T. 1 – Inferencia estadística: muestreo y estimación de parámetros

1. Muestreo

• La inferencia estadística es un tipo de razonamiento que procede de lo concreto a lo general, intentando extraer conclusiones sobre los parámetros de una población a partir de la información contenida en los estadísticos de una muestra de esa población (Pardo y San Martín, 2001).

• Obtener datos de una muestra de la población objeto de interés, en vez de obtenerlos de todas las unidades que componen esa población, entraña importantes ventajas, la más importante la economía de recursos que ello implica pero, no menos relevante, la posibilidad de promover una mejor calidad en los datos recogidos.

• Ahora bien, una condición esencial para que a partir de la información contenida en una muestra se puedan describir las propiedades de la población es que la muestra sea representativa de esa población.

• Pero, ¿qué significa en la práctica que una muestra sea representativa de una población? Pues, como podemos intuir, que las propiedades que caracterizan a la población se distribuyan de forma análoga en la muestra. Por ejemplo, sea la población de los estudiantes de la Facultad: si el 27% son de primero, el 25% de segundo (y así...), que en la muestra ídem; si en la población la mitad son varones y la otra mitad mujeres, pues que en la muestra ídem.; etc.

• A continuación vamos a tratar 3 factores que van a determinar la representatividad de una muestra:

a) El procedimiento de muestreo

• Muestreo: conjunto de técnicas asociadas a la selección de los elementos 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.

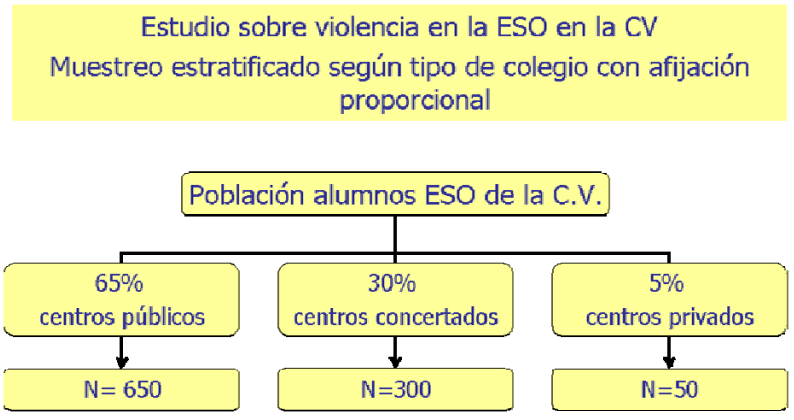
El muestreo aleatorio simple (m.a.s.)

- 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 1: ¿Qué técnica de muestreo fue aplicada en los siguientes ejemplos?

a. 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.
b. 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.
c. 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.

d. Tras discutirse en un programa de radio en torno a la posibilidad de reducir la jornada laboral, se invita a sus oyentes que llamen para expresar su opinión al respecto. Al final del programa, la presentadora afirma que los resultados de la encuesta muestran que el 80% de los que han llamado a la emisora prefieren que se reduzca la jornada laboral.

e. 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.

f. 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.

g. Un psicólogo industrial pretende estudiar la relación entre el turno de trabajo y la productividad de los operarios en las cadenas de montaje de una gran empresa. Dado que algunos estudios previos ponen de manifiesto que existe relación entre la productividad y el sexo, y teniendo en cuenta que en la empresa hay un 50% de hombres y un 50% de mujeres, decide que en la muestra de 100 sujetos haya 50 mujeres y 50 hombres, seleccionados aleatoriamente de entre las mujeres y hombres que trabajan en esa empresa.

h. 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.

i. En la realización del censo de población de España.

b) El 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.

• Criterios orientativos a la hora de definir el tamaño muestral:

1) Un primer criterio determinante en la práctica es el de la cantidad de recursos 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 error muestral) y por el nivel de confianza 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).

c) El 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.

2. La estimación de parámetros

• La inferencia estadística asume que se cuenta con datos de una muestra y que se desea conocer cuáles son las características (ya sea la media, la mediana, la curtosis o cualquier otra que nos pueda interesar), no de esa muestra, sino de la población a la que esa muestra pertenece. A los valores de esas características a nivel poblacional se les conoce como parámetros y se representan simbólicamente con letras griegas (en realidad, sólo algunos de ellos tienen tal privilegio):

$$\mu_X, \sigma_X^2, \sigma_X, \pi_X, \sigma_{XY}, \rho_{XY}, \beta_0, \beta_1 \ldots .$$

• Para conocer los valores de los parámetros podemos plantearnos, bien recoger datos para todos los elementos de la población, algo que puede resultar poco viable en muchas situaciones prácticas, bien realizar una estimación de los mismos a partir de los datos de una muestra. Esta segunda vía es mucho más habitual en la práctica, si bien, supone asumir cierto riesgo de error pues, en cuanto que estimación, el valor que obtengamos no tiene porqué coincidir con el verdadero valor de ese parámetro.

• En la literatura se pueden diferenciar dos grandes aproximaciones a la estimación de parámetros: la estimación puntual y la estimación por intervalos. La diferencia básica entre ambas a la hora de estimar un parámetro es que la primera proporciona una estimación consistente en un valor concreto (puntual), mientras que la segunda ofrece como estimación un rango de valores (intervalo). En realidad, la segunda aproximación consiste en una extensión de la primera, por lo que será la estimación puntal la que se abordará a reglón seguido.

• Señalar en primer lugar que, en el caso que se dispusiese de los datos de una población para una determinada variable (*X*), la obtención de los parámetros que nos pudieran interesar sería inmediata, bastaría con aplicar los índices estadísticos correspondientes para todos los datos de la población. Si, por ejemplo, estuviésemos interesados en conocer los parámetros de la media, de la moda, de la varianza y el índice de asimetría intercuartílico de *X*, los obtendríamos aplicando las fórmulas que representan a estos índices estadísticos:

$$\mu_X = \frac{\sum X_i}{N} \qquad Mo_X = x_i \ \ cuya\ n_i\ es\ maxima \qquad \sigma_X^2 = \frac{\sum (X_i - \mu)^2}{N} \qquad As_{Q_3-Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

• Ahora bien, si lo que disponemos es de datos de una muestra de esa población, ¿cómo se obtiene la estimación de cualquiera de los anteriores parámetros? Ello se lleva a cabo a través de la aplicación de un estimador del parámetro correspondiente, esto es, una función matemática que permite obtener una estimación del valor del parámetro a partir de los datos de la muestra. Pero, ¿cuáles son esas funciones que nos permiten obtener estimaciones de los parámetros?

$$\hat{\mu}_X = ? \qquad \hat{Mo}_X = ? \qquad \hat{\sigma}_X^2 = ? \qquad \hat{As}_{Q_3-Q_1} = ?$$

Como puede observarse en las expresiones anteriores, la estimación de un parámetro se representa con un acento circunflejo sobre la letra del parámetro correspondiente, por ejemplo, $\hat{\sigma}_X$ simboliza el valor estimado de la desviación típica de la variable *X* en la población.

• En realidad, para un determinado parámetro pueden considerarse diferentes funciones matemáticas que nos ofrezcan estimaciones del mismo. Por ejemplo, las siguientes podrían ser hipotéticas candidatas a mejor estimador del parámetro de la media (μ_X):

$$\hat{\mu}_X = \frac{\sum X_i^2}{n} \qquad \hat{\mu}_X = \frac{\sum X_i}{n-2} \qquad \hat{\mu}_X = \frac{\sum X_i}{n^2} \qquad \hat{\mu}_X = \frac{\sqrt{\sum X_i^2}}{n} \qquad \hat{\mu}_X = \frac{\sum X_i}{n} \qquad \hat{\mu}_X = \frac{\sum X_i}{\sqrt{n}}$$

• Es considerada como mejor estimador de un parámetro determinado, aquella función matemática que cumpla las siguientes cuatro propiedades que a continuación se describen:

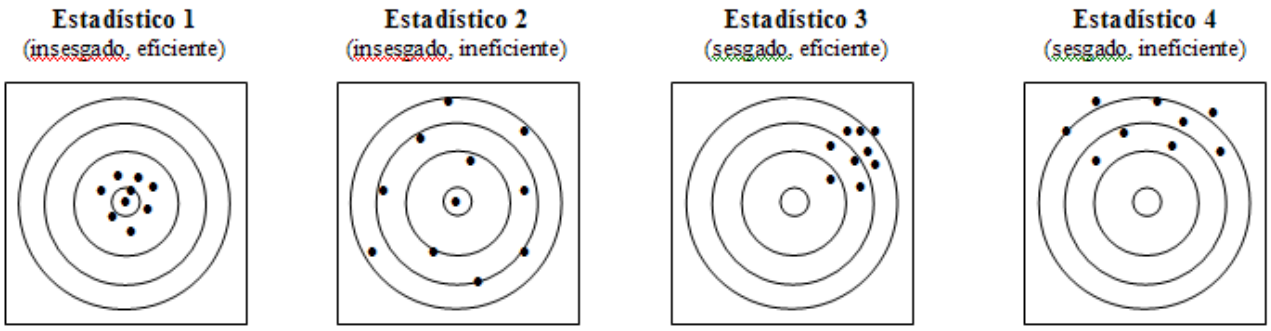
1) Ausencia de sesgo: Un estimador es insesgado cuando el promedio de las estimaciones obtenidas en diferentes muestras es, precisamente, el valor del parámetro que se pretende estimar.

2) Eficiencia: Esta es una propiedad que se establece en términos comparativos, esto es, es más eficiente aquel estimador que, tras ser aplicado a diferentes muestras de una misma población, proporciona estimaciones que tienen una variabilidad menor. Precisamente, una forma de valorar la eficiencia de un estimador es obteniendo la desviación típica de las estimaciones proporcionadas por el mismo, el conocido como *error típico de estimación* del estimador. Así, de entre dos estimadores, será mejor aquél que proporcione un menor error típico de estimación.

3) Consistencia: Un estimador es consistente si la probabilidad de que el valor estimado coincida con el del parámetro aumenta a medida que el tamaño de la muestra crece.

4) Suficiencia: Un estimador es suficiente respecto a un parámetro si agota la información disponible en la muestra aprovechable para la estimación.

