T. 3 – Inferencia estadística: estimación de parámetros

1. La estimación de parámetros
2. La distribución muestral de un estadístico
3. Estimación por intervalos de confianza

• La inferencia estadística es un tipo de razonamiento que procede de lo concreto a lo general: intenta 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, 1998).

1. 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):

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

• 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.

• 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 la variable 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?

$$\hat{\mu}_X = ? \qquad Mo_X = ? \qquad \hat{\sigma}_X^2 = ? \qquad As_{Q_3-Q_1} = ?$$

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 de forma sinóptica:

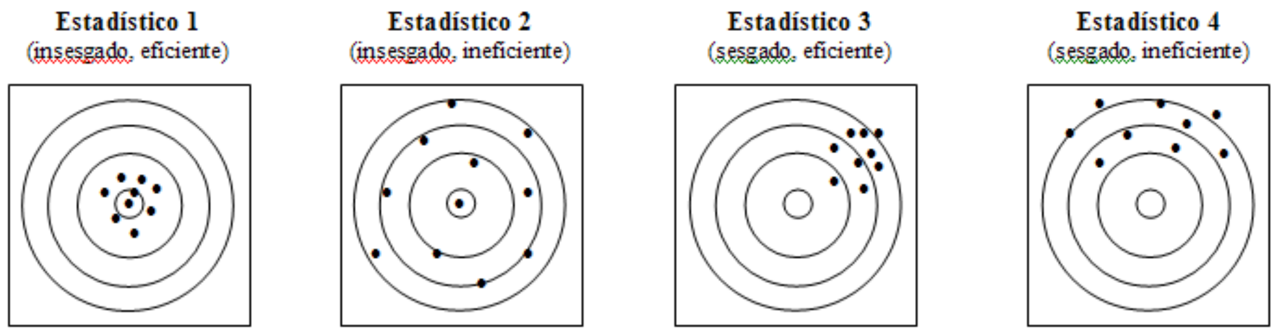
1) Ausencia de sesgo: Un estimador es insesgado cuando el promedio de las estimaciones obtenidas en diferentes muestras es, precisamente, el valor del parámetro que se pretende estimar.
2) Eficiencia: Esta es una propiedad que se establece en términos comparativos, esto es, es más eficiente aquel estimador cuyas estimaciones del verdadero valor del parámetro 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 <u>media</u> (μ_X), el mejor estimador es precisamente el promedio de los datos de la muestra, esto es, el índice estadístico de la media ($\bar{X}$):

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

Y, en general, los <u>mejores estimadores</u> 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:

$$\hat{Mo}_X \rightarrow Mo_X; \quad \hat{RIC}_X \rightarrow RIC_X; \quad \hat{Md}_X \rightarrow Md_X; \quad \hat{\pi}_{Xi} \rightarrow P_{Xi}; \quad \hat{\rho}_{XY} \rightarrow r_{XY} \ \ldots$$

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

- El mejor estimador del parámetro de la varianza (σ_X^2) no es el estadístico de la varianza (s_X^2) sino el de la <u>cuasi-varianza</u> (s'^2_X):

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

Ello es debido a que el índice estadístico de la varianza no cumple el requisito de ser un estimador insesgado del parámetro de la varianza, mientras que la cuasi-varianza sí -de ahí que a este índice estadístico también se le denomine en algunos textos como varianza insesgada.

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

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

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:

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

- Por último, 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}):

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

Otra igualdad que en algún caso nos puede resultar útil es la que relaciona los estadísticos de la covarianza y de la cuasi-covarianza:

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

Ejercicio 1: A partir de los siguientes datos para la variables "Edad" (X) y "Nº de ataques epilépticos durante el último año" (Y) en una muestra de jóvenes con diagnóstico de epilepsia, obtener una estimación de los parámetros de: (1) la media de "Edad"; (2) la mediana y la varianza de "Nº de ataques epilépticos"; (3) la covarianza y el coeficiente de correlación de Pearson entre ambas variables ($\hat{\mu}_X, \hat{Md}_Y, \hat{\sigma}^2_Y, \hat{\sigma}_{XY}, \hat{\rho}_{XY}$).

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

• A modo de resumen, los estimadores tratados 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.

2. La distribución muestral de un estadístico

• La estimación de un parámetro determinado (por ejemplo, la mediana de una determinada variable *X*) 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.

• 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, 1998), 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.

Ejemplo de la construcción empírica de la distribución muestral de un estadístico: en concreto, vamos a obtener las distribuciones muestrales de dos estadísticos, la media y la varianza, en ambos casos para muestras de tamaño 10 (n = 10). Sea el caso de la variable “Nº de horas de estudio al día” (*X*) y la población de referencia los estudiantes de la UVEG.

(Con fines didácticos, vamos a imaginar que desde el *más allá* nos llega una *revelación* estadística: la variable "Nº de horas de estudio al día" en la población de la UVEG se distribuye según la curva normal con $\mu_X = 5{,}63$ y $\sigma_X^2 = 3{,}7$ [$X \rightarrow N$ (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.)

