MANUAL
DE PROCEDIMIENTO
DE MUESTREO

Cámara de Cuentas de Andalucía

Elaborado para la Cámara de Cuentas de Andalucía por Emilio Carrizosa, Catedrático de Estadística e Investigación Operativa de la Universidad de Sevilla.

Sevilla, a 27 de Marzo de 2006

1. Notación y procedimiento

El muestreo estadístico es una técnica que permite extraer una muestra *representativa* de una población, y, a partir de ésta, obtener conclusiones sobre la población (y no sólo sobre los elementos muestreados).

Su importancia queda patente en la Norma Técnica de Auditoría sobre utilización de técnicas de muestreo y de otros procedimientos de comprobación selectiva, pues en el párrafo 2.4.16 establece que *el muestreo estadístico es, en principio, el medio idóneo para expresar en términos cuantitativos el juicio del auditor respecto a la razonabilidad, determinando la extensión de las pruebas y evaluando su resultado.*

En lo que sigue supondremos una *población* $\mathcal{U}$ formada por N individuos, $u_1, \ldots, u_N$. La población $\mathcal{U}$ puede ser, por ejemplo, el conjunto de empresas públicas de la Junta de Andalucía (en cuyo caso, los distintos u_j serían las distintas empresas públicas), el personal de una Consejería (en cuyo caso, los u_j representarían las distintas personas empleadas por la Consejería, posiblemente clasificadas en diferentes categorías, como altos cargos, eventuales, funcionarios, laborales y otros), o el conjunto de contratos de suministros de una Consejería, en cuyo caso los u_j serían tales contratos.

Cada individuo u_i lleva asociada una variable Y_i, cuyo significado dependerá del tipo de prueba que se esté realizando. En pruebas sustantivas, Y_i puede representar, por ejemplo, la cuantía del gasto asociado al individuo u_i. En pruebas de cumplimiento, Y_i puede ser una variable sí/no, con la siguiente interpretación

$$Y_i = \begin{cases} 1, & \text{si el individuo } u_i \text{ cumple la normativa que le es de aplicación} \\ 0, & \text{en caso contrario} \end{cases} \tag{1}$$

El valor Y_i asociado a cada individuo se obtendrá, en ocasiones, directamente de los individuos de la población, mientras que en otras será una variable sintética, obtenida a partir de otras. Por ejemplo, si en cada individuo u_i consideramos r criterios de cumplimiento, y asociamos r variables sí/no, Y_{i1}, $Y_{i2}, \ldots, Y_{ir}$, de modo que cada Y_{ij} tome el valor 1 si hay alguna anomalía *en el criterio* j y 0 en caso contrario, podemos considerar la variable Y_i, también sí/no,

$$Y_i = \begin{cases} 1, & \text{si alguno de los valores } Y_{i1}, Y_{i2}, \ldots, Y_{ir} \text{ es } 1 \\ 0, & \text{si todos los valores } Y_{i1}, Y_{i2}, \ldots, Y_{ir} \text{ son } 0 \end{cases} \tag{2}$$

Se desea estudiar un parámetro ϑ, función de los valores de la variable Y en la población, a partir de los valores de la variable Y en la muestra.

Un caso de interés será aquél en el que queremos estimar la proporción de individuos en $\mathcal{U}$ con una cierta característica. En tal caso,

- cada Y_i se define siguiendo (1), como 1 si u_i presenta la característica, y 0 en caso contrario
- $\vartheta = \frac{1}{N}(Y_1 + Y_2 + \ldots + Y_N)$ (en este caso, denotaremos al parámetro por P, de proporción).

Si, en su lugar, el parámetro de interés es el número de individuos con la característica, tomamos la misma definición de variables Y_i, pero ahora hacemos $\vartheta = Y_1 + \ldots + Y_N$.

En cualquier caso, la estrategia a seguir es la siguiente:

1. `Identificación de la población y de la(s) variable(s) objeto de estudio.`
2. `Determinación del diseño muestral, incluyendo el cálculo del tamaño muestral.`
3. `Generación de la muestra aleatoria y medición de variables.`
4. `Estimación, construcción de intervalos de confianza y contrastes de hipótesis.`

2. Determinación de tamaños muestrales

Como se recoge en la Norma Técnica de Auditoría sobre utilización de técnicas de muestreo y de otros procedimientos de comprobación selectiva, el tamaño de la muestra puede determinarse bien mediante la aplicación de una fórmula estadística o mediante el juicio profesional aplicado objetivamente a las circunstancias.

El principio básico a seguir debe ser el de *tomar el máximo tamaño muestral* permitido por los condicionantes propios del estudio. En particular, en un estudio en el que sea viable analizar el 100 % de los casos, no procede realizar muestreo estadístico, sino estudiar todos los casos.

El procedimiento de determinación del tamaño muestral vendrá dado por las condiciones en las que se plantee el problema de muestreo. Tres escenarios distintos, que dan lugar a tres estrategias distintas de determinación del tamaño muestral, son las siguientes:

1. Escenario con *recursos prefijados.*
2. Escenario con objetivo de *control de cumplimiento.*
3. Escenario con objetivo de *control del error en la estimación.*

2.1. Tamaño muestral con recursos prefijados

En ocasiones, la determinación del tamaño muestral vendrá dada por los recursos disponibles para hacer el estudio. Así, si se estima que el tiempo necesario para realizar el análisis de cada individuo u_i (en otras palabras, el tiempo necesario para obtener el valor de u_i) es c, y se dispone de un total de h horas para hacer el estudio, entonces el número n de unidades a muestrear será, a lo sumo,

$$\boxed{n \;=\; \tfrac{h}{c}.} \tag{3}$$

2.2. Tamaño muestral y control de cumplimiento

En ocasiones el objetivo del muestreo no es *estimar*, a partir de los resultados muestrales de pruebas de cumplimiento, el nivel de cumplimiento en la población, sino *validar* si ciertos controles de cumplimiento están siendo aplicados adecuadamente, y no permiten que la fracción de incumplimientos supere un valor P_0 prefijado.

Asociamos a cada unidad una variable Y_i como se definió en (1), y se desea analizar si la proporción P de incumplimientos, definida como

$$P = \frac{1}{N}\left(Y_1 + \ldots + Y_N\right),$$

está o no por debajo de un cierto umbral P_0.

Realizamos el siguiente proceso:

1. Seleccionar *con reposición* n unidades (n a determinar).

2. Determinar si en la muestra existen o no unidades con incumplimiento. Si existen, rechazamos la hipótesis de que $P \leq P_0$, y si no existen unidades con incumplimiento, aceptamos la hipótesis de que $P \leq P_0$.

El valor de n se toma de modo que la probabilidad de aceptar la hipótesis de que $P \leq P_0$, siendo ésta falsa, sea, a lo sumo, igual a un cierto $\alpha > 0$ dado. Esto se consigue tomando

$$\boxed{n = \frac{\ln \alpha}{\ln(1-P_0)},} \tag{4}$$

donde ln representa el logaritmo neperiano, que se obtiene en Excel con la función `LN`.

Para introducir en Excel la función `LN` basta con escribirla sobre la celda correspondiente, como se muestra, por ejemplo, en la Figura 3. Alternativamente, podemos utilizar la utilidad de Excel para inserción de funciones, posicionando el cursor sobre la opción f_x, como se muestra en la Figura 1. Al pinchar nos aparece una ventana como la de la Figura 2. En la opción `o seleccionar una categoría` seleccionamos `Todas`, como en la Figura 2. En la lista de funciones aparece la función `LN`.

En la Figura 3 se muestra cómo obtener el tamaño muestral n para hacer un control de cumplimiento y validar, con error del 5 %, si la fracción de incumplimientos es inferior al 3 %. El tamaño muestral obtenido es $n = 98,3 \approx 99$.

Podemos representar en una tabla, para distintos valores de P_0 y de α, el tamaño necesario. Los valores aparecen en la Figura 4.

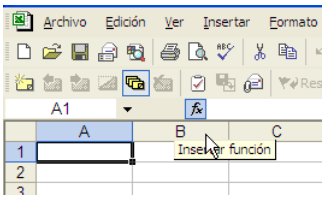


Figura 1: Inserción de funciones en Excel

Si fijamos α, podemos examinar en una gráfica, como la de la Figura 5, cómo varía el tamaño muestral requerido n para refutar (al 95 %) la hipótesis de que la proporción de incumplimientos P verifique $P \leq P_0$.

En la hoja `TAM CONTROL CUMP` del fichero `Ejemplos.xls` aparece

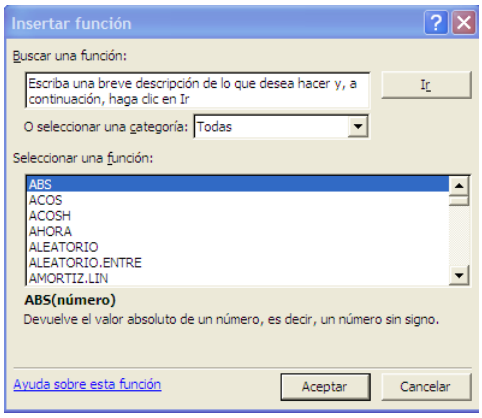


Figura 2: Inserción de funciones en Excel

	A	B	C
1	ALFA	0,05	
2	P_0	0,03	
3	n	=LN(B1)/LN(1-B2)	
4		LN(número)	

Figura 3: Calculando el tamaño muestral en control de incumplimiento

	0,99	0,98	0,97	0,96	0,95	0,94	0,93	0,92	0,91	0,9
0,01	458	389	349	320	298	280	265	251	240	229
0,02	228	194	174	159	148	139	132	125	119	114
0,03	151	128	115	106	98	92	87	83	79	76
0,04	113	96	86	79	73	69	65	62	59	56
0,05	90	76	68	63	58	55	52	49	47	45
0,06	74	63	57	52	48	45	43	41	39	37
0,07	63	54	48	44	41	39	37	35	33	32
0,08	55	47	42	39	36	34	32	30	29	28
0,09	49	41	37	34	32	30	28	27	26	24
0,1	44	37	33	31	28	27	25	24	23	22

Figura 4: Tamaño muestral para el control de incumplimiento

- la tabla que nos permite determinar el tamaño muestral necesario para validar, con una probabilidad de error α, la hipótesis de que la fracción de incumplimento no supera un valor P_0.

- una gráfica en la que se observa la evolución del tamaño muestral n necesario para distintos valores de P_0 y $\alpha = 0,05$.

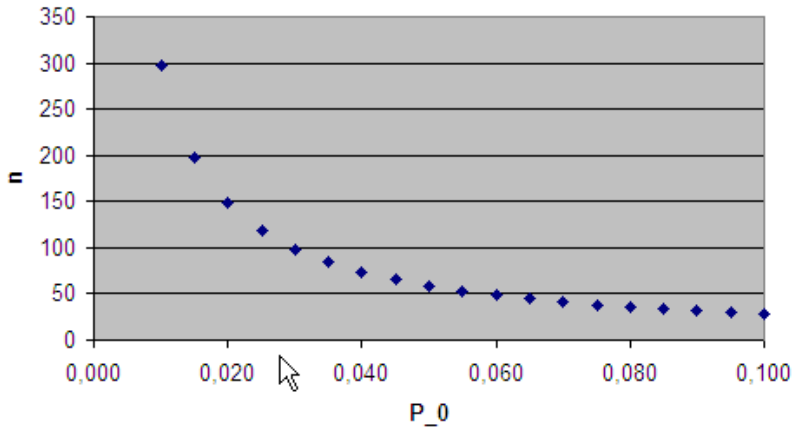


Figura 5: n como función de P_0

Obsérvese que

- El tamaño n obtenido es independiente del tamaño de la población, y se basa en muestreo *con reposición*. Será aplicable al caso en el que el muestreo sea realmente con reposición, o, si es sin reposición, que el tamaño de la población N sea muy grande.

- Tomado este n, tenemos garantías de que, si la hipótesis de que la fracción de incumplimientos P satisface que $P \leq P_0$ es falsa, la probabilidad de que en la muestra no haya ni un solo incumplimiento es a lo sumo α. En otras palabras, la probabilidad de que el test no detecte que la fracción de incumplimiento es superior a P_0 es α.

En el caso en el que el tamaño de la población N sea moderado o pequeño, si el muestreo que se realiza es *sin reposición*, la expresión (4) debe ser modificada: al extraer una muestra aleatoria sin reposición de n de unidades de una población de N unidades en la que una fracción $P \geq P_0$ registran incumplimentos, la probabilidad de que ninguna de las n unidades seleccionadas presente incumplimento es α si tomamos n no como en (4), sino satisfaciendo la ecuación

$$\boxed{\frac{\binom{N(1-P_0)}{n}}{\binom{N}{n}} = \alpha,} \tag{5}$$

donde $\binom{a}{b}$ representa el combinatorio de a sobre b, que se obtiene en Excel con la función `COMBINAT`.

El valor de n se obtiene resolviendo, *por tanteo,* la ecuación (5). En la Figura 6 se muestra el tamaño muestral obtenido resolviendo (5) para distintos tamaños poblacionales N y $P_0 = 0,1$ y $P_0 = 0,05$.

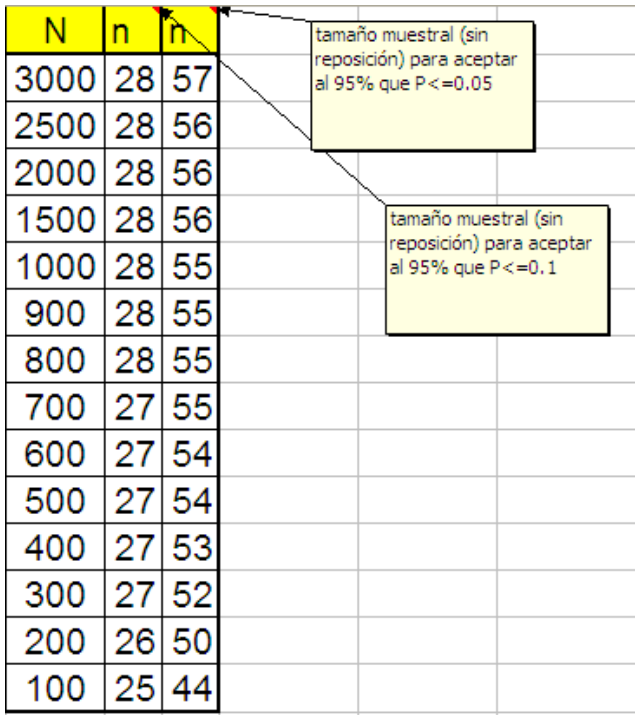


Figura 6: Contraste de hipóteis sobre el nivel de riesgo

En la hoja `TAM CONTROL CUMP SIN REP` del fichero `Ejemplos.xls` aparecen calculados por este procedimiento el tamaño muestral necesario para validar, con una probabilidad de error $\alpha = 0,05$, la hipótesis de que la fracción de incumplimento no supera un valor $P_0 = 0,05$.

En la hoja `TAM CONTROL CUMP SIN REP 2` del fichero `Ejemplos.xls` aparece una tabla con la que determinar, para un valor P_0 dado y una población de tamaño N, el valor de α, es decir: la probabilidad de que, si la fracción de incumplimiento es superior a P_0, analizando una muestra de n registros no se obtenga ni uno solo con incumplimiento.

Como ilustración, en la Figura 7 se muestra, para una población de $N = 5.000$ individuos y $P_0 = 0,03$ el valor de α para distintos valores de n en la fórmula (5).

2.3. Tamaño muestral y control del error de estimación

Tanto si existe una restricción de tiempo rígida como la indicada en la Sección 2.1 como si no existe, *como guía* debe explorarse, con la ayuda de Excel, cuál sería el error cometido para distintos tamaños muestrales posibles. De esta manera debe obtenerse un satisfactorio equilibrio entre esfuerzo de muestreo (tamaño muestral) y error, o bien, si tal equilibrio no existe, descartar el muestreo probabilístico, optar por el subjetivo, justificando tal opción por el error que se cometería

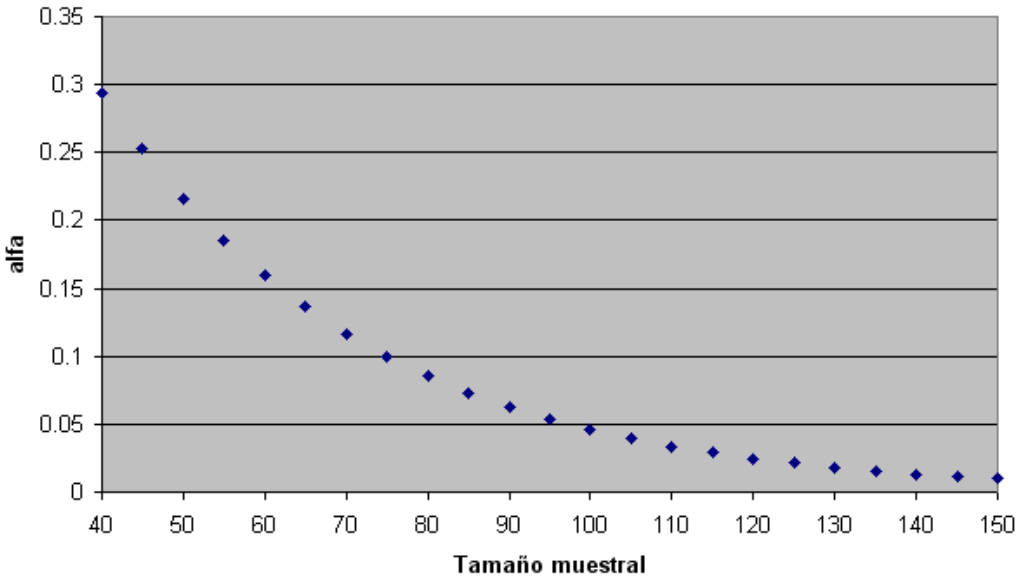


Figura 7: α como función de n ($N = 5.000, P_0 = 0,03$)

con tamaños muestrales asumibles.

El error es función no sólo de la variable que se esté estudiando, sino también del diseño muestral que se esté utilizando.

La **medida del error a priori** de referencia que establecemos es sólo válida para el caso del muestreo aleatorio simple, en el caso en que se vaya a estudiar una única variable, que ésta sea dicotómica (sí/no), y que no haya argumentos en contra de la utilización del teorema del límite central (esencialmente, que la muestra sea de tamaño muy pequeño, o bien que se aborde un problema de sucesos raros).

Cuando no sean válidas las condiciones anteriores, la medida del error que proponemos es una *guía*, si bien puede no representar fielmente el error cometido.

2.3.1. Determinación de tamaños muestrales en el escenario más desfavorable

Denotemos por

- N, el tamaño de la población.
- n, el tamaño de la muestra.
- α, la *significación*. Con esto afirmamos que si el análisis fuera realizado de modo independiente por un número grande de auditores, el $100 \cdot (1-\alpha)\,\%$ de ellos llegarían a la misma conclusión.

El valor habitual para el nivel de significación será del $5\,\%$, esto es: $\alpha = 0,05$, lo que implica que, grosso modo, el $5\,\%$ de los estudios realizados con este procedimiento darán resultados inválidos. Éste será un parámetro que debemos suministrar.

- $Z_{1-\alpha/2}$ es un coeficiente, llamado *cuantil* $1-\alpha/2$ *de la normal estándar*. Éste es un parámetro que se calcula a partir de α. Se obtiene mediante la función Excel `DISTR.NORM.ESTAND.INV`; en la Figura 8 aparece cómo calcular en Excel el coeficiente para el nivel de significación del $5\,\%$, esto es, para $\alpha = 0,05$.

  Para distintos valores de α se obtienen los valores que aparecen en la Figura 9.

En estas circunstancias, a un nivel de significación α, el error ε es

$$\varepsilon = Z_{1-\alpha/2}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{4(N-1)}}. \tag{6}$$

De la fórmula (6), despejando n, obtenemos

$$n = \frac{1}{\frac{1}{N} + \frac{\varepsilon^2 4(N-1)}{Z^2_{1-\alpha/2}N}} \tag{7}$$

En la hoja `TAM MUESTREO ALEAT` del fichero `Ejemplos.xls` aparece

- una tabla en la que, dados el tamaño poblacional N, y el nivel de significación α, determinar para distintos tamaños muestrales n el error cometido.
- Una gráfica en la que se observa la evolución del tamaño muestral n necesario con respecto al error ε admisible.

En la hoja `TAM MUESTREO ALEAT 2` del fichero `Ejemplos.xls` se indica cómo determinar n cuando el error ε viene dado. Describiremos esta hoja más adelante.

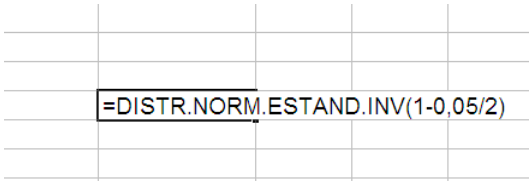


Figura 8: Calculando el cuantil $Z_{1-0.05/2}$

Ejemplo 1 Para una población de $N = 250.000$ individuos, sobre la que se selecciona una muestra aleatoria simple de $n = 1.250$ individuos (es decir, una fracción de muestreo del $0,5\,\%$), a un nivel

ALFA	Z
0.01	2.576
0.02	2.326
0.03	2.170
0.04	2.054
0.05	**1.960**
0.06	1.881
0.07	1.812
0.08	1.751
0.09	1.695
0.1	1.645

Figura 9: $Z_{1-\alpha/2}$ para distintos valores de α

de significación del 5 % se obtiene un error $\varepsilon = 0,02765$. El proceso en la hoja Excel aparece en la Figura 10.

Si nos conformamos con un error de $\varepsilon = 0,05$ usamos la fórmula (7), obteniendo $n = 383,56$ que redondeamos a $n = 384$.

