = 132{,}30;\quad n_1 = 38{,}35 \cong 39;\quad n_2 = 22{,}80 \cong 23;\quad n_3 = 58{,}98 \cong 59;\quad n_4 = 12{,}17 \cong 13 \Rightarrow n = 134$

9. Una escuela desea estimar la calificación media que puede obtener en el examen final de matemáticas en este curso. Los estudiantes de la escuela se agrupan en tres estratos según el tipo de aprendizaje, clasificado como N=Normal, A=Avanzado, L=Lento. En el presente curso, la distribución de los alumnos según el tipo de aprendizaje es 50 normal,

30 avanzado y 20 lento, la calificación media de los estudiantes según el tipo de aprendizaje fue en el primer examen parcial: 75 para el normal, 89 para el avanzado y 70 para el lento, con unas cuasivarianzas de 80, 30 y 40 respectivamente.

Para actualizar esta información, se tomó una muestra aleatoria de estudiantes, se les hizo el examen final de matemáticas y se obtuvieron las siguientes calificaciones (entre paréntesis, el tipo de aprendizaje de cada estudiante):

70(L)	88(A)	72(N)	85(N)	90(N)	82(A)	61(N)	92(N)	65(L)	87(A)
91(A)	81(N)	79(N)	63(L)	82(N)	75(N)	78(A)	71(L)	61(L)	

Se pide:

a. Estime la calificación media en el examen final de matemáticas. De una medida del error de estimación.
b. ¿Qué ocurre si no se tiene en cuenta el tipo de aprendizaje? Compare los resultados de ambos métodos de estimación, así como determine la ganancia en precisión.
c. Se desea mejorar la estimación de la nota media del examen final en matemáticas, teniendo en cuenta más información. Usando estos resultados como muestra previa, qué tamaños muestrales en cada estrato son necesarios para un error máximo admisible de 2 puntos, utilizando asignación Proporcional.
d. Estime, con un intervalo de confianza, el número de estudiantes con aprendizaje normal que han superado los 80 puntos. Si se pudiera planificar de nuevo la muestra, ¿qué tamaño de muestra sería necesario para que esta misma estimación tuviera un error máximo admisible de 10 estudiantes?

Solución: (a) $\hat{\mu} = 78,59;\ B = 3,21$ (b) $\hat{\mu} = 77,53;\quad B = 4,25$

(c) $n = 36,31;\quad n_1 = 18,15 \cong 19;\quad n_2 = 10,89 \cong 11;\quad n_3 = 7,26 \cong 8 \Rightarrow n = 38$

(d) $(11,87,\ 43,69);\quad n = 16,8 \cong 17$

10. Se desea estimar el salario medio de los empleados de una empresa. Se decide clasificarlos en dos estratos: los que tienen contrato fijo y los que poseen un contrato temporal. Los primeros son 143 y su salario varía entre 1500 y 2500 euros mensuales. Los contratos temporales son 320 y su salario está comprendido entre 700 y 1800 euros mensuales. ¿Cuál debe ser el tamaño de la muestra y su asignación para que al estimar el salario medio mensual el error de estimación sea inferior a 100 euros?

Solución: *Neyman* $\quad n = 26,91 \quad n_1 = 7,77 \cong 8 \quad n_2 = 19,14 \cong 20 \quad \Rightarrow \quad n = 28$

3. Muestreo con información auxiliar

1. Una encuesta de consumo fue realizada para determinar la razón de dinero gastado en alimentos sobre el ingreso por año, para las familias de una pequeña comunidad. Una muestra aleatoria de 14 familias fue seleccionada de entre 150. Los datos de la muestra se presentan en la siguiente tabla:

Familia	Ingreso Total	Gasto en alimentos
1	25100	3800
2	32200	5100
3	29600	4200
4	35000	6200
5	34400	5800
6	26500	4100
7	28700	3900
8	28200	3600
9	34600	3800
10	32700	4100
11	31500	4500
12	30600	5100
13	27700	4200
14	28500	4000

Estime la razón poblacional, y establezca un límite para el error de estimación.

Solución: $r = 0{,}1467; \quad B = 0{,}0102$

2. El ingreso nacional para 1981 será estimado con base en una muestra de 10 sectores industriales que declaran sus ingresos de 1981 antes que las 35 restantes. (Existen 45 sectores industriales que se utilizan para determinar el ingreso nacional total). Se dispone de los datos del ingreso de 1980 para los 45 sectores industriales y los totales son 2174,2 (en miles de millones). Los datos se presentan en la tabla adjunta:

Industria	1980	1981
Producto de fábricas textiles	13,6	14,5
Productos químicos y relacionados	37,7	42,7
Madera aserrada y leña	15,2	15,1
Equipo eléctrico y electrónico	48,4	53,6
Vehículos y equipo	19,6	25,4
Transporte y almacenaje	33,5	35,9
Banca	44,4	48,5
Bienes Raíces	198,3	221,2
Servicios de Salud	99,2	114,0
Servicios de Educación	15,4	17,0

(a) Encuentre el estimador de razón del ingreso total de 1981, y establezca un límite para el error de estimación.

(b) Encuentre el estimador de regresión del ingreso total de 1981, y establezca un límite para el error de estimación.

(c) Encuentre el estimador de diferencia del ingreso total de 1981, y establezca un límite para el error de estimación.

(d) ¿Cuál de los tres métodos es el más apropiado en este caso?¿Por qué?

Solución: (a) $\hat{\tau}_Y = 2433,30;\ B = 45,95$ (b) $\hat{\tau}_{YL} = 2432,91;\ B = 48,64$

(c) $\hat{\tau}_Y = 2455,90;\ B = 180,07$

3. Se desea conocer las ventas medias (en euros / habitante) en este año de un determinado producto en un municipio formado por un pueblo A con 291 habitantes y un pueblo B con 200 habitantes. Se sabe que las ventas medias en ese municipio el año pasado fueron de 170 euros / habitante. Tomamos una muestra aleatoria de 4 habitantes del pueblo A y otra de 3 habitantes del pueblo B para los que se conoce su consumo del producto bajo estudio (expresado en euros), este año (Y) y el año pasado (X):

Pueblo A		Pueblo B	
x_i	y_i	x_i	y_i
204	210	137	150
143	160	189	200
82	75	119	125
256	280		

a. Sin hacer distinción entre pueblos, estime las ventas medias para este año utilizando un estimador de razón. Dé un límite para el error de estimación.

b. ¿Qué se obtiene si no se tiene en cuenta los datos del año pasado pero si el pueblo?

c. ¿Qué se obtiene si no se tiene en cuenta los datos del año pasado ni se hace distinción entre pueblos?

d. Compare los estimadores que se obtienen en cada caso justificadamente.

Solución: (a) $\hat{\mu} = 180,53;\quad B = 5,69$ (b) $\hat{\mu} = 171,91;\quad B = 53,81$ (c) $\hat{\mu} = 171,43;\quad B = 49,53$

(d) La mejor estimación es en la que se usa el estimador de razón, por la fuerte relación entre las variables. El muestreo estratificado se comporta mal porque los estratos no son homogéneos.

4. Se está investigando la eficacia de una nueva dieta alimenticia en la crianza de conejos. Los investigadores piensan que hay razones para creer que el comportamiento es diferente dependiendo de la zona de crianza. Por este motivo, deciden formar estratos observándose el peso de los conejos antes de introducir la nueva dieta (X) y el peso resultante al cabo de un mes de tratamiento (Y). Se obtuvieron los siguientes resultados:

$$N_1 = 80; \quad N_2 = 60; \quad N_3 = 40; \quad n_1 = 10; \quad n_2 = 8; \quad n_3 = 6$$

Zona A		Zona B		Zona C	
X	Y	X	Y	X	Y
3,2	4,1	3,1	3,9	2,8	3,8
3,0	4,0	3,0	4,0	2,9	3,7
2,9	4,1	3,1	3,8	2,9	3,8
2,8	3,9	3,2	4,0	3,0	3,6
3,1	3,7	3,0	3,8	3,1	3,8
3,2	4,1	3,2	4,1	3,0	3,7
2,9	4,2	2,9	3,7		
2,8	4,0	3,0	3,8		
3,1	3,9				
2,8	3,8				

a. Estimar el peso medio estratificado de los conejos al principio y al final del tratamiento. Dar una estimación del error.
b. Si se le permite un error de estimación de 0,01 para estimar el peso medio estratificado al final del tratamiento, ¿cuáles deben ser los nuevos tamaños muestrales? Usar asignación Proporcional.
c. Sabiendo que el peso medio de los conejos antes de introducir la nueva dieta era de 3,2 kilogramos, estimar el peso medio de los conejos al final del tratamiento utilizando un estimador de razón. Dar el límite de error de estimación.
d. Estimar el peso medio de los conejos al final del tratamiento utilizando muestreo aleatorio simple. Comentar los resultados.

Solución: (a) $\hat{\mu}_x = 3{,}0008; \quad B = 0{,}0516; \quad \hat{\mu}_y = 3{,}8944; \quad B = 0{,}0523$

(b) $n = 144{,}4; \quad n_1 = 64{,}2 \cong 65; \quad n_2 = 48{,}15 \cong 49; \quad n_3 = 32{,}1 \cong 33 \Rightarrow n = 147$

(c) $\hat{\mu}_y = 4{,}1467; \quad B = 0{,}0793$ (d) $\hat{\mu} = 3{,}8875; \quad B = 0{,}0617$

5. En una escuela de 560 alumnos, se desea estimar la calificación media que puede obtenerse en el examen final de matemáticas en el curso 00/01. Se toma como información auxiliar la calificación de los mismos alumnos en el examen final de matemáticas del curso 99/00 con una nota media de 75. A partir de una muestra aleatoria de estudiantes para los cuales se observó la nota del examen final en el curso 00/01 y la calificación de dicho alumno en la prueba correspondiente al curso 99/00. Los resultados fueron los siguientes:

Nota curso 99/00	Nota curso 00/01
80	87
78	65
98	86
45	47
61	67
83	94
79	67
56	67

Estimar la calificación media del curso 00/01 utilizando como información auxiliar la calificación obtenida en el curso 99/00 mediante un estimador de razón. Dar una estimación del error de muestreo.

Solución: $\hat{\mu}_y = 75; \quad B = 7{,}45$

6. Un director de recursos forestales está interesado en estimar el número de abetos muertos por una plaga en una zona de 300 hectáreas. Usando una fotografía aérea, el director divide la zona en 200 parcelas de hectárea y media. Se toma una muestra aleatoria de 10 parcelas. El número total de abetos muertos, obtenidos según la cantidad en fotografía es 4200.

Parcela	1	2	3	4	5	6	7	8	9	10
Cantidad en fotografía	12	30	24	24	18	30	12	6	36	42
Cantidad en terreno	18	42	24	36	24	36	14	10	48	54

   a. Estime la razón poblacional y obtenga su intervalo de confianza.
   b. Estime el número total de abetos muertos en el área de 300 hectáreas y fije un límite para el error de estimación.
   c. ¿Cuál ha de ser el tamaño de la muestra necesario para estimar el total de abetos muertos, con un límite de error de estimación de 200 abetos?

Solución: (a) $r = 1{,}3077;\ (1{,}2057;\ 1{,}4097)$ (b) $\hat{\tau}_y = 5492{,}31;\ B = 428{,}44$) (c) $n = 38{,}9 \cong 39$

7. De una población de 40 hogares, para la que es conocido que el gasto total general durante un periodo de un año, en general, es de 12000000 um., se obtiene una muestra aleatoria simple de tamaño 4 que proporciona los siguientes valores anuales (en um):

Gasto en alimentación	125000	150000	100000	175000

   a. Estimar el gasto total en alimentación para los 40 hogares mediante un intervalo de confianza.
   b. Supongamos que de esos 4 hogares tenemos también los valores anuales de su gasto general (en um):

Gasto General	250000	300000	200000	350000

Antes de calcular otro estimador, ¿obtendríamos mejores resultados si utilizamos esta información auxiliar?¿Por qué?

c. Estimar mediante un estimador de razón el total de gasto en alimentación, utilizando la información auxiliar del apartado b.

d. Corroborar la respuesta del apartado b indicando qué estimador es mejor, el del apartado a o el del apartado c.