La siguiente figura simboliza, en forma de diana, el cumplimiento de las dos primeras propiedades que debe satisfacer un estimador (figura adaptada de Wonnacott y Wonnacott, 1990):



• Para el caso del parámetro de la media (μ_X), el mejor estimador es precisamente el promedio de los datos de la muestra, esto es, el índice estadístico de la media ($\bar{X}$):

$$\hat{\mu}_X = \frac{\sum X_i}{n} = \bar{X}$$

Y, en general, los mejores estimadores de los parámetros correspondientes a los índices estadísticos tratados a lo largo del curso son esos propios índices estadísticos obtenidos a partir de la muestra, esto es, los estadísticos correspondientes. Así:

$$\hat{Mo}_X = Mo_X \; ; \quad \hat{RIC}_X = RIC_X \; ; \quad \hat{Md}_X = Md_X \; ; \quad \hat{\pi}_{Xi} = p_{Xi} \; ; \quad \hat{\rho}_{XY} = r_{XY} \; ...$$

• Existe, sin embargo, alguna excepción a la anterior generalización. Veamos las tres más relevantes:

1) El mejor estimador del parámetro de la varianza (σ_X^2) no es el estadístico de la varianza (s_X^2) sino el siguiente:

$$\hat{\sigma}_X^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

A este índice estadístico se le conoce como cuasi-varianza ($s_X^{2\prime}$): $s_X^{2\prime} = \frac{\sum (X_i - \bar{X})^2}{n-1}$

Mientras que el índice estadístico de la varianza no cumple el requisito de ser un estimador insesgado del parámetro de la varianza, el de la cuasi-varianza sí que lo cumple -de ahí que a este índice estadístico también se le denomine como varianza insesgada.

2) Análogamente, el mejor estimador del parámetro de la desviación estándar (σ_X) es el estadístico de la cuasi-desviación estándar (s_X'):

$$\hat{\sigma}_X = s_X' = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Señalar que cuando al programa SPSS se le pide que obtenga la desviación estándar, lo que realmente ofrece como resultado es la cuasi-desviación estándar (lo mismo ocurre con la varianza), lo cual puede generar cierta confusión si no se sabe.

Dos igualdades que en algunos casos nos pueden resultar de interés en la práctica son las que ponen en relación varianza y desviación típica con cuasi-varianza y cuasi-desviación típica, respectivamente, pues si conocemos una podremos obtener la otra fácilmente:

$$s_X^{2\prime} = \frac{s_X^2 \cdot n}{n-1} \qquad\qquad s_X' = \frac{s_X \cdot \sqrt{n}}{\sqrt{n-1}}$$

3) Análogamente al caso de la varianza, el mejor estimador del parámetro de la covarianza (σ_{XY}) no es el estadístico de la covarianza, sino el de la cuasi-covarianza (s_{XY}'):

$$\hat{\sigma}_{XY} = s_{XY}' = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1}$$

Otra igualdad que nos puede resultar útil es la que relaciona la covarianza y la cuasi-covarianza, pues si conocemos una de las dos, fácilmente podremos obtener la otra:

$$s_{XY}' = \frac{s_{XY} \cdot n}{n-1}$$

Ejercicio 2: A partir de los siguientes datos para la variables "Edad" (X) y "Nº de ataques epilépticos durante el último año" (Y), recogidos a partir de una muestra de jóvenes con diagnóstico de epilepsia, obtener una estimación de los parámetros de: (1) la media de "Edad" ($\hat{\mu}_X$); (2) la mediana y la varianza de "Nº de ataques epilépticos" ($\hat{Md}_Y, \hat{\sigma}_Y^2$); (3) la covarianza y el coeficiente de correlación de Pearson entre ambas variables ($\hat{\sigma}_{XY}, \hat{\rho}_{XY}$). Se recomienda utilizar alguna calculadora científica o programa informático (Excel, SPSS…) para realizar los cálculos.

X	Y
18	4
19	5
15	3
11	1
17	3
13	2
14	3

• A modo de resumen, los estimadores presentados en esta sección ofrecen una estimación puntual de un parámetro, pues se le atribuye al parámetro el valor concreto (puntual) obtenido a partir de la función matemática utilizada como estimador del mismo. Complementaria a esta estrategia, se abordará en una sección posterior la conocida como estimación por intervalos.

3. La distribución muestral de un estadístico

• La estimación de un parámetro determinado (por ejemplo, la mediana de una determinada variable X en una población) a partir de la aplicación de su mejor estimador sobre los datos de una muestra, supone obtener un valor ($\hat{Md}_X$) que no tiene por qué coincidir exactamente con el verdadero valor del parámetro (Md_X). A esa diferencia se le conoce como error muestral.

> No hay que olvidar que una muestra es un subconjunto (aleatorio, en el mejor de los casos) de la población y que, por tanto, puede no ser perfectamente representativo de la población. Prueba de ese error inherente al muestreo es que para distintas muestras extraídas de una misma población es de esperar que, para un estadístico determinado, se obtenga un resultado distinto en cada una de esas muestras.

• Una limitación importante de los estimadores puntuales es que no ofrecen ningún tipo de información sobre el nivel de error muestral que puede acompañar al valor estimado obtenido. Obviamente, no será igual la incertidumbre asociada a una estimación de un parámetro obtenida a partir de una muestra de 5 sujetos, que a partir de una de 50 o una de 500, pero ¿se podría concretar de algún modo esa incertidumbre?

• Precisamente, el concepto de distribución muestral va a ofrecernos una aproximación a la valoración del error muestral asociado a la estimación estadística. La distribución muestral de un estadístico consiste en la función de probabilidad de un estadístico (Pardo y San Martín, 2001), esto es, la correspondencia entre los distintos valores que tome ese estadístico en todas las posibles muestras de un mismo tamaño extraídas de una determinada población y las probabilidades de que se den esos valores.

• A fin de aproximarnos al concepto de distribución muestral, veamos cómo sería la construcción de la misma con un **ejemplo,** en concreto, vamos a obtener la distribución muestral no de uno, sino de dos estadísticos, la media y la varianza, en ambos casos para muestras de tamaño 10 (n = 10), siendo la variable "Nº de horas de estudio al día" (X) y la población los estudiantes de la *UVEG*.

(Con fines didácticos, vamos a imaginar que desde el *más allá* tenemos el privilegio de recibir la siguiente revelación estadística: la variable "Nº de horas de estudio al día" en la población de la UVEG se distribuye según la curva normal con μ_X = 5,63 y σ_X = 1,92 $[X \rightarrow N(5{,}63;\ 1{,}92]$. Esta información, no conocida habitualmente a priori, nos será útil para comprobar después algunas de las propiedades de una distribución muestral.)

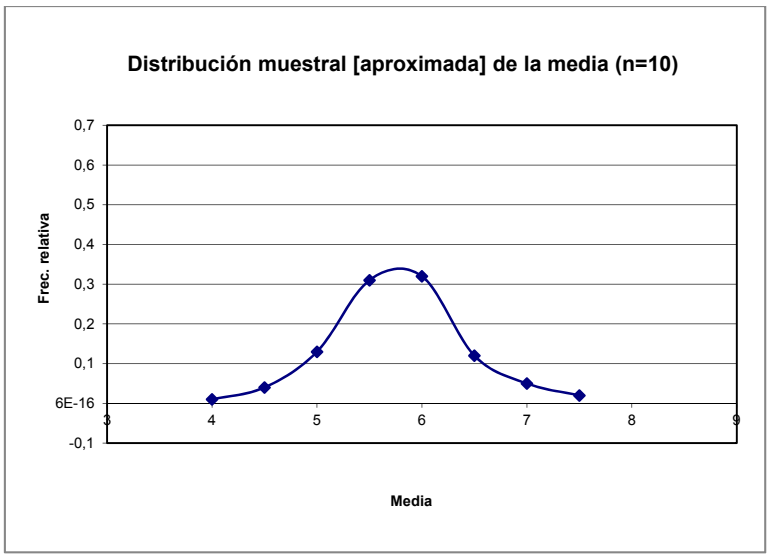
Volviendo a nuestro objetivo, el de obtener la distribución muestral de la media y la distribución muestral de la varianza de la variable "Nº de horas de estudio al día" para muestras de tamaño 10, ello supondría obtener la media y la varianza en todas las muestras posibles (n = 10) de la población de estudiantes de la UVEG. Sin embargo, dada la enorme dificultad práctica de tal cometido, se decide recoger datos en 100 muestras de 10 estudiantes extraídas aleatoriamente de la población de estudiantes de la UVEG. Así, en cada una de esas 100 muestras se calculó la media y la varianza de X, obteniéndose los siguientes resultados (las medias están redondeadas con una precisión de 0,5 unidades y las varianzas de 0,1):

	Media ($\overline{X}$)*	*Varianza* (s_X^2)*
Muestra1	5,5	3,3
Muestra2	4,5	3,8
Muestra3	5	3,6
Muestra4	6,5	3,5
Muestra5	5	3,9
Muestra6	4,5	3,7
.....	.....	.....
.....	.....	.....
Muestra100	6	3,6

Si consideramos a la columna de las medias como una variable y obtenemos la correspondiente distribución de frecuencias relativas, lo que obtendremos será la distribución muestral del estadístico de la media para la variable X en muestras de tamaño $n = 10$ (ver tabla a continuación). En realidad, se trata de una aproximación a la distribución muestral verdadera, dado que se ha obtenido con 100 muestras y no el total de las que se pueden extraer de la población (que son muchísimas).