□

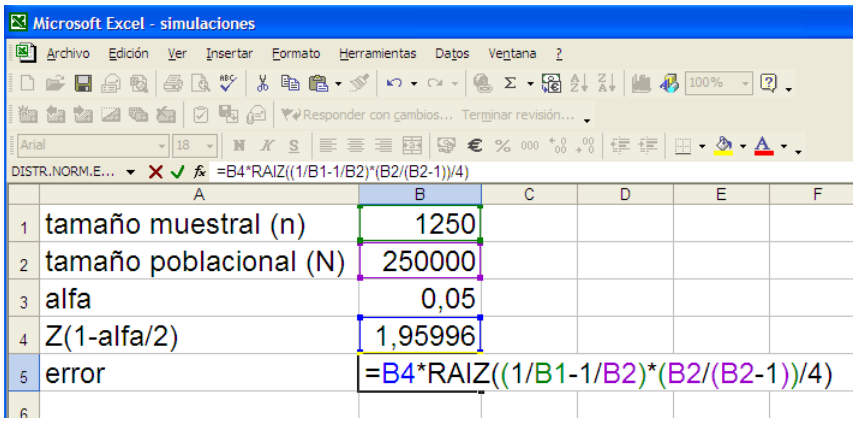


Figura 10: Caculando el error al 95 % para $N = 250.000$, $n = 1.250$

La utilidad del muestreo estadístico será tanto mayor cuanto mayor sea el tamaño de la población, como queda de manifiesto al comparar el resultado del Ejemplo 1 con el siguiente.

Ejemplo 2 Para una población de $N = 250$ individuos, un tamaño muestral de $n = 190$ (es decir, una fracción de muestreo del 76 %) produce, con un nivel de significación del 5 %, un error

$\varepsilon = 0,0349$.

Análogamente, para una fracción de muestreo del 10 %, esto es, $n = 25$, se obtiene un error $\varepsilon = 0,18631$, siendo necesario, según (7), un tamaño muestral de $n = 151,68$ (es decir: $n = 152$) para obtener un error a priori de $\varepsilon = 0,05$.

□

Ejemplo 3 La base de datos de Contratos de 2004 contiene 77.555 registros. Para obtener a un nivel de significación $\alpha = 0,05$ un error $\varepsilon = 0,05$ son necesarios, de acuerdo a la fórmula (7), $n = 382,26 \approx 383$ individuos. En definitiva, necesitamos muestrear algo menos del 0,5 %.

En cambio, si el estudio se reduce a los 410 individuos de la base de Contratos de 2004 con campo `uni-adm` 267, tenemos sólo 410 registros, se necesitan, según la fórmula (7), $n = 198,58 \approx 199$ individuos, lo que representa más del 48 % de los datos.

□

Para poder evaluar el compromiso entre esfuerzo de muestreo y error, debe representarse gráficamente, para distintos valores del tamaño muestral n, el error ε resultante. Como ilustración, en la Figura 11 se muestra el error al 95 %, aplicando la fórmula (6), para una población de $N = 250.000$ individuos.

2.3.2. Determinación de tamaños muestrales con información adicional

La fórmula (6) como medida del error considera el caso más desfavorable (supone máxima indeterminación). En ocasiones parece asumible que, antes de extraer la muestra, se disponga de información (por experiencias previas) sobre el parámetro objeto de estudio. Si el estudio a realizar es continuación de otro ya realizado sobre la misma población, y cabe esperar que la variable no haya sufrido grandes modificaciones, podemos utilizar, en lugar de (6), la expresión

$$\varepsilon = Z_{1-\alpha/2}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{(N-1)}P(1-P)}, \tag{8}$$

con la misma interpretación de los coeficientes que en (6), y P es el valor previsible *a priori* para la proporción.

Despejando, obtenemos el tamaño muestral para un ε fijo: en lugar de la expresión (7), que aplicamos sin conocimiento de información adicional, utilizamos

$$n = \frac{1}{\frac{1}{N} + \frac{\varepsilon^2(N-1)}{P(1-P)Z^2_{1-\alpha/2}N}}. \tag{9}$$

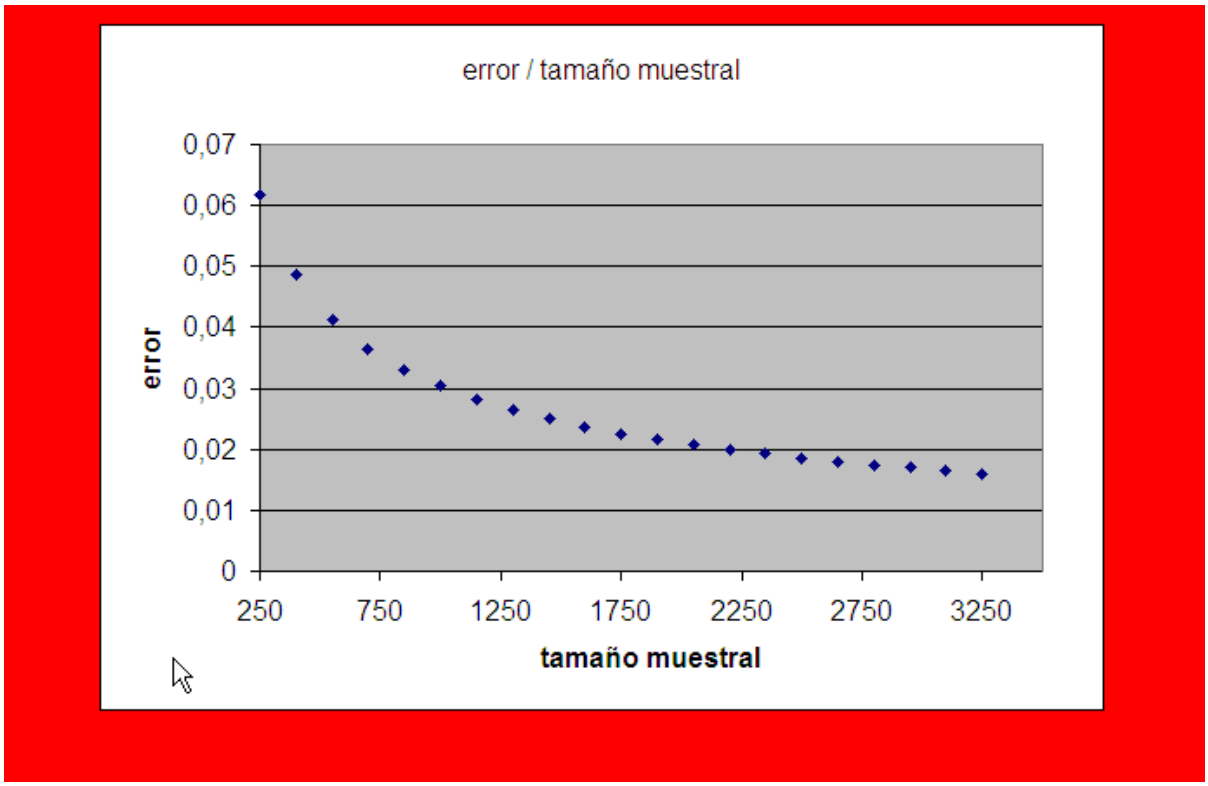


Figura 11: Curva de error al 95 % para $N = 250.000$

En el caso más desfavorable ($P = 0.5$), ambas expresiones coinciden.

Para fijar el valor de P, se sugiere utilizar la tabla siguiente:

Nivel de riesgo	P
Máximo	0.5
Alto	0.2
Medio	0.1
Bajo	0.05

(10)

Si se desea *contrastar* una hipótesis sobre el nivel de riesgo, puede hacerse un análisis como el descrito en la Sección 2.2. Supongamos que queremos comprobar la veracidad de la hipótesis de que el nivel de riesgo es medio o bajo (esto es, que $P \leq 0.1$). Si la población es grande (esto es, si N es grande), hacemos uso de la tabla de la Figura 4, y obtenemos (para $\alpha = 0,05$) un valor $n = 28$. Extraemos pues una muestra de $n = 28$ registros; si en ninguno de ellos observamos incumplimiento, entonces podemos aceptar la hipótesis de que el riesgo es moderado o bajo.

Para valores de N moderados o pequeños, determinamos n mediante la fórmula (5).

En la segunda columna de la Figura 6 se recoge el n que se obtiene al aplicar (5) para distintos tamaños poblacionales N.

Para contrastar la hipótesis de que el riesgo es bajo ($P \leq 0.05$) sería necesario examinar una muestra de $n = 58$ registros (para N grande), o bien el valor de n recogido en la tercera columna de la Figura 6, para valores moderados o pequeños de n.

Una vez valorado el nivel de riesgo (mediante un control como el descrito previamente, o bien por la experiencia del auditor en situaciones similares), podemos determinar, para un nivel de riesgo dado, el error ε a priori asociado a un tamaño muestral dado. En el caso de haberse tomado una muestra inicial para contrastar una hipótesis sobre nivel de riesgo, ésta no debería reutilizarse en el proceso posterior de inferencia.

Ejemplo 4 Con los datos del Ejemplo 2, bajo la hipótesis de máxima indeterminación ($P = 0,5$), se obtuvo, para $N = 250$, $n = 25$, un error $\varepsilon = 0,1863$. Si, en cambio, podemos asumir para el cálculo del error a priori, que $P = 0,02$, entonces el error obtenido, aplicando la fórmula (8), es $\varepsilon = 0,00977$.

El valor del error ε para distintos valores de P, tal como se obtendría en Excel, aparece recogido en la Figura 12.

□

P	0,5	0,25	0,1	0,05	0,02	0,004
error (P conocido)	0,186311	0,16135	0,112	0,081	0,052	0,024

Figura 12: Error con nivel de significación del 5 % para distintas hipótesis sobre P

Ejemplo 5 Para la base de datos de Contratos de 2004, con 77.555 registros, vimos en el Ejemplo 3 que se necesitan $n = 382,26 \approx 383$ registros para obtener un error del $0,05$ y un nivel de significación $\alpha = 0,05$ bajo el principio de máxima indeterminación, es decir: para la hipótesis de *riesgo máximo*. Si es posible admitir hipótesis de riesgo menor, se necesitan, obviamente, menores tamaños muestrales, como se ilustra en la Figura 13.

□

Dado que el hacer hipótesis sobre el valor P de la proporción puede ser determinante para reducir el tamaño muestral necesario, es conveniente representar las curvas con el error representado frente al tamaño muestral *para diferentes escenarios para* P, incluyendo el más desfavorable ($P = Q =$

	A	B	C	D	E
1	tamaño poblacional (N)	77555			
2	n. de atributos (k)	1			
3	alfa	0,05			
4	Z(1-alfa/2)	1,959963			
5	error	0,05			
6					
7	nivel de riesgo	Máximo	Alto	Medio	Bajo
8	P	0,5	0,2	0,1	0,05
9	n	382,2569	250,07	140,8628	74,40809
10	% de esfuerzo muestral	0,492885	0,3224	0,18163	0,095942

Figura 13: Tamaños muestrales en la base de Contratos para distintas hipótesis sobre P

$0,5$) y al mismo tiempo los más plausibles. En otras palabras, en lugar de representar la Figura 11, es deseable representar las curvas de la Figura 14.

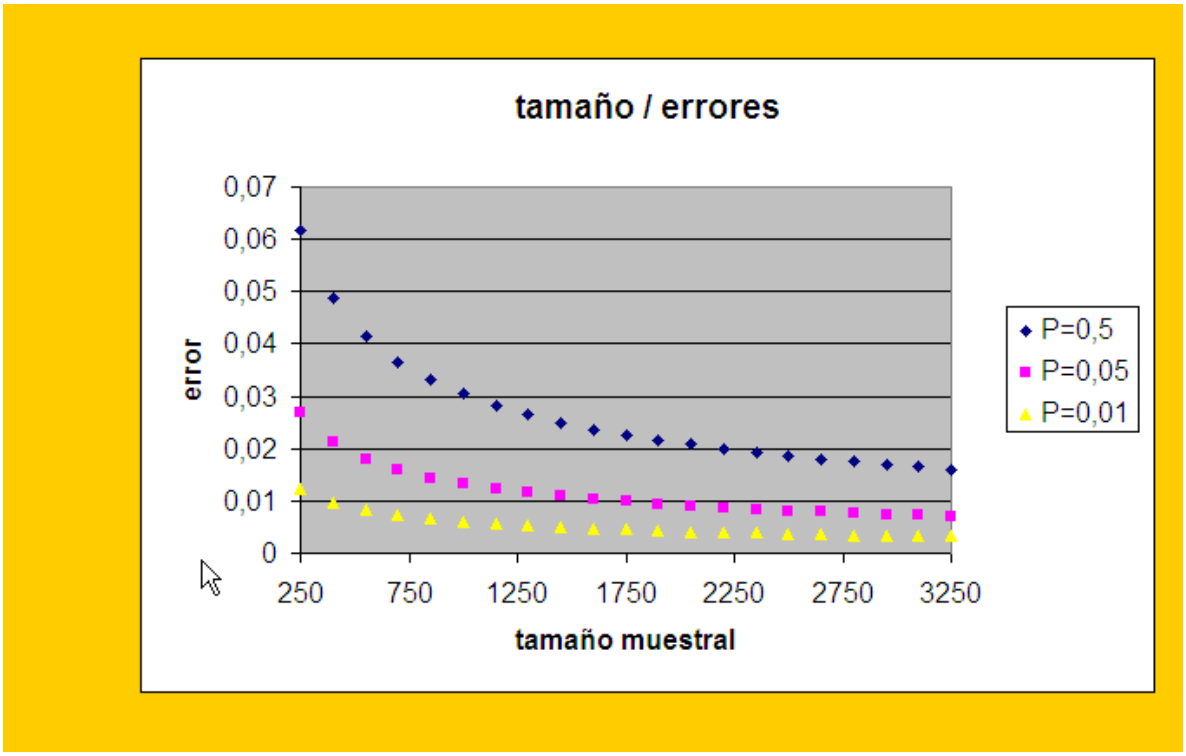


Figura 14: Curva de error al 95 % con $N = 250.000$ en distintos escenarios para P

En la hoja GRAFICA DE ERRORES del fichero Ejemplos.xls aparecen tabulados, para un tamaño poblacional N y un nivel de confianza α fijos, el error que se obtiene para distintos valores de n en tres escenarios distintos. Los resultados se representan gráficamente.

2.3.3. Error para varios atributos

Las expresiones anteriores del error son válidas si el análisis se limita a *una variable*, aunque ésta podría ser una variable compuesta, como la que se proponía en (2).

Si tenemos r atributos, y queremos analizar el error *no para un solo atributo concreto, sino para todos y cada uno de los r atributos*, hay que corregir las fórmulas anteriores. Usando el *método de Bonferroni*, se obtiene para el error

$$\varepsilon = Z_{1-\alpha/(2\cdot r)}\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)\frac{N}{4(N-1)}}. \tag{11}$$

Si las r variables son proporciones, $P_1,\ldots,P_r$, y sobre todas tenemos una cota superior $P \leq 0,5$ entonces la fórmula (8) se corrige, dando

$$\varepsilon = Z_{1-\alpha/(2\cdot r)}\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)\frac{N}{(N-1)}P(1-P)}. \tag{12}$$

En la Figura 15 aparece para una población de $N = 2.500$ y una muestra de tamaño $n = 125$, el error obtenido, para distintas hipótesis sobre la cota $P \leq 0,5$ de r proporciones.

(1 atributo)						
P	0,5	0,25	0,1	0,05	0,02	0,004
error (P conocido)	0,085449889	0,074001775	0,051269933	0,037246743	0,023925969	0,010787012
No. de atributos (r)	2					
Z(1-alfa/2r)	2,241401463					
P	0,5	0,25	0,1	0,05	0,02	0,004
error (P conocido)	0,097719971	0,084627977	0,058631982	0,042595148	0,027361592	0,012335961
No. de atributos (r)	3					
Z(1-alfa/2r)	2,393979431					
P	0,5	0,25	0,1	0,05	0,02	0,004
error (P conocido)	0,104372021	0,090388821	0,062623212	0,045494709	0,029224166	0,013175702
No. de atributos (r)	4					
Z(1-alfa/2r)	2,497706272					
P	0,5	0,25	0,1	0,05	0,02	0,004
error (P conocido)	0,108894273	0,094305207	0,065336564	0,047465913	0,030490397	0,013746581

Figura 15: Errores para tablas con r variables

Recíprocamente, para un error fijado ε, el tamaño muestral n que garantiza el error ε para el conjunto de r estimaciones es

$$n = \frac{1}{\frac{1}{N}+\frac{\varepsilon^2 4(N-1)}{Z^2_{1-\alpha/(2\cdot r)}N}}. \tag{13}$$

En la hoja TAM MUESTRAL VARIOS ATR del fichero Ejemplos.xls se calcula el tamaño muestral a partir del tamaño poblacional N, el nivel de significación α, el error ε y el nivel de riesgo P.

Si, además, podemos asumir que todas las proporciones son inferiores o iguales a un cierto $P_0 \leq \frac{1}{2}$, entonces podemos corregir la expresión (13), obteniendo en su lugar

$$n = \frac{1}{\frac{1}{N} + \frac{\varepsilon^2 (N-1)}{Z^2_{1-\alpha/(2\cdot r)} N P_0 (1-P_0)}}. \tag{14}$$

En la hoja TAM MUESTREO ALEAT 2 del fichero Ejemplos.xls aparece una tabla con la que, fijados el tamaño poblacional N, el tamaño muestral n, el nivel designificación α, y el número de atributos r, se determina, para distintos escenarios sobre P, el error ε cometido.

3. Generación de números aleatorios

El muestreo probabilístico se basa en la generación de la muestra a examinar mediante un mecanismo *aleatorio*.

Para poder simular en ordenador un mecanismo aleatorio, usaremos la utilidad de generación de números *pseudoaleatorios* implementada en Excel, y de características similares a las que aparecen en otros paquetes estadísticos.

Debe tenerse en cuenta que la función que vamos a usar no genera números auténticamente aleatorios: tiene una compleja fórmula matemática implementada, de modo que los números que se van generando *parecen* aleatorios, sin serlo. De esta manera, simulamos la generación aleatoria, pero con la ventaja de que podemos reproducir el experimento tantas veces como queramos con idéntico resultado.

Para reproducir la secuencia de números aleatorios tendremos que generar una *semilla*, esto es, un valor con el que poner en marcha el mecanismo aleatorio. Aunque la semilla propuesta puede quedar a la libre elección del usuario, proponemos un procedimiento estándar, que permita a cualquier usuario, en cualquier ordenador, obtener la misma semilla. Para su aplicación basta introducir los datos que se precisan en la Sección 3.1 en la hoja `SEMILLA` del fichero `Ejemplos.xls`.

3.1. Generación de una semilla

Siempre que usemos la misma semilla obtendremos la misma secuencia de números aleatorios.

La semilla está asociada a una serie de parámetros, que deben hacerse constar en los informes técnicos en los que se describa el proceso de muestreo empleado:

- La fecha en la que se tomó la muestra (e.g. 04/07/2005).
- El código de identificación del técnico (e.g. tarjeta 142).
- El número que ocupa la extracción de ese día por parte de ese técnico (un 1 si fue la primera tanda de números aleatorios del día, 2 si fue la segunda, ...).

De esta manera se obtiene un número de cuatro dígitos, que se usará como semilla para generar los números aleatorios.

La fórmula que se ha implantado es

$$\begin{aligned} \text{semilla} \quad = \quad & 10000 \cdot \text{frac}((198,593 + \text{serie} \cdot \text{tarjeta}) \cdot \log(\text{dia} \cdot \text{serie}) + \\ & +307,476 \cdot \log(\text{mes}) + 1,111 \cdot \log(1 + \text{año} \cdot \text{serie}) + 1000,923 \cdot \log(\text{diasem})), \end{aligned}$$

donde

- *frac* representa la parte fraccionaria. La parte fraccionaria de un número x se obtiene restando al número x su parte entera. La parte entera de un número x se obtiene en Excel con la función `ENTERO`
- *log* es el logaritmo neperiano. Se obtiene en Excel con la función `LN`
- *dia* es el día de la fecha, obtenida en Excel con la función `DIA`
- *mes* es el mes de la fecha, obtenido en Excel con la función `MES`
- *año* es el año de la fecha, obtenido en Excel con la función `AÑO`
- *diasem* se obtiene sobre la fecha con la fución `DIASEM` (vale 1 si el día es domingo, 2 si es lunes, 3 si es martes, etc.
- *tarjeta* es el número de tarjeta del usuario
- *serie* es el lugar que ocupa la semilla en la serie de semillas generadas ese día.

Si el número obtenido es menor que 1000, se corrige el valor obtenido para las decenas de millar (cero) por `DIASEM`.

Ejemplo 6 En la Figura 16 se muestran los tres valores que se obtendrían como semilla generadas el 04/07/2005 por el usuario con número de tarjeta 114 (los números 9014, 4005, 2988, 2124), las dos semillas que obtendría ese mismo día el usuario con número de tarjeta 43 (los números 4745, 1198), y la primera semilla que generaría este último usuario el 05/07/2005 (el 2238).

□

Fecha	Tarjeta	Serie	...	Semilla
04/07/2005	114	1		9014
04/07/2005	114	2		4005
04/07/2005	114	3		2988
04/07/2005	114	4		2124
04/07/2005	43	1		4745
04/07/2005	43	2		1198
05/07/2005	43	1		2238

Figura 16: Generación de semillas

3.2. Generación reproducible de números aleatorios

Para generar números aleatorios usaremos el módulo de `Análisis de Datos`, que aparece en Excel en el menú `Herramientas`, como se muestra en la Figura 17.

