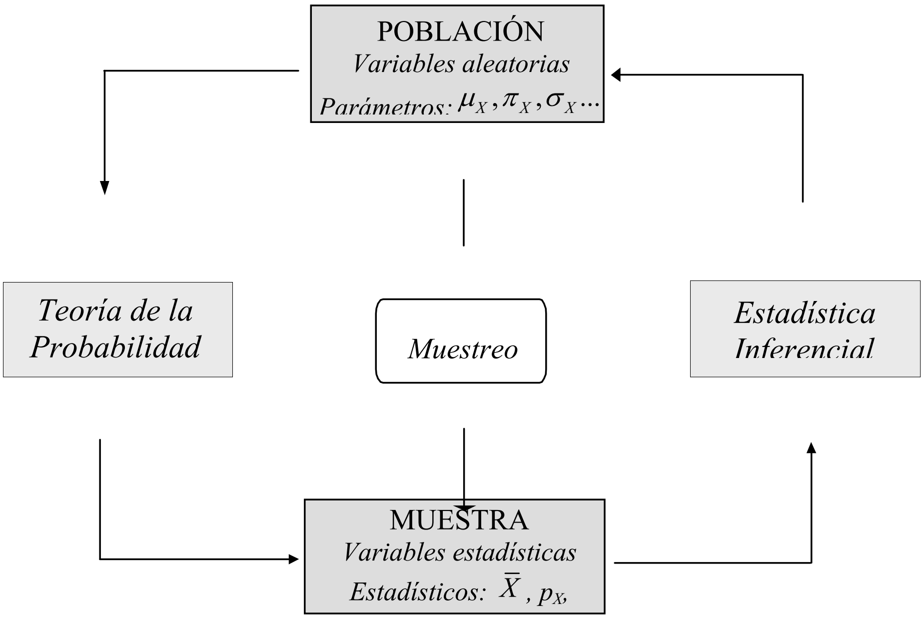
T. 10 – La estadística inferencial: algunos conceptos previos.

• La importancia de la Teoría de la Probabilidad en el ámbito de la estadística se deriva del hecho de constituir ésta uno de los pilares teóricos fundamentales sobre los que se asienta el desarrollo y aplicación de la Estadística Inferencial. Así, mientras que si de una o más variables conocemos sus características (tendencia central, dispersión, asociación...) en la población, la Teoría de la Probabilidad nos permite establecer predicciones de las características que esas variables adoptarán en una muestra de sujetos extraída al azar de esa población, la estadística inferencial -basándose en el conocimiento desarrollado por la Teoría de la Probabilidad en ese camino de la población a la muestra- ha establecido las bases para trazar el camino opuesto, esto es, inferir a partir de los datos de una muestra en una o más variables, cómo serán las características (tendencia central, dispersión, asociación...) de esas variables en la población a la que esa muestra representa. Gráficamente:



1. Teoría de la Probabilidad

• Ante un <u>evento de resultado incierto</u>, el campo de conocimientos de la Teoría de la Probabilidad ha dirigido sus esfuerzos a determinar el grado en qué puede ocurrir cualquiera de los resultados posibles [sucesos] que se pueden derivar de la realización de tal evento [experimento aleatorio].

> **Ejemplos** de evento incierto: (1) el lanzamiento de una moneda (sucesos posibles: que salga cara y que salga cruz); mi estado de salud durante el próximo mes (sucesos posibles: bueno; malo; regular); la práctica religiosa de un estudiante de la Universidad elegido al azar (sucesos posibles: ninguna; católica; protestante; etc.); el *CI* de ese mismo estudiante (sucesos posibles: que sea igual a 120; que sea igual a 85... En este último caso, los sucesos se podrían expresar, no en forma de *sucesos elementales*, sino de *sucesos compuestos*, por ejemplo: que sea menor de 110; que sea mayor o igual a 110.

• El esfuerzo de la Teoría de la Probabilidad por determinar el grado en qué puede ocurrir uno cualquiera de los sucesos asociado a un determinado experimento aleatorio se ha concretado en la asignación de un valor numérico que refleje el grado en que es previsible la ocurrencia de ese suceso.

A este valor numérico se le conoce como probabilidad (*P*) y puede, por convención, oscilar entre 0 y 1 (0: probabilidad nula; 1: probabilidad segura). Así, para el suceso *i* de un experimento aleatorio *X*:

$$0 \leq P(X_i) \leq 1$$

Otra propiedad importante de las probabilidades es que, para los distintos (*n*) sucesos elementales asociados a un experimento aleatorio, la suma de sus probabilidades será igual a 1:

$$\sum_{i=1}^{n} P(X_i) = 1$$

• A continuación se van a describir 3 enfoques en la estimación de las probabilidades asociadas a los resultados posibles de un evento incierto. Dado que normalmente estos enfoques lo que permiten obtener son estimaciones, no los verdaderos valores de probabilidad, haremos referencia a estos valores estimados con el símbolo *P'*, mientras que para el verdadero valor de probabilidad se reserva el símbolo *P*.

(1) Enfoque subjetivo: supone estimar la probabilidad de un suceso en función del grado de confianza personal que se tiene acerca de la ocurrencia del mismo, ya venga esa confianza determinada por nuestra experiencia vital, por nuestras convicciones personales o creencias, o por cualquier otra fuente sobre la que se base el conocimiento que tenemos de nuestro medio. Se trata del procedimiento más utilizado en la práctica desde siempre a la hora de estimar probabilidades, especialmente, cuando no se tienen ciertas nociones sobre otras aproximaciones al cálculo de probabilidades, o bien, cuando aplicar éstas resulta poco operativo. Por ejemplo, cuando me asomo a la ventana antes de salir de casa y veo el cielo, realizo una estimación de la probabilidad de que llueva durante el día. Y como consecuencia de esa estimación, y dependiendo de lo que me importe mojarme, decido qué ponerme o si coger un paraguas. En realidad, hacemos este tipo de estimaciones subjetivas de probabilidad en muchas ocasiones, aunque no siempre de forma muy consciente, constituyendo un elemento determinante de las decisiones que finalmente tomamos.

Ejemplos de estimación subjetiva de probabilidad: (1) a la hora de estimar la probabilidad de que al lanzar dos dados salgan en ambos un seis, muchas personas realizarían una estimación subjetiva de la misma pues, aunque existen otras aproximaciones más precisas a la hora de realizar esa estimación, su aplicación es desconocida para muchos; (2) también las personas suelen realizar estimaciones subjetivas de la probabilidad de que les toque el 'gordo' en un sorteo de lotería -en general, muy al alza- y, curiosamente, suelen ser estimaciones diferentes en función del número considerado; (3) también es habitual realizar estimaciones subjetivas de la probabilidad respecto al resultado de un partido, por ejemplo, de que gane el Valencia CF en su partido del próximo fin de semana.

(2) Enfoque clásico o a priori: consiste en estimar la probabilidad de un suceso (X_i) como la razón entre los resultados favorables a ese suceso y el número total de resultados posibles que se pueden dar en la realización del experimento aleatorio.

$$P'(X_i) = \frac{n^o\, de\, resultados\ favorables}{n^o\, de\, resultados\ posibles}$$

Ejemplo: ¿cuál es la probabilidad de que al lanzar un dado salga un 5?

$$P'(X_i = 5) = \frac{1}{6} = 0{,}167$$

Ejercicio 1: ¿cuál es la probabilidad de que al lanzar un dado salga un 3?; ¿y de que salga número par?; ¿y de que al lanzar dos dados, la suma de los puntos dé igual a 7?; ¿y de que en la lotería de Navidad toque el *gordo* en el número al que juego?