Solución: (a) (4275255, 6724744) (b) $\rho = 1$ (c) $\hat{\tau}_y = 6000000$ (d) $B = 0$ (límite del error de estimación del apartado (c)

8. En una universidad se realizó una prueba de conocimientos matemáticos antes del ingreso a 486 estudiantes. Se consideraron dichas calificaciones como una variable auxiliar de la variable "calificación final en cálculo". Teniendo en cuenta que 291 eran chicos y las calificaciones medias del examen previo fueron de 47 para los chicos y 52 para las chicas, a partir de los datos de la tabla siguiente, se pide:

CHICOS		CHICAS	
Examen previo	Examen de cálculo	Examen previo	Examen de cálculo
39	65	57	92
43	78	47	89
21	52	28	73
64	82	75	98
		34	56
		52	75

a. Sin tener en cuenta el sexo, estima la calificación media en el examen final de cálculo utilizando un estimador de razón. De una medida del error de estimación.

b. ¿Qué ocurre si no se tiene en cuenta la información auxiliar pero si el sexo?

c. ¿Qué ocurre si no se tiene en cuenta la información auxiliar ni el sexo?

d. Compare los estimadores que se obtienen en cada caso justificadamente.

Solución: (a) $\hat{\mu}_y = 80{,}97; \quad B = 10{,}54$ (b) $\hat{\mu} = 73{,}76; \quad B = 9{,}5$ (c) $\hat{\mu} = 76; \quad B = 9{,}46$

4. Muestreo Sistemático

1. La sección de control de calidad de una empresa usa el muestreo sistemático para estimar la cantidad media de llenado en latas de 12 onzas que sale de una línea de producción. Los datos de la tabla adjunta representan una muestra sistemática 1 en 50 de la producción de un día.

Cantidad de llenado (en onzas)					
12,00	11,97	12,01	12,03	12,01	11,80
11,91	11,98	12,03	11,98	12,00	11,83
11,87	12,01	11,98	11,87	11,90	11,88
12,05	11,87	11,91	11,93	11,94	11,89
11,72	11,93	11,95	11,97	11,93	12,05
11,85	11,98	11,87	12,05	12,02	12,04

a. Estime μ, y establezca un límite para el error de estimación. Suponga que N=1800.

b. Determinar el tamaño de muestra requerido para estimar μ dentro de 0,01 unidades.

Solución: (a) $\hat{\mu}_{sy} = 11{,}94; \quad B = 0{,}0259$ (b) $n = 217{,}1 \cong 218$

2. Los funcionarios de cierta sociedad profesional desean determinar la proporción de miembros que apoyan varias enmiendas propuestas en las prácticas de arbitraje. Los funcionarios toman una muestra sistemática de 1 en 10, a partir de una lista en orden alfabético de los 650 miembros registrados. Sea $y_i = 1$ si la i-ésima persona muestreada favorece los cambios propuestos e $y_i = 0$ si se opone a los cambios. Use los siguientes datos de la muestra para estimar la proporción de miembros en favor de los cambios propuestos. Establezca un límite para el error de estimación.

$$\sum_{i=1}^{65} y_i = 48$$

Solución: $\hat{p}_{sy} = 0{,}7385; \quad B = 0{,}1042$

3. La tabla anexa muestra el número de nacimientos y la tasa de natalidad por cada 1000 individuos para Estados Unidos durante seis años seleccionados sistemáticamente.

   (a) Estime el número medio de varones nacidos por año para el periodo 1955-1980, y establezca un límite para el error de estimación.

   (b) Estime la tasa media anual de natalidad para el periodo 1955-1980, y establezca un límite para el error de estimación.

   (c) ¿Cree usted que el muestreo sistemático es mejor que el muestreo aleatorio simple para los problemas de los apartados (a) y (b)?¿Por qué?

Año	Nac.Masculinos	Nac.Femeninos	Total de Nac.	Natalidad
1955	2073719	1973576	4047295	26,0
1960	2179708	2078142	4257850	23,7
1965	1927054	1833304	3760358	19,4
1970	1915378	1816008	3731386	18,4
1975	1613135	1531063	3144198	14,6
1980	1852616	1759642	3612258	15,9

Solución: (a) $\hat{\mu}_{sy} = 1926935; \; B = 139437{,}35$; (b) $\hat{\mu}_{sy} = 19{,}67; \quad B = 3{,}17$;

(c) Si. Observando la tendencia de las muestras se puede decir que las poblaciones en estudio están "ordenadas" de forma decreciente.

4. En la tabla anexa se presentan los datos sobre las tasas de divorcio (por cada 1000 personas) en Estados Unidos para una muestra sistemática de los años de 1900-1980. Estime la tasa media anual de divorcios para tal periodo y establezca un límite para el error de estimación. ¿Es en este caso el muestreo sistemático mejor o peor que el muestreo aleatorio simple?¿Por qué?

Año	Tasa	Año	Tasa
1900	0,7	1945	3,5
1905	0,8	1950	2,6
1910	0,9	1955	2,3
1915	1,0	1960	2,2
1920	1,6	1965	2,5
1925	1,5	1970	3,5
1930	1,6	1975	4,8
1935	1,7	1980	5,2
1940	2,0		

Solución: $\hat{\mu}_{sy} = 2{,}26; \quad B = 0{,}57$. Mejor, se observa, en general, una tendencia creciente en los datos de la muestra, aunque se rompa ese orden parcial en los años 1945-1955.

5. Muestreo por Conglomerados.

1. Un fabricante de sierras quiere estimar el coste medio de reparación mensual para las sierras que ha vendido a ciertas industrias. El fabricante no puede obtener un coste de reparación para cada sierra, pero puede obtener la cantidad total gastada en reparación y el número de sierras que tiene cada industria. Entonces decide usar muestreo por conglomerados, con cada industria como un conglomerado. El fabricante selecciona una muestra aleatoria simple de 20 de 96 industrias a las que da servicio. Los datos sobre coste total de reparaciones por industria y el número de sierras son:

Industria	Nº sierras	Costo total de reparación para el mes pasado (€)	Industria	Nº sierras	Costo total de reparación para el mes pasado (€)
1	3	50	11	8	140
2	7	110	12	6	130
3	11	230	13	3	70
4	9	140	14	2	50
5	2	60	15	1	10
6	12	280	16	4	60
7	14	240	17	12	280
8	3	45	18	6	150
9	5	60	19	5	110
10	9	230	20	8	120

a. Estime el costo medio de reparación por sierra para el mes pasado, y establezca un límite para el error de estimación.
b. Estime la cantidad total gastada por las 96 industrias en la reparación de sierras. Establezca un límite para el error de estimación.
c. Después de verificar sus registros de ventas, el fabricante se percata de que ha vendido un total de 710 sierras a esas industrias. Usando esta información adicional, estime la cantidad total gastada en reparación de sierras por estas industrias, y establezca un límite para el error de estimación.

Solución: (a) $\hat{\mu} = 19{,}73; \quad B = 1{,}78$ (b) $\hat{\tau} = 12312; \; B = 3175{,}07$

(c) $\hat{\tau} = 14008{,}85; \; B = 1110{,}78$

2. Se diseña una encuesta económica para estimar la cantidad media gastada en servicios para los hogares en una ciudad. Ya que no se encuentra disponible una lista de hogares, se usa muestreo por conglomerados, con barrios formando los conglomerados. Se selecciona una muestra aleatoria de 20 barrios de la ciudad de un total de 60. Los entrevistadores obtienen el gasto en servicios de cada hogar en los barrios seleccionados; los gastos totales se muestran en esta tabla:

Barrio	Nº hogares	Cantidad total gastada en servicios (€)
1	55	2210
2	60	2390
3	63	2430
4	58	2380
5	71	2760
6	78	3110
7	69	2780
8	58	2370
9	52	1990
10	71	2810
11	73	2930
12	64	2470
13	69	2830
14	58	2370
15	63	2390
16	75	2870
17	78	3210
18	51	2430
19	67	2730
20	70	2880

a. Estime la cantidad media de gastos en servicios por hogar en la ciudad y establezca un límite para el error de estimación.

b. En la encuesta anterior se desconoce el número de hogares en la ciudad. Estime la cantidad total gastada en servicios por todos los hogares de la ciudad y establezca un límite para el error de estimación.

c. La encuesta económica se va a llevar a cabo en una ciudad vecina de estructura similar. El objetivo es estimar la cantidad total gastada en servicios por los hogares de la ciudad, con un límite de 5000€ para el error de estimación. Use los datos anteriores para encontrar el número aproximado de conglomerados que se necesitan para obtener ese límite.

Solución: (a) $\hat{\mu} = 40{,}17; \quad B = 0{,}64$ (b) $\hat{\tau} = 157020; \; B = 6927{,}88$ (c) $n = 29{,}4 \cong 30$

3. Un inspector quiere estimar el peso medio de llenado para cajas de cereal empaquetadas en una fábrica. El cereal está en paquetes que contienen 12 cajas cada uno. El inspector selecciona aleatoriamente 5 y mide el peso de llenado de cada caja en los paquetes muestreados, con los resultados (en onzas) que se muestran:

Paquete	Onzas de llenado											
1	16,1	15,9	16,1	16,2	15,9	15,8	16,1	16,2	16,0	15,9	15,8	16,0
2	15,9	16,2	15,8	16,0	16,3	16,1	15,8	15,9	16,0	16,1	16,1	15,9
3	16,2	16,0	15,7	16,3	15,8	16,0	15,9	16,0	16,1	16,0	15,9	16,1
4	15,9	16,1	16,2	16,1	16,1	16,3	15,9	16,1	15,9	15,9	16,0	16,0
5	16,0	15,8	16,3	15,7	16,1	15,9	16,0	16,1	15,8	16,0	16,1	15,9

Estime el peso medio de llenado para las cajas empaquetadas por esta fábrica, y establezca un límite para el error de estimación. Suponga que el número total de cajas empaquetadas por la fábrica es lo suficientemente grande para que no se tome en cuenta la corrección por población finita.

Solución: $\hat{\mu} = 16{,}0050; \quad B = 0{,}0215$

4. Un periódico quiere estimar la proporción de votantes que apoyan a cierto candidato A en una elección estatal. La selección y entrevista de una muestra aleatoria simple de votantes registrados es muy costosa por lo que se utiliza muestreo por conglomerados. Se selecciona una muestra aleatoria de 50 distritos (conglomerados) de un total de 497 que tiene el estado. El periódico quiere hacer la estimación el día de la elección, pero antes de que se haya hecho la cuenta final de los votos. Es por eso que los reporteros son enviados a los lugares de votación de cada distrito en la muestra, para obtener la información pertinente directamente de los votantes. Los resultados se muestran en esta tabla:

Nº votantes	Nº votantes A	Nº votantes	Nº votantes A	Nº votantes	Nº votantes A
1290	680	1893	1143	843	321
1170	631	1942	1187	1066	487
840	475	971	542	1171	596
1620	935	1143	973	1213	782
1381	472	2041	1541	1741	980
1492	820	2530	1679	983	693
1785	933	1567	982	1865	1033
2010	1171	1493	863	1888	987
974	542	1271	742	1947	872
832	457	1873	1010	2021	1093
1247	983	2142	1092	2001	1461
1896	1462	2380	1242	1493	1301
1943	873	1693	973	1783	1167
798	372	1661	652	1461	932
1020	621	1555	523	1237	481
1141	642	1492	831	1843	999
1820	975	1957	932		

a. Estime la proporción de votantes que apoyan al candidato A, y establezca un límite para el error de estimación.

b. El periódico quiere realizar una encuesta similar durante la siguiente elección. ¿Cómo de grande debe ser la muestra para estimar la proporción de votantes a favor de un candidato similar con un límite del 5% para el error de estimación?