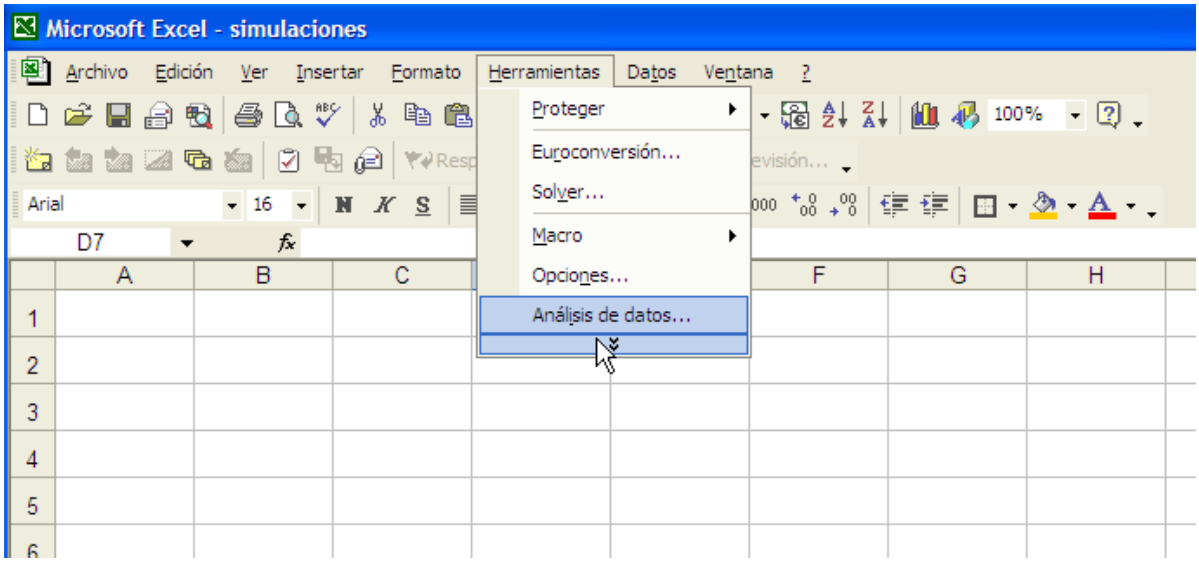


Figura 17: Utilidad de análisis de datos

Si esta utilidad no apareciera en el menú `Herramientas`, entonces es necesario instalarla. Para ello se selecciona en `Herramientas` la opción `Complementos` (ver Figura 18), y en ella, tras seleccionar `Herramientas para análisis`, como se muestra en la Figura 19, pulsamos `Aceptar`. Volviendo de nuevo a la opción `Herramientas`, veremos que sí aparece la opción de `Análisis de Datos`.

Una vez instalada la utilidad de `Análisis de Datos`, seleccionamos la opción de `Generación de números aleatorios`, como se muestra en la Figura 20. Al aceptar, nos aparece un cuadro de diálogo como el de la Figura 21.

La opción `Número de variables` nos pregunta cuántas columnas de números aleatorios deseamos generar. La `cantidad de números aleatorios`, las filas. La `distribución`, el tipo de números aleatorios que vamos a generar. Con las opciones seleccionadas en la Figura 21, se generarán 100 números aleatorios, `Uniformes entre` 0 `y` 1, en una columna de la hoja.

En `Iniciar con`, debemos introducir la semilla que se obtiene con el procedimiento descrito en la Sección 3.1.

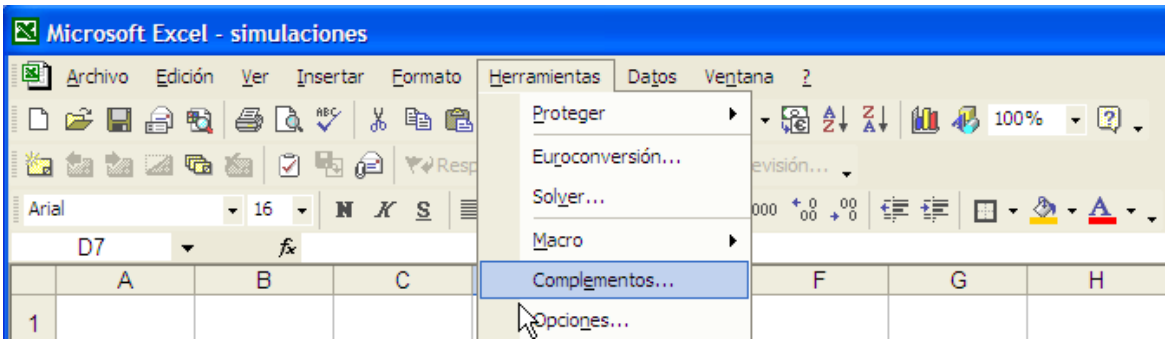


Figura 18: Instalación de la utilidad de análisis de datos

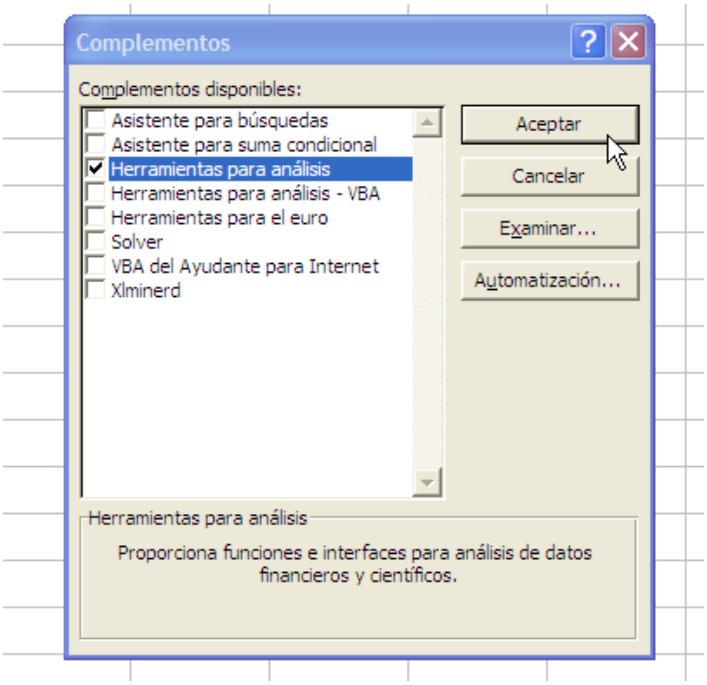


Figura 19: Instalación de la utilidad de análisis de datos

3.3. Generación de etiquetas

En muchas ocasiones los individuos van a venir descritos con un literal (por ejemplo, nombre del municipio), y deseamos asignarle a cada uno un código numérico correlativo. Si bien existen multitud de formas de hacer esto con Excel, una de ellas usa, precisamente, el módulo de generación de números aleatorios.

Si, en lugar de seleccionar en `distribución` la `uniforme` seleccionamos `frecuencia relativa`, nos aparece un cuadro como el de la Figura 22. Los parámetros a indicar son

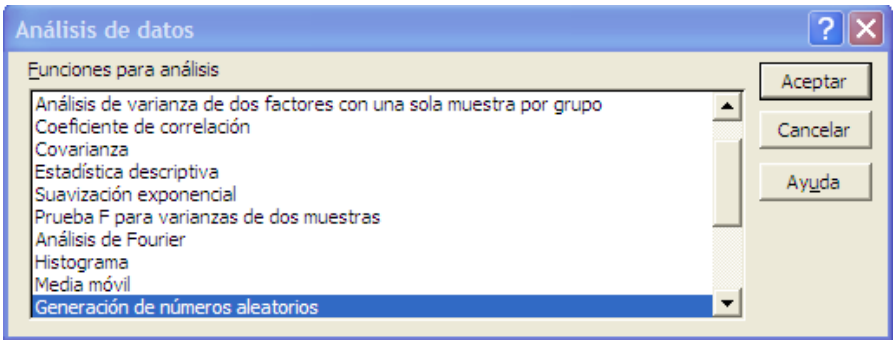


Figura 20: Generación de números aleatorios

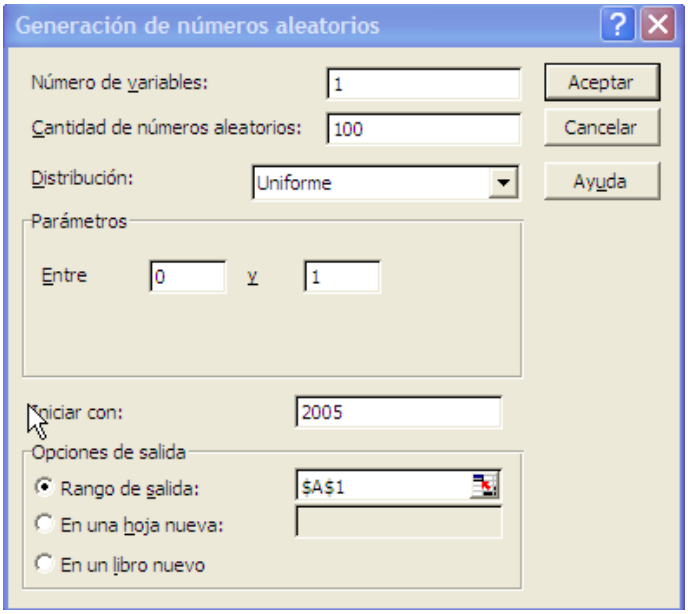


Figura 21: Generación de números aleatorios

- `De`: primer número generado.
- `a`: último número generado.
- `incremento`: el salto. Por ejemplo, `De` 1 a 10 con `incremento` 3, generaría la secuencia $1, 1+3=4, 4+3=7, 7+3=10$.
- `repitiendo cada número`: el número de veces seguidas que se escribe cada número la secuencia obtenida. Por ejemplo, poniendo 2, y con los parámetros anteriores, se genera la secuencia $1, 1, 4, 4, 7, 7, 10, 10$.
- `repitiendo la secuencia`: el número de veces que se repite la secuencia.

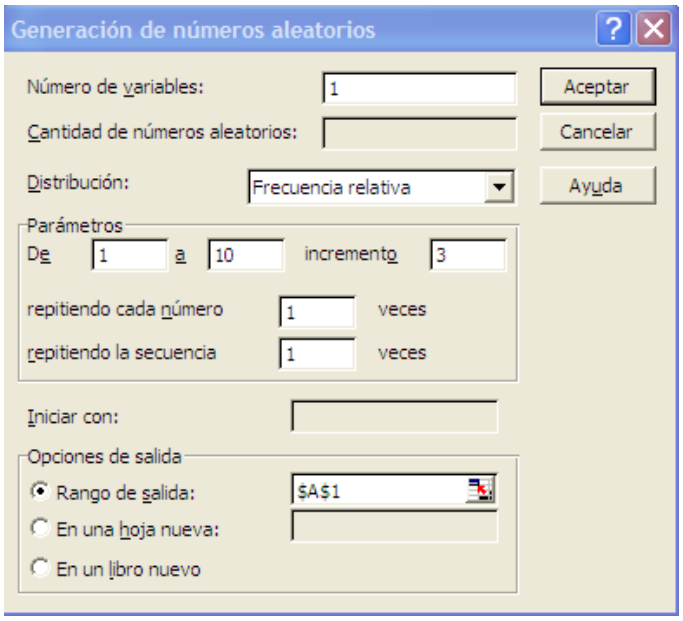


Figura 22: Generando etiquetas

3.4. Ordenación aleatoria

En ocasiones será útil ordenar aleatoriamente las etiquetas $1, 2, \ldots, N$, simulando el proceso en el que cada etiqueta estuviera pintada en un naipe, y pudiéramos barajarlos. Este procedimiento será especialmente útil en el muestreo por unidad monetaria, descrito en la Sección 4.3, pues con esto evitamos que la selección de la muestra venga afectada por la posición que ocupan los registros en la base de datos.

Para obtener una ordenación aleatoria de las etiquetas $1, 2, \ldots, N$, hacemos:

1. En la hoja de cálculo utilizamos la primera columna para anotar los números del 1 al N (las etiquetas). Para ello seguimos el procedimiento descrito en la Sección 3.3.

2. En la siguiente columna se generan, como se ha descrito en la Sección 3.2, N valores de una distribución uniforme entre 0 y 1.

3. Se ordenan las etiquetas en orden creciente de los números aleatorios de la segunda columna.

Ejemplo 7 Ordenemos aleatoriamente, siguiendo el algoritmo anterior, una población de $N = 250$ individuos. Para ello, generamos en la primera columna números correlativos de 1 a 250, y a continuación, en la segunda columna, 250 números uniformes entre 0 y 1. Marcamos las columnas `A` y `B`, y seleccionamos, en `Datos`, la opción `Ordenar`, como aparece en la Figura 23.

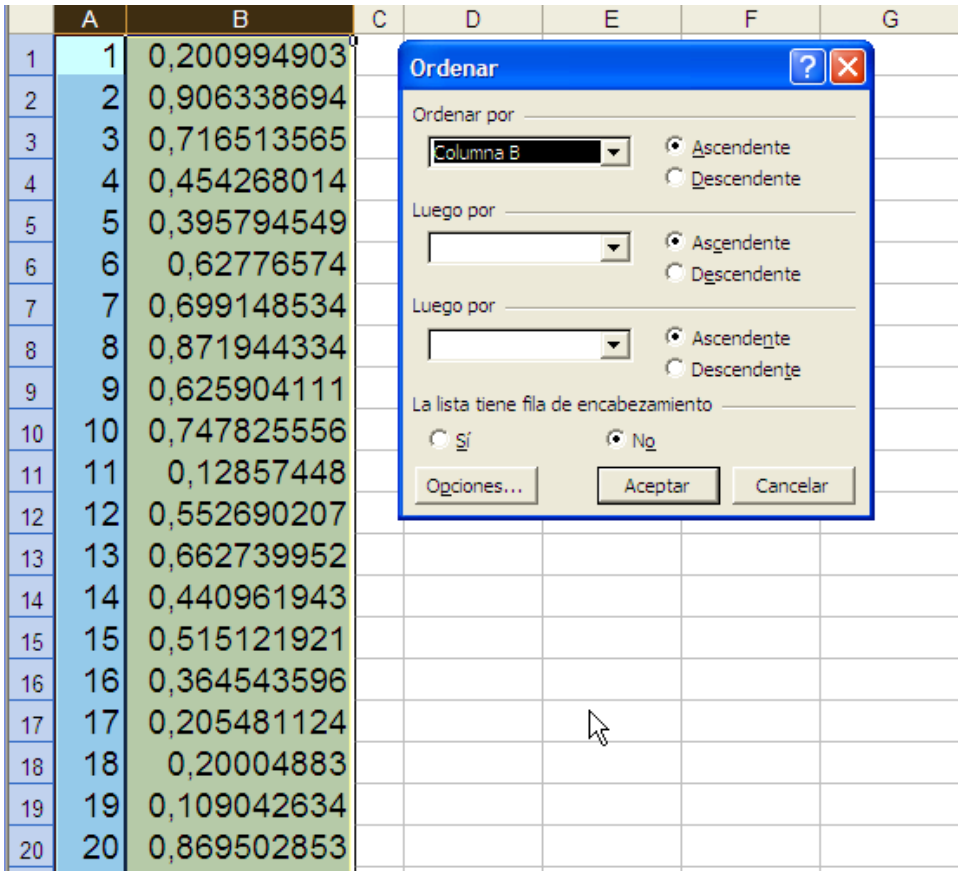


Figura 23: Ordenación aleatoria de una población de 250 individuos

Al ordenar los 250 individuos por el valor de su número aleatorio asociado se obtiene la tabla con la ordenación aleatoria.

□

Ejemplo 8 De la base de datos de contratos de 2004, extraemos, en la hoja de cálculo `DATOS 267` del fichero Excel `ejemplos.xls`, aquellos con el campo `uni-adm` 267, que se corresponde con los registros en los que el campo `den-uni-adm` toma el valor `D.G. ARQUITECTURA Y VIVIENDA`. Se obtienen 410 individuos, como se muestra en la Figura 24.

Con la semilla 2238 generamos 410 valores de una uniforme en $(0, 1)$, como se muestra en la Figura 25.

Al aplicar la opción `Ordenar` de Excel, obtenemos los contratos aleatoriamente permutados, como aparece en la Figura 26.

□

	A	B	C	D	E	F	G	H	I	J
1	entidad	ejercicio	num-cont	cod-contista	tip-cont	den-tip-cont	uni-adm	den-uni-adm	cod-prov	den-prov
2	JA	2004	1994041968	131564110Z 00	05	TRABAJOS ESPEC.(ANT LEY 13/95)	267	D.G. ARQUITECTURA Y VIVIENDA	29	MALAGA
3	JA	2004	2002315058	125975734V 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	23	JAEN
4	JA	2004	2003202035	2A14334411 00	87	OBRAS	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
5	JA	2004	2003385839	2B91217299 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
6	JA	2004	2004045152	128784136G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
7	JA	2004	2004050158	2B41999335 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
8	JA	2004	2004056754	124668169W 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
9	JA	2004	2004056754	124783792G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
10	JA	2004	2004056811	128232034V 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
11	JA	2004	2004056811	128531695B 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
12	JA	2004	2004056849	124668169W 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
13	JA	2004	2004056849	124783792G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
14	JA	2004	2004057327	124078518W 01	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
15	JA	2004	2004057327	124231395K 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
16	JA	2004	2004065555	2B41136193 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41	SEVILLA
17	JA	2004	2004070974	132855706E 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
18	JA	2004	2004071028	127321678R 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
19	JA	2004	2004072186	2B41056193 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
20	JA	2004	2004072211	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
21	JA	2004	2004072242	2B41056193 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
22	JA	2004	2004072262	127900152W 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
23	JA	2004	2004072383	127900152W 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
24	JA	2004	2004072400	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
25	JA	2004	2004072413	127900152W 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
26	JA	2004	2004072438	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
27	JA	2004	2004072590	127320803T 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
28	JA	2004	2004072768	102194118X 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
29	JA	2004	2004072843	2F91108696 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
30	JA	2004	2004072918	2A78357001 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
31	JA	2004	2004072949	2G41698283 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
32	JA	2004	2004073310	2G41698283 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
33	JA	2004	2004073345	2B37407004 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.
34	JA	2004	2004073383	2B91079400 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99	SIN PROV.

Figura 24: Los 410 contratos de la D.G. de Arquitectura y Vivienda

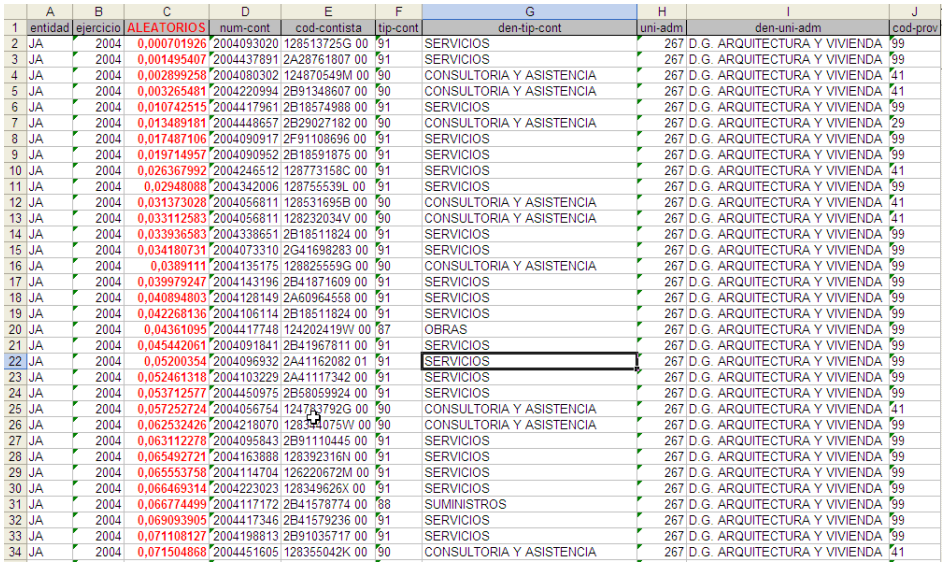
	A	B	C	D	E	F	G	H	I	J
1	entidad	ejercicio	ALEATORIOS	num-cont	cod-contista	tip-cont	den-tip-cont	uni-adm	den-uni-adm	cod-pro
2	JA	2004	0,224188971	1994041968	131564110Z 00	05	TRABAJOS ESPEC.(ANT LEY 13/95)	267	D.G. ARQUITECTURA Y VIVIENDA	29
3	JA	2004	0,333903012	2002315058	125975734V 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	23
4	JA	2004	0,741416669	2003202035	2A14334411 00	87	OBRAS	267	D.G. ARQUITECTURA Y VIVIENDA	41
5	JA	2004	0,55876339	2003385839	2B91217299 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	41
6	JA	2004	0,738547929	2004045152	128784136G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
7	JA	2004	0,54915006	2004050158	2B41999335 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
8	JA	2004	0,830530717	2004056754	124668169W 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
9	JA	2004	0,057252724	2004056754	124783792G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
10	JA	2004	0,033112583	2004056811	128232034V 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
11	JA	2004	0,031373028	2004056811	128531695B 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
12	JA	2004	0,076876125	2004056849	124668169W 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
13	JA	2004	0,865901669	2004056849	124783792G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
14	JA	2004	0,083315531	2004057327	124078518W 01	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
15	JA	2004	0,095187231	2004057327	124231395K 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
16	JA	2004	0,314615314	2004065555	2B41136193 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
17	JA	2004	0,648701437	2004070974	132855706E 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
18	JA	2004	0,897396771	2004071028	127321678R 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	99
19	JA	2004	0,214362011	2004072186	2B41056193 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
20	JA	2004	0,12652974	2004072211	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
21	JA	2004	0,143925291	2004072242	2B41056193 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
22	JA	2004	0,300180059	2004072262	127900152W 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
23	JA	2004	0,295602283	2004072383	127900152W 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
24	JA	2004	0,148625141	2004072400	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
25	JA	2004	0,875972777	2004072413	127900152W 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
26	JA	2004	0,922177801	2004072438	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
27	JA	2004	0,336741234	2004072590	127320803T 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
28	JA	2004	0,219824824	2004072768	102194118X 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
29	JA	2004	0,33027131	2004072843	2F91108696 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
30	JA	2004	0,79955443	2004072918	2A78357001 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
31	JA	2004	0,251167333	2004072949	2G41698283 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
32	JA	2004	0,034180731	2004073310	2G41698283 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
33	JA	2004	0,622547075	2004073345	2B37407004 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
34	JA	2004	0,492110965	2004073383	2B91079400 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99

Figura 25: Insertando los valores aleatorios en los contratos de la D.G. de Arquitectura y Vivienda

3.5. Generación no reproducible de números aleatorios

Si, en el procedimiento de la Sección 3.2, dejamos en blanco el valor de la celda `Iniciar con`, Excel genera internamente una semilla. El proceso no es, por tanto, reproducible.