Este enfoque en la estimación de probabilidades asume el conocido como principio de indiferencia, esto es, que la probabilidad de ocurrencia de todos los sucesos es la misma. Si se cumple este supuesto en la realización de un determinado experimento aleatorio, entonces podremos decir que las estimaciones realizadas de acuerdo a esta aproximación serán los verdaderos valores de probabilidad. Sin embargo, el cumplimiento de este principio resulta difícil de aceptar en muchas situaciones en la práctica. Por ejemplo, si aplicamos la aproximación clásica a la hora de estimar la probabilidad de que un estudiante elegido al azar de la Universidad su estado civil sea viudo/a, nos daría igual a ¼, un resultado poco creíble pero que ha venido motivado por realizar la estimación no cumpliéndose en este caso el principio de indiferencia.

En algunos casos sí que se puede asumir el cumplimiento de este principio -por ejemplo, en juegos de azar-, pero en otros muchos casos puede tenerse serias dudas acerca de la satisfacción del mismo, lo cual cuestionaría la aplicación de este enfoque.

(3) Enfoque frecuencialista, a posteriori o estadístico: dado un suceso X_i asociado a la realización de un determinado experimento aleatorio, la estimación de la probabilidad de X_i a partir de este enfoque se basa en la repetición de una gran cantidad de veces del experimento aleatorio en las mismas condiciones, para así obtener la razón entre el nº de veces que ha ocurrido ese suceso (n_i) y el nº de repeticiones del experimento (n):

$$P'(X_i) = \frac{n_i}{n}$$

Ejercicio 2: ¿cómo se estimaría la probabilidad, de acuerdo a esta aproximación, de que salga un 3 en el lanzamiento de un dado?, ¿y del resto de sucesos planteados en el ejercicio 1?

De acuerdo al enfoque frecuencialista, cuanto mayor sea el número de repeticiones del experimento aleatorio, más cercano será el valor de probabilidad estimado ($P'(X_i)$) al verdadero valor de probabilidad $P(X_i)$. En términos matemáticos:

$$P(X_i) = \lim_{n \to \infty} \frac{n_i}{n}$$

> **Ejemplo**: Si lanzamos una moneda 10 veces a fin de estimar la probabilidad de que salga cara, la probabilidad estimada podría ser, por ejemplo, P'(cara) = 0,6 si nos salieran 6 caras y 4 cruces en esos 10 lanzamientos. Sin embargo, a medida que aumenta el número de lanzamientos (idealmente, hasta infinito) esta estimación se irá acercando a la probabilidad verdadera. Se supone que ese valor será igual a 0,5, pero no tiene por qué necesariamente ser así ya que la moneda podría tener algún tipo de curvatura o ser más pesada por alguno de los dos lados.

Si nos fijamos en la fórmula de la aproximación frecuencialista, la estimación de la probabilidad de un suceso se corresponde con la fórmula de la frecuencia relativa o proporción (p_i) que vimos al construir una distribución de frecuencias:

$$P'(X_i) = p_i$$

En realidad, cualquiera de las variables vistas en los ejemplos de los temas precedentes puede contemplarse como la repetición de un experimento aleatorio concreto, pues consiste en la medición de un atributo determinado en múltiples ocasiones -tantas como diferentes sujetos sean medidos-, sin tener certidumbre a priori de cuál van a ser los valores resultantes (sucesos) de esas mediciones. Y de acuerdo a la aproximación frecuencialista de la probabilidad, las frecuencias relativas que se obtengan al calcular la distribución de frecuencias de esa variable, representarán la estimación de las probabilidades asociadas a esos valores de la variable (experimento aleatorio).

Ejemplo: Queremos estimar la probabilidad de estar casado (X_i = casado) dentro de los estudiantes de la UVEG y disponemos de una muestra de 500 estudiantes de dicha universidad:

- Experimento aleatorio: obtener información del estado civil de un estudiante de la UVEG.
- Repeticiones del experimento aleatorio: se pregunta a 500 estudiantes (n = 500).
- Nº de ocurrencias del suceso de interés: nº de estudiantes casados en esa muestra, supongamos que son 60 (n_{casado} = 60).
- $P'(X$ = casado) = 60/500 = 0,12. Esta estimación se aproximará a la probabilidad verdadera cuanto mayor sea el nº de repeticiones, en este caso, el tamaño de la muestra. Se considerarán como los verdaderos valores de probabilidad en el caso en que se cuente con datos para todos los elementos de la población.

Ejercicio 3 (adaptado a partir de ejemplo de Barón-López, 2005): Se ha repetido en 1000 ocasiones el experimento de elegir a una mujer de la población española de mujeres de entre 45 y 55 años, obteniéndose datos de las variables "Nivel de masa ósea" [*NO*: normal; *OE*: Osteopenia; *OO*: Osteoporosis (Clasificación de la *OMS*)] y "Haber pasado la menopausia" [*N*: No; *S*: Sí]. Los datos obtenidos se muestran resumidos en la siguiente tabla de contingencia:

		MENOPAUSIA		Total
		NO	SI	
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

A partir de los datos recogidos, contéstese a las siguientes cuestiones (entre corchetes aparece el equivalente de la cuestión expresado de forma simbólica):

a) ¿Cuál es la probabilidad (estimada) de que una mujer (extraída al azar de la población española de mujeres de entre 45 y 55 años) tenga osteoporosis? [$P(OO)$]

b) ¿y de que no haya pasado la menopausia? [$P(N)$]

c) ¿y de tenga osteopenia u osteoporosis? [$P(OE \cup OO)$]

d) ¿y de que no haya pasado la menopausia y tenga osteoporosis? [$P(N \cap OO)$]

e) ¿y de que tenga osteoporosis si sabemos que no ha pasado la menopausia? [$P(OO|N)$]

f) Conociendo que una mujer tiene osteopenia ¿cuál es la probabilidad estimada de que haya pasado la menopausia? [$P(S|OO)$]

En este ejercicio se plantea la aplicación práctica de alguno de los teoremas básicos de la probabilidad (más detalles sobre los mismos en, por ejemplo, Botella y cols. (2001, tema 12):

- probabilidad de la intersección de dos sucesos: $P(A \cap B)$
- probabilidad de la unión de dos sucesos: $P(A \cup B)$
- probabilidad condicional: $P(B|A)$ o $P(A|B)$

2. Variables aleatorias

• Frente al concepto de variable estadística, el concepto de variable aleatoria supone contar con información de la probabilidad asociada a cada una de las modalidades de la variable, lo cual implica contar con datos para la población pues, en otro caso, lo que tendríamos serían frecuencias relativas, esto es, estimaciones de las probabilidades, no las probabilidades en sí.