Distribución de frecuencias de la variable $\overline{X}$ ($n = 10$)

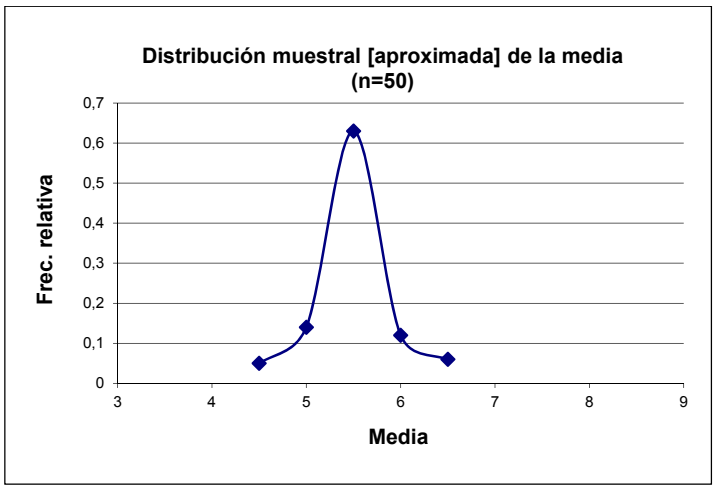
$\overline{X}$	n_i	p_i ($\approx P_i$)
4	1	0,01
4,5	4	0,04
5	13	0,13
5,5	31	0,31
6	32	0,32
6,5	12	0,12
7	5	0,05
7,5	2	0,02
	100	1



La anterior distribución muestral de la media podría haberse obtenido a partir de muestras $n = 50$. Vamos a suponer que lo hemos hecho y que se han obtenido los siguientes resultados:

Distribución de frecuencias de la variable $\overline{X}$ ($n = 50$)

$\overline{X}$	n_i	p_i ($\approx P_i$)
4,5	5	0,05
5	14	0,14
5,5	63	0,63
6	12	0,12
6,5	6	0,06
	100	1

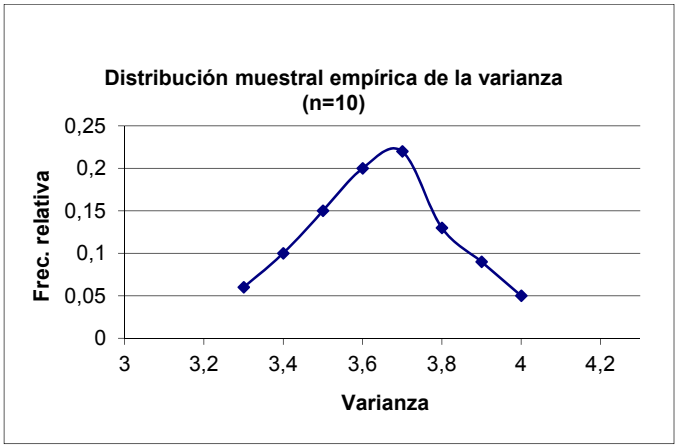


Ejercicio 3: ¿Qué ha cambiado en la distribución muestral de la media al aumentar el tamaño de las muestras de $n = 10$ a $n = 50$?

Si en vez de en la media, nos centramos ahora en el estadístico de la varianza, más concretamente, en la varianza de los datos recogidos con muestras de tamaño $n = 10$ (ver tabla más arriba), al calcular la correspondiente distribución de frecuencias relativas, lo que obtendremos será la distribución muestral (aproximada) del estadístico de la varianza para la variable X en muestras de tamaño $n = 10$.

Distribución de frecuencias de la variable s_X^2

s_X^2	n_i	p_i ($\approx P_i$)
3,3	6	0,06
3,4	10	0,1
3,5	15	0,15
3,6	20	0,2
3,7	22	0,22
3,8	13	0,13
3,9	9	0,09
4	5	0,05
	100	1



Tal como se ha obtenido para la media y para la varianza, podríamos obtener la distribución muestral de otros estadísticos para la variable "Nº de horas de estudio", por ejemplo, de la mediana, del coeficiente de variación... Eso sí, debe tenerse en cuenta que se trataría de aproximaciones a la distribución muestral verdadera de esos estadísticos, dado que las frecuencias relativas son estimaciones de los verdaderos valores de probabilidad que caracterizan la definición de la distribución muestral de un estadístico.

• Las aspectos principales en que se suele centrar la atención a la hora de caracterizar la distribución muestral de un estadístico son: (1) la forma de la distribución; (2) su media (esperanza); y (3) su varianza o la raíz cuadrada de la misma, la desviación estándar, usualmente referida al hablar de una distribución muestral como error típico o error estándar de estimación (en lo sucesivo, utilizaremos habitualmente la expresión más abreviada de *error estándar* o *EE*).

• El error estándar de estimación aporta una información de gran interés sobre la distribución muestral de un estadístico: cuanto menor sea éste, ello supondrá mayor proximidad entre los valores obtenidos por ese estadístico en las posibles muestras que se extraigan de la población. Así, el *EE* representa un concepto clave a la hora de valorar el nivel de error muestral que puede acompañar a las inferencias que realicemos con un determinado estadístico.

• Ahora bien, ¿ello significa que si queremos tener un indicador del grado de precisión de un determinado estadístico obtenido a partir de una muestra, se ha de obtener ese mismo estadístico en 99 muestras más (tantas como posibles, en realidad), a fin de poder conocer el *EE* de la distribución muestral del estadístico aplicado? Afortunadamente, no.

• Un aspecto fundamental del concepto de distribución muestral de un estadístico es que para algunos de los estadísticos más utilizados es conocida la forma de obtener cómo son sus características principales (forma de la distribución, esperanza y *EE*) y ello, independientemente de cuál sea la variable considerada, la población de referencia, o el tamaño elegido para las muestras. A continuación se describen cuáles son esas características para las distribuciones muestrales de los estadísticos de la media y la proporción, dos de los estadísticos más utilizados en la práctica.

3.1. La distribución muestral de la media

• Características de la distribución muestral de la media:

1. Forma de la distribución: (a) si una variable *X* se distribuye normalmente en la población, la distribución muestral del estadístico de la media para esa variable también será normal; (b) en

caso de que *X* no se distribuya normalmente, la distribución muestral de la media de *X* también tiende a distribuirse normalmente cuando ésta se obtiene con muestras de 30 o más casos ($n \geq 30$). Esta es una propiedad derivada del conocido como teorema central del límite, una contribución teórica de gran importancia dentro del campo de la Estadística Inferencial.

2. Media: $\mu_{\bar{X}}\ [E(\bar{X})] = \mu_X$
3. Varianza: $\sigma^2_{\bar{X}}\ [VAR(\bar{X})] = \dfrac{\sigma^2_X}{n}$

   Y, en consecuencia, el error estándar de estimación: $\sigma_{\bar{X}}\ [EE(\bar{X})] = \dfrac{\sigma_X}{\sqrt{n}}$

En resumen, siempre que $n \geq 30$, la distribución muestral del estadístico de la media se distribuye:

$$\bar{X} \rightarrow N\left(\mu_X; \frac{\sigma_X}{\sqrt{n}}\right)$$

• Respecto a la magnitud del *EE,* el cual proporciona la importante información de la precisión de las estimaciones asociadas al estadístico de la media, si nos fijamos en la fórmula, éste será menor: (1) cuanto menor sea la varianza (o desviación típica) de la variable en la población; (2) cuanto mayor sea el tamaño muestral que se considere.

• Podemos comprobar la segunda propiedad, la de la media de la distribución muestral de la media ($\mu_{\bar{X}} = \mu_X$), en nuestro ejemplo de la variable "Nº horas de estudio". Recordemos que, de acuerdo a una revelación que tuvimos la fortuna de recibir, sabemos que la media de esta variable en la población de la *UVEG* es igual a 5,63:

$$\mu_X = 5,63$$

Obsérvese, sin embargo, que si se calcula la media de la distribución muestral obtenida con 100 muestras de $n = 10$ a partir de la distribución de frecuencias presentada más arriba, se obtiene:

$$\mu_{\bar{X}} = 4 \cdot 0,01 + 4,5 \cdot 0,04 + 5 \cdot 0,13 + 5,5 \cdot 0,31 + 6 \cdot 0,32 + 6,5 \cdot 0,12 + 7 \cdot 0,05 + 7,5 \cdot 0,02 = 5,77$$

¿A qué se debe esta inesperada discrepancia? –El resultado obtenido para $\mu_{\bar{X}}$ no coincide exactamente con el de μ_X debido a que aquél se ha obtenido a partir de una distribución muestral construida con un número reducido de todas las muestras posibles (100) y que es, por tanto, una aproximación a la distribución muestral verdadera del estadístico.

Ejercicio 4: Obtener la esperanza de la distribución muestral de la media obtenida con 100 muestras de tamaño $n = 50$ a partir de la distribución de frecuencias correspondiente presentada más arriba. ¿Coincide con el valor *revelado* de la media de la variable "Nº horas de estudio" en la

población (μ_X = 5,63)? ¿A qué puede ser debida esta discrepancia? ¿Es más o menos próximo al valor de la media poblacional que el obtenido a partir de la distribución muestral obtenida con 100 muestras de n = 10? ¿Cuál puede ser el motivo?

• Por lo que respecta a la tercera propiedad de la distribución muestral de la media, teniendo en cuenta el valor revelado para la desviación típica de la variable en la población (σ_X = 1,92), podemos plantearnos el cálculo del error estándar de la distribución muestral de la media tanto para el caso de muestras de n = 10 como para el de n = 50:

$$n = 10 \quad \rightarrow \quad \sigma_{\bar{X}} \ [EE(\bar{X})] = \frac{1,92}{\sqrt{10}} = 0,61$$

$$n = 50 \quad \rightarrow \quad \sigma_{\bar{X}} \ [EE(\bar{X})] = \frac{1,92}{\sqrt{50}} = 0,27$$

Nótese cómo disminuye la dispersión de la distribución muestral de la media a medida que aumenta el tamaño de la muestra, es decir, cómo se obtienen estimaciones puntuales de la media mucho más cercanas al verdadero valor del parámetro de la media en la población.

3.1.1. Aplicaciones del concepto de la distribución muestral de la media

• Una aplicación fundamental que se deriva de saber que la distribución muestral de la media sigue la curva normal es que se puede aprovechar la tabla de la distribución normal estándar para contestar a diferentes preguntas de carácter aplicado. Básicamente, de dos tipos:

1. Obtener la probabilidad asociada a un rango de valores de media → Para una variable (X) de la que se conocen los parámetros de la media (μ_X) y la desviación típica (σ_X), ¿cuál es la probabilidad de que en una muestra extraída al azar de esa población se obtenga una media ($\bar{X}$) menor a un valor determinado (o mayor, o entre tal y tal valor)?