Si buscábamos obtener uno o varios valores (pseudo)aleatorios, sin garantizar la reproducibilidad de la secuencia generada, Excel proporciona un procedimiento más simple: la función `ALEATORIO()`,



	A	B	C	D	E	F	G	H	I	J
1	entidad	ejercicio	ALEATORIOS	num-cont	cod-contista	tip-cont	den-tip-cont	uni-adm	den-uni-adm	cod-prov
2	JA	2004	0,000701926	2004093020	128513725G 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
3	JA	2004	0,001495407	2004437891	2A28761807 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
4	JA	2004	0,002899258	2004080302	124870549M 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
5	JA	2004	0,003265481	2004220994	2B91348607 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
6	JA	2004	0,010742515	2004417961	2B18574988 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
7	JA	2004	0,013489181	2004448657	2B29027182 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	29
8	JA	2004	0,017487106	2004090917	2F91108696 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
9	JA	2004	0,019714957	2004090952	2B18591875 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
10	JA	2004	0,026367992	2004246512	128773158C 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	41
11	JA	2004	0,02948088	2004342006	128755539L 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
12	JA	2004	0,031373028	2004056811	128531695B 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
13	JA	2004	0,033112583	2004056811	128232034V 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
14	JA	2004	0,033936583	2004338651	2B18511824 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
15	JA	2004	0,034180731	2004073310	2G41698283 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
16	JA	2004	0,0389111	2004135175	128825559G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	99
17	JA	2004	0,039979247	2004143196	2B41871609 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
18	JA	2004	0,040894803	2004128149	2A60964558 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
19	JA	2004	0,042268136	2004106114	2B18511824 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
20	JA	2004	0,04361095	2004417748	124202419W 00	87	OBRAS	267	D.G. ARQUITECTURA Y VIVIENDA	99
21	JA	2004	0,045442061	2004091841	2B41967811 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
22	JA	2004	0,05200354	2004096932	2A41162082 01	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
23	JA	2004	0,052461318	2004103229	2A41117342 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
24	JA	2004	0,053712577	2004450975	2B58059924 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
25	JA	2004	0,057252724	2004056754	124783792G 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41
26	JA	2004	0,062532426	2004218070	128314075W 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	99
27	JA	2004	0,063112278	2004095843	2B91110445 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
28	JA	2004	0,065492721	2004163888	128392316N 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
29	JA	2004	0,065553758	2004114704	126220672M 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
30	JA	2004	0,066469314	2004223023	128349626X 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
31	JA	2004	0,066774499	2004117172	2B41578774 00	88	SUMINISTROS	267	D.G. ARQUITECTURA Y VIVIENDA	99
32	JA	2004	0,069093905	2004417346	2B41579236 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
33	JA	2004	0,071108127	2004198813	2B91035717 00	91	SERVICIOS	267	D.G. ARQUITECTURA Y VIVIENDA	99
34	JA	2004	0,071504868	2004451605	128355042K 00	90	CONSULTORIA Y ASISTENCIA	267	D.G. ARQUITECTURA Y VIVIENDA	41

Figura 26: Ordenación aleatoria de los contratos de la D.G. de Arquitectura y Vivienda

que genera, cada vez que se la invoca, un valor pseudoaleatorio entre 0 y 1.

Para introducir en Excel la función `ALEATORIO()` basta con escribirla sobre la celda correspondiente, como se muestra, por ejemplo, en la Figura 27. Alternativamente, podemos utilizar la utilidad de Excel para inserción de funciones, posicionando el cursor sobre la opción f_x, como se mostró en la Figura 1. Al pinchar nos aparece una ventana como la de la Figura 2. En la opción `o seleccionar una categoría` seleccionamos `Todas`, como en la Figura 2. En la lista de funciones aparece la función `ALEATORIO`, como se aprecia en la Figura 28. Aparece la ventana mostrada en la Figura 29. Pulsando `Aceptar` obtenemos en la celda que tuviéramos seleccionada un número aleatorio entre 0 y 1.

Para fijar el contenido de las celdas donde hemos llamado a la función `ALEATORIO()`, seguimos el siguiente proceso:

1. Seleccionar las celdas en las que se haya usado la función `ALEATORIO()`.

2. Utilizar la opción `Copiar`.

3. Seleccionar, en `Edición,` la opción `Pegado especial ...`, como en la Figura 30.

4. Seleccionar la opción `Valores` en `Pegar`, como en la Figura 31.

A partir de ese momento, las celdas en las que aparecían los números aleatorios quedan fijas.

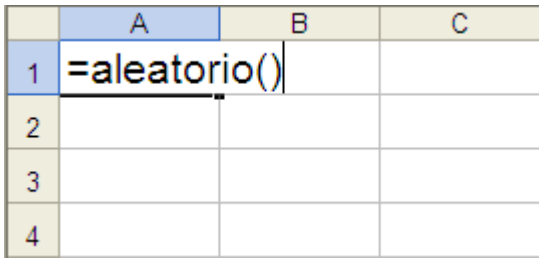


Figura 27: Función `ALEATORIO()`

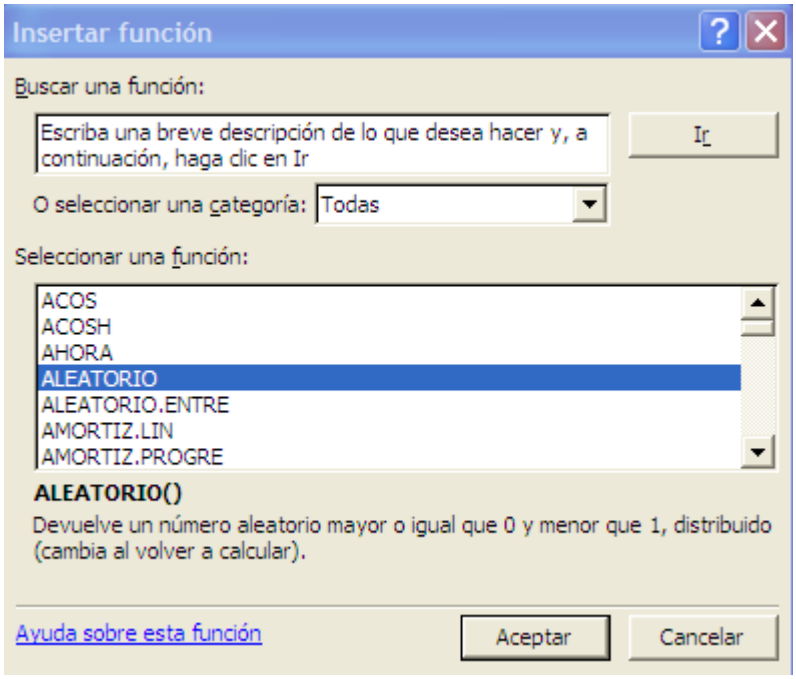


Figura 28: Función `ALEATORIO()`

3.6. Generación de números en un intervalo

De esta forma obtenemos valores pseudoaleatorios entre 0 y 1. Si nuestro propósito es obtener números enteros (pseudo)aleatorios entre 1 y N, aplicamos, a cada U uniforme entre 0 y 1, la fórmula

$$X = 1 + \lfloor N \cdot U \rfloor. \tag{15}$$

En (15), $\lfloor z \rfloor$ denota el mayor natural menor o igual que z, que en Excel aparece como función `ENTERO`.

Por ejemplo, para obtener con Excel $\lfloor 2,8 \rfloor$, usamos `=ENTERO(2,8)`, que nos dará como resultado 2.

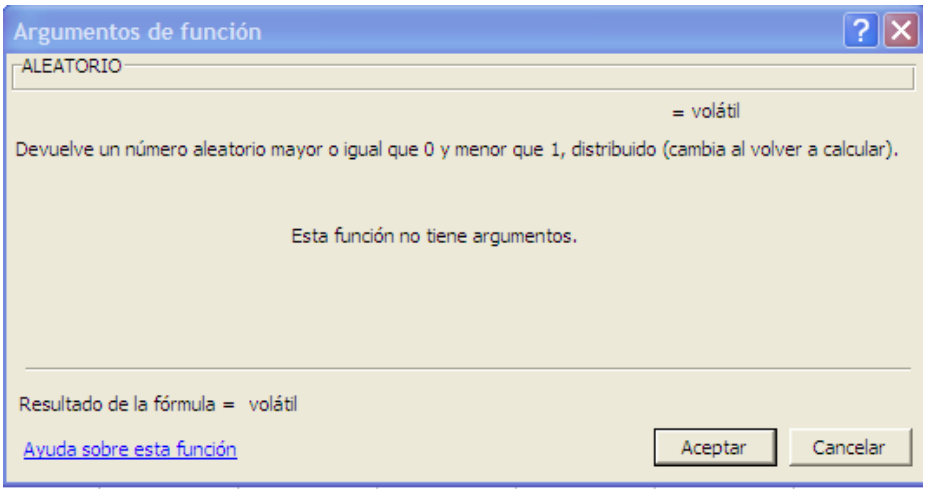


Figura 29: Función `ALEATORIO()`

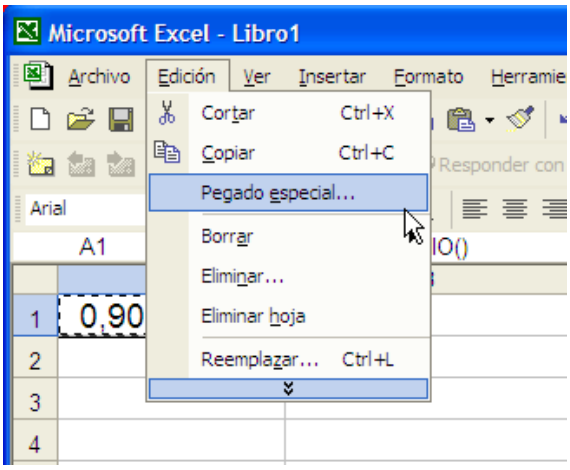


Figura 30: Pegado especial

Ejemplo 9 Generemos, a partir del número $U = 0,2462$ un valor aleatorio entre 1 y 136. Para ello aplicamos la fórmula (15); en Excel: `=1+ENTERO(136*0,2462)`, dando el resultado de 34.

□

En la Figura 32 se generan números aleatorios entre 0 y 1. A partir de éstos, obtenemos en la Figura 33 números aleatorios enteros entre 1 y 250.

En la hoja `ALEAT 1..N` del fichero `Ejemplos.xls` se muestra cómo, a partir de un conjunto de 5.000 valores aleatorios en 0 y 1 se generan, *con reposición*, 5.000 valores aleatorios entre 1 y N, para un valor de N especificado en la hoja.

Obsérvese que este procedimiento *no garantiza que los números obtenidos sean distintos*, co-

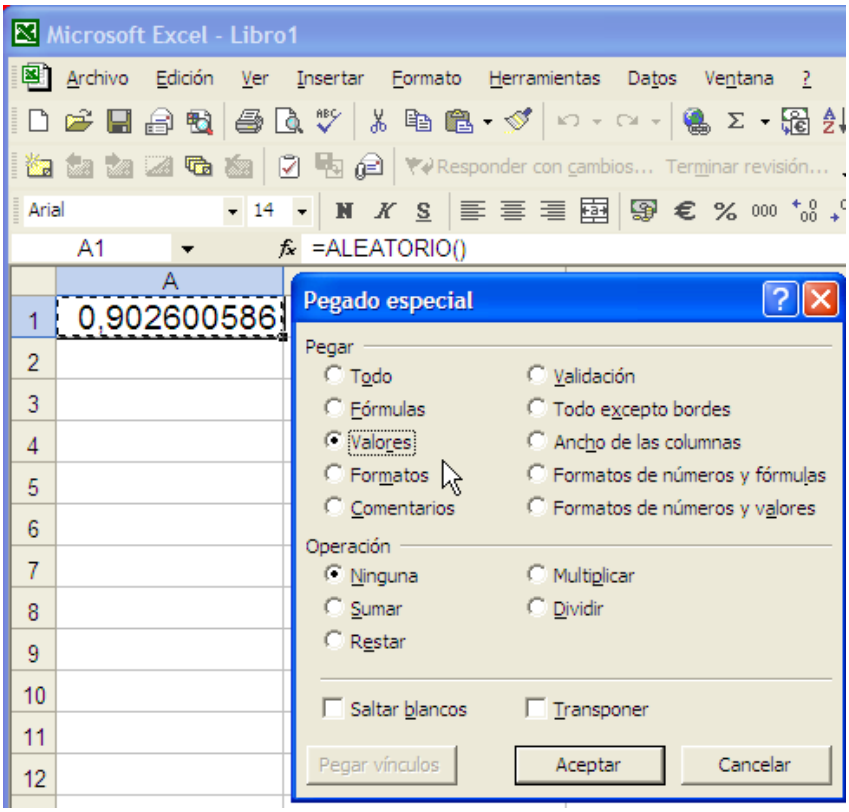


Figura 31: Pegado especial

mo puede verse en la Figura 33 con los valores generados en lugar 6 y 9. Si nuestro objetivo es extraer una muestra aleatoria sin repeticiones, entonces tendremos que usar procedimientos más sofisticados, como los que se describen en la Sección 4.1.2.

1	0,200994903
2	0,906338694
3	0,716513565
4	0,454268014
5	0,395794549
6	0,62776574
7	0,699148534
8	0,871944334
9	0,625904111
10	0,747825556
11	0,12857448
12	0,552690207
13	0,662739952
14	0,440961943
15	0,515121921
16	0,364543596
17	0,205481124
18	0,20004883
19	0,109042634
20	0,869502853

Figura 32: Generación de números aleatorios entre 0 y 1 (semilla fija)

1	0,200994903	51
2	0,906338694	227
3	0,716513565	180
4	0,454268014	114
5	0,395794549	99
6	0,62776574	157
7	0,699148534	175
8	0,871944334	218
9	0,625904111	157
10	0,747825556	187
11	0,12857448	33
12	0,552690207	139
13	0,662739952	166
14	0,440961943	111
15	0,515121921	129
16	0,364543596	92
17	0,205481124	52
18	0,20004883	51
19	0,109042634	28
20	0,869502853	=1+ENTERO(A20*250)

Figura 33: Generación de números aleatorios enteros entre 1 y 250 (semilla fija)

4. Diseños muestrales

Diremos que la población es *homogénea* cuando

- Todos los individuos pueden ser muestreados siguiendo el mismo procedimiento.
- El nivel de riesgo y la prioridad de análisis son los mismos en todos los individuos.
- Las variables objeto de estudio toman valores parecidos en los mismos.

En tales casos, se propone usar un diseño que sea

1. Fácilmente implementable por personal no especializado.
2. Reproducible: a partir de una información de base (la semilla), puede obtenerse, si se desea, la misma muestra.

Las propuestas en esta línea para seleccionar n individuos de una población de tamaño N son:

1. Muestreo aleatorio simple sin reposición.
2. Muestreo sistemático.
3. Muestreo por unidad monetaria.

4.1. Muestreo aleatorio simple sin reposición

4.1.1. Generación de la muestra. Método de eliminación de duplicados

Para generar una muestra aleatoria simple sin reposición de tamaño n de una población de N individuos, tenemos un primer procedimiento basado en la generación sucesiva de números aleatorios en un intervalo, como se describió en la Sección 3.6. Este método, denominado **método de generación de muestras aleatorias basado en la eliminación de duplicados**, actúa como sigue:

1. Fijamos una semilla s.
2. Hacemos $k = 1$.
3. Generamos $k \cdot n$ números aleatorios U entre 0 y 1 a partir de la semilla s.
4. Generamos $k \cdot n$ números aleatorios X en el conjunto $\{1, 2, \ldots, N\}$ aplicando la fórmula (15).
5. Si se han obtenido al menos n números aleatorios distintos en $\{1, 2, \ldots, N\}$, tomar los n primeros obtenidos. Si no, volver al paso 2, incrementando en una unidad el valor de k.

El método de generación de muestras aleatorias basado en la eliminación de duplicados es el recomendado cuando la muestra que se vaya a extraer tenga un tamaño n pequeño, donde sea fácil detectar visualmente duplicados, y cuando sea también pequeña la ratio n/N, para que sea poco probable la aparición de duplicados. Por ejemplo, en una población de $N = 200$ individuos, si extraemos $n = 10$ números aleatorios, la probabilidad de que haya duplicados es del 24 %, mientras que si extraemos $n = 30$ números aleatorios, la probabilidad de que haya duplicados crece hasta el 91 %.

Para identificar los elementos generados, podemos usar la función BUSCAR. Podemos utilizar la utilidad de Excel para inserción de funciones, posicionando el cursor sobre la opción f_x, como se mostró en la Figura 1. Al pinchar nos aparece una ventana como la de la Figura 2. En la opción `o seleccionar una categoría` seleccionamos `Todas`, como en la Figura 2. En la lista de funciones aparece la función BUSCAR, como se aprecia en la Figura 34. Pulsando `Aceptar` obtenemos la ventana de la Figura 35. De las dos opciones presentadas, escogemos la primera (`valor buscado; vector de comparación; resultado`), y obtenemos una ventana como la que aparece en la Figura 36. Debemos indicar a continuación 3 datos:

- `Valor buscado` (el valor que queremos buscar).
- `Vector de comparación` (dónde debemos buscar el valor).
- `Vector resultado` (qué valor queremos que indique sobre el registro en el que se da la coincidencia).

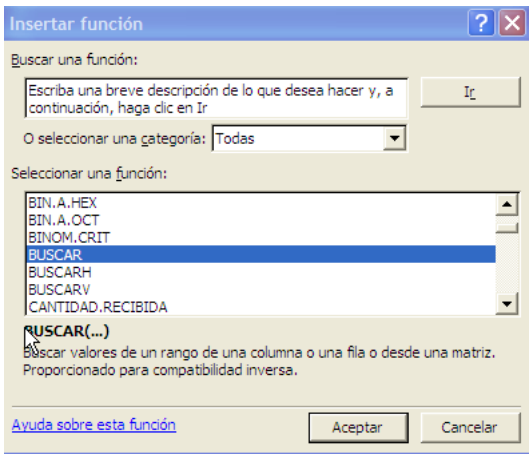


Figura 34: Función `BUSCAR`

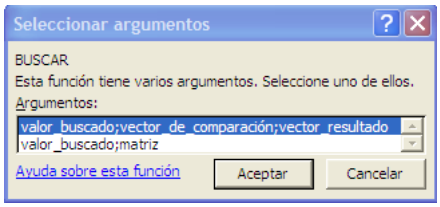


Figura 35: Función `BUSCAR`

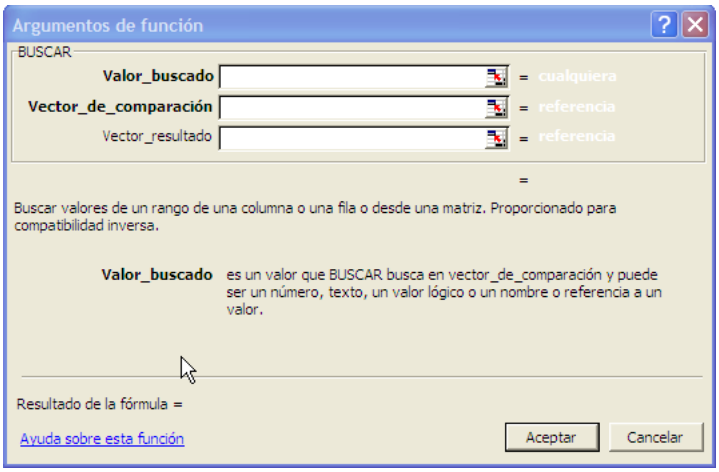


Figura 36: Función `BUSCAR`

Ejemplo 10 Supongamos que queremos generar una muestra aleatoria simple sin reposición de $n = 5$ contratos de entre los 410 contratos de la D.G. de Arquitectura y Vivienda, recogidos en la hoja `DATOS 267` del fichero `ejemplo.xls`

Con la semilla 1190 generamos, en otra hoja, $n = 5$ números aleatorios en $(0, 1)$, y a partir de cada x así obtenido, aplicando la fórmula `=1+ENTERO(410*x)`, obtenemos los valores aleatorios que se muestran en la Figura 37.

	A	B
1	uniforme	aleatorio
2	0,119754631	50
3	0,573839534	236
4	0,40192877	165
5	0,955717643	392
6	0,853480636	350

Figura 37: Generación de 5 números aleatorios enteros entre 1 y 410 (semilla 1190)

Si hubiera repeticiones, en lugar de 5, generaríamos más valores, hasta obtener 5 distintos. Ahora bastaría buscar en la hoja de los contratos cuáles son los contratos número $50, 236$, etc. Esto puede hacerso manualmente, o bien con ayuda de la función BUSCAR, obteniéndose los contratos que se indican en la Figura 38.