Solución: $\hat{p} = 0{,}5701; \quad B = 0{,}0307$ (b) $n = 20{,}1 \cong 21$

5. Un empresario quiere estimar el número de tubos de dentífrico usados por mes en una comunidad de 4000 hogares divididos en 400 bloques. Se selecciona una muestra aleatoria simple de 4 bloques que proporciona los siguientes resultados:

Bloque	tubos gastados por hogar								
1	1	2	1	3	3	2	1	4	
2	1	3	2	2	3	1	4	1	1
3	2	1	1	1	3	2	2		
4	1	1	3	2	1	5	1	3	

Estime de distintas formas el número total de tubos gastados, obtenga el límite para el error de estimación en cada caso y comente los resultados.

Solución: Muestreo por conglomerados $\hat{\tau} = 8000; \quad B = 562{,}85$ Muestreo aleatorio simple $\hat{\tau} = 6400; \quad B = 1077{,}78$

6. En un proceso de control del volumen envasado por una fábrica de bebidas se eligen 5 de los 40 paquetes que tiene la fábrica, cada uno de los cuales contiene 4 envases, y se mide el volumen que cada envase contiene. Las observaciones se presentan en la tabla adjunta:

Paquete nº	Volumen envasado en cm^3			
1	33	32,5	31,7	34,2
2	32	32,6	33,8	32,5
3	30,9	33,1	33	33,4
4	34,1	33,1	32,5	33,2
5	32	32,1	32,6	33,6

Estime el volumen medio por envase y dar la cota de error de estimación.

Solución: $\hat{\mu} = 32{,}80; \quad B = 0{,}22$

7. Cierto tipo de tableros posee 12 microcircuitos cada uno. De un pedido de 50 tableros se seleccionan 10 de ellos para su estudio. El número de microcircuitos defectuosos por tablero fue

2	0	1	3	2	0	0	1	3	4

Estime la proporción de microcircuitos defectuosos en la población y establezca una cota para el error de estimación.

Solución: $\hat{p} = 0{,}1333; \quad B = 0{,}0674$

8. En una pequeña ciudad se quiere estimar el número total de horas diarias que sus residentes dedican a ver el programa ``Gran Hermano", emitido las 24 horas del día por un canal Digital. Dicha ciudad está dividida en 200 manzanas de viviendas. Se extrae una muestra aleatoria simple de 10 manzanas, y se interroga a cada familia acerca de si están conectados a Vía Digital y cuántas horas ven el programa. Los datos de la encuesta se encuentran en la siguiente tabla:

Manzana	*Nº hogares con canal Digital*	*Nº total horas que ven programa*
1	8	13
2	7	13
3	9	14
4	6	13
5	5	0
6	9	10
7	6	6
8	8	14
9	9	16
10	6	4

   a. Estimar el número total de horas que se ve el programa ``Gran Hermano" a través de Canal Digital.
   b. Obtener un intervalo de confianza para el número total de horas.
   c. Determinar cuántas manzanas se deberían muestrear para estimar el total poblacional, con un límite para el error de estimación de magnitud 20. Considere la muestra anterior como una muestra previa para estimar los parámetros necesarios.

Solución: (a) $\hat{\tau} = 2060$; (b) $(1415{,}30,\ 2704{,}70)$ (c) $n = 196{,}4 \cong 197$

9. En un municipio de 5000 familias se pretende estimar el porcentaje de las que poseen ordenador. Se consideran 1000 conglomerados de 5 familias cada uno, y se elige una muestra aleatoria de 10 conglomerados, en los que el número de familias con ordenador es:

2	1	5	3	0	1	4	3	5	0

Estimar la proporción de familias que poseen ordenador y la varianza del estimador usado para estimar dicha proporción.

Solución: $\hat{p} = 0{,}48; \quad \hat{V}(\hat{p}) = 0{,}0143$

10. Se desea conocer la proporción de empleados de una empresa que no están dispuestos a trasladarse a una nueva planta de producción. Realizada una encuesta a los empleados de 5 factorías elegidas al azar entre las 50 que tiene la empresa, los resultados han sido:

Factoría	Nº empleados	Dispuestos
1	250	225
2	190	175
3	210	190
4	400	350
5	150	120

Estimar la proporción de empleados que no están dispuestos a trasladarse a la nueva factoría. Obtenga una estimación de la varianza del estimador empleado.

Solución: $\hat{p} = 0{,}1167; \quad \hat{V}(\hat{p}) = 0{,}0002$

11. Un gran embarque de mariscos congelados es empaquetado en cajas, conteniendo cada una 24 paquetes de 5 kilos. Hay 100 cajas en el embarque. Un inspector del gobierno determina el peso total de mariscos dañados para cada una de las 5 cajas muestreadas. Los datos son:

9	6	3	10	2

   a. Estime el peso total de mariscos dañados en el embarque y establezca un límite para el error de estimación.
   b. Determine el tamaño de la muestra necesario para estimar el peso total de mariscos dañados en el embarque, con un límite de error de 275.

Solución: (a) $\hat{\tau} = 600; \; B = 308,22$ (b) $n = 6,20 \cong 7$

6. Estimación del Tamaño de la Población.

1. Un club deportivo se interesa por el número de truchas de río en un arroyo. Durante un periodo de varios días, sea atrapan 100 truchas, se marcan y se devuelven al arroyo. Obsérvese que la muestra representa 100 peces diferentes, ya que cualquier pez atrapado

en esos días, que ya había sido marcado, se devolvía inmediatamente. Varias semanas después se atrapó una muestra de 120 peces y se observó el número de peces marcados. Supongamos que este número fue de 27 en la segunda muestra. Estime el tamaño total de la población de truchas y dé un límite de error de estimación.

Solución: $\hat{N} = 444{,}4; \quad B = 150{,}60$

2. Ciertos biólogos de poblaciones salvajes desean estimar el tamaño total de la población de codorniz común en una sección del sur de Florida. Se usa una serie de 50 trampas. En la primera muestra se atrapan 320 codornices. Después de ser capturadas, cada ave es retirada de la trampa y marcada con una banda de metal en su pata izquierda. Luego se sueltan todas las aves. Varios meses después se obtiene una segunda muestra de 515 codornices. Suponga que 91 de estos pájaros están marcados. Estimar el tamaño total de la población de codornices y dar un límite de error de estimación.

   Solución: $\hat{N} = 1810{,}99; \quad B = 344{,}51$

3. Expertos en pesca están interesados en estimar el número de salmones de una reserva. Se atrapa una muestra aleatoria de 2876 salmones. Cada uno es marcado y soltado. Un mes después se atrapa una segunda muestra de 2562. Supongamos que 678 tienen marcas en la segunda muestra. Estime el tamaño de la población total y establezca un límite del error de estimación.

   Solución: $\hat{N} = 10867,72; \; B = 715,82$

4. Los regentes de una ciudad están preocupados por las molestias que causan las palomas alrededor del ayuntamiento. A fin de cuantificar el problema contratan un equipo de investigadores para que estime el número de palomas que ocupan el edificio. Con varias trampas se captura una muestra de 60 palomas, se marcan y se sueltan. Un mes después se repite el proceso, usando 60 palomas, de las que 18 están marcadas. Estimar el tamaño total de la población de palomas y dar un límite de error de estimación.

   Solución: $\hat{N} = 200; \quad B = 78{,}88$

5. Una zoóloga desea estimar el tamaño de la población de tortugas en determinada área geográfica. Ella cree que el tamaño de la población está entre 500 y 1000; por lo que una muestra inicial de 100 parece ser suficiente. Las 100 tortugas son capturadas, marcadas y liberadas. Toma una segunda muestra un mes después y decide continuar muestreando hasta que se recapturen 15 tortugas marcadas. Atrapa 160 tortugas antes de obtener las 15 marcadas. Estime el tamaño total de la población de tortugas y establezca un límite de error de estimación.

   Solución: $\hat{N} = 1066,67; \; B = 507,72$

6. En una plantación de pinos de 200 acres, se va a estimar la densidad de árboles que presentan hongos parásitos. Se toma una muestra de 10 cuadros de 0,5 acres cada uno. Las diez parcelas muestreadas tuvieron una media de 2,8 árboles infectados por cuadro.
    a) Estime la densidad de árboles infectados y establezca un límite de error de estimación.
    b) Estime el total de árboles infectados en los 200 acres de la plantación y establezca un límite de error de estimación.

    Solución: (a) $\hat{\lambda} = 5{,}6; \quad B = 2{,}1$ (b) $\hat{M} = 1120; \; B = 423{,}32$

7. Se desea estimar el número total de personas que diariamente solicitan información en una oficina turística. Se observa que 114 personas solicitan información, durante 12 intervalos de 5 minutos cada uno, repartidos aleatoriamente entre las 8 horas que permanece abierta la oficina. Estimar el total de personas que visitan la oficina diariamente y dar la cota de error de estimación.

    Solución: $\hat{M} = 912; \quad B = 170{,}8$

8. Un alumno de A.T.C. desea estimar el número de alumnos que una determinada mañana han ido a la Facultad. Para ello se basa en que dicho día una conocida marca comercial ha repartido a primeras horas de la mañana en la entrada de la Facultad 500 carpetas. En un intercambio de clase, sentado en un banco del pasillo, decide contar los alumnos que pasan hasta observar a 100 que portan la carpeta, para lo que fue necesario contar hasta 382 alumnos.

    Estime con un intervalo de confianza el número de alumnos que asistieron esa mañana a la Facultad.

    Solución: muestreo inverso $\left(1910 \mp 326{,}58\right)$

9. El hermano de un alumno de T.A.M. está pensando en abrir una farmacia de 24 horas. Para saber si los ingresos compensarían los gastos de esta inversión deciden observar un establecimiento similar para estimar los ingresos diarios. Este asiduo alumno de T.A.M. conoce perfectamente que es una pérdida de tiempo innecesaria observar el flujo de clientes las 24 horas del día por lo que decide observar de forma sistemática media hora cada 3 horas, obteniendo los datos de la siguiente tabla

	clientes
10:00-10:30	35
13:00-13:30	20
16:00-16:30	19
19:00-19:30	30
22:00-22:30	25
01:00-01:30	9
04:00-04:30	12
07:00-07:30	18

Sabiendo que el gasto medio por cliente es de $20 \in$, estime los ingresos diarios de la farmacia observada y el correspondiente límite para el error de estimación utilizando diferentes métodos.

Solución: Muestreo por cuadros $Ingresos = 20160;\ B = 3110,76$; Muestreo aleatorio simple $Ingresos = 20160;\ B = 5402,22$

10. Se desea estimar el número total de palomas en la glorieta de una ciudad. Se capturan 80 palomas, se marcan y se devuelven a la población. Se realiza una segunda muestra hasta encontrar 30 palomas marcadas, se han tenido que capturar para ello 300 aves. Estimar el tamaño total y el límite del error de estimación.

    Solución: $\hat{N} = 800;\quad B = 272,62$

11. Se desea estimar el número total de pingüinos en una determinada zona. Se obtiene una muestra de tamaño 60, se marcan y se devuelven a la población. Al día siguiente se elige otra muestra de tamaño 400 y en ella se encuentran 12 marcados. Estimar el número total de pingüinos y dar la cota de error de estimación.