Ejemplo: Si medimos la variable "estado civil" en toda la población de estudiantes de la *UVEG* (n = 45000) y obtenemos que 350 son viudos, entonces la frecuencia relativa correspondiente a la modalidad 'ser viudo' (p_{viudo}= 350/45000 = 0,008) será precisamente la probabilidad de 'ser viudo' (P_{viudo}) en la población de estudiantes de la *UVEG*, y no una estimación de la misma. Análogamente, si obtenemos las probabilidades asociadas a las otras modalidades (sucesos) de la variable estado civil, tendremos la distribución de probabilidades asociada a esta variable en la citada población:

X_i	$P(X_i)$
soltero/a	0,884
casado/a	0,105
separado/a	0,009
viudo/a	0,002
	1,00

• La <u>distribución de probabilidad</u> de una variable aleatoria -de forma análoga a la distribución de frecuencias de una variable estadística- consiste en la correspondencia entre los distintos valores que toma la variable y las probabilidades asociadas a esos valores.

- A esa correspondencia entre las modalidades de una variable y sus probabilidades se le suele llamar <u>función de probabilidad</u> en el caso de tratarse de una variable aleatoria discreta (variables categóricas, ordinales o cuantitativas discretas) y <u>función de densidad de probabilidad</u> en el caso de que sea continua (variables cuantitativas continuas).

Ejemplo: función de probabilidad correspondiente a la variable 'Nº de contratos laborales en los 2 últimos años' para la población de personas en edad laboral de la comarca del Camp de Morvedre:

X_i	$P(X_i)$
0	0,08
1	0,31
2	0,35
3	0,18
4	0,07
5	0,01
	1

- Y se llama <u>función de distribución</u>, ya sea la variable ordinal o cuantitativa, a la que hace corresponder a cada valor de la variable, la probabilidad de que se dé un valor como ese o inferior (concepto análogo al de frecuencia relativa acumulada). Este concepto no es aplicable si la variable es categórica.

Ejemplo: función de distribución para el ejemplo anterior:

X_i	$P_a(X_i)$
0	0,08
1	0,39
2	0,74
3	0,92
4	0,99
5	1

• La <u>distribución de probabilidad</u> de una variable no suele ser conocida dado que, normalmente, no es factible contar con los datos de todas las entidades de la población de interés para una determinada variable- Una aproximación a la misma consiste en estimar las probabilidades correspondientes a partir de los datos recogidos para una muestra de esa población y la aplicación de la aproximación frecuencialista al cálculo de probabilidades presentada antes. Otra vía de aproximación a la distribución de probabilidad de una variable supone asumir, atendiendo a razones sustantivas o a la experiencia práctica acumulada, que dicha variable se distribuye de acuerdo a algún modelo teórico de características conocidas, tal como algunos de los que se presentarán en la siguiente sección.

• Cuando se obtiene un índice cualquiera -por ejemplo, la media- a partir de la distribución de probabilidad de una variable, al valor resultante se le denomina parámetro (mientras que si fuera a partir de una distribución de frecuencias, sería un estadístico).

• Se suelen utilizar letras griegas minúsculas para representar a los parámetros. Así, por ejemplo, dada una variable *X*, se utiliza μ_X para representar la media, σ_X para la desviación típica, π_X para la proporción... Es lo mismo para el caso de los índices estadísticos bivariados, por ejemplo, σ_{XY} para la covarianza, ρ_{XY} para el coeficiente de correlación de Pearson, α para la constante de la ecuación de regresión... Hay algún caso especial, siendo el más notable el de la media aritmética que, como parámetro, aparece también representada como *E(X)* y denominada, de forma alternativa, como 'valor esperado' y también 'esperanza matemática'. Por último, para algunos índices no existe doble asignación simbólica en función de que se traten de parámetros o de estadísticos, sea el caso de la mediana (*Md*) o el coeficiente de variación (*CV*), entre otros muchos.

• La aplicación sobre la distribución de probabilidad de una variable aleatoria de los índices de tendencia central, dispersión, etc. presentados en temas previos, así como los que se van a presentar en temas sucesivos, implica ciertas adaptaciones en las fórmulas presentadas para los mismos. A título ilustrativo, se muestran a continuación las de la media (valor esperado) y la varianza para el caso en que la variable sea ordinal o cuantitativa discreta:

$$E(X) = \mu_X = \sum X_i \cdot P(X_i)$$

$$\sigma_X^2 = \sum (X_i - \mu_X)^2 \cdot P(X_i)$$

Ejercicio 4: Obtener a partir del ejemplo presentado antes de la distribución de probabilidad de la variable "Nº de contratos laborales en los 2 últimos años":

1. ¿Cuál es la probabilidad de que una persona seleccionada al azar de la población anterior haya tenido 3 contratos en los 2 últimos años?
2. ¿Y de que haya tenido más de 2 contratos?
3. Obtener la mediana y la moda de la variable.
4. Obtener el valor esperado (media) y la varianza de la variable.

Ejercicio 5: Sea el caso que contamos con datos de dos variables aleatorias en la población de trabajadores del sector de la construcción de Gandía (n = 1000): "Nº accidentes laborales en el último año" (X) y "Tipo de contrato" (Y) [Fijo; Temporal]:

Y_j \ X_i	0	1	2	3	
Fijo	250	90	50	40	430
Temporal	150	160	160	100	570
	400	250	210	140	1000

1. Obtener la distribución de probabilidad (función de probabilidad) de la variable X [$P(X_i)$]
2. Obtener la función de distribución de X [$P_a(X_i)$]
3. Obtener la distribución de probabilidad conjunta de ambas variables [$P(X_i, Y_j)$]
4. ¿Cuál es la probabilidad de que un trabajador (extraído al azar de dicha población)…
   a. haya tenido 1 o más accidentes? [$P(X \geq 1)$]
   b. tenga contrato fijo y haya tenido 0 accidentes? [P(Fijo$\cap$0)]
   c. haya tenido 2 o 3 accidentes? [$P(2 \cup 3)$]
   d. tenga un contrato fijo? [P(Fijo)]
   e. haya tenido 0 accidentes, sabiendo que tiene un contrato fijo? [$P(0/Fijo)$]
5. Obtener la moda, la mediana, la esperanza matemática y la varianza de la variable X.

• La siguiente tabla resume algunos de los conceptos planteados hasta ahora en función de que hagan referencia a una muestra o a una población:

MUESTRA	POBLACIÓN
1. Variable estadística	1. Variable aleatoria
2. Frecuencia relativa (p_i)	2. Probabilidad [$P(X_i)$]
3. Distribución de frecuencias relativas	3. Distribución de probabilidad - Dos tipos: Función de probabilidad Función de densidad de probabilidad
4. Frecuencia relativa acumulada (p_a)	4. Probabilidad acumulada [$P_a(X_i)$]
5. Distribución de frec. relativas acumuladas	5. Función de distribución
6. Estadísticos	6. Parámetros

3. Modelos teóricos de distribución de probabilidad

• El conocimiento acumulado en Psicología ha permitido evidenciar como algunas variables de interés en este campo se distribuyen de un modo característico, esto es, tienen una distribución de probabilidad particular que se repite a lo largo del tiempo y para diferentes muestras. De muchos de estos patrones regulares de distribución de probabilidad han sido planteados los modelos teóricos que representan matemáticamente a esas distribuciones y que, por tanto, permiten obtener fácilmente, a partir de una función matemática, cuál será la probabilidad (o probabilidad acumulada) asociada a un valor cualquiera de la variable.