	A	B	C
1	uniforme	aleatorio	num-cont
2	0,119754631	50	2004219176
3	0,573839534	236	2004091020
4	0,40192877	165	2004451727
5	0,955717643	392	2004441427
6	0,853480636	350	2004143332

Figura 38: Generación de muestra aleatoria de 5 contratos de la D.G. de Arquitectura y Vivienda

□

4.1.2. Generación de la muestra. Método de ordenación aleatoria

Cuando el tamaño muestral vaya a ser grande en términos absolutos, o relativos con respecto a n, en lugar de usar el método de generación de muestras aleatorias basado en la eliminación de duplicados, usaremos el **método de generación de muestras aleatorias basado en ordenación aleatoria:**

1. En la hoja de cálculo utilizamos la primera columna para anotar los números del 1 al N (las etiquetas). Para ello seguimos el procedimiento descrito en la Sección 3.3.

2. Ordenamos aleatoriamente las etiquetas, como se describió en la Sección 3.4.

3. Se toma como muestra aleatoria simple el conjunto de n etiquetas con menores números aleatorios.

Ejemplo 11 Generemos, siguiendo el algoritmo anterior, una muestra de $n = 10$ individuos de una población de $N = 250$ individuos. Para ello, generamos una ordenación aleatoria, como se descibió en el Ejemplo 7, y se mostró en la Figura 23. Los 10 primeros elementos, marcados en otro tono en la Figura 39, constituyen la muestra aleatoria simple.

□

	A	B	C
1	118	0,002441481	
2	45	0,008270516	
3	132	0,015991699	
4	143	0,022034364	
5	80	0,024231697	
6	235	0,034546953	
7	245	0,03604236	
8	96	0,044526505	
9	231	0,054231391	
10	157	0,056642354	
11	189	0,058992279	
12	240	0,063447981	
13	129	0,074159978	
14	161	0,075807978	
15	97	0,081423383	
16	25	0,086336863	
17	190	0,089388714	
18	197	0,090029603	
19	93	0,09726249	
20	61	0,100711081	

Figura 39: Generación de muestra de 10 en una población de 250 individuos

Ejemplo 12 Generemos, siguiendo el método de la permutación aleatoria, una muestra aleatoria simple de 25 contratos de la hoja `DATOS 267` del fichero `ejemplo.xls`.

Construimos una ordenación aleatoria de todos los contratos, como se ilustró en la Figura 26. Los 25 primeros (marcados en otro tono en la Figura 40) forman la muestra.

□

La limitación de este procedimiento, en cuanto al tamaño de n, está en la limitación de Excel para generar números aleatorios. Por ejemplo, en la Figura 41 se muestran los primeros elementos de la muestra que se obtendrían por el método de la ordenación aleatoria para $N = 30.000$.

4.1.3. Estimación

En el muestreo aleatorio simple, estimamos el parámetro ϑ como sigue:

- **la media de una variable** Y mediante la media $\overline{y}$ de la muestra.

	A	B	C	D	E
1	entidad	ejercicio	ALEATORIOS	num-cont	cod-contista
2	JA	2004	0,000701926	2004093020	128513725G 00
3	JA	2004	0,001495407	2004437891	2A28761807 00
4	JA	2004	0,002899258	2004080302	124870549M 00
5	JA	2004	0,003265481	2004220994	2B91348607 00
6	JA	2004	0,010742515	2004417961	2B18574988 00
7	JA	2004	0,013489181	2004448657	2B29027182 00
8	JA	2004	0,017487106	2004090917	2F91108696 00
9	JA	2004	0,019714957	2004090952	2B18591875 00
10	JA	2004	0,026367992	2004246512	128773158C 00
11	JA	2004	0,02948088	2004342006	128755539L 00
12	JA	2004	0,031373028	2004056811	128531695B 00
13	JA	2004	0,033112583	2004056811	128232034V 00
14	JA	2004	0,033936583	2004338651	2B18511824 00
15	JA	2004	0,034180731	2004073310	2G41698283 00
16	JA	2004	0,0389111	2004135175	128825559G 00
17	JA	2004	0,039979247	2004143196	2B41871609 00
18	JA	2004	0,040894803	2004128149	2A60964558 00
19	JA	2004	0,042268136	2004106114	2B18511824 00
20	JA	2004	0,04361095	2004417748	124202419W 00
21	JA	2004	0,045442061	2004091841	2B41967811 00
22	JA	2004	0,05200354	2004096932	2A41162082 01
23	JA	2004	0,052461318	2004103229	2A41117342 00
24	JA	2004	0,053712577	2004450975	2B58059924 00
25	JA	2004	0,057252724	2004056754	124783792G 00
26	JA	2004	0,062532426	2004218070	128344075W 00
27	JA	2004	0,063112278	2004095843	2B91[illegible]445 00
28	JA	2004	0,065492721	2004163888	128392316N 00
29	JA	2004	0,065553758	2004114704	126220672M 00
30	JA	2004	0,066469314	2004223023	128349626X 00
31	JA	2004	0,066774499	2004117172	2B41578774 00

Figura 40: Generación de muestra de 25 contratos

- **la proporción de individuos que tienen una característica** mediante la proporción p de individuos con la característica en la muestra.

- **el total de una variable** Y mediante una regla de tres: $N\overline{y}$.

- **el total de individuos que tienen una característica** mediante la fórmula $N \cdot p$, siendo p la proporción de individuos con la característica en la muestra.

Los errores no se pueden calcular, pues si se pudiese conocer el error cometido, se corregiría el estimador, anulando el error. En cambio, sí es posible construir a posteriori *estimadores del error* a un nivel de significación determinado (el 5 %, por ejemplo). Así,

	A	B
1	3652	3,05185E-05
2	21639	0,000152593
3	12991	0,000244148
4	15840	0,000274667
5	8927	0,000305185
6	8618	0,000366222
7	9856	0,000366222
8	20198	0,000366222
9	26950	0,000366222
10	8217	0,000396741
11	23505	0,000396741
12	29045	0,000396741
13	24737	0,000427259
14	26384	0,000427259
15	462	0,000457778
16	12539	0,00061037
17	12866	0,00061037

Figura 41: Generación de muestra de en una población de 30.000 individuos

- el estimador del error al nivel $1-\alpha$ al estimar la media se obtiene mediante la fórmula

$$\widehat{\varepsilon} = Z_{1-\alpha/2}\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)S^2},$$

donde S^2 es la cuasivarianza de Y en la muestra, que se obtiene en Excel con la función VAR

- el estimador del error al nivel $1-\alpha$ al estimar la proporción de individuos que tienen una característica mediante la proporción p de individuos con la característica en la muestra se obtiene mediante la fórmula

$$\widehat{\varepsilon} = Z_{1-\alpha/2}\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)\frac{n}{n-1}p(1-p)} \tag{16}$$

- el estimador del error al nivel $1-\alpha$ al estimar el total de una variable Y por $N\overline{y}$ es

$$\widehat{\varepsilon} = Z_{1-\alpha/2}N\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)S^2} \tag{17}$$

- el estimador del error al nivel $1-\alpha$ al estimar el total de individuos que tienen una característica por $N \cdot p$, siendo p la proporción de individuos con la característica en la muestra,

es

$$\widehat{\varepsilon} = Z_{1-\alpha/2} N \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{n-1} p(1-p)}.$$

La manera habitual de presentar la información es en la forma

$$\text{estimador} \pm \text{estimador del error},$$

que se conoce como *intervalo de confianza*.

Ejemplo 13 Supongamos que, sobre una población de $N = 620$ individuos, queremos estimar la proporción de individuos con una característica. Para ello seleccionamos una muestra aleatoria simple de $n = 31$ individuos, de los que 4 presentan la característica.

La proporción muestral es, por tanto, $p = 4/31 = 0,12903$. En otras palabras, estimamos que el 12.9 % de los individuos tienen la característica en cuestión. Si queremos construir un intervalo de confianza al 95 %, debemos estimar el error para $\alpha = 0,05$. Aplicando la fórmula (21), el estimador del error es $\widehat{\varepsilon}$,

$$\widehat{\varepsilon} = Z_{1-\alpha/2} \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{n-1} p(1-p)} = 0,11692.$$

Por tanto, el intervalo de confianza al 95 % para la proporción es $0,12903 \pm 0,11692 = [0,01211\,;\,0,24596]$.

□

4.2. Muestreo sistemático

Para utilizar sólo excepcionalmente, cuando existan circunstancias que justifiquen sus ventajas sobre el aleatorio simple. La mayor ventaja es que, para generar la muestra, se necesita *sólo un número aleatorio*, que se puede incluir en el informe correspondiente, facilitando la trazabilidad del proceso de generación de la muestra.

En cambio, tiene el inconveniente de que, si bien todos los elementos pueden ser seleccionados, no todas las parejas de elementos pueden ser seleccionados. Por ello es conveniente ordenar aleatoriamente los datos antes de aplicar el muestreo.

4.2.1. Generación de la muestra

1. Se etiquetan los individuos del 1 al N.

2. Se determina el *paso* $k = \frac{N}{n}$, redondeado al entero más próximo, con la función `REDONDEO`.

3. Se genera un número aleatorio entre 1 y k. Si no se requiere reproducibilidad, usar la función `ALEATORIO.ENTRE`; si se requiere reproducibilidad, con la utilidad de *generación de números aleatorios*, generar un valor U de una distribución uniforme entre 0 y 1, (la semilla, como se especificó en la Sección 3.2), y aplicar la fórmula $k^* = 1 + \lfloor k \cdot U \rfloor$, donde $\lfloor z \rfloor$ denota el mayor natural menor o igual que z.

4. La muestra es la de etiquetas $k^*, k^* + k, k^* + 2 \cdot k \ldots$

Ejemplo 14 Extraigamos una muestra sistemática de $n = 5$ individuos de una población de $N = 60$ individuos. Para ello,

1. Etiquetamos los individuos del 1 al N.

2. Determinamos el *paso* $k = \frac{N}{n} = 12$.

3. Se genera un número aleatorio entre 1 y 12. Usamos la función `ALEATORIO.ENTRE`, como `ALEATORIO.ENTRE(1;12)`. Obtenemos el valor $k^* = 8$ (obsérvese que, al usar un mecanismo irreproducible, no tenemos garantías de obtener idéntico valor al repetir este ejercicio).

4. La muestra es la de etiquetas $8, 20, 32, 44, 56$, como se muestra en la Figura 42.

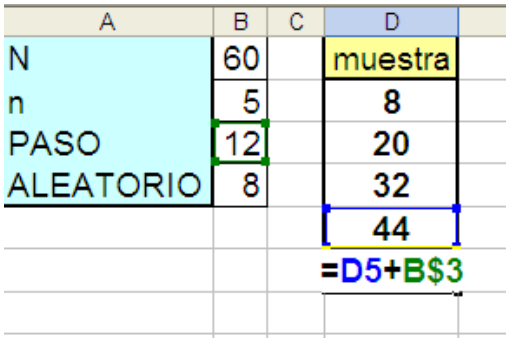


Figura 42: Generación de muestra sistemática

□

Con este procedimiento, no tenemos garantías de generar una muestra de tamaño *exactamente* n : puede diferir, si n no es divisor de N, en una unidad.

Ejemplo 15 Extraigamos una muestra sistemática de $n = 5$ individuos de una población de $N = 128$ individuos. Para ello,

1. Etiquetamos los individuos del 1 al $N = 128$.

2. Redondeamos $\frac{N}{n} = 25,6$ al entero más próximo, haciendo `REDONDEAR(25,6;0)`.

3. Generamos un número aleatorio entre 1 y 16, con `ALEATORIO.ENTRE(1;26)`. Obtenemos el valor $k^* = 5$.

4. La muestra es la de etiquetas $5, 31, 57, 83, 109$.

En cambio, si el número aleatorio generado con la función `ALEATORIO.ENTRE` fuera, por ejemplo, el 25, la muestra sería la formada por las etiquetas $25, 51, 77, 103$, pues la quinta (129) ya cae fuera del rango.

□

Por la naturaleza del método, la inclusión en la muestra de algunos individuos automáticamente excluye la de otros. Por ejemplo, dos unidades con etiquetas consecutivas nunca podrían estar simultáneamente en la muestra. para evitar este fenómeno, *los datos deben ser previamente reordenados, mediante el método de la permutación aleatoria.*

Ejemplo 16 De la base de datos de contratos de 2004, extraemos, en la hoja `DATOS 267` del fichero Excel `ejemplos.xls`, los 410 registros con el campo `uni-adm` 267.

Extraigamos una muestra sistemática de tamaño 25 (aproximadamente).

Realizamos, en primer lugar, una reordenación de los datos mediante el método de la permutación aleatoria. Tomando como semilla el número 2238 obtenemos los contratos aleatoriamente permutados, como aparece en la Figura 26.

Determinamos el paso: $410/25 = 16,4 \approx 16$.

Generamos ahora un número aleatorio en el conjunto $\{1, 2, \ldots, 16\}$. Con la semilla 2244 se obtiene el $0,224799$ y a partir de él el valor en $\{1, 2, \ldots, 16\}$ obtenido como

$$1 + \texttt{ENTERO}(16 * 0,224799) = 4.$$

La muestra sistemática es, por tanto, la formada por los contratos con etiqueta $4, 4 + 16 = 20, 4 + 2 \cdot 16 = 36, \ldots, 4 + 25 \cdot 16 = 404$ (obsérvese que el tamaño final de la muestra es 26 en vez de los 25 pretendidos). La muestra aparece en la Figura 43.

□

1	etiquetas	num-cont	cod-contista	tip-cont	den-tip-cont
2	4	2004220994	2B91348607 00	90	CONSULTORIA Y ASISTENCIA
3	20	2004091841	2B41967811 00	91	SERVICIOS
4	36	2004301425	129752254Y 00	90	CONSULTORIA Y ASISTENCIA
5	52	2004451680	128861334Z 00	90	CONSULTORIA Y ASISTENCIA
6	68	2004222338	127830607D 00	90	CONSULTORIA Y ASISTENCIA
7	84	2004195403	124870549M 00	90	CONSULTORIA Y ASISTENCIA
8	100	2004440832	175367658Q 00	90	CONSULTORIA Y ASISTENCIA
9	116	2004194952	2B92282888 00	87	OBRAS
10	132	2004205068	129752254Y 00	90	CONSULTORIA Y ASISTENCIA
11	148	2004065555	2B41136193 00	90	CONSULTORIA Y ASISTENCIA
12	164	2004193994	2B91110445 00	91	SERVICIOS
13	180	2004307234	2F91202077 00	90	CONSULTORIA Y ASISTENCIA
14	196	2004313146	2B11531993 00	91	SERVICIOS
15	212	2004117130	139670066B 00	91	SERVICIOS
16	228	2004164000	128862277Z 00	91	SERVICIOS
17	244	2004328328	2B41390592 00	91	SERVICIOS
18	260	2004437877	2A28761807 00	91	SERVICIOS
19	276	2004322194	2A41010521 00	91	SERVICIOS
20	292	2004275617	128731153J 00	90	CONSULTORIA Y ASISTENCIA
21	308	2004045152	128784136G 00	90	CONSULTORIA Y ASISTENCIA
22	324	2004420684	2B91397935 00	91	SERVICIOS
23	340	2004272718	2A28229813 00	91	SERVICIOS
24	356	2004397052	2A41117342 00	91	SERVICIOS
25	372	2004231447	2G41373903 00	91	SERVICIOS
26	388	2004317030	2B41705492 00	91	SERVICIOS
27	404	2004163873	2A41117342 00	91	SERVICIOS

Figura 43: Muestra sistemática de contratos de la D.G. de Arquitectura y Vivienda

4.2.2. Estimación

En el muestreo sistemático las fórmulas de estimación (aunque no los errores) son prácticamente idénticos a los del muestreo aleatorio simple.

Así, estimamos el parámetro ϑ como sigue:

- **la media de una variable** Y mediante $\frac{1}{k}\sum_{i\in m} Y_i$.
- **la proporción de individuos que tienen una característica** mediante el cociente $\frac{t}{k}$, donde t representa el número de individuos con la característica *en la muestra.*
- **el total de una variable** Y a partir del estimador de la media mediante una regla de tres: $\frac{N}{k}\sum_{i\in m} Y_i$.
- **el total de individuos que tienen una característica** mediante la fórmula $\frac{Nt}{k}$ siendo t

el número de individuos con la característica en la muestra.

4.3. Muestreo por unidad monetaria con reposición

El muestreo aleatorio simple, descrito en la Sección 4.1, puede ser utilizado en cualquier tipo de pruebas (tanto sustantivas como de cumplimiento). Las dos características fundamentales del muestreo aleatorio simple son:

1. *Simplicidad* de los procedimientos para generar la muestra.

2. *Equiprobabilidad:* todas las unidades tienen idéntica probabilidad de ser incluidas en la muestra.

Esto no implica necesariamente que el muestreo aleatorio simple sea el más eficiente: puede ocurrir que otros procedimientos de muestreo requieran menores tamaños muestrales para tener unos niveles de error aceptables. Tal es el caso del *muestreo por unidad monetaria*, que describimos a continuación, y que está especialmente indicado para pruebas sustantivas.

Para poder aplicar el muestreo por unidad monetaria es necesario disponer de una variable auxiliar positiva, que identificamos con unidades monetarias. Usaremos este procedimiento sólo cuando se dé alguno de los dos supuestos siguientes:

- Que pueda suponerse que la variable objeto de estudio es aproximadamente proporcional a la variable auxiliar. Por ejemplo, si cabe esperar que la cuantía en el pago no debidamente justificada (variable objeto de estudio) es, aproximadamente, proporcional a la variable auxiliar (cuantía del pago).

- Que se desee muestrear de modo equiprobable cada unidad monetaria (aunque eso suponga que unos individuos serán seleccionados con mayor probabilidad que otros).

4.3.1. Generación de la muestra

Para obtener una muestra de tamaño n, se generan con reposición n números aleatorios del conjunto $\{1, 2, \ldots, N\}$, de modo que, en cada extracción, el valor j es generado con probabilidad $\frac{X_j}{T(X)}$. Se consigue así que, en cada etapa, la probabilidad de inclusión del individuo j sea proporcional a X_j.

El algoritmo es el siguiente:

1. Se crean, en la columna `A`, las etiquetas o identificadores de las unidades muestrales (código de los gastos, etc.)

2. En la columna B, anotamos los valores de la variable auxiliar X.

3. En la columna C, anotamos las probabilidades asociadas a las unidades muestrales (dividiendo el valor correspondiente de la columna B por la suma de la columna B).

4. En la columna D, calculamos las probabilidades acumuladas; por ejemplo, en D4 se anotará C1+C2+C3.

5. En la columna E anotamos los valores de la columna D, desplazados una fila, con E1 igual a 0. Por ejemplo, en E4 ponemos D3.

6. Generamos, en la columna F, n números aleatorios en el intervalo $[0, 1]$.

7. Determinamos en G, para cada uno de los n números aleatorios, el individuo u_i tal que el número aleatorio está entre Ei y Di. Esto se consigue con la función BUSCAR. Por ejemplo, haciendo en G1 =BUSCAR(F1;E:E;A:A), escribe en G1 la etiqueta del primer individuo seleccionado.

Ejemplo 17 De la base de datos de contratos de 2004, extraemos, en la hoja DATOS 267 del fichero Excel ejemplos.xls, aquellos con el campo uni-adm 267. Seleccionemos con reposición una muestra por unidad monetaria, siendo la variable auxiliar imp-contrato.

En la Figura 44 aparecen, en la columna A, la identificación del registro (número del contrato / código de contratista), y el importe del contrato (columna B), que es la variable auxiliar X.

Calculamos, como se ha descrito,

1. Las probabilidades asociadas a cada contrato (columna C).

2. Las probabilidades acumuladas (columna D).

3. Extremos inferiores de los intervalos (columna E).

El resultado se muestra en la Figura 44 y en la hoja MUM 267.

Por ejemplo, el contrato con identificación 2004056754 / 124783792G 00 (Fila 9 en Figura 44), tiene un importe de 20.009 euros, por lo que, en cada etapa, será seleccionado con una probabilidad $0,005461027$, es decir: algo superior al $0,5\,\%$. Para que sea seleccionado, el número aleatorio generado debe estar entre $0,229632777$ (Columna E) y $0,235093804$ (Columna D).

Generamos 20 números aleatorios (semilla 1030) en la columna G, y buscamos los contratos a los que se corresponden, como se muestra en la Figura 45).

Por ejemplo, el primer número aleatorio obtenido es el $0,103823969$. Como al contrato de la Fila 4 (2003202035/2A14334411 00) le corresponde el intervalo que va del $0,012060438$ al $0,210615937$,

	A	B	C	D	E	F
1	num-cont/cod-contista	imp-contrato	Probabilidades	Acumuladas	Extremos inf.	
2	1994041968/1315641102 00	28.149 €	0,007682665	0,007682665	0	
3	2002315058/125975734V 00	16.040 €	0,004377773	0,012060438	0,007682665	
4	2003202035/2A14334411 00	727.500 €	0,198555499	0,210615937	0,012060438	
5	2003385839/2B91217299 00	28.500 €	0,007778463	0,2183944	0,210615937	
6	2004045152/128784136G 00	13.959 €	0,003809809	0,222204209	0,2183944	
7	2004050158/2B41999335 00	5.080 €	0,001386477	0,223590686	0,222204209	
8	2004056754/124668169W 00	22.138 €	0,006042092	0,229632777	0,223590686	
9	2004056754/124783792G 00	20.009 €	0,005461027	0,235093804	0,229632777	
10	2004056811/128232034V 00	23.070 €	0,006296461	0,241390265	0,235093804	
11	2004056811/128531695B 00	20.851 €	0,005690833	0,247081098	0,241390265	
12	2004056849/124668169W 00	6.911 €	0,001886209	0,248967307	0,247081098	
13	2004056849/124783792G 00	6.358 €	0,00173528	0,250702586	0,248967307	
14	2004057327/124078518W 01	10.051 €	0,002743205	0,253445791	0,250702586	
15	2004057327/124231395K 00	11.059 €	0,003018317	0,256464107	0,253445791	
16	2004065555/2B41136193 00	1.280 €	0,000349349	0,256813456	0,256464107	
17	2004070974/132855706E 00	11.980 €	0,003269684	0,260083139	0,256813456	
18	2004071028/127321678R 00	11.379 €	0,003105654	0,263188793	0,260083139	
19	2004072186/2B41056193 00	4.876 €	0,001330799	0,264519593	0,263188793	
20	2004072211/2B41579236 00	1.995 €	0,000544492	0,265064085	0,264519593	
21	2004072242/2B41056193 00	3.200 €	0,000873371	0,265937456	0,265064085	
22	2004072262/127900152W 00	10.500 €	0,002865749	0,268803206	0,265937456	
23	2004072383/127900152W 00	7.100 €	0,001937792	0,270740998	0,268803206	

Figura 44: Cálculo de probabilidades para un muestreo de contratos por unidad monetaria

éste es el contrato seleccionado en primer lugar. Se repite el proceso con los 20 números aleatorios, generándose los contratos de la columna H.