Ejemplo: sabiendo que las puntuaciones en un test de rendimiento verbal se distribuyen según $N(5; 1,8)$ en la población de adultos, ¿cuál es la probabilidad de que en una muestra de 25 adultos, la media de las puntuaciones en el test sea inferior o igual a 4?
En este caso sabemos que la distribución muestral del estadístico media obtenida en muestras de n = 25 de dicha población de adultos se ajustará a una distribución normal con parámetros:

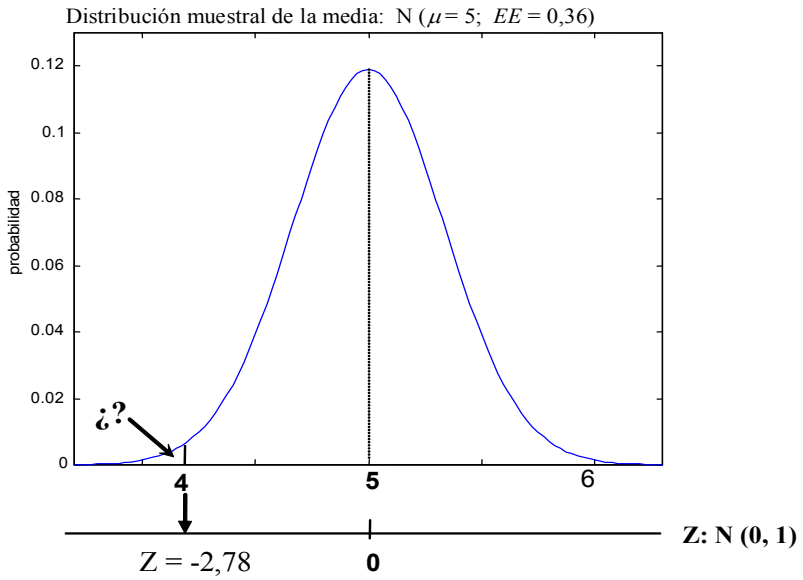
$$\mu_{\bar{X}} = \mu_X = 5 \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1,8}{\sqrt{25}} = 0,36$$

esto es, $N(5; 0,36)$

Utilizar la tabla de la curva normal estandarizada implica que antes tendremos que tipificar el valor de la media a consultar:

$$z_{\bar{X}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{4-5}{0,36} = -2,78$$

El proceso ilustrado gráficamente es:



Y, por tanto, la probabilidad buscada es:

$$P(\bar{X} \leq 4) = P(z \leq -2,78) = 0,003$$

De forma análoga, la probabilidad de que en dicha muestra de 25 adultos la media de las puntuaciones sea superior a 4 es: 1 − 0,003 = 0,997

2. Obtener una media asociada a un determinado valor de probabilidad o, lo que es más habitual, un rango de medias central (intervalo de probabilidad) → Para una variable (X) de la que se conocen los parámetros de la media (μx) y la desviación típica (σx), ¿entre qué valores se encontrará, con un determinado nivel de probabilidad, la media de una muestra extraída al azar de esa población?

(A ese nivel de probabilidad se le conoce como "nivel de confianza" y se representa simbólicamente como "$1-\alpha$")

Ejemplo: sabiendo que las puntuaciones en un test de rendimiento verbal se distribuyen según $N(5; 1,8)$ en la población de adultos, ¿entre qué rango de valores central es de esperar que se encuentre, con un 90% de probabilidades ($1-\alpha = 0,90$), la puntuación media de una muestra de 100 adultos extraída al azar de esa población?

En este caso sabemos que la distribución muestral del estadístico media obtenida en muestras de $n = 100$ de dicha población de adultos se ajustará a una distribución normal con parámetros:

$$\mu_{\bar{X}} = \mu_X = 5 \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1{,}8}{\sqrt{100}} = 0{,}18$$

esto es, $N(5; 0{,}18)$

Utilizar la tabla de la curva normal estandarizada implica saber que los valores z que delimitan el intervalo de medias que nos interesa son:

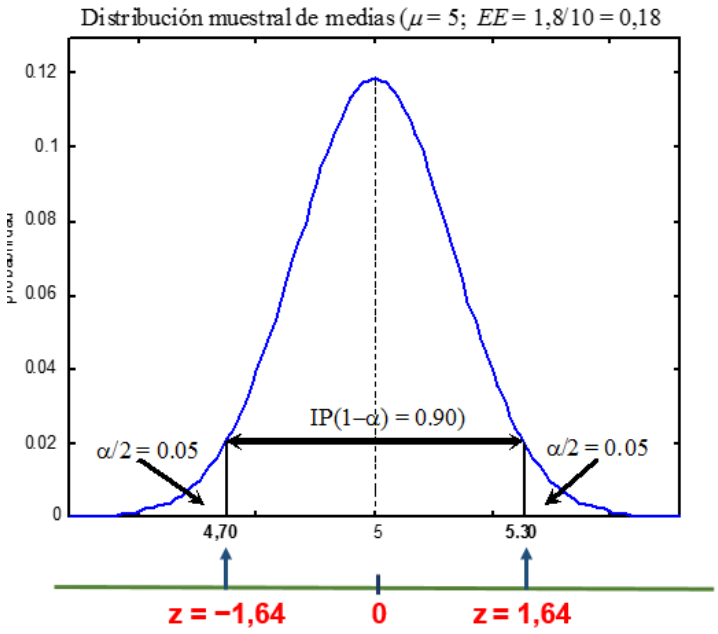
$$z_{0{,}05} = -1{,}64 \text{ y } z_{0{,}95} = 1{,}64,$$

de manera que, despejando el valor de las medias, tenemos:

$$\mathbf{-1,64 = \frac{\bar{X} - 5}{0,18} \quad \rightarrow \quad \bar{X} = 5 - (1,64 \times 0,18) = 4,70}$$

$$\mathbf{1,64 = \frac{\bar{X} - 5}{0,18} \quad \rightarrow \quad \bar{X} = 5 + (1,64 \times 0,18) = 5,30}$$

El proceso ilustrado gráficamente:



Expresión formal de <u>cálculo del intervalo de probabilidad</u> (*IP*) de la media muestral ($\bar{X}$) para un determinado nivel de confianza (1-α):

$$IP(1-\alpha)(\bar{X}) = \left[l_{\text{inf}}\,; l_{\text{sup}}\right] = \left[E\left(\bar{X}\right) + z_{(\alpha/2)} \cdot EE\left(\bar{X}\right); E\left(\bar{X}\right) + z_{(1-\alpha/2)} \cdot EE\left(\bar{X}\right)\right]$$

$$= \left[\mu_X + z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma_X}{\sqrt{n}}; \mu_X + z_{\left(1-\frac{\alpha}{2}\right)} \cdot \frac{\sigma_X}{\sqrt{n}}\right]$$

Así, para nuestro **ejemplo**:

$$IP(0,90)(\bar{X}) = \left[5 - 1,64 \cdot \frac{1,8}{\sqrt{100}} ; 5 + 1,64 \cdot \frac{1,8}{\sqrt{100}} \right] = \left[4,70 ; 5,30 \right]$$

3.1.2. Acerca de (1-α) y de los valores z asociados

• Como ya se ha señalado, se utiliza la expresión (*1-α*) o nivel de confianza para hacer referencia a la probabilidad de que el *IP* que obtengamos contenga el valor de interés. En cuanto que probabilidad, el nivel de confianza siempre será un valor que oscile entre 0 y 1 ($0 \leq (1-\alpha) \leq 1$), si bien, es también habitual que se exprese como %.

• También se suele utilizar en la práctica el término complementario, nivel de riesgo (α), para hacer referencia a la probabilidad de que el *IP* no contenga el valor de la media de una muestra extraída al azar de la población –por ejemplo, en el *IP* de la media que fue construido anteriormente, 0,10 representa ese nivel de riesgo o α.

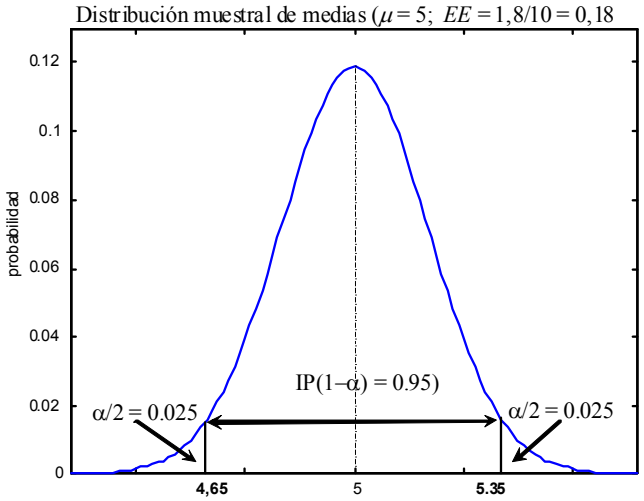
• Valores de la distribución normal estandarizada asociados a niveles de confianza/riesgo concretos:

(1- α)	*α*	$z_{(\alpha/2)}$	$z_{(1-\alpha/2)}$
0,68 [68%]	0,32 [32%]	-1	1
0,90 [90%]	**0,10 [10%]**	**-1,64**	**1,64**
0,95 [95%]	**0,05 [5%]**	**-1,96**	**1,96**
0,954 [95,4%]	0,046 [4,6%]	-2	2
0,99 [99%]	**0,01 [1%]**	**-2,58**	**2,58**
0,9974 [99,74%]	0,0026 [0,26%]	-3	3

Ejemplo: si obtenemos de nuevo el *IP* del ejemplo anterior pero considerando un nivel de riesgo del 5% ($\alpha = 0,05$) o, lo que es lo mismo, un nivel de confianza del 95%, será igual a:

$$IP(0,95)(\bar{X}) = \left[5 - 1,96 \cdot \frac{1,8}{\sqrt{100}} ; 5 + 1,96 \cdot \frac{1,8}{\sqrt{100}} \right] = \left[4,65; 5,35 \right]$$

Gráficamente:



3.1.3. Acerca de la precisión de los intervalos

• Los valores de z van a determinar cuan probable es que el *IP* contenga la media muestral. Cuanto mayor se desee que sea esa probabilidad (nivel de confianza), mayores en valor absoluto serán los valores de z y, en consecuencia, la amplitud del intervalo. Ello implica también que el intervalo será menos informativo, menos preciso. El establecimiento de un *IP* supone un compromiso entre el nivel de confianza y la precisión de la información ofrecida.

• A modo de resumen, un *IP* será más preciso (más informativo) cuanto más estrecho sea, esto es, cuanto menor sea la distancia entre l_{inf} y l_{sup}. De la expresión de cálculo del *IP* se deriva que éste será más estrecho cuanto más bajos sean, bien el nivel de confianza -o sea, los valores de z (lo cual implica menor probabilidad de que se encuentra la $\bar{X}$ en el *IP*)-, bien el valor de (σ_x/n). En este segundo caso, al tratarse de un cociente, éste será menor cuanto mayor sea n o cuanto menor sea σ_x. Esta última, σ_x, es un parámetro intrínseco a la variable de interés, no dependiendo en principio de ninguna decisión externa, cosa que no ocurre con n, el tamaño de la muestra, que sí que es una decisión que puede venir determinada por nosotros.