• Dos de los modelos más relevantes en el contexto de la Psicología son el de la distribución binomial, para variables ordinales y cuantitativas discretas, y el de la distribución normal, para variables cuantitativas continuas. En los siguientes apartados se describen las características de estos dos modelos teóricos de distribución de probabilidad y se muestra su aplicación en la práctica.

• Otros modelos teóricos de distribución de probabilidad como la distribución *t* de Student, la distribución *ji-cuadrado* y la distribución *F* de Snedecor son también especialmente importantes en el campo de la Estadística, ello debido a que la 'distribución muestral' de algunos estadísticos se ajusta a estos modelos teóricos de distribución de probabilidad. La distribución muestral de un estadístico es un concepto clave en la estadística inferencial que será introducido en un capítulo posterior.

3.1. La distribución binomial

• Comencemos con un **ejemplo**: supongamos que se elige al azar una muestra de 6 personas para formar parte de un jurado popular que ha de juzgar a una persona inmigrante y sabemos que en la población de la que se ha extraído esa muestra un 30% de las personas son racistas. A partir de estos datos, ¿cuál es la probabilidad de que la mitad o más de los miembros del tribunal sean racistas?

• La respuesta a la pregunta anterior se puede resolver fácilmente sabiendo que la distribución de probabilidad de la variable que nos ocupa se corresponde con la de la distribución binomial. Veamos más formalmente las condiciones para que la distribución de probabilidad de una variable se ajuste al modelo teórico de la distribución binomial:

1. Partimos de una variable categórica dicotómica de la que conocemos su distribución de probabilidad en una determinada población.

Sea la variable X [X_1; X_2], entonces en esa variable dicotómica se debe decidir cuál de las dos modalidades de la misma es la que nos interesa a fin de poder buscar la respuesta a aquello que nos interese saber (sea, por ejemplo, X_1), pues es la probabilidad asociada a esa modalidad a la que se conoce simbólicamente como π (y, complementariamente, a la de X_2 como $1-\pi$):

X_i	$P(X_i)$
X_1	π
X_2	$1-\pi$
	1

Ejemplo: Variable “ser racista” [Si; No] => modalidad de interés para contestar a la pregunta planteada en el ejemplo: ‘Si’ => $P(X_1) = \pi = 0{,}30$ => $P(X_2) = 1-\pi = 0{,}70$

Señalar que en la práctica del análisis de datos es bastante habitual contar con datos de variables dicotómicas (o dicotomizadas), por ejemplo, variables en que se han recogida datos del tipo correcto/incorrecto, a favor/en contra, de acuerdo/en desacuerdo, bien/mal, si/no, curado/no curado, tratamiento/no tratamiento, etc. Es frecuente en la literatura de la distribución binomial considerar a esas dos modalidades, de forma genérica, como ‘Acierto’ y ‘Error’, de modo que $P(Acierto) = \pi$ y $P(Error) = 1-\pi$

2. Se realiza n veces esa variable aleatoria (por ejemplo, se extrae una muestra de n casos de una población en los que no sabemos a priori que va a ocurrir con la variable X), debiéndose satisfacer la condición de que se mantenga π constante para todas las realizaciones.

Ejemplo: $n = 6$ (tamaño de la muestra, pues cada extracción de un miembro del jurado representa una realización de la variable aleatoria “Ser racista”) y en cada extracción se asume que se mantiene constante π (= 0,30)

3. Si se cumplen las dos condiciones anteriores, la variable aleatoria "número de experimentos (casos) en los que se verifica la condición X_1" se distribuye según el modelo teórico de la distribución binomial. Si la distribución de probabilidad de una variable *X* se ajusta al modelo binomial se expresa simbólicamente como: $X \rightarrow B(n;\pi)$

Ejemplo. Variable número de miembros del tribunal que son racistas $\rightarrow B(6;\ 0{,}30)$

• La distribución de probabilidad (función de probabilidad) de una variable binomial *X* viene definida por la siguiente función matemática, donde X_i puede oscilar entre 0 y *n*:

$$P(X_i) = \frac{n!}{X_i!(n-X_i)!} \cdot \pi^{X_i} \cdot (1-\pi)^{n-X_i}$$

Ejemplo: si queremos obtener la probabilidad de que 2 miembros del tribunal sean racistas:

$$P(2) = \frac{6!}{2!(6-2)!} \cdot 0{,}30^2 \cdot (1-0{,}30)^{6-2} = \frac{720}{2 \cdot 24} \cdot 0{,}30^2 \cdot 0{,}70^4 = 0{,}324$$

Si sustituyéramos en la fórmula del modelo binomial para los distintos valores que puede tomar *X* en este ejemplo, obtendríamos la distribución de probabilidad de *X* (nº de miembros del tribunal que son racistas):

X_i	$P(X_i)$
0	0,118
1	0,303
2	0,324
3	0,185
4	0,060
5	0,010
6	0,001
	1

• Otro procedimiento alternativo para obtener la distribución de probabilidad de una variable que se ajuste al modelo binomial es acudir a la tabla de la distribución de probabilidad de este modelo, la cual se puede encontrar en el apéndice de tablas de muchos libros de estadística. En la misma se pueden encontrar tabuladas las distribuciones de probabilidad de una variable binomial para diferentes valores de π y de *n*.

Fragmento de la tabla de la distribución binomial

n	x	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	x
2	0	980	902	810	640	490	360	250	160	090	040	010	002	0+	0
	1	020	095	180	320	420	480	500	480	420	320	180	095	020	1
	2	0+	002	010	040	090	160	250	360	490	640	810	902	980	2
3	0	970	857	729	512	343	216	125	064	027	008	001	0+	0+	0
	1	029	135	243	384	441	432	375	288	189	096	027	007	0+	1
	2	0+	007	027	096	189	288	375	432	441	384	243	135	029	2
	3	0+	0+	001	008	027	064	125	216	343	512	729	857	970	3
4	0	961	815	656	410	240	130	062	026	008	002	0+	0+	0+	0
	1	039	171	292	410	412	346	250	154	076	026	004	0+	0+	1
	2	001	014	049	154	265	346	375	346	265	154	049	014	001	2
	3	0+	0+	004	026	076	154	250	346	412	410	292	171	039	3
	4	0+	0+	0+	002	008	026	062	130	240	410	656	815	961	4
5	0	951	774	590	328	168	078	031	010	002	0+	0+	0+	0+	0
	1	048	204	328	410	360	259	156	077	028	006	0+	0+	0+	1
	2	001	021	073	205	309	346	312	230	132	051	008	001	0+	2
	3	0+	001	008	051	132	230	312	346	309	205	073	021	001	3
	4	0+	0+	0+	006	028	077	156	259	360	410	328	204	048	4
	5	0+	0+	0+	0+	002	010	031	078	168	328	590	774	951	5
6	0	941	735	531	262	118	047	016	004	001	0+	0+	0+	0+	0
	1	057	232	354	393	303	187	094	037	010	002	0+	0+	0+	1
	2	001	031	098	246	324	311	234	138	060	015	001	0+	0+	2
	3	0+	002	015	082	185	276	312	276	185	082	015	002	0+	3
	4	0+	0+	001	015	060	138	234	311	324	246	098	031	001	4
	5	0+	0+	0+	002	010	037	094	187	303	393	354	232	057	5
	6	0+	0+	0+	0+	001	004	016	047	118	262	531	735	941	6
7	0	932	698	478	210	082	028	008	002	0+	0+	0+	0+	0+	0
	1	066	257	372	367	247	131	055	017	004	0+	0+	0+	0+	1
	2	002	041	124	275	318	261	164	077	025	004	0+	0+	0+	2
	3	0+	004	023	115	227	290	273	194	097	029	003	0+	0+	3
	4	0+	0+	003	029	097	194	273	290	227	115	023	004	0+	4