    Solución: $\hat{N} = 2000;\ B = 1137,25$

12. Se desea estimar el número de vehículos de un modelo determinado que el mes próximo utilizarán el aparcamiento de Puerta Real. Durante las 720 horas del mes se van a establecer 5 controles aleatorios de 1 hora de duración cada uno. Transcurrido el mes, se ha observado en los 5 controles los siguientes resultados:

Control	Número de vehículos de ese modelo que usan el aparcamiento
1	1
2	1
3	2
4	1
5	3

Estimar el número total de vehículos del modelo en estudio que utilizaron el aparcamiento.

Solución: $\hat{M} = 1152;\quad B = 814,59$

13. El ayuntamiento de Madrid está interesado en conocer el número de aficionados que acudieron al aeropuerto a vitorear al equipo campeón de la Champion League. Para ello, dividieron la sala de espera, de dimensiones 100 metros de largo por 35 metros de ancho, en 100 cuadros de igual tamaño y seleccionaron 40, observando que el número de personas era 2100.
    a. Estime la densidad de asistentes por metro cuadrado mediante un intervalo de confianza del 95%.
    b. Estime el número total de asistentes, y fije un límite para el error de estimación.

    Solución: (a) $(1{,}4,\ 1{,}6)$ (b) $\hat{M} = 5250;\ B = 229{,}13 \cong 229$

14. Se toman periódicamente muestras del aire en un área industrial de la ciudad. La densidad de cierto tipo de partículas dañinas es el parámetro de interés para el sector industrial. A partir de 15 muestras de 1 cm^3, se obtuvo un promedio de 210 partículas/ cm^3. Estimar la densidad de las partículas dañinas en dicha zona, así como dar una estimación del error de dicha estimación.

    Solución: $\hat{\lambda} = 210\ part/cm^3;\quad B = 7{,}48$

15. Se desea conocer cuántas personas asistieron a la inauguración del pabellón de Portugal en la Expo de Lisboa. Se sabe que el pabellón tiene forma cuadrada de 35 metros de lado y se traza una malla que divide el área total en 100 cuadros de igual tamaño. Se selecciona una muestra aleatoria de 40 cuadros, observando que el número de personas es de 750.
    a. Estime la densidad de asistentes por metro cuadrado y obtenga su intervalo de confianza.
    b. Estime el número total de asistentes a la inauguración y fije un límite para el error de estimación.

    Solución: (a) $\hat{\lambda} = 1{,}5306;\ (1{,}4188;\ 1{,}6424)$ (b) $\hat{M} = 1875;\quad B = 136{,}9$

16. Un equipo de ecólogos quiere medir la efectividad de un fármaco para controlar el crecimiento de la población de palomas. Se quiere conocer el tamaño de la población de este año para compararlo con el del año pasado. Se atrapa una muestra inicial de 600 palomas y se les da el fármaco, a la vez que se aprovecha para marcarlas en una pata. En fechas posteriores se atrapa otra muestra de 100 palomas de las cuales 48 tienen marca.
    a. Estime el tamaño de la población con un intervalo del 95% de confianza.
    b. Para reducir el límite de error de estimación a la mitad, ¿en qué proporción deben ser mayores las cantidades 100 y 48 observadas en la segunda muestra?, ¿se

deberían observar el doble de las cantidades anteriores, es decir, 200 y 96?, ¿el triple?, ¿el cuádruplo?,...

Solución: (a) $(989,79;\ 1510,21)$ (b) el cuádruplo.

7. Muestreo con probabilidades desiguales.

1. Una empresa encargada de realizar publicidad para empresas quiere conocer el éxito de sus campañas de publicidad. Durante un determinado periodo la empresa ha trabajado con 70 empresas de unos 30 empleados (Empresas pequeñas), aproximadamente, en cada una de ellas, y con otras 30 empresas con unos 200 empleados (Empresas grandes), aproximadamente, en cada una de estas empresas más grandes. La empresa desea realizar un muestreo aleatorio para recabar información sobre los beneficios de las empresas antes y después de sus respectivas campañas publicitarias, pero con una representación mayor de las empresas más grandes, por lo que decide aplicar un muestreo con probabilidades proporcionales al número de empleados de las empresas, obteniéndose los siguientes resultados medidos en miles de euros:

Tipo de Empresa	*Beneficios (Antes de publicidad)*	*Beneficios (Después de publicidad)*	*Gasto en publicidad*
Pequeña	15	19	3
Grande	25	24	6
Grande	30	45	10
Pequeña	10	12	5
Grande	40	45	2
Grande	50	60	8
Pequeña	12	10	3
Pequeña	10	15	3
Grande	25	35	5
Grande	36	40	3

Si la empresa considera que una campaña en publicidad ha tenido éxito si las diferencias en ganancias entre después y antes de la campaña son superiores a los gastos en publicidad realizados por dicha empresa, obtenga una estimación mediante un intervalo de confianza del porcentaje de éxito de la empresa encarga de realizar la publicidad.

8. Decisión en ambiente de incertidumbre.

1. Se considera la siguiente tabla de beneficios:

	e_1	e_2	e_3
a_1	50	30	20
a_2	40	20	40

Elige la alternativa óptima según los criterios de Laplace, Wald, Hurwicz (α=1/2) y Savage.

Solución: Laplace=*(a_1, a_2),* Wald=*(a_1, a_2)*, Hurwicz= a_1, Savage= a_2.

2. Disponemos de la siguiente tabla de decisión (con beneficios):

	e_1	e_2	e_3	e_4
a_1	65	110	70	90
a_2	60	125	50	105

Tome la decisión óptima utilizando todos los criterios de decisión en ambiente de incertidumbre (Para el criterio de Hurwicz use α=2/3).

Solución: Laplace= a_2, Wald= a_1, Hurwicz= a_2, Savage= a_1.

3. Disponemos de la siguiente tabla de decisión (con gastos):

	e_1	e_2	e_3	e_4
a_1	65	110	70	90
a_2	60	125	50	105

Tome la decisión óptima utilizando todos los criterios de decisión en ambiente de incertidumbre (Para el criterio de Hurwicz use α=2/3).

Solución: Laplace= a_1, Wald= a_1, Hurwicz= a_2, Savage= a_2.

4. Un inversor está ansioso por iniciar un nuevo negocio. Actualmente tiene tres posibilidades y el beneficio que obtendrá de cada una de ellas depende de su aceptación por los consumidores. Este inversor ha considerado para las condiciones del mercado tres posibles estados y el beneficio que espera de cada posible alternativa es:

	e_1	e_2	e_3
a_1	15	0	39
a_2	18	21	21
a_3	6	12	27

¿Qué decisión tomaría utilizando los criterios de Wald, de Hurwicz (α=2/3), de Savage y de Laplace?

Solución: Laplace= a_2, Wald= a_2, Hurwicz= a_1, Savage= a_3.

5. Un decisor con un índice de optimismo de 0,4 dispone de tres alternativas entre las que elegir, y que en función de la incertidumbre del entorno, le pueden proporcionar en una situación buena 150, 67 y 123 euros respectivamente, y en una situación mala, 57, 93 y 76

euros respectivamente, ¿cuál de las alternativas debería elegir según su índice de optimismo?

Solución: a_3 (94,8€).

6. Una empresa está analizando tres posibles estrategias de futuro, las cuales a su vez dependerían de la situación económica del país. Las ganancias estimadas por la empresa para cada estrategia son las siguientes:

	Ralentización	*Crecimiento sostenido*	*Fuerte crecimiento*
Estrategia 1	160	160	173
Estrategia 2	150	150	210
Estrategia 3	42	81	248

Establezca la estrategia que seguiría la empresa según adoptase los criterios de decisión de Wald, Hurwicz (α=0,8) o Savage.

Solución: Wald=Estrategia 1(160), Hurwicz (α=0,8)= Estrategia 3 (206,8), Savage= Estrategia 2 (38, pérdida de oportunidad).

7. Una empresa pretende comercializar un nuevo producto. La demanda que espera recibir depende del precio de otro producto análogo lanzado por una empresa competidora. La empresa desea fijar su propio precio para lo que efectúa una estimación de la demanda que conseguiría captar en cada supuesto, conforme sean su precio y el de la competencia:

		Precio competencia		
	de unidades demandadas ↘	*Alto*	*Medio*	*Bajo*
opio precio	*Alto*	2100	1800	900
	Medio	3700	3400	2500
	Bajo	4500	3600	3000

Suponiendo que el coste unitario del producto es de 20€, y que los precios se fijan en:

Alto: 40€ Medio:30€ Bajo: 24€

Determine la decisión óptima y los beneficios que se podrían conseguir utilizando los distintos métodos de decisión con incertidumbre (α=0,6).

Solución: Laplace: *propio p*recio alto ó medio (beneficio=32000€), Wald: *propio p*recio medio (beneficio=25000€), *H*urwicz(α=0,6): *propio p*recio alto (beneficio=32400€), Savage: *propio p*recio medio (pérdida de oportunidad=5000€).

8. Una sociedad financiera quiere realizar una inversión de medio millón de euros adquiriendo valores en bolsa para el próximo ejercicio económico. Puede adquirir tres tipos diferentes de valores: X, Y y Z. El precio para todos en el momento de realizar la inversión es de 100€ por título. La sociedad financiera estima que al finalizar el ejercicio, los precios de cotización de dichos títulos pueden haber alcanzado diferentes valores según haya sido el comportamiento de la bolsa:

	Subida moderada	*Subida selectiva*	*Recesión*	*Fuerte subida*
Valor X	120	150	40	120
Valor Y	100	120	80	120
Valor Z	80	20	80	160

Estudie la decisión que adoptaría según el criterio de Hurwicz.

Solución: Valor Y si $\alpha<0{,}571428572$, valor X si $0{,}571428572<\alpha<0{,}666666666$ y valor Z si $0{,}666666666<\alpha$

9. La decisión de una empresa de lanzar un nuevo producto ha planteado la necesidad de construir unas nuevas instalaciones que pueden ser de $600m^2$ o de $900m^2$. La decisión sobre el tamaño de la nueva planta depende de cómo vaya a reaccionar el mercado al nuevo producto. La empresa considera que se pueden dar tres posibilidades en cuanto a la demanda que derivarían en los siguientes rendimientos:

	Demanda baja	*Demanda media*	*Demanda alta*
Planta con $600m^2$	30	40	40
Planta con $900m^2$	10	40	100

¿Qué decisión tomaría la empresa, si considera como criterio de decisión la pérdida de oportunidad que asume en su decisión?

Solución: *Planta con $900m^2$* (pérdida de oportunidad: 20).

10. Una empresa agrícola considera la decisión de plantar soja, trigo o centeno. Los resultados de la cosecha dependerán del clima: seco, húmedo o lluvioso. Se estima que los rendimientos netos en miles de euros para cada cultivo son los siguientes:

	Seco	*Húmedo*	*Lluvioso*
Soja	96	56	-20
Trigo	40	60	30
Centeno	50	40	30

Indique cuál sería el cultivo más adecuado a los intereses de la empresa agrícola.

Solución: Laplace: soja(44) , Wald: trigo(30) ó centeno(30), Hurwicz (α=0,5): trigo(45) , Savage: centeno(46, pérdida de oportunidad).

9. Decisión en ambiente de riesgo.

1. Considere la tabla de decisión con beneficios:

Probabilidades de los estados de la naturaleza	0,25	0,25	0,25	0,25
	e_1	e_2	e_3	e_4
a_1	X	3	4	6
a_2	2	2	2	4
a_3	3	2	1	9
a_4	6	6	1	3

donde X es un número real. Encuentre qué decisión debe ser tomada en función de X.

Solución: Si $X>3$, elegiría a_1 (VME=4+(X-3)/4). Si $X<3$, elegiría a_4 (VME=4). Si $X=3$, elegiría a_1 ó a_4 (VME=4).