	A	B	C	D	E	F	G	H
1	num-cont/cod-contista	imp-contrato	Probabilidades	Acumuladas	Extremos inf.		Aleat	num-cont/cod-contista
2	1994041968/1315641102 00	28.149 €	0,007682665	0,007682665	0		0,103823969	2003202035/2A14334411 00
3	2002315058/125975734V 00	16.040 €	0,004377773	0,012060438	0,007682665		0,091372417	2003202035/2A14334411 00
4	2003202035/2A14334411 00	727.500 €	0,198555499	0,210615937	0,012060438		0,174504837	2003202035/2A14334411 00
5	2003385839/2B91217299 00	28.500 €	0,007778463	0,2183944	0,210615937		0,45905942	2004203388/2B92282888 00
6	2004045152/128784136G 00	13.959 €	0,003809809	0,222204209	0,2183944		0,489333781	2004218070/128344075W 00
7	2004050158/2B41999335 00	5.080 €	0,001386477	0,223590686	0,222204209		0,789361248	2004393632/124125822H 00
8	2004056754/124668169W 00	22.138 €	0,006042092	0,229632777	0,223590686		0,089297159	2003202035/2A14334411 00
9	2004056754/124783792G 00	20.009 €	0,005461027	0,235093804	0,229632777		0,97195349	2004450619/129696030V 00
10	2004056811/128232034V 00	23.070 €	0,006296461	0,241390265	0,235093804		0,329477828	2004095928/2A80140122 00
11	2004056811/128531695B 00	20.851 €	0,005690833	0,247081098	0,241390265		0,681356243	2004338714/2B41526401 00
12	2004056849/124668169W 00	6.911 €	0,001886209	0,248967307	0,247081098		0,529496139	2004232635/2B14573620 00
13	2004056849/124783792G 00	6.358 €	0,00173528	0,250702586	0,248967307		0,242103336	2004056811/128531695B 00
14	2004057327/124078518W 01	10.051 €	0,002743205	0,253445791	0,250702586		0,928495132	2004443714/2B18211326 00
15	2004057327/124231395K 00	11.059 €	0,003018317	0,256464107	0,253445791		0,110995819	2003202035/2A14334411 00
16	2004065555/2B41136193 00	1.280 €	0,000349349	0,256813456	0,256464107		0,195318461	2003202035/2A14334411 00
17	2004070974/132855706E 00	11.980 €	0,003269684	0,260083139	0,256813456		0,475905637	2004212898/128906669Q 00
18	2004071028/127321678R 00	11.379 €	0,003105654	0,263188793	0,260083139		0,245796075	2004056811/128531695B 00
19	2004072186/2B41056193 00	4.876 €	0,001330799	0,264519593	0,263188793		0,96417127	2004449382/2B41833344 00
20	2004072211/2B41579236 00	1.995 €	0,000544492	0,265064085	0,264519593		0,636646626	2004313146/2B11531993 00
21	2004072242/2B41056193 00	3.200 €	0,000873371	0,265937456	0,265064085		0,901455733	2004437838/2A28761807 00
22	2004072262/127900152W 00	10.500 €	0,002865749	0,268803206	0,265937456			

Figura 45: Muestra de contratos por unidad monetaria con reposición

□

Obviamente, por ser el muestreo con reposición, pueden aparecer elementos repetidos en la muestra. Debemos anotar los elementos obtenidos y cuántas veces aparece cada elemento en la muestra.

Por construcción, si en la muestra aparecen elementos en los que la variable auxiliar monetaria es mucho mayor que en otros, es presumible que la muestra esté formada mayoritariamente por elementos con valor alto de la variable, en ocasiones repetidos.

Esto se aprecia con claridad en el Ejemplo 17, de la hoja MUM 267. En efecto, el contrato 2003202035/2A14334411 00 tiene un importe de 727.500 euros, que supone el $19,86\,\%$ del total. La probabilidad de que sea éste el contrato generado en cada extracción del muestreo por unidad monetaria es, pues, del $19,86\,\%$. Como cabe esperar, al extraer la muestra, este contrato aparece representado varias (6 de los 20).

Debe reflexionarse sobre si ese efecto es asumible o no. Si no es el caso, y se desea al mismo tiempo que aparezcan más los elementos con mayor valor de la variable, pero al mismo tiempo que aparezcan *también* representados elementos con bajo valor de la variable monetaria, es necesario *estratificar* la población, dejando en un grupo los individuos con mayor valor de la variable, y el resto en uno o dos estratos.

Los detalles sobre la estratificación se darán en la Sección 4.5.

Ejemplo 18 En el fichero de contratos de 2004, con campo uni-adm 267, el campo uni-adm den-prov toma 10 valores distintos, que se corresponden a las 8 provincias andaluzas, en 3 aparece el valor uni-adm SS.CC., en 307 contratos el valor uni-adm SIN PROV. y en 1 el valor uni-adm VALLADOLID. Excluimos el registro con el valor uni-adm VALLADOLID.

Para extraer una muestra sobre los 409 registros restantes, podemos hacer un *muestreo por conglomerados*, como el descrito en la Sección 4.6, seleccionando los registros sólo de *algunas* provincias. De hecho, la muestra de este ejmplo será usada en el Ejemplo 22 para extraer una muestra de registros de varias (que no de todas) las provincias.

En este caso, procedemos a extraer una muestra con reposición de 3 unidades (valores de uni-adm den-prov), proporcional a la cuantía total de los contratos en las mismas.

Generamos los números aleatorios (semilla 2020), obteniéndose las unidades uni-adm SEVILLA y uni-adm SIN PROVINCIA (2 veces). Véase la Figura 46 y la hoja MUM PROVINCIAS.

□

4.3.2. Estimación

En el muestreo por unidad monetaria con reposición, estimamos

- **la media de una variable** Y mediante la fórmula $\overline{X} \cdot \frac{\overline{y}}{x}$, donde
  - $\overline{X}$ denota la media de X *en la población.*

	A	B	C	D	E	F	G	H	I
1	den-prov	número de contratos	importe contratos	probabilidades	acumuladas	extremo inferior		aleatorio	provincia
2	ALMERIA	3	15475	0,00422678	0,00422678	0		0,20249031	SEVILLA
3	CÁDIZ	5	27501	0,007511515	0,011738295	0,00423		0,826563311	SIN PROVINCIA
4	CÓRDOBA	2	15311	0,004181986	0,015920281	0,01174		0,894070254	SIN PROVINCIA
5	GRANADA	6	292468	0,079883557	0,095803838	0,01592			
6	HUELVA	3	38823	0,010603961	0,106407799	0,0958			
7	JAÉN	5	44543	0,012166299	0,118574099	0,10641			
8	MÁLAGA	7	62235	0,016998623	0,135572721	0,11857			
9	SEVILLA	68	1319134	0,360303061	0,495875782	0,13557			
10	SS.CC.	3	47125	0,012871537	0,508747319	0,49588			
11	SIN PROVINCIA	307	1798564	0,491252681	1	0,50875			

Figura 46: Generando muestra de provincias

- $\frac{\overline{y}}{x}$ denota la media en la muestra de los cocientes entre la variable Y y las unidades monetarias X.

Como el muestreo es con reposición, puede haber elementos repetidos. En la fórmula anterior, cada elemento debe ser contabilizado tantas veces como veces aparezca en la muestra.

- **la proporción P de individuos que tienen una característica** mediante la fórmula

$$\widehat{P} = \overline{X}\frac{1}{n} \cdot \sum_{i \in m^*} \frac{1}{X_i}, \tag{18}$$

donde $\sum_{i \in m^*} \frac{1}{X_i}$ representa la suma de los recíprocos de la variable X en los individuos de la muestra con la característica. En la fórmula anterior, cada elemento debe ser contabilizado tantas veces como veces aparezca en la muestra.

- **el total de una variable Y** mediante la fórmula

$$T(X) \cdot \frac{\overline{y}}{x}, \tag{19}$$

donde

- $T(X)$ es el total *en la población* de la variable X.
- $\frac{\overline{y}}{x}$ denota la media en la muestra de los cocientes entre la variable Y y las unidades monetarias X.

En la fórmula anterior, cada elemento debe ser contabilizado tantas veces como veces aparezca en la muestra.

- **el total de individuos que tienen una característica** mediante la fórmula $\frac{T(X)}{n} \cdot \sum_{i \in m^*} \frac{1}{X_i}$, donde $\sum_{i \in m^*} \frac{1}{X_i}$ representa la suma de los recíprocos de la variable X en los individuos de la muestra con la característica. En la fórmula anterior, cada elemento debe ser contabilizado tantas veces como veces aparezca en la muestra.

4.4. Muestreo por unidad monetaria sin reposición. Método de Madow

Es una variante del método por unidad monetaria descrito en la Sección 4.3, aunque en este caso sin reposición.

Al ser sin reposición, es más preciso que el muestreo por unidad monetaria descrito en la Sección 4.3, sobre todo cuando el tamaño poblacional es pequeño. No obstante, obtener la muestra es más complejo que en el caso con reposición.

Las hipótesis bajo las que se puede usar este método, y cuándo es éste superior al muestreo aleatorio simple, son idénticas a las que se describieron en la Sección 4.3 para el muestreo con reposición: es necesario disponer de una variable auxiliar positiva (unidades monetarias), que debe ser aproximadamente proporcional a la variable objeto de estudio.

4.4.1. Generación de la muestra

El método para generar muestras por unidad monetaria sin reposición es, en esencia, el método sistemático, descrito en la Sección 4.2, pero no aplicado sobre los individuos, sino sobre las unidades monetarias. El procedimiento es el siguiente:

1. Analizar si existe alguna relación entre la variable auxiliar y la etiqueta. Si existe tal relación, ordenar aleatoriamente los datos, siguiendo el procedimiento descrito en la Sección 3.4, y trabajar en lo que sigue con la tabla obtenida después de la reordenación.

2. Se multiplica la variable auxiliar por la unidad seguida de ceros, hasta convertir la variable en otra X, que sólo tome valores enteros.

3. Se calcula la suma $T(X)$ de la variable X sobre todos los individuos restantes en la población (inicialmente, todos).

4. Se determina el *paso* $k = \frac{T(X)}{n}$, redondeado al entero más próximo, con la función `REDONDEAR`.

5. Pasamos a otro estrato, para análisis exhaustivo, aquellos individuos u_i cuya X_i es *demasiado grande*: $X_i > k$. Estos individuos formarán otro estrato. Volver al Paso 3, con los individuos que quedan en el estrato, y el tamaño muestral que proceda.

6. A cada individuo u_i se le asocia un intervalo $\{\underline{A}_i, \overline{A}_i\}$, con

$$\begin{aligned} \underline{A}_1 &= 1, & \overline{A}_1 &= X_1 \\ \underline{A}_2 &= X_1 + 1, & \overline{A}_2 &= X_1 + X_2 \\ \underline{A}_3 &= X_1 + X_2 + 1, & \overline{A}_3 &= X_1 + X_2 + X_3 \\ &\vdots & &\vdots \\ \underline{A}_N &= X_1 + X_2 + \ldots + X_{N-1} + 1, & \overline{A}_N &= X_1 + X_2 + \ldots + X_N = T(X) \end{aligned}$$

7. Se genera un número aleatorio entre 1 y k. Si no se requiere reproducibilidad, usar la función `ALEATORIO.ENTRE`; si se requiere reproducibilidad, con la utilidad de *generación de números aleatorios*, generar un valor U de una distribución uniforme entre 0 y 1, (la semilla, como se especificó en la Sección 3.2), y aplicar la fórmula $k^* = 1 + \lfloor k \cdot U \rfloor$, donde $\lfloor z \rfloor$ denota el mayor natural menor o igual que z.

8. La muestra está formada por aquellos individuos u_i en los que

$$\frac{\underline{A}_i - k^*}{k} \leq \lfloor \frac{\overline{A}_i - k^*}{k} \rfloor$$

Ejemplo 19 Supongamos la población de $N = 15$ individuos, con variable auxiliar entera positiva X, representada la Figura 47, de la que queremos extraer una muestra, por el método de Madow, de $n = 5$ individuos.

1. No encontramos relación entre la variable auxiliar y la etiqueta, por lo que no es necesario ordenar aleatoriamente los datos.

2. Como la variable auxiliar ya es entera, no hay que multiplicar por la unidad seguida de ceros, y tomamos la variable auxiliar como X.

3. Calculamos su suma, $T(X) = 7.268$.

4. Se determina el *paso* $k = \frac{T(X)}{n} = 1.453,60$. Redondeando al entero más próximo, obtenemos el paso $k = 1.454$ como `REDONDEAR(1453,60;0)`.

5. Eliminar de la población aquellos individuos u_i cuya X_i es *demasiado grande*: $X_i > 1.454$.

   En nuestro caso, hay que considerar en otro estrato a u_2, pues $X_2 = 4.140 > 1.454$.

   Volvemos al Paso 3, pero excluyendo a u_2. El nuevo tamaño muestral es $n = 4$. Se calcula ahora la nueva suma $T(X) = 3.128$, el paso, $k = \frac{T(X)}{n} = 782,0$.

   Ningún individuo tiene un valor de X mayor que el paso, por lo que se va al paso siguiente

	A	B	C
1	INDIVIDUO	X	
2	1	185	
3	2	4140	
4	3	347	
5	4	435	
6	5	95	
7	6	235	
8	7	160	
9	8	123	
10	9	459	
11	10	227	
12	11	175	
13	12	139	
14	13	66	
15	14	432	
16	15	50	

Figura 47: Población del Ejemplo 19

6. A cada individuo u_i se le asocia el intervalo $\{\underline{A}_i, \overline{A}_i\}$, como se muestra en la Figura 48.

7. Se genera un número aleatorio entre 1 y 782. Sin requerir reproducibilidad, usamos la función `ALEATORIO.ENTRE(1;782)`, y se obtiene el valor $k^* = 405$.

8. La muestra está formada por aquellos individuos u_i en los que

$$\frac{\underline{A}_i - k^*}{k} \leq \lfloor \frac{\overline{A}_i - k^*}{k} \rfloor,$$

es decir: u_3, u_6, u_9, u_{14}, como se muestra en la Figura 49.

□

Ejemplo 20 En la hoja `DATOS 267` del fichero `ejemplos.xls`, el campo `uni-adm den-prov` toma 10 valores distintos, que se corresponden a las 8 provincias andaluzas, en 3 aparece el valor `uni-adm SS.CC.`, en 307 contratos el valor `uni-adm SIN PROV.` y en 1 el valor `uni-adm VALLADOLID`.

Excluimos el registro con el valor `uni-adm VALLADOLID`, y extraemos una muestra sobre los 409 registros restantes, extrayendo una muestra sin reposición de 3 unidades (valores de `uni-adm den-prov`), y proporcional a la cuantía total de los contratos en las mismas.

	A	B	C	D	E
1	INDIVIDUO	X	desde	hasta	
2	1	185	1	185	
3	3	347	186	532	
4	4	435	533	967	
5	5	95	968	1062	
6	6	235	1063	1297	
7	7	160	1298	1457	
8	8	123	1458	1580	
9	9	459	1581	2039	
10	10	227	2040	2266	
11	11	175	2267	2441	
12	12	139	2442	2580	
13	13	66	2581	2646	
14	14	432	2647	3078	
15	15	50	3079	3128	

Figura 48: Generando intervalos para Madow

Las unidades de `uni-adm SEVILLA` y `uni-adm SIN PROVINCIA` resultan demasiado grandes, luego deben pasar al estrato superior, donde censamos. Véase la Figura 50.

Excluidas éstas, nos queda extraer una provincia, que, de acuerdo al número aleatorio obtenido, es Granada, como se observa en la Figura 51.

□

4.4.2. Estimación

En el muestreo por unidad monetaria sin reposición, estimamos

- **la media de una variable** Y mediante la fórmula $\overline{X} \cdot \frac{\overline{y}}{x}$, donde $\frac{\overline{y}}{x}$ denota la media en la muestra de los cocientes entre la variable Y y las unidades monetarias X.
- **la proporción de individuos que tienen una característica** mediante la fórmula $\frac{\overline{X}}{n} \cdot \sum_{i \in m^*} \frac{1}{X_i}$, donde $\sum_{i \in m^*} \frac{1}{X_i}$ representa la suma de los recíprocos de la variable X en los individuos de la muestra con la característica.
- **el total de una variable** Y mediante la fórmula $T(X) \cdot \frac{\overline{y}}{x}$, donde $\frac{\overline{y}}{x}$ denota la media en la muestra de los cocientes entre la variable Y y las unidades monetarias X.

	A	B	C	D	E	F	G	H	I
1	INDIVIDUO	X	desde	hasta		muestra		T(X)	3128
2	1	185	1	185				n	4
3	3	347	186	532		*		paso	782
4	4	435	533	967				aleatorio	405
5	5	95	968	1062					
6	6	235	1063	1297		*			
7	7	160	1298	1457					
8	8	123	1458	1580					
9	9	459	1581	2039		*			
10	10	227	2040	2266					
11	11	175	2267	2441					
12	12	139	2442	2580					
13	13	66	2581	2646					
14	14	432	2647	3078		*			
15	15	=SI(ENTERO((D15-I$4)/I$3) >= (C15-I$4)/I$3;"*";"")							
16									

Figura 49: Generando la muestra por Madow

provincia	n. de contratos	importe total	empieza	acaba	demasiado grande?	muestra?
ALMERIA	3	15475	1	15475		
CÁDIZ	5	27501	15476	42976		
CÓRDOBA	2	15311	42977	58287		
GRANADA	6	292468	58288	350755		
HUELVA	3	38823	350756	389578		
JAÉN	5	44543	389579	434121		
MÁLAGA	7	62235	434122	496356		
SEVILLA	68	1319134	496357	1815490	*	*
SS.CC.	3	47125	1815491	1862615		
SIN PROVINCIA	307	1798564	1815491	3661179	*	*

total	3661179
tamaño	3
paso	1220393
aleatorio 0.1	0,48743
aleatorio	594857

Figura 50: Muestra por Madow

- **el total de individuos que tienen una característica** mediante la fórmula $\frac{T(X)}{n} \cdot \sum_{i \in m^*} \frac{1}{X_i}$, donde $\sum_{i \in m^*} \frac{1}{X_i}$ representa la suma de los recíprocos de la variable X en los individuos de la muestra con la característica.

4.5. Estratificación

Si la población puede dividirse en grupos, de modo que dentro de un grupo

provincia	n. de contratos	importe total	empieza	acaba	demasiado grande?	muestra?			
ALMERIA	3	15475	1	15475				total	543481
CÁDIZ	5	27501	15476	42976				tamaño	1
CÓRDOBA	2	15311	42977	58287				paso	543481
GRANADA	6	292468	58288	350755		*		aleatorio 0.1	0,33661
HUELVA	3	38823	350756	389578				aleatorio	182942
JAÉN	5	44543	389579	434121					
MÁLAGA	7	62235	434122	496356					
SS.CC.	3	47125	496357	543481					

Figura 51: Muestra por Madow

- La subpoblación objeto de estudio tenga un comportamiento similar, o bien su estudio sea igualmente relevante,
- El riesgo de los individuos de un mismo grupo es similar,
- Puede utilizarse un único mecanismo de generación de la muestra en cada subpoblación,

entonces podemos considerar estos grupos como **estratos**, y aplicar, *en cada estrato*, los procedimientos anteriores (muestreo aleatorio simple o por unidad monetaria).

La construcción de estratos será conveniente en el caso en el que la población sea de gran tamaño. Debe observarse que, en ocasiones, el tamaño de la base de datos es superior a la capacidad que soporta Excel, por lo que, salvo que se utilicen herramientas más potentes (como la aplicación Access), será necesario dividir la población en subconjuntos homogéneos, de tamaño más manejable.

4.5.1. Construcción de estratos

No existe un procedimiento universal para diseñar los estratos. En cualquier caso, deben tenerse en cuenta que, cuanto mayor sea el número de estratos, tanto menor será el esfuerzo muestral en cada estrato; como los errores muestrales son función, como se describió en la Sección 2, no sólo de la fracción de muestreo, sino también del tamaño muestral, si, por haber muchos estratos, se toman muestras muy reducidas en los distintos estratos, el error total será grande aunque los estratos sean homogéneos. Como regla general *evitaremos estratificar en más de* 5 *estratos.*

Para construir los estratos, podemos utilizar

- procedimientos a medida, basándonos en la información que tengamos sobre el problema y las características de las unidades muestrales (estratificación por provincias, etc.),
- procedimientos automatizados, que intenten minimizar el error en las mediciones.

En caso de usar un procedimiento automatizado, la estratificación propuesta debe ser revisada, y en su caso, corregida, por el técnico.

Cuando se dispone de una variable auxiliar X positiva, que represente unidades monetarias, una posible estrategia de estratificación consiste en dividir la población en K estratos (por ejemplo, $K = 3$) de modo que cada estrato contenga, aproximadamente, las mismas unidades monetarias.