Ejemplo: Observar el fragmento adjunto de la tabla de la distribución binomial. Buscando en la misma podríamos obtener fácilmente, por ejemplo, la distribución de probabilidad de una variable X que se distribuya según el modelo binomial con parámetros $n = 4$ y $\pi = 0{,}50$: $X \rightarrow B(4;0{,}50)$

X_i	$P(X_i)$
0	0,062
1	0,250
2	0,375
3	0,250
4	0,062

Otro ejemplo de variable que se distribuye según ésta distribución de probabilidad es el "nº de veces que sale cara al lanzar una moneda al aire 4 veces". Tras comprobar que se cumplen los requisitos para que se pueda decir que esta variable se distribuye según el modelo binomial, obtener la distribución de probabilidad correspondiente a la misma.

Ejercicio 6: Para la variable "Nº de miembros del tribunal que son racistas" presentada antes, obtener las siguientes probabilidades: (1) de que 4 miembros del tribunal sean racistas; (2) de que, como máximo, 5 sean racistas; (3) de que más de la mitad sean racistas; (4) obtener la esperanza matemática y la varianza de esta variable aleatoria.

Ejercicio 7: Sabiendo que en la población española la proporción de mujeres es de 0,60, (1) ¿cuál es la probabilidad de que al extraer una muestra aleatoria de 7 personas de esa población, ninguna de ellas sea mujer? (2) ¿y cuál la de que todas fuesen mujeres? (3) construir la distribución de probabilidad correspondiente a la variable "nº de mujeres al extraer al azar una muestra de 7 personas de la población española" (*X*); (4) obtener la media (o valor esperado), la mediana y la moda de la variable aleatoria *X*; (5) representar gráficamente la función de probabilidad de *X*; (6) *ídem* de la función de distribución de *X*; (7) obtener la distribución de probabilidad de la variable aleatoria complementaria, esto es, la de la variable "nº de hombres al extraer al azar una muestra de 7 personas de la población española".

• Es una propiedad de cualquier variable que se distribuye según la distribución binomial que su esperanza matemática y su varianza se pueden obtener según las siguientes fórmulas:

$$E(X) = \mu = n \cdot \pi$$

$$\sigma^2_X = n \cdot \pi \cdot (1-\pi)$$

Ejercicio 8: Obtener con las anteriores fórmulas la esperanza matemática y la varianza de la variable "Nº de miembros del tribunal que son racistas". Comprobar que coincide con los valores ya obtenidos anteriormente para estos dos índices.

Ejercicio 9: Suponiendo que se contesta completamente al azar a un examen de 10 preguntas de verdadero/falso y de que se corrige puntuando con un 1 los aciertos y con un 0 los errores, obtener la probabilidad de que se saque un 5 en el examen; ¿y cuál es la probabilidad de sacar un 10?; ¿y la de sacar un 5 o más? Obtener el valor esperado de la variable "puntuación en el examen" e interpretarla.

3.2. La distribución o curva normal

• Se trata de un modelo teórico de distribución de probabilidad para variables aleatorias cuantitativas continuas que se caracteriza, gráficamente, por tener forma similar a la de una campana. Por ello, y por haber sido estudiada inicialmente por el matemático Karl Gauss, se le denomine también como curva o campana de Gauss.

• La importancia de esta distribución reside en el hecho de que diversas variables, como los caracteres fisiológicos y morfológicos de individuos —altura, peso o longevidad—, atributos sociológicos, psicológicos y, en general, variables que vienen determinados por muchos factores, se distribuyen según el modelo de la curva normal.

• Abajo se muestra la representación gráfica de una variable aleatoria que se distribuye según el modelo teórico de la distribución normal (más usualmente dicho, "que se distribuye normalmente") Como puede observarse, algunas de las características distintivas de este tipo de distribución son:

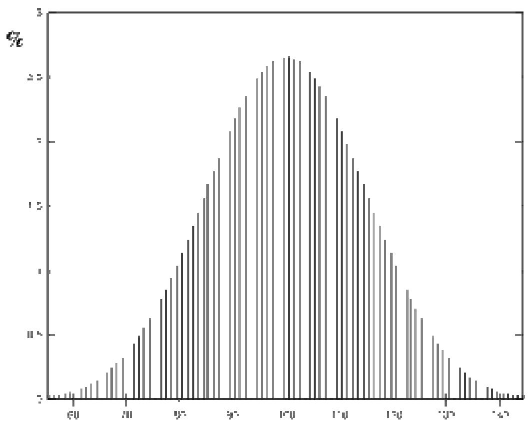
(1) que es unimodal;

(2) que es simétrica, situándose el eje de simetría sobre el valor de la media (mediana, moda) de la distribución de la variable;

(3) que es asintótica;

(4) que -hace corresponder valores de probabilidad altos para los valores centrales de la variable, mientras que esas probabilidades decaen de forma progresiva a medida que nos alejamos del centro de la distribución, más aceleradamente en la zona central, menos en los extremos.

Ejemplo para una variable con las puntuaciones en un test de inteligencia (escala *CI*):



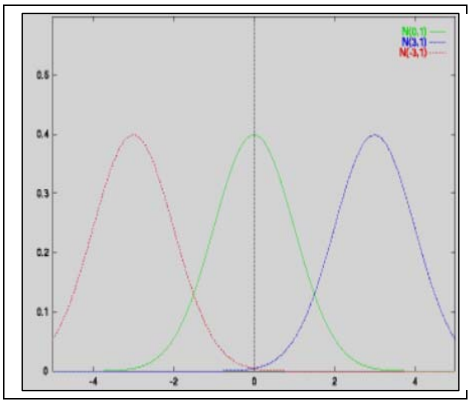
• La distribución de probabilidad (función de densidad de probabilidad) de la curva normal viene definida matemáticamente por la siguiente <u>función matemática</u>, originalmente planteada por Abraham de Moivre en 1733:

$$P(X_i) = \frac{1}{2{,}507 \cdot \sigma_X} \cdot e^{-0{,}5 \cdot \frac{(X_i - \mu_X)^2}{\sigma_X^2}}$$

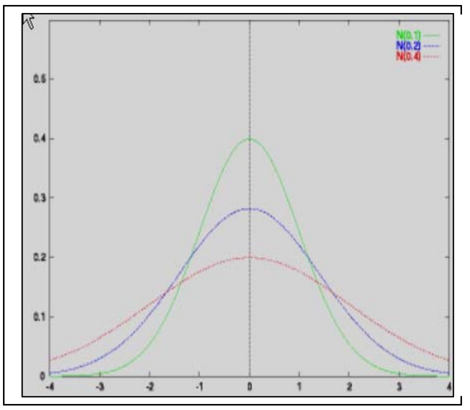
donde X_i es un valor concreto de la variable X, e es una constante matemática (=2,72), y μ y σ son dos parámetros de la función que se corresponden, precisamente, con la media y la desviación típica de X.