2. Una empresa estudia la posibilidad de lanzar un nuevo producto al mercado. Considera que las probabilidades de que el producto tenga éxito o no son del 70% y 30%, respectivamente.
    a) Obtenga la decisión óptima según el VME y el POE, e indique ambos valores.
    b) ¿Cuánto estaría dispuesto a pagar por conocer el verdadero estado de la naturaleza?
    c) ¿Cuál es el beneficio esperado si se conoce el verdadero estado de la naturaleza?

Beneficios	Éxito	Fracaso
Lanzar el producto A	4000	-1000
Lanzar el producto B	7000	-5000
Lanzar ambos	5000	-3000
No lanzar	0	0

Solución: a) Lanzar B. VME=3400, POE=1500. b) 1500. c) 4900.

3. La siguiente tabla representa las pérdidas de una operación de adquisición y posterior venta de cuatro automóviles:

Probabilidad	0,7	0,2	0,1
	Mercado baja	Mercado se mantiene	Mercado aumenta
Automóvil 1	20000	10000	5000
Automóvil 2	85000	7000	0
Automóvil 3	50000	15000	1000
Automóvil 4	15000	15000	15000

    a) ¿Cuál es la decisión óptima según el VME y el POE? ¿cuales son dichos valores?
    b) ¿Cuánto estaría dispuesto a pagar por conocer el verdadero estado de la naturaleza?

c) ¿Cuál es la pérdida de oportunidad esperada si conoce el verdadero estado de la naturaleza?

Solución: a) Automóvil 4. VME=15000, POE=3100. b) 3100. c) 0.

4. Una empresa está confeccionando su plan estratégico para la próxima década y dispone de cuatro posibles estrategias de desarrollo comercial: autónomo, franquicias, red de agentes o apertura de sucursales. Los beneficios previstos de estas cuatro estrategias en millones de euros dependen de cuatro posibles situaciones económicas y son los siguientes:

	e_1	e_2	e_3	e_4
Autónomo	70	70	0	35
Franquicias	35	35	35	35
Red de agentes	0	140	0	0
Apertura de sucursales	35	105	0	0

Si las probabilidades de que se dé cada una de las situaciones económicas son respectivamente: 0,30, 0,15, 0,35 y 0,20

a) Determine cuál sería la estrategia adecuada para la empresa y el beneficio esperado.
b) Indique si resultaría conveniente para la empresa tratar de mejorar su información sobre la incertidumbre generada por el entorno, supuesto que el coste del equipo que haría la investigación de mercado durante seis meses fuera de veinte millones de euros.

Solución: a) Autónomo, VME=38,5. b) Si, VIP=22,75>20.

5. Una distribuidora europea de un fabricante chino tiene que firmar un contrato de suministro sobre la base de que el producto adquirido debe ser pagado a 130€ la unidad, se venda o no por la distribuidora, y la distribuidora lo vende a 225€ la unidad. La demanda pude ser de 500, 600 o 700 unidades diarias (esas mismas cantidades son las que podrían aparecer en el contrato de suministro diario).

En los últimos 50 días la distribuidora ha tenido las siguientes ventas que considera fiables para la decisión a tomar:

Número de unidades vendidas	500	600	700
Número de días	15	30	5

a) A la vista de esta información, ¿qué decisión debe tomar la empresa respecto del número de unidades diarias a adquirir en la renovación del contrato con el fabricante chino? ¿cuál sería el valor monetario esperado?
b) ¿Cuál será el límite del coste correspondiente a la información adicional que hipotéticamente facilitara el número exacto de unidades diarias que demandaría la clientela a la distribuidora europea?

Solución: a) 600 unidades/día, VME=50250. b) VIP=4850.

6. Un juego consiste en lanzar dos monedas simultáneamente y apostar sobre el resultado. La apuesta se realiza sobre la suma de cruces resultante del lanzamiento de ambas monedas. Si el resultado del lanzamiento coincide con la apuesta del sujeto, este gana 10€. Si el número de cruces del lanzamiento supera al número de cruces de la apuesta, gana 5€. En cualquier otro caso, pierde 10€.
   a) ¿Qué apuesta aconsejaría al jugador? ¿cuál sería el valor monetario esperado?
   b) Si el jugador tuviera dudas acerca de que las monedas estén cargadas, según el criterio de Savage, ¿cómo debería apostar?
   c) Si el camarero del casino donde se está apostando le filtra al jugador la información de que las monedas están trucadas para que salga cruz en un 75% de las ocasiones, ¿qué apuesta debería hacer? ¿cuál sería el valor monetario esperado?

   Solución: a) 0 cruces, VME=6,25. b) 0 cruces. c) 1 cruz, VME=5,9375.

7. Un agricultor compra semillas en paquetes por un valor de 5€. Cada paquete sirve para plantar una hectárea que produce 500kg de hortalizas. Cada kilo de hortalizas puede ser vendido a 85 céntimos de euro en el mercado mayorista. El agricultor tiene arrendadas 20 hectáreas por un total de 1000€. Los salarios que tiene que abonar a temporeros para la plantación y recogida son de 165€ por hectárea cultivada.

   El agricultor tiene que decidir cuántas hectáreas debe plantar, debido a que no tiene claro que el mercado vaya a asumir toda su producción. Las alternativas que se plantea son: plantar de 15 a 20 hectáreas. En el mercado mayorista le pueden hacer pedidos de 7, 8 , 9 o 10 toneladas.
   a) ¿Cuántas hectáreas debería plantar, si el agricultor es una persona pesimista? ¿qué beneficios obtendría como mínimo?
   b) ¿Cuántas hectáreas debería plantar, si el agricultor es una persona optimista? ¿qué beneficios obtendría como máximo?
   c) Después de recapacitar sobre los resultados anteriores, el agricultor piensa que el valor más probable de demanda del mercado será de 8000kg, existiendo sólo un 20% de posibilidades de que alcance cada uno de los otros valores. ¿Cuántas hectáreas entonces debería plantar? ¿cuál será el valor monetario esperado?
   d) ¿Cuánto podría ganar el agricultor si, manteniendo la anterior distribución de probabilidad, tuviera información que le permitiera conocer lo que le va a demandar el mercado?

   Solución: a) 15ha, 2400€. b) 20ha, 4100€. c) 16ha, 17ha ó 18ha, 2910. d) VIP=340€.

10. Decisión bayesiana.

1. Un ahorrador ha de decidir en qué tipo de fondo de inversión va a depositar su dinero. Puede optar por un fondo de inversión muy agresivo, un fondo mixto que entraña un riesgo moderado o decidirse por títulos de renta fija. En función de la evolución de los mercados (buena, regular o mala) las ganancias que puede obtener con cada uno de los fondos son:

	Evolución buena	Evolución regular	Evolución mala
Fondo agresivo	140	56	-140
Fondo mixto	84	42	-28
Títulos renta fija	56	35	7

La probabilidad estimada de que la evolución de los mercados sea buena, regular o mala es, respectivamente, de un 0,5, 0,3 y 0,2.

c) ¿Cuánto estaría dispuesto a pagar, como máximo, por conocer la evolución de los mercados?

d) Suponga que el inversor puede recurrir a una consultora que pronostica bastante acertadamente cuál va a ser la marcha del mercado. Las probabilidades de si el mercado sigue una determinada evolución, este hecho haya sido pronosticado por la consultora, están recogidas en la siguiente tabla:

	La consultora pronostica una evolución		
	buena	*regular*	*mala*
Evolución buena	0,7	0,2	0,1
Evolución regular	0,2	0,8	0
Evolución mala	0,1	0,3	0,6

¿Cuál es el valor máximo de esta información?

Solución: a) VIP=29,4. b) VII=13,44.

2. Una empresa desea lanzar un nuevo producto. Realiza estudios de mercado, obteniendo los siguientes datos:

Proporción de personas dispuestas a adquirir el nuevo producto	*Probabilidad*	*Resultados esperados en millones de euros*
e_1: 30%	0,1	15
e_2: 20%	0,4	3,5
e_3: 10%	0,3	-2,5
e_4: 5%	0,2	-15

a) Según los anteriores datos, ¿decidiría la empresa lanzar el producto?

Para reducir la incertidumbre, la empresa puede recibir información complementaria por dos procedimientos:

I. Realizar un estudio de mercado con un coste de medio millón de euros, cuyos posibles resultados serían:

c_1: la proporción de clientes estará comprendida entre el 20% y el 35%.

c_2: la proporción de clientes estará comprendida entre el 10% y el 20%.

c_3: la proporción de clientes será inferior al 10%.

Con datos de productos análogos se han valorado las correspondientes verosimilitudes:

	Estudio de mercado		
Estados	$P(c_1/e_j)$	$P(c_2/e_j)$	$P(c_3/e_j)$
e_1: 30%	0,7	0,3	0
e_2: 20%	0,45	0,45	0,1
e_3: 10%	0	0,5	0.5
e_4: 5%	0	0,1	0,9

II. Utilizar un panel de consumidores, con coste 275000€, que ofrece los siguientes resultados:

c_1^*: la proporción de clientes será superior al 30%.

c_2^*: la proporción de clientes estará comprendida entre el 20% y el 30%.

c_3^*: la proporción de clientes estará comprendida entre el 10% y el 20%.

c_4^*: la proporción de clientes será inferior al 10%.

	Panel de consumidores			
Estados	$P(c_1^*/e_j)$	$P(c_2^*/e_j)$	$P(c_3^*/e_j)$	$P(c_4^*/e_j)$
e_1: 30%	0,5	0,4	0,1	0
e_2: 20%	0,2	0,5	0,2	0,1
e_3: 10%	0,1	0,2	0,45	0,25
e_4: 5%	0	0,1	0,25	0,65

b) Si lo lanzara, ¿cómo lo haría, con un estudio de mercado o con el panel de consumidores? ¿cuál sería el beneficio esperado?

Solución: a) No lanzaría el nuevo producto. b) Con un estudio de mercado, 1585000€.

3. Una empresa de consultoría desea incorporar un nuevo consultor a su departamento de ingeniería. Después de realizar los primeros filtros, llega a la selección final de dos candidatos, los cuales tienen diferente perfil. Según se desarrollen uno de los tres proyectos industriales cuya aprobación está pendiente en estos momentos, pueden ser más o menos necesarios.

   El coste para la empresa de consultoría dependerá de si se concretan o no estos proyectos y de lo que tendría que pagarles para retenerles y que no se fueran a otra empresa de la competencia. Los costes en euros serían:

	Proyecto S	***Proyecto T***	***Proyecto U***
Candidato X	160000	225000	80000
Candidato Y	70000	250000	170000

a) Si el Director de Recursos Humanos tuviese un grado de optimismo del 80% en relación al desarrollo de los proyectos, ¿cuál sería el candidato más adecuado y el coste asociado?

b) ¿Cuál sería el candidato adecuado si se utiliza el criterio de Savage? ¿cuál sería la pérdida de oportunidad?

c) Suponga que el Director de Recursos Humanos estima que las probabilidades de que se den cada uno de los proyectos anteriores (*S, T o U*) son respectivamente 0,25, 0,60 y 0,15. ¿Cuál sería en este caso el candidato idóneo y el coste monetario esperado?

d) El Director Financiero puede aportar un informe económico en relación a la situación económica que va a vivir el país. Se sabe que si se aprobara la realización del proyecto *S*, la situación del país sería buena con una probabilidad del 85%; si se aprobara el proyecto *T*, la situación del país sería buena con una probabilidad de 65% y si se aprobara el proyecto *U*, la situación del país sería mala con una probabilidad del 80%. Según el informe económico que presente el Director Financiero, ¿cuál sería el candidato que se debería contratar y el coste monetario esperado en cada caso?

e) ¿Qué valor, como máximo, se le puede otorgar a la información contenida en el informe económico?

Solución: a) *Y*, 106000€. b) *X o Y*, 90000€. c) *X,* 187000€. d) Si se prevé una situación económica buena: *Y,* 185731,23€. Si se prevé una situación económica mala: *X,* 171020,41€. d) VII=6675€.