Ejercicio 5: Una variable X sabemos que, en una determinada población, se distribuye normalmente con media μ=100 y desviación estándar σ=15 [$X \rightarrow N(100; 15)$].

a) Si de esa población extraemos al azar infinitas muestras de n sujetos, ¿cuál será la forma de la distribución muestral del estadístico $\overline{\mathbf{X}}$ y cuáles sus características?
b) ¿Cuál es la probabilidad de que un sujeto extraído al azar de esta población tenga un valor en X igual o superior a 110 [$P(X \geq 110)$]?

c) ¿Cuál es la probabilidad de que una muestra de 20 sujetos extraída al azar de dicha población tenga una media superior o igual a 110 [$P(\overline{X} \geq 110)$]?

d) ¿Cuál es la probabilidad al extraer una muestra aleatoria de 20 sujetos de dicha población de que la media observada en dicha muestra sea inferior o igual a 95?, ¿y si la muestra fuera de 100 sujetos?

e) En una distribución normal ¿entre qué puntuaciones típicas (z) se encuentra el 90% central de las observaciones?, ¿y el 95%?, ¿y el 99%?, ¿a que valores de la variable *X* se corresponden esas puntuaciones típicas?

f) [Intervalos de probabilidad] ¿Entre qué valores se encontrará, con una probabilidad del 95%, la media de *X* en una muestra de 20 sujetos extraída aleatoriamente de dicha población?, ¿y si la muestra fuera de 50 sujetos?

g) [Intervalos de probabilidad] ¿Entre qué valores se encontrará, con una probabilidad del 90%, la media de una muestra de 20 sujetos extraída al azar de dicha población?

3.2. La distribución muestral de la proporción

• Características de la distribución muestral de la proporción:

1. Forma de la distribución: La de la distribución binomial $B(n,\ \pi_{X_i})$, donde π_{X_i} es la proporción asociada a la categoría *i* de la variable categórica *X* en la población, y *n* es el tamaño de muestra con que se construya la distribución muestral.
   Si el tamaño de muestra es suficientemente grande, la forma de la distribución muestral de la proporción puede considerarse como normal. → Criterio de *muestra suficientemente grande* que se suele considerar en la práctica: que se cumpla que $n \cdot \pi_{Xi} \geq 5$ y $n \cdot (1-\pi_{Xi}) \geq 5$
2. Media: $\mu_{P_{Xi}}\ [E(p_{Xi})] = \pi_{Xi}$
3. Varianza: $\sigma^2_{P_{Xi}}\ [VAR(p_{Xi})] = \dfrac{\pi_{Xi} \cdot (1-\pi_{Xi})}{n}$

   Y, en consecuencia, el error estándar de estimación: $\sigma_{P_{Xi}}\ [EE(p_{Xi})] = \sqrt{\dfrac{\pi_{Xi} \cdot (1-\pi_{Xi})}{n}}$

En resumen, siempre que la muestra sea suficientemente grande, la distribución muestral del estadístico de la proporción se distribuye:

$$p_{X_i} \rightarrow N\left(\pi_{Xi};\ \sqrt{\frac{\pi_{Xi} \cdot (1-\pi_{Xi})}{n}}\right)$$

• **Ejemplo** de la construcción empírica de la distribución muestral del estadístico proporción: Del mismo modo en que se construyó más arriba la distribución muestral de la media para la variable "Nº horas...", imagina el proceso de construcción de la distribución muestral de la proporción de mujeres entre los estudiantes de la UVEG (X = "Sexo"; X_i = "Mujer") para muestras de tamaño n = 20 sabiendo que el porcentaje de mujeres en esa población es del 60% ($\pi_{Xi} = 0{,}60$).

Obtener la distribución muestral supondría obtener la proporción de mujeres en todas las muestras posibles (n = 20) de la población de estudiantes de la UVEG. Supongamos que se seleccionan 1000 muestras y, tras calcularse la proporción de mujeres en cada una de ellas, se obtiene la distribución de frecuencias siguiente:

p_{mujer}	n_i	p_i
0	15	0,015
0,125	34	0,034
0,25	53	0,053
0,375	74	0,074
0,5	220	0,22
0,675	375	0,375
0,75	152	0,152
0,875	54	0,054
1	23	0,023
	1000	1

La media aritmética de la distribución muestral obtenida es:

$$\mu_{P_{mujer}} = (0 \cdot 15 + 0{,}125 \cdot 34 + 0{,}25 \cdot 53 + 0{,}375 \cdot 74 + ...)/1000 = 0{,}593$$

Este resultado sólo se puede considerar una aproximación al verdadero valor del parámetro ($\pi_{Xi} = 0{,}60$) porque la distribución muestral a partir de la que ha sido calculado es también una aproximación a la verdadera distribución muestral, pues sólo se ha obtenido a partir de 1000 muestras y no a partir de todas las posibles de tamaño n = 20.

La verdadera distribución muestral del estadístico proporción en este ejemplo, es decir, si se hubieran obtenido todas las posibles muestras de n = 20 de esta población, se ajustaría a la curva normal dado que:

$$20 \cdot 0{,}60\ (=12) > 5 \quad \text{y} \quad 20 \cdot 0{,}40\ (=8) > 5$$

con parámetros:

$$\mu_{P_{Xi}} = 0{,}60$$

$$\sigma_{P_{Xi}} = \sqrt{\frac{0{,}60 \cdot 0{,}40}{20}} = 0{,}11$$

esto es, podemos asumir que esta distribución muestral se distribuye según $N(0{,}60;\ 0{,}11)$

Respecto a la magnitud del *EE,* informativo de la precisión de las estimaciones asociadas al estadístico de la proporción, éste será menor: (1) cuanto más pequeño sea el numerador que aparece en la fórmula del *EE* ($=\pi_{Xi}\cdot(1-\pi_{Xi})$), en consecuencia, cuanto más alejado esté π_{Xi} de 0,5; (2) complementariamente, cuanto mayor sea el tamaño muestral (n) que se considere.

Así, siguiendo con el ejemplo anterior, si las muestras hubieran sido de 100 estudiantes, el error estándar disminuiría a:

$$\sigma_{P_{Xi}}[EE(p_{Xi})] = \sqrt{\frac{0{,}60\cdot 0{,}40}{100}} = 0{,}05$$

3.2.1. Aplicaciones del concepto de la distribución muestral de la proporción

• Una aplicación fundamental que se deriva de saber que la distribución muestral de la proporción sigue la curva normal (en caso contrario, habría que recurrir a la tabla de la distribución binomial) es que se puede aprovechar la tabla de la distribución normal estándar para contestar a diferentes preguntas de carácter aplicado. Se trata, en esencia, de dos tipos de preguntas:

1. Obtener la probabilidad asociada a un valor o a un rango de valores de proporción → Para una variable categórica (X) de la que se conoce a nivel poblacional la proporción para una determinada categoría de la misma π_{Xi}, ¿cuál es la probabilidad de que para una muestra extraída al azar de esa población se obtenga un valor de proporción (p_{Xi}) menor a un valor determinado (o mayor, o entre tal y tal valor)?

Ejemplo: sabiendo que en la población de estudiantes de la UVEG la proporción de estudiantes que tienen su residencia habitual en la ciudad de Valencia es de 0,68 ($\pi_{Valencia}$ = 0,68), ¿cuál es la probabilidad de extraer una muestra de 20 estudiantes y que sólo la mitad (o menos) tengan su residencia habitual en la ciudad de Valencia ($p_{Valencia} \leq 0{,}50$)?

Primero, ¿se puede asumir que la distribución muestral de la proporción en este caso se ajusta a la curva normal? Criterios: $0{,}68\cdot 20 = 13{,}6\ (\geq 5)$ y $0{,}32\cdot 20 = 6{,}4\ (\geq 5)$ → Sí se puede.

Por tanto, sabemos que la distribución muestral del estadístico proporción obtenida en muestras de $n = 20$ de dicha población se ajustará a una distribución normal con parámetros:

$$\mu_{P_{Xi}} = 0{,}68; \qquad \sigma_{P_{Xi}} = \sqrt{\frac{0{,}68\cdot 0{,}32}{20}} = 0{,}104$$

esto es: $p_{Xi} \rightarrow N(0,68;\ 0,104)$

Por otra parte, utilizar la tabla de la curva normal estandarizada implica que antes tendremos que tipificar el valor de la proporción a consultar => $z_{p_{Xi}} = \frac{p_{Xi} - \mu_{p_{Xi}}}{\sigma_{p_{Xi}}} = \frac{0,50 - 0,68}{0,104} = -1,73$

Así, para nuestro ejemplo: $P(p_{Valencia} \leq 0,50) = P\,(z \leq -1,73) = 0,042$

Complementariamente, la probabilidad de que en dicha muestra de 20 estudiantes más de la mitad vivan en Valencia será: $1 - 0,042 = 0,958$

2. Obtener una proporción asociada a un determinado valor de probabilidad o, más comúnmente, un rango de proporciones central (intervalo de probabilidad): Para la categoría *i* de una variable nominal *X* de la que se conoce su proporción en la población de interés (π_{Xi}), ¿entre qué rango de valores central se encontrará, con un determinado valor de probabilidad (nivel de confianza), la proporción de esa categoría en una muestra extraída al azar de esa población (p_{Xi})?

Ejemplo: siguiendo con el ejemplo de la variable "Lugar de residencia habitual" [Valencia; fuera de Valencia] en la población de estudiantes de la UVEG ($\pi_{Valencia} = 0,68$), ¿entre que valores cabe esperar que se encuentre, con una probabilidad del 99%, la proporción de estudiantes que residen en Valencia en una muestra aleatoria de 120 estudiantes de la UVEG?