• Teniendo en cuenta la fórmula presentada, la distribución normal puede adoptar diferentes formas, tantas como distintos valores de μ y σ se consideren (o sea, infinitas). Cada uno de estos posibles modelos integran la conocida como <u>familia de la distribución normal</u> y para representar, a nivel simbólico, a cada uno de los miembros de esa familia se utiliza la expresión $N(\mu; \sigma)$. A continuación se muestra la representación gráfica de diversos modelos de la familia de la distribución normal en los que se han considerado distintos valores de μ y σ:

- Modelos de distribución normal con distinto valor de μ pero el mismo de σ:



- Modelos de distribución normal con el mismo valor de μ pero distinto de σ:



• Entre los modelos de la familia de la distribución normal, el más relevante en la práctica es el denominado como distribución normal estándar o unitaria , esto es, el modelo de la familia de la distribución normal que tiene $\mu = 0$ y $\sigma = 1$ [$N(0; 1)$]. Así, es común que en los libros de estadística se incluya en un apéndice final de tablas estadísticas, la correspondiente a la curva normal estándar o unitaria. Aunque hay variaciones en la forma de presentar esta tabla en los libros, en la misma nos será posible consultar para un rango de valores comprendido habitualmente entre -3 y 3, cuál es el valor de probabilidad y de probabilidad acumulada correspondiente a esos valores. En suma, se trata de una representación tabular de la función de probabilidad o de la función de distribución (o de ambas, como en el fragmento que se muestra más abajo) de la curva normal con media 0 y desviación típica 1. Así, si se tiene una variable X que se distribuye según la distribución normal estándar [$X \rightarrow N(0; 1)$], en esta tabla podemos consultar para distintos valores de X, cuál es el valor de probabilidad ("*y ordenada*") [$P(X=X_i)$] y de probabilidad acumulada ("*Area*") [$P(X \leq X_i)$] correspondiente a los mismos.

• Una vez conocida la probabilidad asociada a un determinado valor, de forma inmediata se puede dar respuesta fácilmente a diferentes tipos de preguntas, por ejemplo, el porcentaje acumulado o percentil correspondiente a ese valor, el nº de sujetos que es de esperar que tengan un valor inferior o igual a ése, o superior a ése, o entre dos valores determinados, etc. (ver preguntas del ejercicio siguiente).

Ejercicio 10: Dada una variable X que para un determinado grupo de sujetos se distribuye según $N(0;1)$: (a) ¿cuál es la probabilidad acumulada correspondiente a un valor de 1,18 [$P(X \leq 1{,}18) = P_a(X=1{,}18)$]?; (b) ¿qué porcentaje de sujetos tendrán una puntuación inferior o igual a 1,18?; (c) sabiendo que el grupo de sujetos era de 1000 personas (n =1000), ¿cuántos tendrán una puntuación inferior o igual a 1,18?; (d) ¿cuál es la probabilidad de que al extraer un sujeto al azar, éste tenga una puntuación inferior o igual a 1 [$P(X \leq 1)$]?; (e) ¿y de que sea superior a 1 [$P(X>1)$]?; (f) ¿y de qué esté entre 1 y 2 [$P(1 \leq X \leq 2)$], sabiendo que $P(X \leq 2) = 0{,}9772$? ; (g) ¿y de qué esté entre la media de la distribución y 1 [$P(0 \leq X \leq 1)$]?; (h) ¿a qué valor de la variable X le corresponde una probabilidad acumulada de 0,75 (esto es, el 75% de los sujetos obtienen una puntuación inferior o igual a ese valor en la variable = Q_3)?; (i) ¿qué valor de la variable X será sólo superado por el 25% de los sujetos? (j) ¿qué valor corresponde al percentil 25?; (k) ¿cuál es la probabilidad de que al extraer un sujeto al azar, éste tenga una puntuación inferior o igual a -1 [$P(X \leq -1)$]?; (l) ¿y de que la puntuación sea superior a -1 [$P(X>-1)$]?

Fragmento de la tabla de la curva normal estándar

z	Área	y ordenada	z	Área	y ordenada
0.31	.6217	.3802	0.76	.7764	.2989
0.32	.6255	.3790	0.77	.7794	.2966
0.33	.6293	.3778	0.78	.7823	.2943
0.34	.6331	.3765	0.79	.7852	.2920
0.35	.6368	.3752	0.80	.7881	.2897
0.36	.6406	.3739	0.81	.7910	.2874
0.37	.6443	.3725	0.82	.7939	.2850
0.38	.6480	.3712	0.83	.7967	.2827
0.39	.6517	.3697	0.84	.7995	.2803
0.40	.6554	.3683	0.85	.8023	.2780
0.41	.6591	.3668	0.86	.8051	.2756
0.42	.6628	.3653	0.87	.8078	.2732
0.43	.6664	.3637	0.88	.8106	.2709
0.44	.6700	.3621	0.89	.8133	.2685
0.45	.6736	.3605	0.90	.8159	.2661
0.46	.6772	.3589	0.91	.8186	.2637
0.47	.6808	.3572	0.92	.8212	.2613
0.48	.6844	.3555	0.93	.8238	.2589
0.49	.6879	.3538	0.94	.8264	.2565
0.50	.6915	.3521	0.95	.8289	.2541
0.51	.6950	.3503	0.96	.8315	.2516
0.52	.6985	.3485	0.97	.8340	.2492
0.53	.7019	.3467	0.98	.8365	.2468
0.54	.7054	.3448	0.99	.8389	.2444
0.55	.7088	.3429	1.00	.8413	.2420
0.56	.7123	.3410	1.01	.8438	.2396
0.57	.7157	.3391	1.02	.8461	.2371
0.58	.7190	.3372	1.03	.8485	.2347
0.59	.7224	.3352	1.04	.8508	.2323
0.60	.7257	.3332	1.05	.8531	.2299
0.61	.7291	.3312	1.06	.8554	.2275
0.62	.7324	.3292	1.07	.8577	.2251
0.63	.7357	.3271	1.08	.8599	.2227
0.64	.7389	.3251	1.09	.8621	.2203
0.65	.7422	.3230	1.10	.8643	.2179
0.66	.7454	.3209	1.11	.8665	.2155
0.67	.7486	.3187	1.12	.8686	.2131
0.68	.7517	.3166	1.13	.8708	.2107
0.69	.7549	.3144	1.14	.8729	.2083
0.70	.7580	.3123	1.15	.8749	.2059
0.71	.7611	.3101	1.16	.8770	.2036
0.72	.7642	.3079	1.17	.8790	.2012
0.73	.7673	.3056	1.18	.8810	.1989
0.74	.7704	.3034	1.19	.8830	.1965
0.75	.7734	.3011	1.20	.8849	.1942

• Una consecuencia práctica derivada de la aplicación del modelo teórico de la distribución normal y, en general, de cualquier modelo teórico de distribución de probabilidad es que, si sabemos o podemos asumir que una variable se distribuye según un modelo teórico, entonces se pueden obtener las probabilidades asociadas a cualquier valor de esa variable y, en consecuencia, la correspondiente distribución de probabilidad. Será suficiente para ello con aplicar la fórmula matemática del modelo de probabilidad correspondiente o, más sencillo, recurrir a una tabla estadística de ese modelo y consultar en la misma los valores que nos interesen.