4. Un empresario está planteándose el futuro de su negocio, y ha considerado la posibilidad de ampliarlo, acometer una reforma modesta, dejarlo como está o venderlo. La evolución de la demanda del sector podría ser alta, media o baja. En función de la evolución de la demanda, las ganancias que piensa que puede obtener con cada una de las alternativas descritas son las siguientes:

	Ampliar negocio	Reformar negocio	No modificarlo	Venderlo
Demanda alta	210	140	105	140
Demanda media	84	154	126	140
Demanda baja	35	70	168	140

La probabilidad estimada de que la demanda sea alta, media o baja es, respectivamente, 0,2, 0,3 y 0,5.

a) ¿Cuánto estaría dispuesto a pagar el empresario, como máximo, por conocer la evolución de la demanda?

b) Supongamos que el empresario puede recurrir a un estudio de mercado que le informe acerca de la evolución futura de la demanda. La probabilidad de que si la demanda va a ser alta o media, este hecho haya sido pronosticado por el estudio de mercado , es de 0,8, repartiéndose por igual la probabilidad de error en las otras dos posibilidades. Si la demanda va a ser baja, es seguro que lo habrá pronosticado así. ¿Cuál será , como máximo, el valor de la anterior información? ¿qué acción tomaría el empresario en función del informe?

Solución: a) VIP=29,4. b) VII=22,96. Si el informe es demanda alta, ampliaría el negocio. Si el informe es demanda media, lo reformaría. Si el informe es demanda baja, lo dejaría como está. En ningún caso lo vendería.

5. Antes de que se inicie la final de la Champions League, Real Madrid-Barcelona FC, un grupo de hinchas del Real Madrid que están en un bar tienen que decidir si participan o no en una apuesta organizada por el dueño del local. El juego propuesto consiste en una apuesta de 50€, con un premio de 150€ si aciertan quién ganará. Los expertos están divididos y piensan que los dos equipos tienen las mismas posibilidades de ganar.

   a) ¿Les interesaría apostar a su equipo en el juego que les propone el dueño del bar?

   Antes de tomar la decisión, deciden consultar a unos muy buenos aficionados al futbol, quienes a cambio de ser invitados a una ronda de cerveza (10€) estarían dispuestos a decirles su opinión sobre el resultado del partido. Otro cliente del bar les avisa que los citados aficionados son muy proclives a que ganará el Real Madrid y que cuando gana el Barcelona FC sólo aciertan en un 70% de las ocasiones, mientras que cuando gana el Real Madrid suelen acertar en un 90% de las veces.

   b) ¿Les interesaría pagar la ronda de cerveza para conocer la opinión de los aficionados?

   Solución: a) Sí. VME=25€>0€. b) Sí. VII=12,5€>10€.

6. Una empresa de servicios de una zona montañosa está considerando la posibilidad de invertir en una máquina quitanieves para el próximo invierno. La empresa ha analizado la situación cuidadosamente y cree que si nieva mucho el próximo invierno se trataría de una inversión rentable, considera que podría obtener una pequeña rentabilidad si la nieve es moderada, pero perdería dinero si las nevadas son débiles. Concretamente, la empresa prevé un beneficio de 42000€ si las nevadas son fuertes, 12000€ si son moderadas y una pérdida de 50000€ si las nevadas son débiles.

   Según el pronóstico del instituto de meteorología, las probabilidades de que se produzcan nevadas fuertes, moderadas o débiles son 0,5, 0,3 y 0,2, respectivamente.

a) ¿Cuál es la mejor decisión que podría adoptar la empresa? ¿cuáles serían los beneficios esperados?
b) Supongamos que la empresa decide esperar para poder conocer las temperaturas vividas a mediados de diciembre, antes de tomar la decisión final. La probabilidad de que el mes de diciembre sea frío cuando van a haber fuertes nevadas es de 0,8, mientras que cuando las nevadas van a ser moderadas es de 0,5 y cuando serán débiles es de 0,4. Si la empresa se encuentra con que diciembre no es frío, ¿cuál sería la decisión recomendada? ¿qué beneficios esperaría?

Solución: a) Comprar el quitanieves, VME=14600. b) Indiferente comprar o no el quitanieves. VME=0.

7. El ayuntamiento de una ciudad está considerando la posibilidad de reemplazar la flota de autobuses municipales por una red de tranvías. El concejal que defiende la idea, indica que la ciudad se ahorrará 7 millones de euros en gasóleo, contra un gasto de un millón en electricidad más el coste de las obras del tranvía (3 millones de euros). No obstante, si la idea se pone en marcha y los ciudadanos no usan suficientemente los tranvías, habría que volver a usar los autobuses con un coste por las obras de tres millones de euros.

   Según experiencias ocurridas en otras ciudades hay una probabilidad 0,50 de que los ciudadanos **sigan usando los tranvías igual que usaban los autobuses**, una probabilidad 0,30 de que **disminuya en un 20% la utilización del nuevo transporte**, en cuyo caso el ahorro de energía es igual a la pérdida en la recaudación por viajeros y una probabilidad 0,20 de que **disminuya más del 20%**, en cuyo caso habría que poner en marcha de nuevo los autobuses y habrían unas pérdidas iguales al coste de las obras realizadas.

   El Ayuntamiento finalmente acuerda incorporar al estudio la posibilidad de realizar un programa piloto que consiste en poner en marcha un tranvía durante seis meses. El coste de este programa piloto sería de 300000€. Basándose en la experiencia de otras ciudades que han realizado programas piloto similares, se ha efectuado una estimación empírica del resultado de dicho programa. En la siguiente tabla se recogen las probabilidades de cada resultado del programa piloto condicionadas a cada resultado obtenido con los nuevos tranvías:

	$p\left(c_i / e_j\right)$	***Resultados según el programa piloto***		
		c_1 *Igual utilización*	c_2 *Disminuye un 20% la utilización*	c_3 *Disminuye más del 20% la utilización*
Resultados reales que se obtendrían con los nuevos tranvías	e_1 *Igual utilización*	0,65	0,25	0,10
	e_2 *Disminuye un 20% la utilización*	0,35	0,35	0,30
	e_3 *Disminuye más del 20% la utilización*	0,15	0,45	0,40

¿Qué decisión debería tomar el Ayuntamiento si desea maximizar el ahorro? ¿Cuál sería el ahorro esperado?

a) Si no se hace el programa piloto.

b) Si según el programa piloto habrá un 20% menos de utilización del transporte.

Solución:

$p\left(e_j\right)$	*0,50*	*0,30*	*0,20*
AHORRO	*Igual utilización*	*Disminuye un 20% la utilización*	*Disminuye más del 20% la utilización*
Cambiar autobuses por tranvías	7-1-3=3	0	-3
Seguir con autobuses	0	0	0

a) Cambiar los autobuses por tranvías. VME=0,9 millones de euros=900000€. b) Cambiar los autobuses por tranvías. VME= 328125€-300000€=28125€.

FORMULARIO de MUESTREO

(*95% de confianza,* $z_c = 1{,}96 \cong 2$)

(*90% de confianza,* $z_c = 1{,}645$)

(*99% de confianza,* $z_c = 2{,}576$)

MUESTREO ALEATORIO SIMPLE EN POBLACIONES INFINITAS.

	VARIABLES NUMÉRICAS	**VARIABLES DICOTÓMICAS**
ESTIMADOR	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad y_i = 0,1$
VARIANZA MUESTRAL *(apenas se utiliza en muestreo)*	$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \bar{y}^2$	$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \hat{p}\hat{q}$
CUASIVARIANZA MUESTRAL	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{n\hat{p}\hat{q}}{n-1}$
VARIANZA DEL ESTIMADOR	$\hat{V}(\bar{y}) = \frac{S^2}{n}$	$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}$
B **LIMITE DEL ERROR DE ESTIMACIÓN**	$B_\mu = z_c\sqrt{\hat{V}(\bar{y})} = z_c\frac{S}{\sqrt{n}}$	$B_p = z_c\sqrt{\hat{V}(\hat{p})} = z_c\sqrt{\frac{\hat{p}\hat{q}}{n-1}}$
INTERVALO DE CONFIANZA	$\left(\bar{y} - B_\mu \ , \ \bar{y} + B_\mu\right)$	$\left(\hat{p} - B_p \ , \ \hat{p} + B_p\right)$
TAMAÑO MUESTRAL	$n = \frac{\sigma^2}{\frac{B_\mu^2}{z_c^2}} = \frac{\sigma^2}{D} \qquad D = \frac{B_\mu^2}{z_c^2}$ $\sigma^2 = S^2 \quad o \quad \sigma^2 = \left(\frac{R}{4}\right)^2$	$n = \frac{pq}{\frac{B_p^2}{z_c^2}} = \frac{pq}{D} \qquad D = \frac{B_p^2}{z_c^2}$ $p = \hat{p} \quad o \quad p = \frac{1}{2}$

MUESTREO ALEATORIO SIMPLE EN POBLACIONES FINITAS.

	VARIABLES NUMÉRICAS	**VARIABLES DICOTÓMICAS**
ESTIMADOR	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ $\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i$	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad y_i = 0,1$ $\hat{\tau} = N\hat{p}$
VARIANZA DEL ESTIMADOR	$\hat{V}(\bar{y}) = \frac{S^2}{n}\frac{N-n}{N}$ $\hat{V}(\hat{\tau}) = N^2\hat{V}(\bar{y}) = N(N-n)\frac{S^2}{n}$	$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N}$ $\hat{V}(\hat{\tau}) = N^2\hat{V}(\hat{p}) = N(N-n)\frac{\hat{p}\hat{q}}{n-1}$
B **LIMITE DEL ERROR DE ESTIMACIÓN**	$B_\mu = z_c\sqrt{\hat{V}(\bar{y})}$ $B_\tau = z_c\sqrt{\hat{V}(\hat{\tau})} = NB_\mu$	$B_p = z_c\sqrt{\hat{V}(\hat{p})}$ $B_\tau = z_c\sqrt{\hat{V}(\hat{\tau})} = NB_p$
INTERVALO DE CONFIANZA	$\left(\bar{y} - B_\mu\,,\,\bar{y} + B_\mu\right)$ $\left(\hat{\tau} - B_\tau\,,\,\hat{\tau} + B_\tau\right) = N\left(\bar{y} - B_\mu\,,\,\bar{y} + B_\mu\right)$	$\left(\hat{p} - B_p\,,\,\hat{p} + B_p\right)$ $\left(\hat{\tau} - B_\tau\,,\,\hat{\tau} + B_\tau\right) = N\left(\hat{p} - B_p\,,\,\hat{p} + B_p\right)$
TAMAÑO MUESTRAL	$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$ $D = \frac{B_\mu^2}{z_c^2} \quad (media)$ $D = \frac{B_\tau^2}{z_c^2 N^2} \quad (total)$ $\sigma^2 = S^2 \quad o \quad \sigma^2 = \left(\frac{R}{4}\right)^2$	$n = \frac{Npq}{(N-1)D + pq}$ $D = \frac{B_p^2}{z_c^2} \quad (proporcion)$ $D = \frac{B_\tau^2}{z_c^2 N^2} \quad (total)$ $p = \hat{p} \quad o \quad p = \frac{1}{2}$

MUESTREO ALEATORIO ESTRATIFICADO: ESTIMACIÓN.