4.5.2. Reparto del esfuerzo muestral por estratos

Disponemos de un tamaño muestral total n, a repartir entre los diferentes estratos. Cómo efectuemos tal reparto repercutirá notablemente en la precisión de nuestros estimadores.

El procedimiento de reparto propuesto, que dependerá del criterio que se haya usado para estratificar, es el siguiente:

1. Detectar si existe un estrato en el que se vaya a censar. En caso afirmativo, apartarlo del estudio, y restar al tamaño muestral el tamaño del estrato a censar.

2. Sea n el tamaño muestral restante al salir del paso anterior.

   - Si la estratificación se ha hecho por consideraciones externas, repartir el tamaño muestral restante n entre los distintos estratos proporcionalmente al tamaño de los mismos, redondeando al entero más próximo.
   - Si la estratificación se ha hecho de acuerdo a una variable X, supuestamente proporcional a la variable de interés,
     a) Calcular, en cada estrato, la desviación estándar de X, usando la función de Excel DESVEST.
     b) Repartir el tamaño muestral restante n entre los distintos estratos proporcionalmente al producto de su tamaño y su desviación estándar, redondeando al entero más próximo.
     c) Si, para algún estrato E_i, el valor obtenido de n_i supera el tamaño del estrato, entonces añadir los elementos de E_i al estrato a censar, y repetir el procedimiento.
     d) Si, para algún estrato E_i, el valor obtenido de n_i es cero, entonces hacer $n_i = 1$.

4.5.3. Estimación en el muestreo estratificado

Para la estimación, usamos el siguiente estimador:

- **la media de una variable** Y mediante la media ponderada de los estimadores de la media en los distintos estratos. Los pesos de ponderación son los tamaños poblacionales de los estratos.

- **la proporción de individuos que tienen una característica** mediante la media ponderada de los estimadores de la proporción en los distintos estratos. Los pesos de ponderación son los tamaños poblacionales de los estratos.
- **el total de una variable** Y como la suma de los estimadores del total en cada estrato.
- **el total de individuos que tienen una característica** como la suma de los estimadores de individuos con la característica en cada estrato.

Ejemplo 21 Supongamos una población de $N = 200$ unidades, dividida en dos estratos, E_1, E_2, de $N_1 = 25$ y $N_2 = 175$ unidades, respectivamente, y se desea estimar la proporción P de individuos que presentan alguna irregularidad.

Se extraen mmuestras aleatorias simples de tamaño 15 y 25 respectivamente, obteniéndose 3 y 8 individuos con irregularidades.

Para estimar P, hacemos la estimación separadamente en cada estrato como se detalló en la Sección 4.1.3:

- E_1 : la proporción P_1 se estima mediante la proporción muestral, $\widehat{P_1} = 3/15 = 0.2$
- E_2 : la proporción P_2 se estima mediante la proporción muestral, $\widehat{P_2} = 8/25 = 0.32$

Calculamos los pesos W_i de los estratos:

$$\begin{aligned} W_1 &= N_1/(N_1 + N_2) = 0.125 \\ W_2 &= N_2/(N_1 + N_2) = 0.875 \end{aligned}$$

El estimador $\widehat{P}$ de la proproción se obtiene como

$$\widehat{P} = W_1 \cdot \widehat{P_1} + W_2 \cdot \widehat{P_2} = 0.305$$

El proceso aparece en la figura 52 y en la hoja ESTRATIFICADO.

□

Las expresiones del error en el caso en el que en cada estrato h (de N_h individuos) se realice un muestreo aleatorio simple sin reposición de n_h individuos, son las siguientes:

- el estimador del error al nivel $1 - \alpha$ al estimar la media *en la población* se obtiene mediante la fórmula

$$\widehat{\varepsilon} = Z_{1-\alpha/2}\sqrt{\sum_h W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2}, \tag{20}$$

  donde W_h es el peso del estrato h, es decir, $W_h = N_h / \sum_j N_j$, y S_h^2 es la cuasivarianza de Y en la muestra del estrato h, que se obtiene en Excel con la función VAR

	A	B	C	D
1		ESTRATO1	ESTRATO2	
2	tamaño	25	175	200
3	muestra	15	25	
4	irreg. en muestra	3	8	
5				
6	proporciones muest	0,2	0,32	
7	pesos	0,125	0,875	
8				
9	ESTIMADOR	0,305		
10				

Figura 52: Estimación en muestreo estratificado

- el estimador del error al nivel $1-\alpha$ al estimar la proporción de individuos *en la población* que tienen una característica, con la fórmula

$$\widehat{\varepsilon} = Z_{1-\alpha/2}\sqrt{\sum_h W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \frac{n_h}{n_h - 1} p_h(1-p_h)} \tag{21}$$

- el estimador del error al nivel $1-\alpha$ al estimar el total de una variable Y es

$$\widehat{\varepsilon} = Z_{1-\alpha/2}\sqrt{\sum_h N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2} \tag{22}$$

- el estimador del error al nivel $1-\alpha$ al estimar el total de individuos que tienen una característica,

$$\widehat{\varepsilon} = Z_{1-\alpha/2}\sqrt{\sum_h N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \frac{n_h}{n_h - 1} p_h(1-p_h)}. \tag{23}$$

4.6. Conglomerados

En ocasiones las unidades muestrales quedan agrupadas de manera natural en grupos, que pensamos representativos del total, en tanto en cuanto no parecen existir diferencias significativas entre unos grupos y otros con respecto al comportamiento de la variable objeto de estudio.

Obsérvese que estos grupos no desempeñan el papel de estratos, descritos en la Sección 4.5: al dividir una población en estratos se supone que los individuos de un mismo estrato se parecen, mientras que los de estratos diferentes no; por el contrario, al dividir una población en conglomerados, se supone que cada conglomerado tiene toda la diversidad del total, y que seleccionando uno (o varios) conglomerados, tenemos representado el total.

4.6.1. Construcción de conglomerados

Aunque existen procedimientos estadísticos de diseño automático de conglomerados, aquí supondremos que los conglomerados vienen dados por características externas. Por ejemplo, por provincias (si puede suponerse, basándose en la experiencia previa del auditor, que el comportamiento en las distintas provincias será similar, y que no hay, por tanto, necesidad alguna de seleccionar elementos de todas las provincias).

4.6.2. Generación de la muestra

Describimos un muestreo por conglomerados en el que *todos los individuos tienen idéntica probabilidad de ser incluidos en la muestra*, sea cual fuere el conglomerado al que pertencen.

Supongamos que hay G conglomerados, de modo que el conglomerado g tiene N_g individuos. Se desea seleccionar una muestra de tamaño n, que pertenezcan sólo a k conglomerados diferentes (por ejemplo, a 2 provincias).

1. Seleccionamos con reposición una muestra de k conglomerados, con probabilidades proporcionales al número de individuos de los mismos.

2. De cada conglomerado seleccionado, se extrae una muestra aleatoria simple sin reposición de n/k individuos.

Ejemplo 22 Generemos una muestra por conglomerados de 30 contratos de la hoja `DATOS 267` del fichero Excel `ejemplos.xls`, pertenecientes a lo sumo a 3 unidades distintas del `den-prov` , excluido el registro con valor de `den-prov` VALLADOLID.

Para ello, extraemos con reposición una muestra de 3 unidades primarias (provincias). Esto se hizo en el Ejemplo 18, obteniéndose `SEVILLA` (1 vez) y `SIN PROVINCIA` (2 veces).

De cada uno de estas tres unidades, extraemos, sin reposición, muestras aleatorias simples de tamaño $30/3 = 10$.

Filtramos los registros de `SEVILLA`, y los ordenamos por el método de la permutación aleatoria. Los 10 primeros forman la muestra del primer conglomerado. Véase la Figura 53.

Del mismo modo, repetimos, 2 veces, el proceso de selección de muestras de tamaño 10 en el grupo `SIN PROVINCIA`. Ordenando los 307 registros mediante 2 permutaciones aleatorias, y seleccionando en cada caso los 10 primeros, se obtienen los $10 + 10 = 20$ registros que aparecen en la Figura 54.

□

aleatorio	num-cont	cod-contista
0,00332652	2004226060	127216741J 00
0,00412	2004246512	128773158C 00
0,01525925	2004080248	124870549M 00
0,0343028	2004272862	127527855Y 00
0,03442488	2004246419	2B92282888 00
0,06088443	2004057327	124231395K 00
0,06939909	2004222553	145655838X 00
0,0880459	2004450694	131257521S 00
0,09039583	2004451680	128861334Z 00
0,11740471	2004056754	124783792G 00
0,12390515	2004056811	128531695B 00
0,12442396	2004440832	175367658Q 00
0,18320261	2004410880	128784136G 00
0,18427076	2004056849	124783792G 00

Figura 53: Generando muestra del primer conglomerado

aleatorio	num-cont	cod-contista
0,000488296	2004231447	2G41373903 00
0,00573748	2004256391	128731153J 00
0,011963256	2004095928	2A80140122 00
0,014069033	2004338948	2B91110445 00
0,017151402	2004175416	2B41967811 00
0,019287698	2004123937	2A41162082 01
0,025147252	2004117130	139670066B 00
0,028199103	2004076534	2B41518879 00
0,033387249	2004072438	2B41579236 00
0,038178655	2004119466	142862845E 00
0,040253914	2004209958	2191255083 00
0,041047395	2004444529	2A78357001 00
0,041993469	2004328356	2B41871609 00
0,043427839	2004125437	2B91311886 00

aleatorio	num-cont	cod-contista
0,000885	2004256408	2B41958521 00
0,0038759	2004441511	105087975F 00
0,0059206	2004072438	2B41579236 00
0,0133061	2004091841	2B41967811 00
0,0162358	2004382834	2B41402413 00
0,0167547	2004219157	2A41117342 00
0,0200201	2004310516	2B18618983 00
0,0217597	2004096897	100678070F 00
0,0261849	2004163873	2A41117342 00
0,0394299	2004322047	175052072J 00
0,0440382	2004085108	2B11531993 00
0,0461745	2004117130	139670066B 00
0,0523087	2004186942	3070129275 00
0,0528886	2004229357	2B41999335 00

Figura 54: Generando muestra de los restantes conglomerados

5. Ficha técnica

El muestreo debe venir acompañado de una ficha técnica, en la que se describa el proceso seguido y el nivel de error *global* del estudio.

En el caso que en que se haya realizado un muestreo estratificado, y la trascendencia o naturaleza jurídica de los distintos estratos así lo requiera, se detallarán también los niveles de error en los distintos estratos. Un posible modelo de ficha es el siguiente:

Título:

Objetivo:

Universo:

Periodo de referencia:

Periodo de realización:

Diseño muestral:

Tamaño muestral:

Error muestral: El conjunto de la muestra es representativo con un error de muestreo del $\pm***$ % para un nivel de significación del $***$%, en el supuesto de muestreo aleatorio simple, y de máxima indeterminación ($P = Q = 0.5$).

(En el caso en que haya distintos estratos:) Las muestras en cada uno de los estratos suponen errores de muestreo distintos en función de los correspondientes tamaños muestrales. Como referencia, en submuestras de tamaño $***$, el error de muestreo, para un nivel de significación del 5 %, en el supuesto de muestreo aleatorio simple y máxima indeterminación ($P = Q = 0,5$), es de $***$%,; en submuestras de $***$, de $***$%.

6. Un ejemplo

Sobre el Mayor de Gastos se ha aplicado el filtro:

- agrupación tipo = 01
- organismo autónomo = 00
- capítulo: 4
- artículo: 7

De los 130.617 registros obtenidos, nos quedamos con los documentos `op` o `adop`, correspondientes a las secciones que van de la 01.00 a la 20.00. Se obtienen 605 registros, en los que tenemos las variables `Ident.` (números correlativos de 1 a 605), `Aplicacion`, `DOCObliO`, como aparece en la Figura 55, y en la hoja DATOS MAYOR.

A	B	C
Ident.	Aplicacion	DOCObliO
1	0101000100 47001 11A 0	45.000,00
2	0101000100 47001 11A 0	15.000,00
3	0101000100 47101 52C 5	45.525,00
4	0101000100 47101 52C 5	7.500,00
5	0101000100 47101 52C 5	28.247,56
6	0101000100 47101 52C 5	11.933,89
7	0101000100 47101 52C 5	67.500,00
8	0101000100 47101 52C 5	10.862,52
9	0101000100 47101 52C 5	18.030,12
10	0101000100 47101 52C 5	45.525,00
11	0101000100 47101 52C 5	52.200,00
12	0101000100 47101 52C 5	11.933,89
13	0101000100 47101 52C 5	21.906,71
14	0101000100 47801 82A 2	1.810,00
15	0101000100 47801 82A 2	1.500,00
16	0101000100 47801 82A 2	2.500,00
17	0101000100 47801 82A 2	6.000,00
18	0101000100 47801 82A 2	3.700,00
19	0101000100 47801 82A 2	6.000,00
20	0101000100 47801 82A 2	3.800,00
21	0101000100 47801 82A 2	6.000,00
22	0101000100 47801 82A 2	4.200,00
23	0101000100 47801 82A 2	5.700,00
24	0101000100 47801 82A 2	13.500,00
25	0101000100 47801 82A 2	4.500,00
26	0109000100 47100 75B 5	11.795,00
27	0109000100 47100 75B 5	14.144,32

Figura 55: Datos del ejemplo

Tenemos una población de $N = 605$ registros, y queremos estudiar, a partir de una muestra,

1. la fracción de registros con incumplimientos. Por simplificar, sólo se analizaron tres tipos de incumplimiento:

a) Incumplimiento sobre publicidad debida (incumpliento tipo 1).

b) Incumplimiento sobre finalidad del gasto (incumpliento tipo 2).

c) Incumplimento sobre publicidad y/o finalidad (incumpliento tipo 3).

2. La cuantía total no debidamente justificada.

Sigamos los pasos descritos en la Sección 1. Abordamos, en primer lugar, el caso de las proporciones de incumplimiento, y posponemos a la Sección 6.4 los detalles para la estimación del total.

6.1. Identificación de la población y de la(s) variable(s) objeto de estudio

La población está formada por los $N = 605$ registros. Vamos a estudiar tres parámetros, que se corresponden a los tres tipos de incumplimento descritos previamente.

Los dos primeros indicadores de incumplimiento son primitivos, mientras que el tercero se deriva de éstos. Siguiendo el esquema de la Sección 1, asociamos, a cada registro u_i dos variables Y_{ij} binarias, con

$$\begin{aligned} Y_{i1} &= \begin{cases} 1, & \text{si en } u_i \text{ existe un incumplimiento sobre publicidad} \\ 0, & \text{si no} \end{cases} \\ Y_{i2} &= \begin{cases} 1, & \text{si en } u_i \text{ existe un incumplimiento sobre finalidad del gasto} \\ 0, & \text{si no} \end{cases} \end{aligned} \tag{24}$$

A partir de Y_{i1}, Y_{i2}, definimos Y_{i3},

$$Y_{i3} = \text{máx}(Y_{i1}, Y_{i2}) = \begin{cases} 1, & \text{si en } u_i \text{ existe un incumplimiento sobre publicidad y/o finalidad del gasto} \\ 0, & \text{si no} \end{cases} \tag{25}$$

Tenemos 3 parámetros de interés, P_1, P_2, P_3,

$$\begin{aligned} P_1 &= \text{fracción de registros con incumplimiento sobre publicidad} \\ P_2 &= \text{fracción de registros con incumplimiento sobre finalidad del gasto} \\ P_3 &= \text{fracción de registros con incumplimiento sobre publicidad y/o finalidad del gasto,} \end{aligned}$$

es decir,

$$\begin{aligned} P_1 &= \frac{1}{N}\left(Y_{11} + \ldots + Y_{N1}\right) \\ P_2 &= \frac{1}{N}\left(Y_{12} + \ldots + Y_{N2}\right) \\ P_3 &= \frac{1}{N}\left(Y_{13} + \ldots + Y_{N3}\right). \end{aligned}$$

6.2. Determinación del diseño muestral, incluyendo el cálculo del tamaño muestral

Por simplicidad, aplicaremos como diseño muestral un muestreo aleatorio simple sin reposición, como se describió en la Sección 4.1.

Si admitimos, *para cada uno de los tres parámetros separadamente*, errores respectivos ε_1, ε_2 y ε_3 a un nivel de significación $\alpha = 0.05$, los intervalos de confianza con nivel de significación α serán de la forma

$$\begin{aligned} \widehat{P_1} &\pm \varepsilon_1 \\ \widehat{P_2} &\pm \varepsilon_2 \\ \widehat{P_3} &\pm \varepsilon_3 \end{aligned}$$

Con esto afirmamos que la probabilidad de que P_1 no se encuentre en la horquilla $\widehat{P_1} \pm \varepsilon_1$ es inferior al 5 %, que la probabilidad de que P_2 no se encuentre en la horquilla $\widehat{P_2} \pm \varepsilon_2$ es inferior al 5 %, y que la probabilidad de que P_3 no se encuentre en la horquilla $\widehat{P_3} \pm \varepsilon_3$ es inferior al 5 %.

Para determinar el tamaño muestral tenemos dos opciones:

1. Centrarnos en uno de los tres parámetros, fijando el error admisible ϵ, y calculando el valor de n que lo produce, o bien

2. Considerar un intervalo simultáneo para los tres parámetros, como se describió en la Sección 2.3.3.

Nosotros seguiremos en este ejemplo la primera estrategia, que requiere un tamaño muestral menor (aunque pagando el precio de que no dispondremos de un intervalo de confianza conjunto al 95 %, sino tres intervalos: uno por variable). Calculemos el tamaño muestral con respecto a cada parámetro P_i, para un error ε y un nivel de riesgo medio: $P \leq 0,10$ según la fórmula (10).

Con la metodología descrita en la Sección 2, en lugar de fijar de antemano un valor de ε y determinar el n aplicando (7), tabulamos, con la ayuda de Excel, para distintos valores de ε, el n correspondiente. Los resultados aparecen en las Figuras 56 y 57.

Consideramos asumible un valor $n = 56$, que provoca para cada P_i, bajo la hipótesis de riesgo medio, un error a priori $\varepsilon = 0,075$.

Procedemos a extraer una muestra aleatoria simple sin reposición de 56 registros.

6.2.1. Generación de la muestra aleatoria y medición de variables

Generamos la muestra de 56 registros con el método de la permutación aleatoria, descrito en la Sección 3.4. Para reproducir los números aleatorios debemos generar una semilla, como se describe

N	605,00	error	n
alfa	0,05	0,010	515
Z(1-alfa/2)	1,96	0,015	434
P	0,10	0,020	356
		0,025	289
		0,030	235
		0,035	193
		0,040	159
		0,045	133
		0,050	113
		0,055	96
		0,060	83
		0,065	72
		0,070	63
		0,075	56
		0,080	50
		0,085	44
		0,090	40
		0,095	36
		0,100	33

Figura 56: Determinación del tamaño muestral

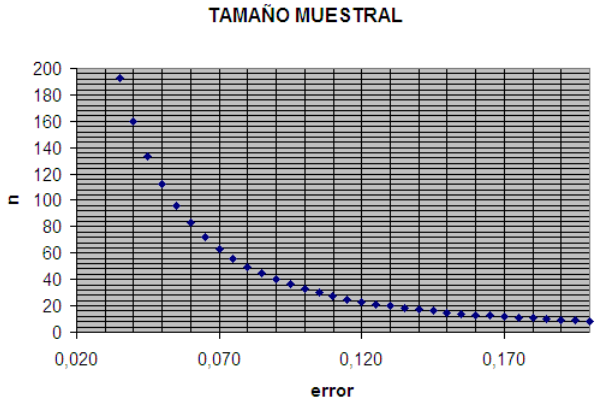


Figura 57: Determinación del tamaño muestral

en la Sección 3.1. Tomando la fecha `16/02/2006`, número de tarjeta `43` y serie `1`, se obtiene la semilla 3360.

Generamos a continuación $N = 605$ números aleatorios, con semilla 3360, como se describe en la Figura 58.

Al ordenar los 605 registros en orden creciente de los números aleatorios obtenidos se obtiene la tabla Excel de la Figura 59.

Seleccionamos los $n = 56$ primeros registros, que, ordenados, aparecen en la Figura 60.