• Ahora bien, ¿cómo aprovechar la tabla de la curva normal unitaria si tengo una variable que, aunque se pueda asumir que se distribuye normalmente, su media y su desviación típica no son precisamente 0 y 1? La respuesta está en transformar los valores de la variable a puntuaciones típicas (*Z*), con lo que la variable se seguirá distribuyendo según la curva normal, si bien, pasa a tener media 0 y desviación típica 1, haciendo así factible la utilización de la tabla de la curva normal unitaria.

Ejercicio 11: Supongamos que un conocido nos dice que ha obtenido en un test de inteligencia una puntuación *CI* igual a 95. Asumiendo que las puntuaciones en un test de inteligencia se distribuyen normalmente y sabiendo que las puntuaciones *CI* tienen media 100 y desviación típica 15, ¿qué le podemos decir acerca de su puntuación?, más concretamente, (a) ¿qué porcentaje de sujetos es de esperar que obtengan un valor inferior o igual a 95?, o (b) ¿qué porcentaje de sujetos es de esperar que obtengan un valor superior a 95?; (c) Supongamos también que nos pregunta qué puntuación CI habría que sacar en el test de inteligencia para estar en el 30% inferior (puntuación de *CI* que deja el 30% de sujetos por debajo); (d) ¿y para estar en el 10% superior? (puntuación de *CI* que es superada solo por el 10% de los sujetos) (e) ¿entre qué valores de *CI* se encuentra el 50% central de los sujetos?

4. La selección de la muestra

• Condición para que a partir de la información contenida en una muestra se puedan describir las propiedades de la población:

→ Que la muestra sea representativa de esa población.

• Pero, ¿qué significa en la práctica que una muestra sea representativa de una población?

Que las propiedades que caracterizan a la población se distribuyan de forma análoga en la muestra. Por ejemplo, sea la población los estudiantes de la Facultad: si el 25% son de primero (y así...), que en la muestra ídem; si en la población la mitad son varones y la otra mitad mujeres, pues en la muestra ídem.; etc.

• Factores que van a determinar la representatividad de una muestra:

(1) El procedimiento de selección de los elementos de la población.

(2) El tamaño de la muestra.

(3) El contenido de la investigación.

4.1. Respecto al procedimiento de selección de los elementos de la población.

• Muestreo: conjunto de técnicas asociadas a la selección de una muestra de una población.

• Subrayar que el muestreo no es algo exclusivo de la estadística; es, en realidad, una estrategia muy utilizada en la vida cotidiana. El conocimiento que nos formamos del mundo está con frecuencia basado en el muestreo, y es razonable que así lo hagamos por economía de recursos. A modo de ejemplo, la cata de una cucharada de un guiso que estamos realizando es una forma de muestreo del guiso. Ello nos va a permitir tener una idea del sabor del mismo sin tener que probarlo todo.

• Existen diferentes procedimientos o técnicas que satisfacen -con mayor o menor éxito- los dos objetivos que podemos considerar básicos del muestreo: (1) obtener una muestra que sea tan representativa de la población como sea posible; y (2) plantear una forma de recogida de datos que se ajuste a los recursos (económicos, temporales…) con que se cuente.

• A las distintas estrategias que se puede seguir en la selección de los elementos de una población se les conoce como técnicas de muestreo, existiendo un extenso repertorio de ellas, algunas de gran sofisticación. Entorno a su estudio y aplicación se ha desarrollado un área de conocimiento conocida como Teoría del Muestreo.

• En términos matemáticos, ¿la satisfacción de qué criterio va a determinar la selección de una muestra representativa?

→ El que todos los elementos de la población tengan la misma probabilidad de formar parte de la muestra o, si no la misma, que sea conocida esa probabilidad para todos los elementos de la población.

• A las técnicas de muestreo que satisfacen el criterio anteriormente planteado se les conoce como técnicas de muestreo probabilístico, siendo la más conocida por su sencillez y eficacia, el muestreo aleatorio simple. Algunas variantes de la anterior de mayor utilización en la práctica son el muestreo aleatorio estratificado y el muestreo por conglomerados.

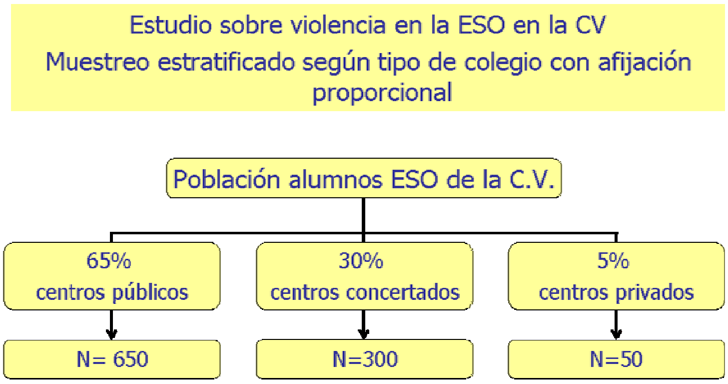
El muestreo aleatorio simple (m.a.s.)

- Los elementos de la muestra son elegidos al azar de de entre todos los de la población. Utilizando este procedimiento, todos los elementos de la población tienen la misma probabilidad de formar parte de la muestra.
- Requiere la identificación y listado de todos los elementos de la población, algo no siempre factible, por lo que su utilización resulta limitada en la práctica.

El muestreo aleatorio estratificado

- Supone forzar que, para una determinada variable(s), se mantenga en la muestra la misma distribución que la misma tiene en la población. Por ejemplo, si en la población de estudiantes de la UVEG hay un 60% de mujeres y un 40% de varones, en una muestra de la misma se forzaría para que se mantuviesen esos porcentajes. Se llaman estratos a las categorías de la variable en función de la que se estratifique el muestreo –mujeres y varones para la variable sexo, en nuestro ejemplo. Por supuesto, cada elemento de la población debe pertenecer a un único estrato.
- Pasos principales: determinar la proporción de cada estrato en la población para la variable de estratificación; fijar el número de elementos que se deben seleccionar de cada estrato en la muestra (afijación); extraer mediante m.a.s. de cada estrato de la población el nº de casos establecido en el paso anterior.

Ejemplo de aplicación del muestreo aleatorio estratificado en la obtención de una muestra (n = 1000) de la población de estudiantes de la ESO en la Comunitat Valenciana (CV) a fin de realizar un estudio sobre la incidencia de la violencia en este tipo de centros:



El muestreo por conglomerados

- Se trata de una forma de m.a.s. en que la unidad de muestreo no son los elementos de la población, sino agrupaciones de éstos que de forma natural existan en aquélla (conglomerados), por ejemplo, colegios, hospitales, distritos postales, calles de una población, secciones del censo electoral...
- Supone seleccionar al azar uno o más conglomerados, recogiéndose datos de todos los elementos de esos conglomerados.
- Suele resultar mucho más fácil acceder a las unidades de los conglomerados, normalmente próximos entre sí, que a elementos individuales dispersos geográficamente.