	VARIABLES NUMÉRICAS	**VARIABLES DICOTÓMICAS**
ESTIMADOR	$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$ $\hat{\tau}_{st} = N\bar{y}_{st} = \sum_{i=1}^{L} N_i \bar{y}_i$	$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i = \sum_{i=1}^{L} \frac{N_i}{N} \hat{p}_i$ $\hat{\tau}_{st} = N\hat{p}_{st} = \sum_{i=1}^{L} N_i \hat{p}_i$
VARIANZA DEL ESTIMADOR	$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \hat{V}(\bar{y}_i) =$ $= \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i} =$ $= \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i}$ $\frac{N_i - n_i}{N_i} \cong 1$ *en poblaciones infinitas* $\hat{V}(\hat{\tau}_{st}) = N^2\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i}$	$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \hat{V}(\hat{p}_i) =$ $= \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i} =$ $= \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i}$ $\frac{N_i - n_i}{N_i} \cong 1$ *en poblaciones infinitas* $\hat{V}(\hat{\tau}_{st}) = N^2\hat{V}(\hat{p}_{st}) = \sum_{i=1}^{L} N_i^2 \frac{\hat{p}_i\hat{q}_i}{n_i - 1}\frac{N_i - n_i}{N_i}$

MUESTREO ALEATORIO ESTRATIFICADO: ASIGNACIÓN MUESTRAL. POBLACIONES FINITAS.

	VARIABLES NUMÉRICAS	**VARIABLES DICOTÓMICAS**
ASIGNACIÓN ÓPTIMA	*(error fijo B)* $n=\dfrac{\sum_{i=1}^{L} N_i\,\sigma_i\,\sqrt{c_i}\quad\sum_{i=1}^{L}\dfrac{N_i\,\sigma_i}{\sqrt{c_i}}}{N^2D+\sum_{i=1}^{L}N_i\sigma_i^2}$ *(coste fijo C)* $n=\dfrac{C\sum_{i=1}^{L}\dfrac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L}N_i\sigma_i\sqrt{c_i}}$	*(error fijo B)* $n=\dfrac{\sum_{i=1}^{L} N_i\sqrt{p_iq_ic_i}\quad\sum_{i=1}^{L}N_i\sqrt{\dfrac{p_iq_i}{c_i}}}{N^2D+\sum_{i=1}^{L}N_ip_iq_i}$ *(coste fijo C)* $n=\dfrac{C\sum_{i=1}^{L}N_i\sqrt{\dfrac{p_iq_i}{c_i}}}{\sum_{i=1}^{L}N_i\sqrt{p_iq_ic_i}}$
	$n_i=n\omega_i \qquad \omega_i=\dfrac{\dfrac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{j=1}^{L}\dfrac{N_j\sigma_j}{\sqrt{c_j}}}$	$n_i=n\omega_i \qquad \omega_i=\dfrac{N_i\sqrt{\dfrac{p_iq_i}{c_i}}}{\sum_{j=1}^{L}N_j\sqrt{\dfrac{p_jq_j}{c_j}}}$
ASIGNACIÓN DE NEYMAN *(error fijo B)*	$n=\dfrac{\left(\sum_{i=1}^{L}N_i\,\sigma_i\right)^2}{N^2D+\sum_{i=1}^{L}N_i\sigma_i^2}$	$n=\dfrac{\left(\sum_{i=1}^{L}N_i\sqrt{p_iq_i}\right)^2}{N^2D+\sum_{i=1}^{L}N_ip_iq_i}$
	$n_i=n\omega_i \qquad \omega_i=\dfrac{N_i\sigma_i}{\sum_{j=1}^{L}N_j\sigma_j}$	$n_i=n\omega_i \qquad \omega_i=\dfrac{N_i\sqrt{p_iq_i}}{\sum_{j=1}^{L}N_j\sqrt{p_jq_j}}$
ASIGNACIÓN PROPORCIONAL *(error fijo B)*	$n=\dfrac{\sum_{i=1}^{L}N_i\,\sigma_i^2}{ND+\dfrac{1}{N}\sum_{i=1}^{L}N_i\sigma_i^2}$	$n=\dfrac{\sum_{i=1}^{L}N_ip_iq_i}{ND+\dfrac{1}{N}\sum_{i=1}^{L}N_ip_iq_i}$
	$n_i=n\omega_i \qquad \omega_i=\dfrac{N_i}{N}$	$n_i=n\omega_i \qquad \omega_i=\dfrac{N_i}{N}$
	$D=\dfrac{B_\mu^2}{z_c^2}\quad(media)$ $D=\dfrac{B_\tau^2}{z_c^2N^2}\quad(total)$ $\sigma_i^2=S_i^2 \quad o \quad \sigma_i^2=\left(\dfrac{R_i}{4}\right)^2$	$D=\dfrac{B_p^2}{z_c^2}\quad(proporcion)$ $D=\dfrac{B_\tau^2}{z_c^2N^2}\quad(total)$ $p_i=\hat{p}_i \quad o \quad p_i=\dfrac{1}{2}$

MUESTREO ALEATORIO ESTRATIFICADO: ASIGNACIÓN MUESTRAL. POBLACIONES INFINITAS. Pesos de los estratos conocidos: W_i $(\cong N_i / N)$

	VARIABLES NUMÉRICAS	**VARIABLES DICOTÓMICAS**
ASIGNACIÓN ÓPTIMA	*(error fijo B)* $n = \dfrac{\sum_{i=1}^{L} W_i \sigma_i \sqrt{c_i} \quad \sum_{i=1}^{L} \dfrac{W_i \sigma_i}{\sqrt{c_i}}}{D}$	*(error fijo B)* $n = \dfrac{\sum_{i=1}^{L} W_i \sqrt{p_i q_i c_i} \quad \sum_{i=1}^{L} W_i \sqrt{\dfrac{p_i q_i}{c_i}}}{D}$
	(coste fijo C) $\quad n = \dfrac{C \sum_{i=1}^{L} \dfrac{W_i \sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i \sigma_i \sqrt{c_i}}$	*(coste fijo C)* $\quad n = \dfrac{C \sum_{i=1}^{L} W_i \sqrt{\dfrac{p_i q_i}{c_i}}}{\sum_{i=1}^{L} W_i \sqrt{p_i q_i c_i}}$
	$n_i = n\omega_i \qquad \omega_i = \dfrac{\dfrac{W_i \sigma_i}{\sqrt{c_i}}}{\sum_{j=1}^{L} \dfrac{W_j \sigma_j}{\sqrt{c_j}}}$	$n_i = n\omega_i \qquad \omega_i = \dfrac{W_i \sqrt{\dfrac{p_i q_i}{c_i}}}{\sum_{j=1}^{L} W_j \sqrt{\dfrac{p_j q_j}{c_j}}}$
ASIGNACIÓN DE NEYMAN *(error fijo B)*	$n = \dfrac{\left(\sum_{i=1}^{L} W_i \sigma_i\right)^2}{D}$	$n = \dfrac{\left(\sum_{i=1}^{L} W_i \sqrt{p_i q_i}\right)^2}{D}$
	$n_i = n\omega_i \qquad \omega_i = \dfrac{W_i \sigma_i}{\sum_{j=1}^{L} W_j \sigma_j}$	$n_i = n\omega_i \qquad \omega_i = \dfrac{W_i \sqrt{p_i q_i}}{\sum_{j=1}^{L} W_j \sqrt{p_j q_j}}$
ASIGNACIÓN PROPORCIONAL *(error fijo B)*	$n = \dfrac{\sum_{i=1}^{L} W_i \sigma_i^2}{D}$	$n = \dfrac{\sum_{i=1}^{L} W_i p_i q_i}{D}$
	$n_i = n\omega_i \qquad \omega_i = W_i$	$n_i = n\omega_i \qquad \omega_i = W_i$
	$D = \dfrac{B_\mu^2}{z_c^2} \quad (media)$ $\sigma_i^2 = S_i^2 \quad o \quad \sigma_i^2 = \left(\dfrac{R_i}{4}\right)^2$	$D = \dfrac{B_p^2}{z_c^2} \quad (proporcion)$ $p_i = \hat{p}_i \quad o \quad p_i = \dfrac{1}{2}$

ESTIMACIÓN DE RAZÓN.

	RAZÓN	MEDIA TOTAL
ESTIMADOR	$r = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \dfrac{\bar{y}}{\bar{x}}$	$\hat{\mu}_y = r\mu_x$ $\hat{\tau}_y = r\tau_x$
VARIANZA RESIDUAL	$S_r^{\,2} = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(y_i - rx_i\right)^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + r^2\sum_{i=1}^{n} x_i^2 - 2r\sum_{i=1}^{n} x_i y_i\right)$	
VARIANZA DEL ESTIMADOR	$\hat{V}(r) = \dfrac{1}{\mu_x^2}\dfrac{N-n}{N}\dfrac{S_r^2}{n} \cong \dfrac{1}{\bar{x}^2}\dfrac{N-n}{N}\dfrac{S_r^2}{n}$	$\hat{V}(\hat{\mu}_y) = \mu_x^2\hat{V}(r) = \dfrac{N-n}{N}\dfrac{S_r^2}{n}$ $\hat{V}(\hat{\tau}_y) = \tau_x^2\hat{V}(r) = N^2\dfrac{N-n}{N}\dfrac{S_r^2}{n}$ $\hat{V}(\hat{\tau}_y) \cong \dfrac{\tau_x^2}{\bar{x}^2}\dfrac{S_r^2}{n}$ *en poblaciones infinitas*
TAMAÑO MUESTRAL	$n = \dfrac{N\sigma_r^2}{ND + \sigma_r^2}$ $n = \dfrac{\sigma_r^2}{D}$ *en poblaciones infinitas* $\widehat{\sigma_r^2} = S_r^2$ *de una muestra previa* $D = \dfrac{B_R^2\mu_x^2}{z_c^2}$ (*para estimar* R) $D = \dfrac{B_\mu^2}{z_c^2}$ (*para estimar* μ_y) $D = \dfrac{B_\tau^2}{z_c^2 N^2}$ (*para estimar* τ_y)	

ESTIMACIÓN DE REGRESIÓN.

	MEDIA TOTAL
VARIANZA, COVARIANZA Y COEF. DE CORRELACIÓN MUESTRALES	$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$ *(análogamente para la variable Y)* $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}$ $r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$
ESTIMADOR	$\hat{\mu}_{yL} = \bar{y} + b(\mu_x - \bar{x})$ $\qquad b = \frac{s_{xy}}{s_x^2}$ $\hat{\tau}_{yL} = N\hat{\mu}_{yL}$
VARIANZA RESIDUAL	$S_L^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \left(\bar{y} + b(x_i - \bar{x})\right)\right)^2 = \frac{n}{n-2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = \frac{n}{n-2} s_y^2\left(1 - r_{xy}^2\right)$
VARIANZA DEL ESTIMADOR	$\hat{V}(\hat{\mu}_{yL}) = \frac{N-n}{N}\frac{S_L^2}{n}$ $\hat{V}(\hat{\tau}_{yL}) = N^2\hat{V}(\hat{\mu}_{yL})$
TAMAÑO MUESTRAL	$n = \frac{N\sigma_L^2}{ND + \sigma_L^2}$ $n = \frac{\sigma_L^2}{D}$ *en poblaciones infinitas* $\widehat{\sigma_L^2} = S_L^2$ *de una muestra previa* $D = \frac{B_\mu^2}{z_c^2}$ $\quad$ (*para estimar* μ_y) $D = \frac{B_\tau^2}{z_c^2 N^2}$ $\quad$ (*para estimar* τ_y)

ESTIMACIÓN DE DIFERENCIA.