En este caso sabemos que la distribución muestral del estadístico proporción obtenida en muestras de $n = 120$ de dicha población de adultos se ajustará a una distribución normal con parámetros:

$$\mu_{P_{Xi}} = 0,68\,; \qquad \sigma_{P_{Xi}} = \sqrt{\frac{0,68 \cdot 0,32}{120}} = 0,043$$

esto es: $p_{Xi} \rightarrow N(0,68;\ 0,043)$

Utilizar la tabla de la curva normal estandarizada implica saber que los valores z que delimitan el intervalo de medias que nos interesa son: $z_{0,005} = -2,58$ y $z_{0,995} = 2,58$ de manera que, despejando el valor de las proporciones, tenemos:

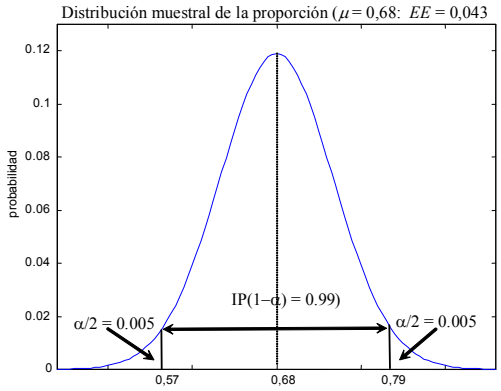
$$-2,58 = \frac{p - 0,68}{0,043} \rightarrow \quad p = 0,57$$

$$2,58 = \frac{p - 0,68}{0,043} \rightarrow \quad p = 0,79$$

Así, el *IP* buscado será igual a:

$$IP(0,99)(p_{\text{Valencia}}) = [0,57;0,79]$$

Gráficamente:



Otra forma de obtenerlo es directamente a partir de la siguiente expresión de cálculo del *IP* de la proporción muestral (p_{Xi}) para un determinado nivel de confianza (1-α):

$$IP(1-\alpha)(p_{X_i}) = \left[E\left(p_{X_i}\right) + z_{(\alpha/2)} \cdot EE\left(p_{X_i}\right) ;\ E\left(p_{X_i}\right) + z_{(1-\alpha/2)} \cdot EE\left(p_{X_i}\right) \right]$$

$$= \left[\pi_{X_i} + z_{(\alpha/2)} \cdot \sqrt{\frac{\pi_{X_i} \cdot \left(1-\pi_{X_i}\right)}{n}} \ ;\ \pi_{X_i} + z_{(1-\alpha/2)} \cdot \sqrt{\frac{\pi_{X_i} \cdot \left(1-\pi_{X_i}\right)}{n}} \right]$$

Así, para el **ejemplo** anterior:

$$IP(0,99)(p_{\text{Valencia}}) = \left[0,68 - 2,58 \cdot \sqrt{\frac{0,68 \cdot 0,32}{120}} \ ;\ 0,68 + 2,58 \cdot \sqrt{\frac{0,68 \cdot 0,42}{120}} \right] = [0,57;0,79]$$

Ejercicio 6: En la población de personas que viven en la Comunidad Valenciana con edad comprendida entre 30 y 40 años hay un 40% de fumadores.

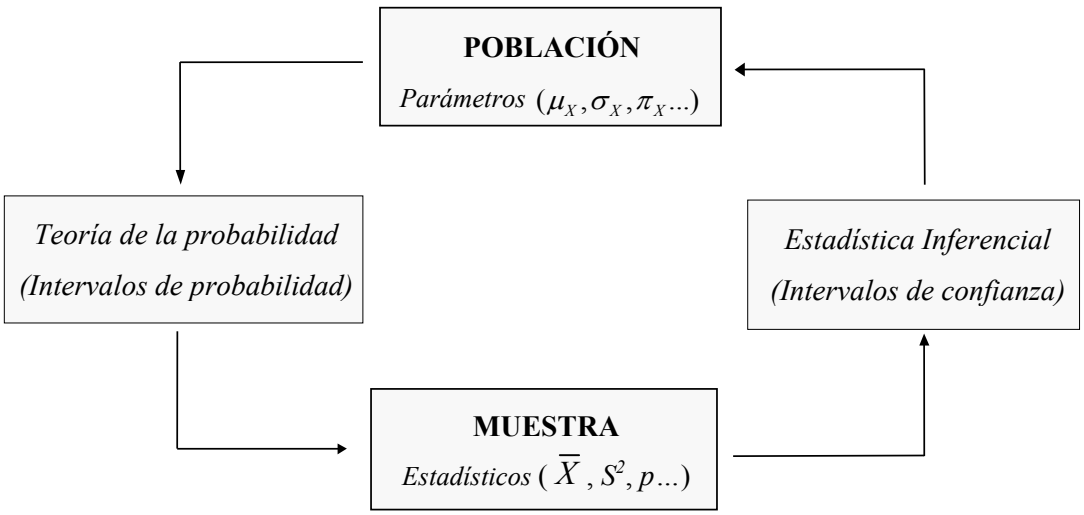
a) Si extraemos al azar de esa población infinitas muestras de 25 sujetos, ¿cuál será la forma de la distribución muestral del estadístico p_x y cuáles sus características?
b) Al extraer una muestra aleatoria de 25 sujetos de dicha población, ¿cuál es la probabilidad de que haya una proporción de fumadores superior a 0,5?
c) ¿Cuál sería la probabilidad de obtener un porcentaje de fumadores superior al 50% ($p_x >$ 0,50) si la muestra fuera de 40 sujetos?
d) ¿Entre qué valores se encontrará, con una probabilidad del 95%, la proporción observada de fumadores (p) en una muestra de 20 sujetos extraída aleatoriamente de dicha población?
e) ¿Y si la muestra fuese de 40 sujetos?

4. Estimación basada en intervalos de confianza

4.1. Intervalos de probabilidad vs. intervalos de confianza

Ambos conceptos reflejan la complementariedad de la Probabilidad y de la Estadística:

- La teoría de la probabilidad establece los procedimientos que permiten realizar predicciones acerca de las características de una muestra (estadísticos) extraída al azar de una población en que esas características (parámetros) son conocidas. Un procedimiento básico para realizar tal tipo de predicción es el **intervalo de probabilidad** (*IP*), un intervalo de valores que, con un determinado nivel de confianza, contendrá el valor del estadístico. En la sección anterior se trató como obtener los *IP* de la media y la proporción.
- La teoría estadística estudia la realización de inferencias acerca de las características de una población (parámetros) a partir de las características de una muestra extraída al azar de esa población (estadísticos). Un procedimiento básico para realizar tal tipo de inferencia es el **intervalo de confianza** (*IC*), un intervalo de valores que tiene un determinado nivel de confianza de contener el valor del parámetro.



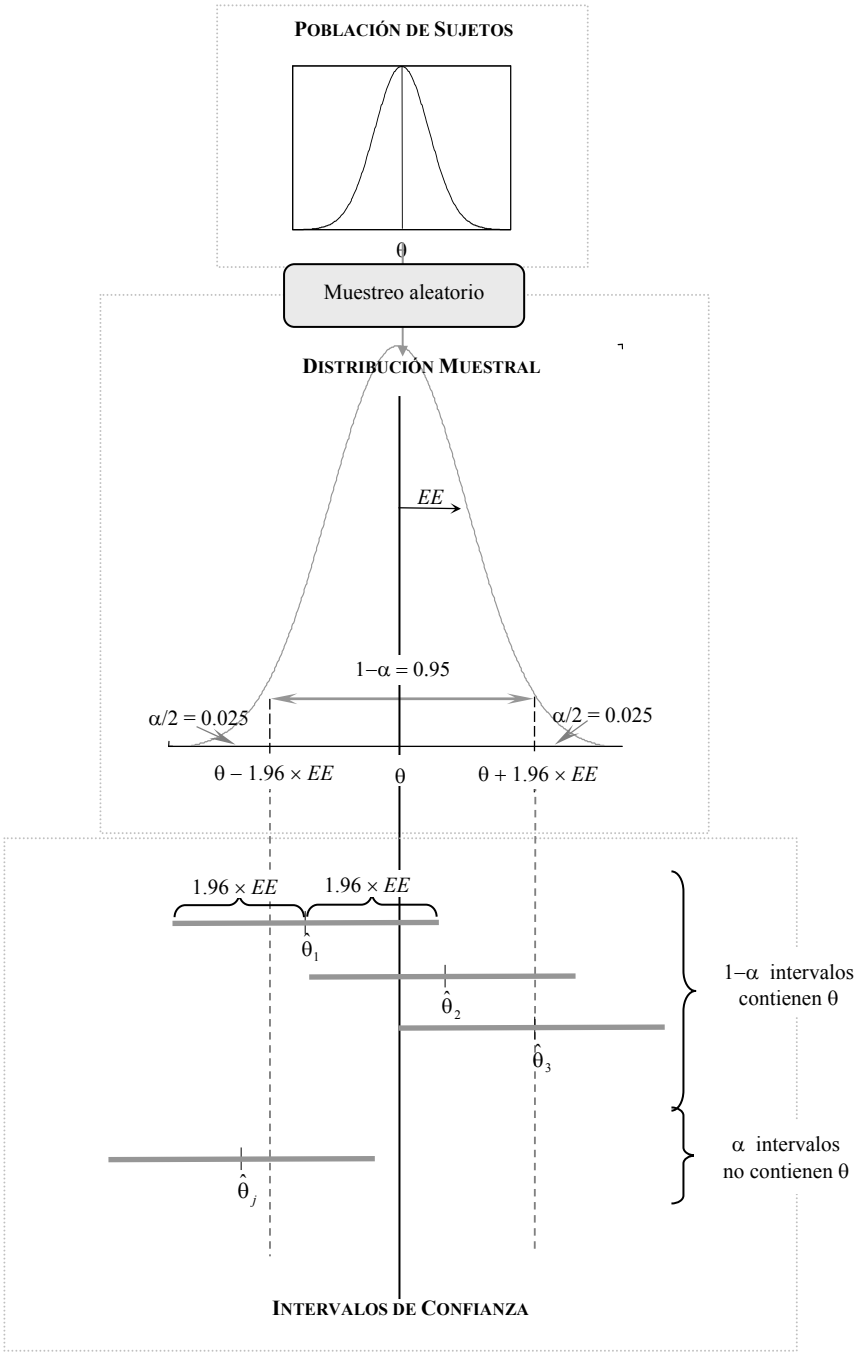
• La estimación por intervalos de confianza (*IC*) de un parámetro cualquiera (θ) consiste en obtener un intervalo de valores a partir de los datos de una muestra de modo que, con una determinada probabilidad (conocida como *nivel de confianza* o, también, *1-α*), el *IC* construido contendrá el verdadero valor del parámetro.