• Existen algunas técnicas de muestreo no probabilístico, como es el caso del muestreo accidental (o casual), el muestreo intencional y el muestreo por cuotas que, aunque no cumplen los requisitos del muestreo probabilístico, son utilizadas con frecuencia en la práctica debido a la mayor facilidad de aplicación de las mismas.

• Los diferentes procedimientos de muestreo anteriores no son excluyentes, se puede secuenciar la utilización de diferentes técnicas de muestreo dando lugar a lo que se conoce como un muestreo polietápico. Un ejemplo típico es el muestreo polietápico conglomerados/m.a.s., esto es, en primer lugar se aplica un muestreo por conglomerados y, a continuación, se lleva a cabo un m.a.s. de los elementos dentro de cada conglomerado. Por supuesto, otras combinaciones son posibles, así como la consideración de más de dos etapas.

• Si en una investigación la muestra es representativa de la población, es lícito generalizar los resultados obtenidos en la muestra a la población origen y se dice entonces, en términos metodológicos, que la investigación tiene validez externa.

Ejercicio 12: ¿Qué técnica de muestreo fue aplicada en los siguientes ejemplos?

1. Se solicita a las personas que entran y salen a un centro comercial que contestan a una serie de cuestiones relativas a un estudio sobre comunicación social.
2. De entre los asegurados de una compañía de seguros se selecciona al azar a 500 de ellos, a los que se envía una carta con un cuestionario para evaluar una campaña de seguridad promovida por esa compañía entre sus asegurados.
3. Un estudiante realiza una encuesta con el fin de explorar las actitudes solidarias de los alumnos de psicología de su facultad. Sabe a partir de un estudio previo que, de los estudiantes de su facultad, el 25% está comprometido con alguna ONG, mientras que el 75% restante no lo está. Así, decide seleccionar al azar a 200 estudiantes, 50 de entre los comprometidos con una ONG y 150 de entre los que no.
4. En un estudio que pretende poner de manifiesto la influencia de la deseabilidad social al contestar tests psicológicos, el investigador pasa a sus conocidos y familiares una serie de cuestionarios.
5. Un psicólogo está interesado en explorar los hábitos de sociabilidad de las personas de la tercera edad que viven en residencias. Para ello, obtiene una muestra aleatoria de diez residencias de su comunidad autónoma y en cada una de ellas selecciona al azar a 30 personas.
6. En un estudio a nivel nacional sobre las estrategias de aprendizaje de estudiantes de la ESO se seleccionan al azar a 30 centros y dentro de éstos a 4 grupos de cada centro, realizándose una entrevista a todos los estudiantes de los grupos seleccionados.
7. En la realización del censo en España.

4.2. Respecto al tamaño de la muestra

• Obviamente, cuanto mayor sea el tamaño de la muestra mayor será la probabilidad de que ésta sea representativa de la población, sin embargo, no hay que olvidar que también será mayor el esfuerzo implicado en la recogida de los datos y, por lo tanto, se perderá la principal ventaja inherente al muestreo: la economía de recursos a la hora de obtener los datos.

• <u>Criterios orientativos</u> a la hora de definir el tamaño muestral:

1) Un primer criterio determinante en la práctica es el de la cantidad de <u>recursos</u> con que se cuenta para llevar a cabo la recogida de los datos.
2) Un segundo criterio viene determinado por el margen de error que estamos dispuestos a asumir en nuestras inferencias (el <u>error muestral</u>) y por el <u>nivel de confianza</u> con que se

establecerán esas inferencias.

Este segundo criterio puede concretarse a través de fórmulas específicas que permiten obtener el tamaño muestral en función de esos dos criterios, si bien, otros aspectos suelen aparecer implicados en la aplicación de esas fórmulas:

a) el tamaño de la población;
b) el índice estadístico que vaya a aplicarse (pero, ¿cuál de todos?);
c) el valor del índice estadístico en la población para la variable objeto de interés (- "¿en la población?, pero si eso es precisamente lo que queremos conocer...") (¿variable objeto de interés?, en mi estudio no tengo una única varia variable objeto de interés...").

• A fin de facilitar los cálculos en la aplicación de la citada fórmula, diversos autores han elaborado tablas que, para un índice estadístico concreto, para un valor concreto del mismo en la población, y para un nivel de confianza determinado (habitualmente, el 95 o el 99%), permiten obtener el valor del tamaño muestral a considerar en función de los dos criterios restantes: el tamaño de la población y el valor de error muestral que se está dispuesto a asumir.

Ejemplo: La tabla que aparece a continuación permite obtener el tamaño de la muestra que deberemos considerar en la recogida de datos para el caso en que se vaya a aplicar el estadístico de la proporción y que, además, se asuma que el valor de ese estadístico en la población para la variable objeto de interés es 0,5 ($\pi = 0,5$) y un nivel de confianza del 95%.

Un caso concreto:, ¿cuál sería el tamaño de la muestra a considerar en un estudio en que se quiere conocer la proporción (porcentaje) de la población española que está a favor de algún tipo de selectividad para acceder a la Universidad, asumiendo un nivel de confianza del 95% y un error muestral del 2%?

POBLACIÓN	Errores muestrales					
	±1%	±2%	±3%	±4%	±5%	±10%
500	—	—	—	—	222	83
1.000	—	—	—	385	286	91
2.500	—	1.250	769	500	345	96
5.000	—	1.667	909	556	370	98
10.000	5.000	2.000	1.000	588	385	99
25.000	7.143	2.273	1.064	610	394	100
50.000	8.333	2.381	1.087	617	397	100
100.000	9.091	2.439	1.099	621	398	100
infinito	10.000	2.500	1.111	625	400	100

Estos datos están tomados de las tablas publicadas por Arkin y Colton (1962).

4.3. Respecto al contenido de la investigación

• Una premisa básica cuando se trabaja con una muestra a fin de hacer inferencias acerca de una población es que la muestra sea representativa de esa población. Ahora bien, ¿esa representatividad debe darse para todas y cada una de las variables que caracterizan a la población?

• Una condición menos exigente de representatividad, que puede ser considerada razonable en la práctica, es que la representatividad de la muestra lo sea para aquellas variables realmente relevantes en el estudio que se plantee, pero no necesariamente para aquéllas que se considere que no tengan ningún tipo de influencia sobre aquello que se vaya a estudiar.

• La contestación a las dos siguientes cuestiones va a ayudarnos a contrastar si se satisface esta condición en el caso de un muestreo no probabilístico:

- ¿En qué es diferente mi muestra de lo que sería una muestra representativa?
- En caso de considerarse la existencia de diferencias, ¿pueden éstas tener repercusión sobre la(s) variable(s) que se va a medir en mi estudio?

Por **ejemplo**, la investigación de procesos psicológicos básicos (atención, percepción,...) puede en algunos estudios no verse muy influida por el tipo de participantes en la misma.

Referencias:

Barón-López, J. (2005). Bioestadística: métodos y aplicaciones. Apuntes y material disponible en http://www.bioestadistica.uma.es/baron/apuntes/

Botella, J., León, O. G., San Martín, R. y Barriopedro, M. I. (2001). *Análisis de datos en psicología I: teoría y ejercicios.* Madrid: Pirámide.