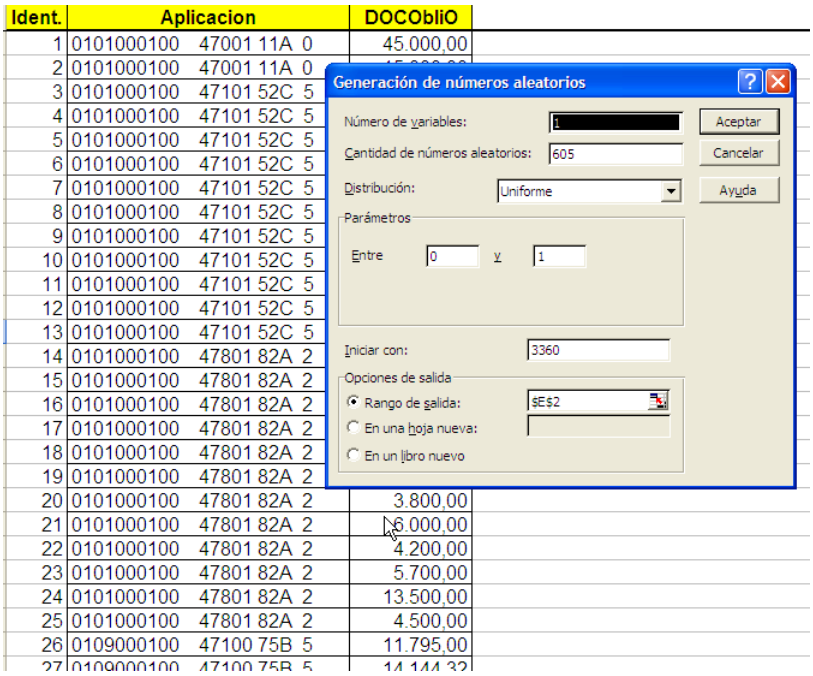


Figura 58: Generando números aleatorios

Ident.	Aplicacion	DOCObliO	aleatorios
86	0113000123 47700 32C 2	10.090,00	0,002227851
6	0101000100 47101 52C 5	11.933,89	0,003204443
249	0113001814 47700 32C 9	6.000,00	0,004669332
349	0113001821 47700 32C 2	3.000,00	0,004669332
345	0113001821 47700 32C 2	6.000,00	0,004730369
570	0113001841 47700 32C 4	2.025,00	0,00573748
83	0113000123 47700 32C 2	4.713,00	0,005951109
503	0113001841 47700 32C 4	12.150,00	0,007324442
573	0113001841 47700 32C 4	2.025,00	0,008667257
252	0113001814 47700 32C 9	3.000,00	0,010620441
557	0113001841 47700 32C 4	2.700,00	0,011627552
11	0101000100 47101 52C 5	52.200,00	0,012451552
422	0113001829 47700 32C 5	3.000,00	0,012665181
483	0113001841 47700 32C 4	2.700,00	0,013183996
359	0113001821 47700 32C 2	6.000,00	0,013641774
579	0113001841 47700 32C 4	2.700,00	0,013794366
60	0113000118 47700 32C 0	384	0,014618366
353	0113001821 47700 32C 2	2.250,00	0,016571551
65	0113000118 47700 32C 0	420,49	0,021423994
245	0113001814 47700 32C 9	3.000,00	0,021454512
198	0113001811 47700 32C 1	3.000,00	0,021576586
347	0113001821 47700 32C 2	2.250,00	0,021881771
355	0113001821 47700 32C 2	6.000,00	0,022461623
99	0113000141 47000 31O 7	1.200,00	0,023804437
307	0113001818 47700 32C 5	3.000,00	0,024628437
374	0113001823 47700 32C 0	3.000,00	0,026337474
519	0113001841 47700 32C 4	5.400,00	0,030365917

Figura 59: Ordenando los números aleatorios

A	B	C
Ident.	Aplicacion	DOCObliO
6	0101000100 47101 52C 5	11.933,89
11	0101000100 47101 52C 5	52.200,00
24	0101000100 47801 82A 2	13.500,00
60	0113000118 47700 32C 0	384
65	0113000118 47700 32C 0	420,49
83	0113000123 47700 32C 2	4.713,00
86	0113000123 47700 32C 2	10.090,00
99	0113000141 47000 31O 7	1.200,00
127	0113001804 47700 32C 8	6.000,00
135	0113001804 47700 32C 8	3.000,00
165	0113001804 47700 32C 8	6.000,00
198	0113001811 47700 32C 1	3.000,00
208	0113001811 47700 32C 1	9.000,00
230	0113001814 47700 32C 9	2.250,00
245	0113001814 47700 32C 9	3.000,00
249	0113001814 47700 32C 9	6.000,00
252	0113001814 47700 32C 9	3.000,00
265	0113001818 47700 32C 5	6.000,00
267	0113001818 47700 32C 5	3.000,00
272	0113001818 47700 32C 5	6.750,00
283	0113001818 47700 32C 5	6.000,00
297	0113001818 47700 32C 5	6.000,00
307	0113001818 47700 32C 5	3.000,00
308	0113001818 47700 32C 5	6.000,00
345	0113001821 47700 32C 2	6.000,00
347	0113001821 47700 32C 2	2.250,00
349	0113001821 47700 32C 2	3.000,00

Figura 60: Muestra aleatoria de 80 registros

6.3. Estimación, construcción de intervalos de confianza y contrastes de hipótesis

La muestra de $n = 56$ registros nos sirve para estimar simultáneamente los tres parámetros de interés. Estimamos cada P_j como se indicó en la Sección 4.1.3: por la proporción muestral.

Analizados los 56 registros generados, se obtuvo que

- En 3 había incumplimientos de publicidad, pero no de finalidad
- En 1 registro había incumplimiento de fiinalidad, pero no de publicidad
- En 52 registros no había incumplimiento.

En otras palabras, el total en la muestra de registros con incumplimientos de tipo 1, 2 y 3 era, respectivamente, de 3, 1 y 4, y, por tanto, los estimadores son

$$\begin{aligned} \widehat{P_1} &= \frac{3}{56} = 0,05 \\ \widehat{P_2} &= \frac{1}{56} = 0,02 \\ \widehat{P_3} &= \frac{4}{56} = 0,07 \end{aligned}$$

Los intervalos de confianza, $\widehat{P}_i \pm 0,075$, es decir,

$$\begin{array}{ll} (-0,021; 0,129) & \text{para } P_1 \\ (-0,057; 0,093) & \text{para } P_2 \\ (-0,004; 0,146) & \text{para } P_3 \end{array}$$

No obstante, como estamos hablando de proporciones, que han de ser valores entre 0 y 1, podemos eliminar los valores negativos, y queda

$$\begin{array}{ll} (0; 0,129) & \text{para } P_1 \\ (0; 0,093) & \text{para } P_2 \\ (0; 0,146) & \text{para } P_3 \end{array}$$

Decimos entonces que

- Con una probabilidad del 95 %, es $P_1 \leq 0,129$
- Con una probabilidad del 95 %, es $P_2 \leq 0,093$
- Con una probabilidad del 95 %, es $P_3 \leq 0,146$

6.4. Estimación de un total

Supongamos ahora que el objetivo del estudio no es estimar la fracción de registros con algún tipo de incumplimento, sino una variable cuantitativa. Estudiamos, en este caso, la cuantía ϑ que *no* está adecuadamente justificada según el procedimiento previsto.

La variable Y que asociamos ahora a cada registro no es una variable binaria, sino la que indica la cantidad no justificada, y el parámetro a estimar ϑ es el total de la variable Y.

Podemos generar la muestra, como se describió en la Sección 6.2. El estimador del total se obtiene como se explicó en la Sección 4.1.3:

$$\widehat{\vartheta} = N\overline{y},$$

donde $\overline{y}$ es el promedio de la variable en la muestra obtenida. Se ha obtenido en la muestra un total de 72.344, 55 euros no adecuadamente justificados. Por lo tanto

$$\begin{aligned} \overline{y} &= \frac{72.344,55}{56} = 1.291,87 \\ \widehat{\vartheta} &= N\overline{y} = 781.579,52 \end{aligned}$$

El error ε de la estimación viene estimado mediante la fórmula (22):

$$\begin{aligned} \widehat{\varepsilon} &= Z_{1-\alpha/2} N \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S^2} \\ &= 355.730,96 \end{aligned}$$

Por tanto, el intervalo de confianza al 95 % para ϑ será $N\overline{y} \pm \widehat{\vartheta}$, es decir: $(425.848,56; 1.137.310,49)$

Alternativamente, si se dispone de la variable `DOCObliO`, podemos realizar, en lugar de un muestreo aleatorio simple, un muestreo por unidad monetaria, como se describió en la Sección 4.3. Construimos las probabilidades (los valores de `DOCObliO` divididos entre el total de dicha variable), y calculamos los extremos superior e inferior que se corresponden con los distintos registros. Así se obtiene la Figura 61.

	A	B	C	D	E	F
1	Ident.	DOCObliO	Probabilidades	Acum.	Extremos inf	
2	1	45.000,00	0,01085	0,01085	0,00000	
3	2	15.000,00	0,00362	0,01447	0,01085	
4	3	45.525,00	0,01098	0,02545	0,01447	
5	4	7.500,00	0,00181	0,02726	0,02545	
6	5	28.247,56	0,00681	0,03407	0,02726	
7	6	11.933,89	0,00288	0,03695	0,03407	
8	7	67.500,00	0,01628	0,05323	0,03695	
9	8	10.862,52	0,00262	0,05585	0,05323	
10	9	18.030,12	0,00435	0,06020	0,05585	
11	10	45.525,00	0,01098	0,07118	0,06020	
12	11	52.200,00	0,01259	0,08377	0,07118	
13	12	11.933,89	0,00288	0,08665	0,08377	
14	13	21.906,71	0,00528	0,09194	0,08665	
15	14	1.810,00	0,00044	0,09237	0,09194	
16	15	1.500,00	0,00036	0,09273	0,09237	
17	16	2.500,00	0,00060	0,09334	0,09273	
18	17	6.000,00	0,00145	0,09479	0,09334	
19	18	3.700,00	0,00089	0,09568	0,09479	
20	19	6.000,00	0,00145	0,09712	0,09568	
21	20	3.800,00	0,00092	0,09804	0,09712	
22	21	6.000,00	0,00145	0,09949	0,09804	
23	22	4.200,00	0,00101	0,10050	0,09949	
24	23	5.700,00	0,00137	0,10188	0,10050	
25	24	13.500,00	0,00326	0,10513	0,10188	
26	25	4.500,00	0,00109	0,10622	0,10513	
27	26	11.795,00	0,00284	0,10906	0,10622	
28	27	14.144,32	0,00341	0,11247	0,10906	

Figura 61: Preparando proceso de generación de muestra por unidad monetaria

Generamos los $n = 56$ números aleatorios uniformes, y detectamos los registros a los que se corresponden, como se aprecia en la Figura 62. Se ha tomado la semilla 3360.

La estimación se realizará como se detalló en la Sección 4.3.2: se calcula, *sobre la población*, el total de la variable `DOCObliO`: $8.291.909,66$; se determina, en cada individuo de la muestra, el valor

	A	B	C	D	E	F	G	H
1	Ident.	DOCObliO	Probabilidades	Acum.	Extremos inf		Aleat	
2	1	45.000,00	0,01085	0,01085	0,00000		0,336008789	148
3	2	15.000,00	0,00362	0,01447	0,01085		0,36710715	175
4	3	45.525,00	0,01098	0,02545	0,01447		0,423535875	217
5	4	7.500,00	0,00181	0,02726	0,02545		0,504013184	289
6	5	28.247,56	0,00681	0,03407	0,02726		0,917050691	586
7	6	11.933,89	0,00288	0,03695	0,03407		0,003204443	1
8	7	67.500,00	0,01628	0,05323	0,03695		0,403088473	201
9	8	10.862,52	0,00262	0,05585	0,05323		0,825281533	551
10	9	18.030,12	0,00435	0,06020	0,05585		0,401898251	200
11	10	45.525,00	0,01098	0,07118	0,06020		0,517105625	302
12	11	52.200,00	0,01259	0,08377	0,07118		0,012451552	2
13	12	11.933,89	0,00288	0,08665	0,08377		0,374065371	180
14	13	21.906,71	0,00528	0,09194	0,08665		0,134434034	36
15	14	1.810,00	0,00044	0,09237	0,09194		0,653340251	415
16	15	1.500,00	0,00036	0,09273	0,09237		0,190282907	39
17	16	2.500,00	0,00060	0,09334	0,09273		0,317453536	129
18	17	6.000,00	0,00145	0,09479	0,09334		0,16483047	39
19	18	3.700,00	0,00089	0,09568	0,09479		0,107333598	26
20	19	6.000,00	0,00145	0,09712	0,09568		0,811700797	536
21	20	3.800,00	0,00092	0,09804	0,09712		0,645619068	409
22	21	6.000,00	0,00145	0,09949	0,09804		0,175328837	39
23	22	4.200,00	0,00101	0,10050	0,09949		0,374095889	180
24	23	5.700,00	0,00137	0,10188	0,10050		0,267159032	91
25	24	13.500,00	0,00326	0,10513	0,10188		0,060823389	10
26	25	4.500,00	0,00109	0,10622	0,10513		0,93856624	586
27	26	11.795,00	0,00284	0,10906	0,10622		0,997436445	605
28	27	14.144,32	0,00341	0,11247	0,10906		0,839930418	566

Figura 62: Generación de muestra por unidad monetaria

de la variable Y (cuantía mal justificada), y se calcula, *sobre la muestra,* la media de los cocientes entre la variable Y y $X =$ `DOCObliO`. Se obtiene

$$\overline{y/x} = 0,35$$

Usando la expresión (19), obtenemos el estimador $\widehat{\vartheta}$,

$$\widehat{\vartheta} = 8.291.909,66 \cdot 0,35 = 2.882.205,61$$

7. La hoja de cálculo

En este manual se sugiere el uso sistemático de la hoja de cálculo Excel para realizar, entre otras, las operaciones de

- Determinación de tamaños muestrales.
- Generación de la muestra aleatoria (los núeros aleatorios generados con una semilla).
- Cálculo de las expresiones de estimadores.

Se adjunta a este manual un fichero Excel, en el que aparecen recogidos, con fórmulas y gráficas, los ejemplos desarrollados en el texto. A continuación se da una breve descripción de las distintas hojas.

`ALEAT 1..N`**:** Dado un valor N, se generan *con reposición* 5.000 valores aleatorios entre 1 y N a partir de un conjunto de 5.000 valores aleatorios en 0 y 1.

`DATOS DOC`**:** Contiene 605 registros.

`DATOS 267`**:** Contiene los registros de contratos de 2004 con el campo `uni-adm` 267, que se corresponde con los registros en los que el campo `den-uni-adm` toma el valor `D.G. ARQUITECTURA Y VIVIENDA`.

`ESTRATIFICADO`**:** Contiene el desarrollo del Ejemplo 21.

`GRAFICA DE ERRORES`**:** Contiene tabulados, para un tamaño poblacional N y un nivel de confianza α fijos, el error que se obtiene para distintos valores de n en tres escenarios distintos. Los resultados se representan gráficamente.

`MUM PROVINCIAS`**:** Contiene el desarrollo de la selección de muestra por unidad monetaria descrito en el Ejemplo 18.

`MUM 267`**:** Contiene el desarrollo de la selección de muestra por unidad monetaria descrito en el Ejemplo 17.

`SEMILLA`**:** Contiene el procedimiento de construcción de la semilla para la generación de números aleatorios. Los datos de entrada son

- Fecha en la que se realiza la generación.
- Tarjeta. Número identificativo del técnico.

- Serie. Tanda de números aleatorios extraída por el técnico en esa fecha.

TAM CONTROL CUMP: Contiene una tabla para determinar el tamaño muestral necesario para validar, con una significación α, la hipótesis de que la fracción de incumplimento no supera un valor P_0. El proceso se describe en la Sección 2.2. Se añade también una gráfica en la que se observa la evolución del tamaño muestral n necesario para distintos valores de P_0 y $\alpha = 0,05$.

TAM CONTROL CUMP SIN REP: Contiene una tabla con el tamaño muestral necesario para validar, con una significación $\alpha = 0,05$, la hipótesis de que la fracción de incumplimento no supera el valor $P_0 = 0,05$. Los tamaños muestrales aparecen para distintos tamaños poblacionales, que oscilan entre $N = 3.000$ y $N = 100$. Para valores de N mayores, puede usarse la aproximación descrita en la hoja TAM CONTROL CUMP.

TAM CONTROL CUMP SIN REP 2: Contiene una tabla y gráfica con la siguiente información: para N y P_0 fijos, determinar, para los distintos tamaños muestrales n, la probabilidad α de no encontrar ningún incumplimiento en los n documentos muestreados, suponiendo que la fracción de documentos con incumplimiento es al menos P_0.

TAM MUESTREO ALEAT: Contiene

- una tabla en la que, dados el tamaño poblacional N, y el nivel de significación α, determinar para distintos tamaños muestrales n el error cometido
- una gráfica en la que se observa la evolución del tamaño muestral n necesario con respecto al error ε admisible.

TAM MUESTREO ALEAT 2: Contiene una tabla con la que, fijados el tamaño poblacional N, el tamaño muestral n, el nivel designificación α, y el número de atributos r, se determina, para distintos escenarios sobre P, el error ε cometido.

TAM MUESTRAL VARIOS ATR: calcula el tamaño muestral a partir del tamaño poblacional N, el nivel de significación α, el error ε y el nivel de riesgo P.

8. Glosario

Conglomerado: Como se describe en la Sección 4.6, los conglomerados son grupos en los que podemos dividir la población, de modo que cada grupo (conglomerado) sea una representación fiel del total. Por ejemplo, si estamos estudiando las subvenciones a agrupaciones, y pensamos que el comportamiento es similar en las 8 provincias, podemos tomar 8 conglomerados (uno por provincia).

Definidos los conglomerados, Se seleccionan uno o varios conglomerados, y de éstos se seleccionan unidades muestrales (por muestreo aleatorio simple, por muestrero por unidad monetaria, etc.)

Cuantil: El cuantil β de la normal estándar es el valor x tal que una normal de media 0 y varianza 1 toma un valor menor o igual que x con probabilidad β. El más usado en este texto es el cuantil para $\beta = 0.975$, cuyo valor es 1.960, como se mostró en la Figura 9.

Estimador: Al extraer una muestra aleatoria, el estimador es una función de dicha muestra (la media aritmética, su media aritmética multiplicada por el tamaño poblacional, etc.) que aproxima al parámetro que estamos estudiando sobre la población. El estimador dependerá tanto del parámetro que vayamos a estudiar como del procedimiento seguido para generar la muestra.

Estrato: Como se describe en la Sección 4.5, un estrato es cada uno de los grupos en los que se divide la población, de modo que se pueda suponer que los individuos de un mismo grupo tienen un comportamiento homogéneo con respecto a la variable que se desea estudiar. Por ejemplo, si la población la forman las nóminas de un determinado colectivo de funcionarios, podemos dividir tales nóminas en grupos (estratos) de acuerdo a su cuantía.

Intervalo de confianza: Cuando estimamos un parámetro ϑ (por ejemplo, la desviación total en el gasto) mediante una función de la muestra $\widehat{\vartheta}$, un intervalo de confianza con nivel de significación α es un intervalo $\widehat{\vartheta} \pm \delta$, de modo que la probabilidad de errar (es decir, que el verdadero valor ϑ quede fuera del rango $\widehat{\vartheta} \pm \delta$) no supere el $100(1-\alpha)\,\%$

Nivel de riesgo: Si la variable objeto de estudio es de respuesta "sí / no"(por ejemplo, si hay o no incumplimiento grave en un gasto), la proporción P de síes se encuentra, obviamente, entre 0 y 1. El tamaño muestral necesario va a venir dado por una estimación a priori del valor de P. El caso más desfavorable es $P = 0.5$ Cuanto más cerca esté P de 0 (o de 1,), tanto menor será el tamaño muestral necesario para obtener un intervalo de confianza con un error predeterminado. En la Sección 2.3.2 se definen los valores críticos: En caso de riesgo

máximo tomamos $P = 0.5$, y calculamos el tamaño muestral aplicando la fórmula (7). Si el riesgo es **alto**, tomamos $P = 0.2$ y calculamos el tamaño muestral aplicando la fórmula (9) con $P = 0.2$; si el riesgo es **medio**, aplicamos (9) con $P = 0.1$, y si es bajo, con $P = 0.05$

Nivel de significación: Al afirmar que los resultados obtenidos tienen un nivel de significación α, estamos diciendo que, si repitiéramos de modo independiente el proceso de inferencia un número suficiente grande de veces, el $100 \cdot (1 - \alpha)\,\%$ de las veces llegaríamos a la misma conclusión.

Semilla: un número entero, de 4 cifras, que, al fijarlo en la `Generación de números aleatorios`, nos generará siempre *la misma secuencia de números aleatorios*. La semilla se genera siguiendo el procedimiento descrito en la Sección 3.1.

Si no tenemos necesidad en reproducir los resultados, no es necesario usar semillas, y Excel generará los números aleatorios automáticamente.

Índice

1. Notación y procedimiento **1**

2. Determinación de tamaños muestrales **3**

2.1. Tamaño muestral con recursos prefijados 3

2.2. Tamaño muestral y control de cumplimiento 3

2.3. Tamaño muestral y control del error de estimación 7

2.3.1. Determinación de tamaños muestrales en el escenario más desfavorable 8

2.3.2. Determinación de tamaños muestrales con información adicional 11

2.3.3. Error para varios atributos 15

3. Generación de números aleatorios **17**

3.1. Generación de una semilla 17

3.2. Generación reproducible de números aleatorios 19

3.3. Generación de etiquetas 20

3.4. Ordenación aleatoria 22

3.5. Generación no reproducible de números aleatorios 24

3.6. Generación de números en un intervalo 26

4. Diseños muestrales **30**

4.1. Muestreo aleatorio simple sin reposición 30

4.1.1. Generación de la muestra. Método de eliminación de duplicados 30

4.1.2. Generación de la muestra. Método de ordenación aleatoria 33

4.1.3. Estimación 34

4.2. Muestreo sistemático 37

4.2.1. Generación de la muestra 37

4.2.2. Estimación 40

4.3. Muestreo por unidad monetaria con reposición 41

4.3.1. Generación de la muestra 41

4.3.2. Estimación 44

4.4. Muestreo por unidad monetaria sin reposición. Método de Madow 46

4.4.1. Generación de la muestra 46

4.4.2. Estimación 49

4.5. Estratificación 50

4.5.1. Construcción de estratos 51

4.5.2. Reparto del esfuerzo muestral por estratos 52
4.5.3. Estimación en el muestreo estratificado 52
4.6. Conglomerados 54
4.6.1. Construcción de conglomerados 55
4.6.2. Generación de la muestra 55

5. Ficha técnica **56**

6. Un ejemplo **58**
6.1. Identificación de la población y de la(s) variable(s) objeto de estudio 59
6.2. Determinación del diseño muestral, incluyendo el cálculo del tamaño muestral 60
6.2.1. Generación de la muestra aleatoria y medición de variables 60
6.3. Estimación, construcción de intervalos de confianza y contrastes de hipótesis 63
6.4. Estimación de un total 64

7. La hoja de cálculo **67**

8. Glosario **69**