• La obtención de los dos límites del *IC* de un parámetro θ supondrá sumar al estadístico obtenido en una muestra ($\hat{\theta}$, esto es, la estimación puntual de θ), dos términos (ambos del mismo valor, pero de distinto signo). En ese término aparecen multiplicados: (1) un valor que depende del nivel

de confianza asumido en la construcción del *IC*, y (2) el error estándar de la distribución muestral del estadístico en cuestión; Así, la expresión general del *IC* para un determinado parámetro θ es:

$$IC(1-\alpha)(\theta) = \left[\hat{\theta} + z_{(\alpha/2)} \cdot EE(\hat{\theta}) \; ; \; \hat{\theta} + z_{(1-\alpha/2)} \cdot EE(\hat{\theta}) \right]$$

Los dos valores que delimitan un *IC* suelen denominarse como *límite inferior* y *límite superior* del *IC*. La siguiente figura ilustra la construcción del intervalo de confianza de un parámetro con un nivel de confianza del 95%.



Construcción de intervalos de confianza de un parámetro (adaptada de Wonnacott y Wonnacott, 1991, p. 128).

• El nivel de confianza de un *IC* no se ha de interpretar como la probabilidad de que el valor del parámetro de interés se encuentre dentro del *IC* que hayamos construido, sino como el porcentaje de éxito del procedimiento de cálculo que se utiliza en la obtención de un *IC*. Por ejemplo, si creamos un *IC* en que (1−α) es igual a 0,95, ello supone que si calculamos un mismo *IC* en distintas muestras, un 95% de los *ICs* contendría el verdadero valor del parámetro. Es incorrecto interpretar que un *IC* en concreto tiene una probabilidad de 0,95 de contener el valor del parámetro. La figura adjunta, adaptada de Wonnacott y Wonnacott (1991, p. 125-131), permite ilustrar el concepto de intervalo de confianza de un parámetro θ, a partir de la distribución muestral de un estadístico que se distribuye según la curva normal y asumiendo un riesgo de error del 5%:

4.1.1. Acerca de (1-α) y de los valores z asociados

• Como ya se ha señalado, se utiliza la expresión (*1-α*) o nivel de confianza para hacer referencia a la probabilidad de que el *IC* que hayamos construido contenga el valor del parámetro de interés. En cuanto que probabilidad, su valor oscilará entre 0 y 1 ($0 \leq (1-\alpha) \leq 1$), si bien, suele expresarse también como % y, por lo tanto, oscilaría entre 0 y 100. También se suele utilizar en la práctica el término antónimo, nivel de riesgo (*α*), para hacer referencia a la probabilidad complementaria al nivel de confianza.

• A la hora de construir un *IC*, debemos traducir el nivel de confianza en dos valores ($z_{(\alpha/2)}$ y $z_{(1-\alpha/2)}$) cuya obtención depende de la forma de la distribución muestral del estadístico a partir del cual vayamos a construir el *IC*. Si esa forma es la distribución normal, como es el caso de los *ICs* cuya obtención se va a tratar en las dos secciones siguientes, los valores de $z_{(\alpha/2)}$ y $z_{(1-\alpha/2)}$ se obtendrán a partir de la distribución normal estandarizada, en concreto, como los valores *z* correspondientes a las probabilidades acumuladas *α/2* y *1- α/2*, respectivamente. A continuación se muestras algunos valores de la distribución normal estandarizada asociados a niveles de confianza concretos:

(1- α)	α	$z_{(\alpha/2)}$	$z_{(1-\alpha/2)}$
0,68 [68%]	0,32 [32%]	-1	1
0,90 [90%]	**0,10 [10%]**	**-1,64**	**1,64**
0,95 [95%]	**0,05 [5%]**	**-1,96**	**1,96**
0,954 [95,4%]	0,046 [4,6%]	-2	2
0,99 [99%]	**0,01 [1%]**	**-2,58**	**2,58**
0,997 [99,7%]	0,003 [0,3%]	-3	3

• Los valores $z_{(\alpha/2)}$ y $z_{(1-\alpha/2)}$ correspondientes a los niveles de confianza/riesgo más utilizados en la práctica están subrayados en negrita en la tabla anterior, recomendándose su memorización a fin de evitar su búsqueda repetida en la tabla de la distribución normal estandarizada.

4.1.2. Acerca de la precisión de los intervalos

• Cuanto mayor se desee que sea el nivel de confianza a la hora de construir un *IC*, mayores en valor absoluto serán los valores de $z_{(\alpha/2)}$ y $z_{(1-\alpha/2)}$ y, por tanto, la amplitud del intervalo que creemos. Ello implica que el intervalo será menos informativo, menos preciso, pues un *IC* será más preciso cuanto más estrecho sea, esto es, cuanto menor sea la distancia entre los l_{inf} y l_{sup} del *IC*. En consecuencia, no debe olvidarse que el establecimiento de un *IC* supone un compromiso entre el nivel de confianza y la precisión de la información ofrecida.

4.2. Intervalo de confianza de la media (μ_X)

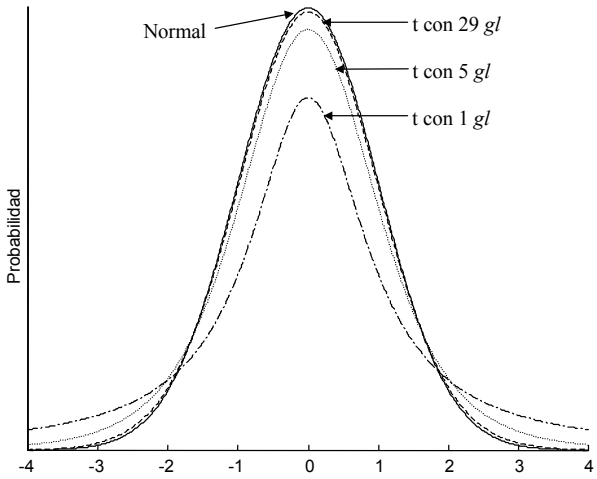
• Dada una muestra de la que se hayan obtenido datos para una variable *X* y en que se conozca la desviación típica de esa variable en la población (algo no habitual), el *IC* de la media se obtiene de acuerdo a la siguiente expresión:

$$IC(1-\alpha)(\mu_X) = \left[\bar{X} + z_{(\alpha/2)} \cdot \frac{\sigma_X}{\sqrt{n}}; \bar{X} + z_{(1-\alpha/2)} \cdot \frac{\sigma_X}{\sqrt{n}} \right]$$

• Dada una muestra de la que se hayan obtenido datos para una variable *X* y en que no sea conocida la desviación típica de esa variable en la población, se sustituye la desviación típica poblacional por su mejor estimador: la cuasi-desviación típica obtenida en la muestra (s'_X):

$$IC(1-\alpha)(\mu_X) = \left[\bar{X} + t_{(n-1)(\alpha/2)} \cdot \frac{s'_X}{\sqrt{n}}; \bar{X} + t_{(n-1)(1-\alpha/2)} \cdot \frac{s'_X}{\sqrt{n}} \right]$$

• La aplicación de la anterior expresión del *IC* de la media implica conocer dos valores de la distribución *t* ($t_{(n-1)(\alpha/2)}$ y $t_{(n-1)(1-\alpha/2)}$), algo que podremos conseguir consultando la tabla de esta distribución en un libro de estadística, o bien, a través de algún programa informático que lo permita (p. ej., *Excel*). Ahora bien, a medida que se considera un mayor número de grados de libertad en la distribución *t*, ésta converge con la distribución normal: las diferencias son ya prácticamente inexistentes para la distribución *t* con 30 grados de libertad (ver figura abajo).



Convergencia de la distribución *t* de Student-Fisher a la Normal

En consecuencia, para muestras de 30 o más sujetos, se puede utilizar la curva normal para obtener los valores *z* asociados al nivel de confianza elegido:

$$IC(1-\alpha)(\mu_X) = \left[\overline{X} + z_{(\alpha/2)} \cdot \frac{s'_X}{\sqrt{n}} ; \overline{X} + z_{(1-\alpha/2)} \cdot \frac{s'_X}{\sqrt{n}} \right]$$

Ejemplo: el gobierno del país pretende realizar una reforma de la jubilación que ha suscitado una gran polémica a nivel nacional. Para sondear la opinión pública sobre dicha propuesta, encarga a una empresa de demoscopia que realice un sondeo. Esta empresa entrevista al azar a 1000 personas de la población y les pide que evalúen en una escala de 0 a 10 en qué medida están de acuerdo con dicha propuesta (siendo 0: totalmente en desacuerdo y 10: totalmente de acuerdo). Se obtiene una media de 4,5 y una cuasi desviación típica de 2,7. ¿Entré qué valores se encontrará la media de la población española con una confianza del 95%? ¿Y con una confianza del 99%?

En este caso sabemos que la distribución muestral de la media obtenida en muestras de n = 1000 de la población española se ajustará a una distribución normal y estimamos que el *EE* de dicha distribución será:

$$\sigma_{\overline{X}} = \frac{s'_X}{\sqrt{n}} = \frac{2,7}{\sqrt{1000}} = 0,085$$

Por tanto, el *IC* del 95% es:

$$IC(0,95)(\mu) = \left[4,5 + (-1,96 \cdot 0,085) \; ; \; 4,5 + (1,96 \cdot 0,085) \right] = \left[4,33 \; ; \; 4,67 \right]$$

El *IC* obtenido contendrá la media poblacional con una probabilidad del 95%. Si se disminuye el riesgo de error a α=0,01, el *IC* del 99% tiene una probabilidad mayor de contener a la media poblacional pero, por el contrario, será más amplio y, por lo tanto, menos preciso:

$$IC(0,99)(\mu) = \left[4,5 - 2,58 \cdot 0,085 \; ; \; 4,5 + 2,58 \cdot 0,085 \right] = \left[4,28 ; 4,72 \right]$$

• De la expresión de cálculo del *IC* se deriva que éste será más estrecho cuanto más bajo sean, bien el nivel de confianza (siendo lo habitual en la literatura científica que se fijen al 90, al 95 o al 99%), bien el valor de (σ_x/n). En este segundo caso, al tratarse de un cociente, éste será menor cuanto mayor sea n o cuanto menor sea σ_x. Esta última, σ_x, es un parámetro intrínseco a la variable de interés, no dependiendo en principio de ninguna decisión externa, cosa que no ocurre con n, el tamaño de la muestra, que sí que es una decisión que puede venir determinada por nosotros.