	MEDIA TOTAL
ESTIMADOR	$\widehat{\mu}_{yD} = \bar{y} + (\mu_x - \bar{x}) = \mu_x + \bar{d}$ $\qquad \bar{d} = \bar{y} - \bar{x} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad d_i = y_i - x_i$ $\hat{\tau}_{yD} = N\widehat{\mu}_{yD}$
VARIANZA RESIDUAL	$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - (x_i + \bar{d})\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2 = \frac{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}}{n-1}$
VARIANZA DEL ESTIMADOR	$\widehat{V}(\widehat{\mu}_{yD}) = \frac{N-n}{N}\frac{S_D^2}{n}$ $\widehat{V}(\hat{\tau}_{yD}) = N^2\widehat{V}(\widehat{\mu}_{yD})$
TAMAÑO MUESTRAL	$n = \frac{N\sigma_D^2}{ND + \sigma_D^2}$ $n = \frac{\sigma_D^2}{D}$ *en poblaciones infinitas* $\widehat{\sigma_D^2} = S_D^2$ *de una muestra previa* $D = \frac{B_\mu^2}{z_c^2}$ (*para estimar* μ_y) $D = \frac{B_\tau^2}{z_c^2 N^2}$ (*para estimar* τ_y)

MUESTREO POR CONGLOMERADOS.

	MEDIA o PROPORCIÓN **TOTAL** *(M conocido)*	**TOTAL**
ESTIMADOR	$\hat{p} \quad o \quad \hat{\mu} = \bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \qquad \hat{\tau} = M\bar{y}$	$\hat{\tau}_t = N\bar{y}_t \qquad \left(\bar{y}_t = \dfrac{1}{n}\sum_{i=1}^{n} y_i\right)$
VARIANZA DEL ESTIMADOR	$\hat{V}(\bar{y}) = \dfrac{1}{\overline{M}^2}\dfrac{N-n}{N}\dfrac{S_c^2}{n}$ $\hat{V}(\hat{\tau}) = M^2\hat{V}(\bar{y}) = N(N-n)\dfrac{S_c^2}{n}$	$\hat{V}(\hat{\tau}_t) = N^2\hat{V}(\bar{y}_t) = N(N-n)\dfrac{S_t^2}{n}$
	$S_c^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2 =$ $= \dfrac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \bar{y}^2\sum_{i=1}^{n} m_i^2 - 2\bar{y}\sum_{i=1}^{n} m_i y_i\right)$	$S_t^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}_t\right)^2 = \dfrac{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$
TAMAÑO MUESTRAL	$n = \dfrac{N\sigma_c^2}{ND + \sigma_c^2}$ $n = \dfrac{\sigma_c^2}{D}$ *en poblaciones infinitas* $\widehat{\sigma_c^2} = S_c^2$ *de una muestra previa* $D = \dfrac{B_\mu^2\overline{M}^2}{z_c^2}$ *(media)* $D = \dfrac{B_\tau^2}{z_c^2 N^2}$ *(total)*	$n = \dfrac{N\sigma_t^2}{ND + \sigma_t^2}$ $n = \dfrac{\sigma_t^2}{D}$ *en poblaciones infinitas* $\widehat{\sigma_t^2} = S_t^2$ *de una muestra previa* $D = \dfrac{B_\tau^2}{z_c^2 N^2}$ *(total)*

NOTACIÓN:

$N =$ *conglomerados en la población (habitualmente conocido)*

$n =$ *conglomerados en la muestra*

$m_i =$ *elementos en el conglomerado i*

$y_i =$ *suma de las observaciones del conglomerado i*

$M = \sum_{i=1}^{N} m_i =$ *elementos en la población (habitualmente desconocido)*

$m = \sum_{i=1}^{n} m_i =$ *elementos en la muestra*

$\overline{M} = \dfrac{1}{N}\sum_{i=1}^{N} m_i = \dfrac{M}{N} =$ *tamaño medio de los conglomerados de la población (habitualmente desconocido)*

$\overline{m} = \dfrac{1}{n}\sum_{i=1}^{n} m_i = \dfrac{m}{n} =$ *tamaño medio de los conglomerados de la muestra . Este valor* $\overline{m}$ *se usa para estimar el anterior,* $\overline{M}$ *.*

ESTIMACIÓN DEL TAMAÑO DE LA POBLACIÓN

	MUESTREO DIRECTO	**MUESTREO INVERSO**
NOTACIÓN	$t =$ *elementos marcados* $n =$ *total de elementos en la muestra de recaptura* $s =$ *elementos marcados en la muestra de recaptura*	
ESTIMADOR	$\widehat{N} = \frac{t}{\widehat{p}} = \frac{nt}{s}$	$\widehat{N} = \frac{t}{\widehat{p}} = \frac{nt}{s}$
PROPIEDADES DEL ESTIMADOR	$E\left(\widehat{N}\right) = N + \frac{N(N-t)}{nt}$ $\widehat{V}\left(\widehat{N}\right) = \frac{t^2 n(n-s)}{s^3}$	$E\left(\widehat{N}\right) = N$ $\widehat{V}\left(\widehat{N}\right) = \frac{t^2 n(n-s)}{s^2(s+1)}$

ESTIMACIÓN DEL TAMAÑO DE LA POBLACIÓN

MUESTREO POR CUADROS		
	DENSIDAD	**TOTAL**
NOTACIÓN	$A =$ *área total* $a =$ *área de cada cuadro* $n =$ *número de cuadros en la muestra* $\overline{m} =$ *número medio de elementos por cuadro en la muestra*	
ESTIMADOR	$\hat{\lambda} = \dfrac{\overline{m}}{a}$	$\widehat{M} = \hat{\lambda} A$
VARIANZA DEL ESTIMADOR	$\hat{V}\left(\hat{\lambda}\right) = \dfrac{\hat{\lambda}}{an} = \dfrac{\overline{m}}{a^2 n}$	$\hat{V}\left(\widehat{M}\right) = A^2 \hat{V}\left(\hat{\lambda}\right) = \dfrac{A^2 \hat{\lambda}}{an} = \dfrac{A^2 \overline{m}}{a^2 n}$
TAMAÑO MUESTRAL	$n = \dfrac{\lambda}{aD}$ $D = \dfrac{B_\lambda^2}{z_c^2}$ (*para estimar* λ) $\qquad$ $D = \dfrac{B_M^2}{z_c^2 A^2}$ (*para estimar* M) λ *debe estimarse con una muestra previa*	

CUADROS CARGADOS		
	DENSIDAD	**TOTAL**
NOTACIÓN	$A =$ *área total* $a =$ *área de cada cuadro* $n =$ *número de cuadros en la muestra* $y =$ *número total de cuadros no cargados en la muestra*	
ESTIMADOR	$\hat{\lambda} = -\dfrac{1}{a}\ln\left(\dfrac{y}{n}\right)$	$\widehat{M} = A\hat{\lambda} = -\dfrac{A}{a}\ln\left(\dfrac{y}{n}\right)$
VARIANZA DEL ESTIMADOR	$\hat{V}\left(\hat{\lambda}\right) = \dfrac{1}{a^2}\dfrac{n-y}{ny}$	$\hat{V}\left(\widehat{M}\right) = A^2 \hat{V}\left(\hat{\lambda}\right) = \dfrac{A^2}{a^2}\dfrac{n-y}{ny}$

MUESTREO CON PROBABILIDADES DESIGUALES.

	MEDIA, PROPORCIÓN y TOTAL
PROBABILIDADES DE INCLUSIÓN	$\pi_i = \sum_{s \ni i} p(s)$ $\qquad$ $\pi_{ij} = \sum_{s \ni i \& j} p(s)$
PESOS MUESTRALES	$d_i = \dfrac{1}{\pi_i}$
PROBABILIDADES DE INCLUSIÓN EN UN DISEÑO PPT	$\pi_i = n\dfrac{x_i}{\tau_x}$
PROBABILIDADES DE INCLUSIÓN EN M. A. SIMPLE	$\pi_i = \dfrac{n}{N}$ $\qquad$ $\pi_{ij} = \dfrac{n}{N}\dfrac{n-1}{N-1}$
PROBABILIDADES DE INCLUSIÓN EN M. A. ESTRATIFICADO	$\pi_i = \dfrac{n_h}{N_h}$ si el individuo i pertenece al estrato h. $\pi_{ij} = \begin{cases} \dfrac{n_h}{N_h}\dfrac{n_h-1}{N_h-1} & \text{si ambos individuos } i \text{ y } j \text{ pertenecen al estrato } h. \\ \dfrac{n_h}{N_h}\dfrac{n_k}{N_k} & \text{si el individuo } i \text{ pertenece al estrato } h, \text{y el individuo } j \text{ al estrato } k \end{cases}$
ESTIMADOR DE TIPO HORVITZ-THOMPSON	$\bar{y}_{HT} = \dfrac{1}{N}\sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$ $\hat{p}_{HT} = \dfrac{1}{N}\sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$ $\qquad$ $y_i = 0 \quad o \quad y_i = 1$ $\hat{\tau}_{HT} = N\,\bar{y}_{HT} = N\,\hat{p}_{HT} = \sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$
VARIANZA DEL ESTIMADOR DE HORVITZ-THOMPSON	$\hat{V}_{HT}\left(\bar{y}_{HT}\right) = \dfrac{1}{N^2}\sum_{i=1}^{n}\left(1-\pi_i\right)\dfrac{y_i^2}{\pi_i^2} + \dfrac{2}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\dfrac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\dfrac{y_i}{\pi_i}\dfrac{y_j}{\pi_j}$ $\hat{V}_{SYG}\left(\bar{y}_{HT}\right) = \dfrac{1}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\dfrac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\dfrac{y_i}{\pi_i}-\dfrac{y_j}{\pi_j}\right)^2$ $\hat{V}_{HT}\left(\hat{\tau}_{HT}\right) = N^2\hat{V}_{HT}\left(\bar{y}_{HT}\right) = \sum_{i=1}^{n}\left(1-\pi_i\right)\dfrac{y_i^2}{\pi_i^2} + 2\sum_{i=1}^{n}\sum_{j>i}^{n}\dfrac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\dfrac{y_i}{\pi_i}\dfrac{y_j}{\pi_j}$ $\hat{V}_{SYG}\left(\hat{\tau}_{HT}\right) = N^2\hat{V}_{SYG}\left(\bar{y}_{HT}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n}\dfrac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\dfrac{y_i}{\pi_i}-\dfrac{y_j}{\pi_j}\right)^2$
ESTIMADOR DE TIPO HÁJEK	$\bar{y}_H = \dfrac{1}{\widehat{N}}\sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$ $\qquad$ $\widehat{N} = \sum_{i=1}^{n}\dfrac{1}{\pi_i}$ $\hat{p}_H = \dfrac{1}{\widehat{N}}\sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$ $\qquad$ $y_i = 0 \quad o \quad y_i = 1$ $\hat{\tau}_H = N\,\bar{y}_H = \dfrac{N}{\widehat{N}}\sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$

VARIANZA DEL ESTIMADOR DE HÁJEK	$$\hat{V}_J\left(\bar{y}_H\right)=\frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{y}_{H(i)}-\bar{y}_H\right)^2$$ $$\bar{y}_{H(i)}=\frac{1}{\hat{N}_{(i)}}\sum_{j\in s,j\neq i}\frac{y_j}{\pi_j}\quad,\quad \hat{N}_{(i)}=\sum_{j\in s,j\neq i}\frac{1}{\pi_j}$$ $$\hat{V}_J\left(\hat{\tau}_H\right)=N^2\hat{V}_J\left(\bar{y}_H\right)=\frac{N-n}{N}\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\tau}_{H(i)}-\hat{\tau}_H\right)^2$$ $$\hat{\tau}_{H(i)}=\frac{N}{\hat{N}_{(i)}}\sum_{j\in s,j\neq i}\frac{y_j}{\pi_j}\quad,\quad \hat{N}_{(i)}=\sum_{j\in s,j\neq i}\frac{1}{\pi_j}$$
VARIANZA DE UN ESTIMADOR $\hat{\theta}$ USANDO BOOTSTRAP	$$\hat{V}_B\left(\hat{\theta}\right)=\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{(b)}-\bar{\theta}_B\right)^2 \quad ; \quad \bar{\theta}_B=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{(b)}$$