Ejemplo con *SPSS* a partir de los datos obtenidos con el Cuestionario de Vida Académica: Estimar con un nivel de confianza del 95% la edad media de los estudiantes de Estadística en Psicología de la *UVEG*, asumiendo que los datos obtenidos provienen de una muestra representativa de estudiantes (n = 174) de dicha materia y titulación. En dicha muestra la media se situó en 21,15 años y la cuasi-desviación típica en 5,06 años.

$$EE(\overline{X}) = \frac{5{,}06}{\sqrt{174}} = 0{,}384$$

$$IC\ (0{,}95)(\mu) = 21{,}15 \pm 1{,}96 \cdot 0{,}384 = [20{,}39\ ;\ 21{,}91]$$

Obsérvese la equivalencia con los resultados obtenidos con *SPSS*:

SPSS: Analizar | Estadísticos descriptivos | Explorar:

Descriptivos

			Estadístico	Error típ.
edad	Media		21,15	,384
	Intervalo de confianza para la media al 95%	Límite inferior	20,39	
		Límite superior	21,91	
	Media recortada al 5%		20,30	
	Mediana		20,00	
	Varianza		25,608	
	Desv. típ.		5,060	
	Mínimo		17	
	Máximo		50	
	Rango		33	
	Amplitud intercuartil		2	
	Asimetría		3,561	,184
	Curtosis		13,922	,366

Nota: el botón *Estadísticos*, en el cuadro de diálogo de la función *Explorar* de *SPSS*, permite modificar el nivel de confianza con el que se crea el *IC*.

Ejercicio 7: En una muestra de 40 estudiantes se mide el ritmo cardiaco al comienzo de un examen, obteniéndose un valor medio de 123 p.p.m. (media: 123; varianza = 47). ¿Entre qué valores se hallará el verdadero valor de la media de ritmo cardiaco para la población de estudiantes con un nivel de confianza del 90%? ¿Y con una confianza del 95%? %? (Una pista para empezar a resolver

el problema: dado que no se conoce el valor de la desviación típica de la variable en la población, hay que estimarla a partir de la cuasi-desviación típica obtenida en la muestra).
¿Y cuáles hubiesen sido esos *IC* si la muestra hubiera sido de 20 sujetos? (Datos de interés para la obtención del *IC* (95%): ($_{0,025}t_{19} = -2,09$; $_{0,975}t_{19} = 2,09$) y del IC (90%): ($_{0,05}t_{19} = -1,73$; $_{0,95}t_{19} = 1,73$).

Ejercicio 8: Una cadena hotelera desea conocer el nivel de satisfacción de sus clientes sobre la calidad del servicio. Para ello seleccionan una muestra aleatoria de 100 clientes y obtienen una puntuación media de 7 en una escala de satisfacción del cliente (rango posible de la puntuación: 0 a 10) con una cuasi-desviación típica de 2,3.

a) ¿Entre qué valores se encuentra la media de satisfacción en la población de clientes de esa cadena hotelera con un nivel de confianza del 95% ($\alpha = 0,05$)?
b) ¿Cuál sería el intervalo de confianza si, con el mismo nivel de riesgo, la muestra hubiera sido de 300 clientes?
c) ¿Y cuál sería el *IC* con $\alpha = 0,10$ y n = 300?

Ejercicio 9: En una muestra aleatoria de 20 estudiantes de Psicología encontramos que la media de satisfacción con la carrera es de 6,5 puntos (en una escala de 0 a 10) con una cuasi-desviación típica de 2,4.

a) ¿Cuál será el valor de la media de satisfacción en la población de estudiantes de Psicología? Realiza la estimación por *IC* con un $\alpha = 0,05$ ($_{0,025}t_{19} = -2,09$; $_{0,975}t_{19} = 2,09$)
b) ¿Cuál sería esta estimación si considerásemos un $\alpha = 0,10$? ($_{0,05}t_{19} = -1,73$; $_{0,95}t_{19} = 1,73$)

4.3. El intervalo de confianza de la proporción (π_{X_i})

• Si se han obtenido datos para una variable categórica *X* en una muestra de tamaño grande, el <u>*IC* del parámetro de la proporción</u> para una categoría *i* de esa variable (π_{Xi}) se obtiene según:

$$IC(1-\alpha)\left(\pi_{X_i}\right) = \left[p_{X_i} + z_{(\alpha/2)} \cdot \sqrt{\frac{p_{X_i} \cdot (1-p_{X_i})}{n}} \; ; \; p_{X_i} + z_{(1-\alpha/2)} \cdot \sqrt{\frac{p_{X_i} \cdot (1-p_{X_i})}{n}} \right]$$

Nótese que para la obtención del *EE* de la distribución muestral de la proporción se ha sustituido el valor del parámetro proporción (π_{Xi}) por el de la estimación de éste obtenida en la muestra (p_{Xi}).

• La consideración de <u>muestra de tamaño grande</u> se basa en los dos siguientes criterios:

$$n \cdot \pi_{Xi} \geq 5 \text{ y } n \cdot (1-\pi_{Xi}) \geq 5,$$

Ahora bien, dado que no se conoce el valor de π_{Xi}, se utilizan los límites del *IC* en el que se estima que está π_{Xi} y que deberemos haber calculado antes de poner a prueba estos criterios. Así, los criterios a satisfacer pasan a ser cuatro:

$$n \cdot L_{\inf}(IC) \geq 5;\ \ n \cdot L_{\sup}(IC) \geq 5;\ \ n \cdot (1 - L_{\inf}(IC)) \geq 5;\ \ n \cdot (1 - L_{\sup}(IC)) \geq 5$$

Ejemplo: para la obtención de un certificado de calidad en la producción, una empresa de fabricación de faros para coche debe demostrar que el nº de piezas defectuosas que produce y que pueden salir al mercado es inferior al 5%. Para ello se seleccionaron al azar 200 piezas de las fabricadas en la última semana y se obtiene que 14 de ellas presentan algún defecto de fabricación. ¿Entre qué valores se encontraría la proporción de piezas defectuosas entre todas las fabricadas la última semana? (considera α=0,05)

En esta muestra $p = 0{,}07$ y estimamos que el *EE* de la distribución muestral de la proporción obtenida en muestras de $n = 200$ es:

$$\sigma_{P_{Xi}} = \sqrt{\frac{0{,}07 \cdot 0{,}93}{200}} = 0{,}018$$

Por tanto, el *IC* del 95% es:

$$IC(0{,}95)(\pi) = \left[0{,}07 - 1{,}96 \cdot 0{,}018\ ;\ 0{,}07 + 1{,}96 \cdot 0{,}018\right] = \left[0{,}035; 0{,}105\right]$$

Comprobación del cumplimiento de los criterios de muestra grande:

$0{,}035 \cdot 200 = 7 \quad (\geq 5)$

$0{,}105 \cdot 200 = 21 \quad (\geq 5)$

$(1\text{-}0{,}035) = 0{,}965 \cdot 200 = 193 \quad (\geq 5)$

$(1\text{-}0{,}105) = 0{,}895 \cdot 200 = 179 \quad (\geq 5)$

Dado que se cumplen los cuatro, podemos considerar como adecuado el *IC* construido.

Ejemplo con *SPSS* a partir de los datos obtenidos con el Cuestionario de Vida Académica, se estimó con una confianza del 95%, la proporción de mujeres en la población de estudiantes de la Facultad de Psicología de la *UVEG*, sabiendo que en la muestra de $n = 174$ había 142 mujeres. Nota: la variable Sexo fue codificada como: 0, Hombre; 1, Mujer.

$$p_{mujer} = 142/174 = 0{,}816 \qquad EE(p_{mujer}) = \sqrt{\frac{0{,}816 \cdot 0{,}184}{174}} = 0{,}029$$

$$IC(0{,}95)(\pi_{mujer}) = 0{,}816 \pm 1{,}96 \cdot 0{,}029 = [0{,}76;\ 0{,}87]$$

(Al ser la muestra tan grande, los criterios de muestra grande se satisfacen sin duda)

Obsérvese la equivalencia con los resultados obtenidos con *SPSS* (El *IC* de la proporción se obtiene en *SPSS* igual que el *IC* de una media dado que la media de una variable dicotómica codificada con los valores 0 y 1 es igual a la proporción de casos en la categoría codificada con el valor 1).

SPSS: Analizar | Estadísticos descriptivos | Explorar:

Descriptivos

			Estadístico	Error típ.
sexo	Media		,816	,029
	Intervalo de confianza para la media al 95%	Límite inferior	,76	
		Límite superior	,87	
	Media recortada al 5%		,85	
	Mediana		1,00	
	Varianza		,151	
	Desv. típ.		,389	
	Mínimo		0	
	Máximo		1	
	Rango		1	
	Amplitud intercuartil		0	
	Asimetría		-1,646	,184
	Curtosis		,718	,366

¿Y cuál será el *IC* del 95% para la proporción de hombres?

$$IC(95\%)(\pi_{hombre}) = 0{,}184 \pm 1{,}96 \cdot 0{,}029 = [0{,}13;\ 0{,}24]$$

Al tratarse de una variable dicotómica, se podría haber obtenido como el complementario del *IC* obtenido para las mujeres:

$$IC(95\%)(\pi_{hombre}) = [1 - 0{,}87;\ 1 - 0{,}76] = [0{,}13;\ 0{,}24]$$

Ejercicio 10: A la misma muestra del ejercicio 6 (n = 40 estudiantes) se le preguntó si utilizaban alguna técnica de relajación, siendo 18 los que contestaron afirmativamente. Obtener el *IC* de la proporción de estudiantes que utilizan alguna técnica de relajación ($1 - \alpha = 0{,}95$).

Ejercicio 11: Seleccionamos una muestra aleatoria de 80 adolescentes con problemas relacionados con la alimentación y encontramos que un 60% tienen baja autoestima.

a) En la población de adolescentes con problemas relacionados con la alimentación, estima qué porcentaje tienen autoestima baja. Realiza la estimación basada en un *IC* del 99%.

b) Comprobar si se puede considerar que la distribución muestral de la proporción se ajusta a la distribución normal en este caso.

c) ¿Cuál sería el intervalo de confianza si el nivel de riesgo aumentara a un $\alpha = 0{,}10$?

d) ¿Cuál sería el intervalo de confianza obtenido si la muestra hubiera sido de 150 adolescentes y se considerase un $\alpha = 0{,}10$?

Referencias:

Losilla, J.M.; Navarro, J.B.; Palmer, A.; Rodrigo, M.F. y Ato, M. (2005). *Del contraste de hipótesis al modelado estadístico*. Girona: Documenta Universitaria.

Pardo, A., Ruiz, M.A. y San Martín, R. (2009). *Análisis de datos en ciencias sociales y de la salud I.* Madrid: Síntesis.

Wonnacott, T. H. y Wonnacott, R. J. (1990). *Introductory Statistics.* New York: Wiley.