- Obtener la distribución muestral de la media o la distribución muestral de la varianza 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:

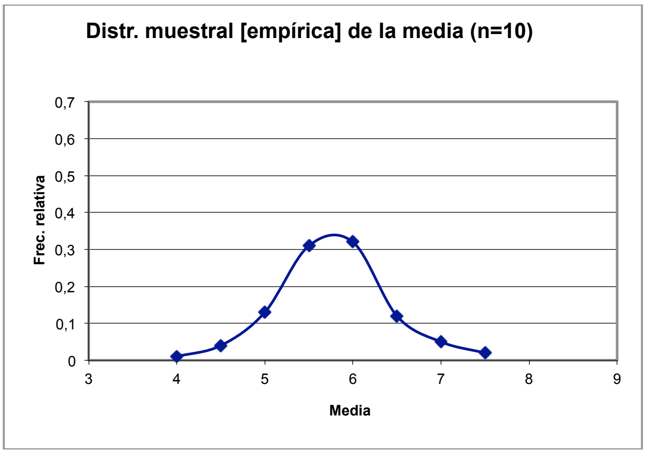
	Media ($\bar{X}$)*	*Varianza* (s_X^2)*
Muestra1	5,5	3,3
Muestra2	4,5	3,8
Muestra3	5	3,6
Muestra4	6,5	3,5
Muestra5	5	3,9
Muestra6	4,5	3,7
....	....	....
....	....	....
Muestra100	6	3,6

* Las medias están redondeadas con una precisión de 0,5 unidades y las varianzas de 0,1.

- 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. 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.

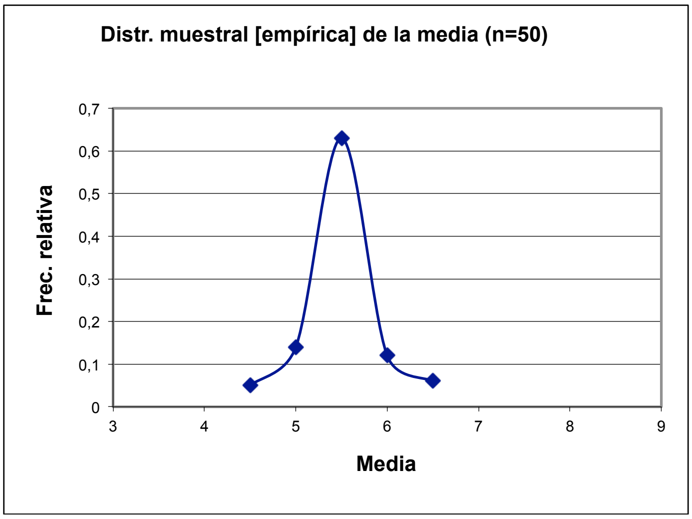
Distr. de frecuencias de la variable $\bar{X}$ (n = 10)

	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. Tras hacerlo se obtuvieron los siguientes resultados:

Distr. de frecuencias de la variable $\bar{X}$ (*n* = 50)		
	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

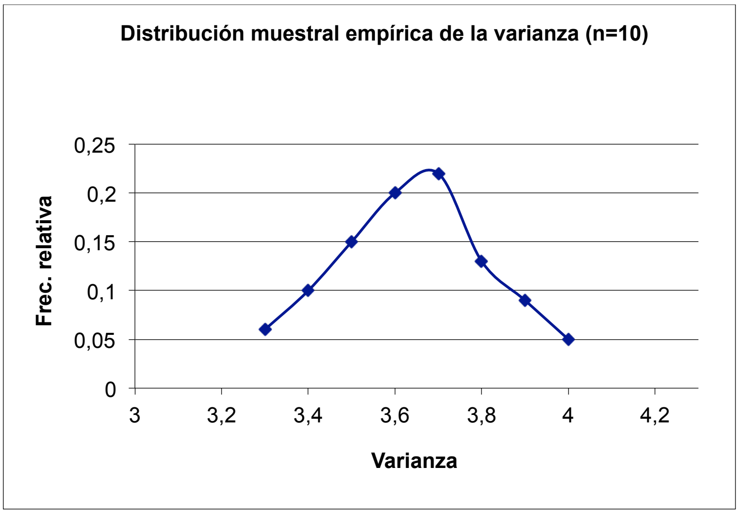


¿Qué ha cambiado al aumentar el tamaño de muestra?

- Por su parte, si en los datos recogidos con muestras de tamaño n = 10 nos centramos ahora en la columna de las varianzas y obtenemos la correspondiente distribución de frecuencias relativas, lo que obtendremos será la distribución muestral (estimada) del estadístico de la varianza para la variable X en muestras de tamaño n = 10.

Distr. de frecuencias de la variable "varianza"

	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 típica/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*).

• La última aporta un tipo de información de gran interés, pues cuanto menor sea el error estándar de estimación de la distribución muestral de un estadístico, ello supondrá mayor proximidad entre los valores obtenidos para 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 estadísticas que realicemos.

• Ahora bien, ¿ello significa que si queremos tener un indicador del grado de precisión de un determinado estadístico obtenido a partir de una muestra como estimación del parámetro poblacional, 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 son conocidas sus características principales (forma de la distribución, esperanza y error estándar) y, lo más importante, estas características se mantienen 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.

2.1. 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, de acuerdo al conocido como teorema central del límite, la distribución muestral de la media de X también tiende a distribuirse normalmente cuando ésta se obtiene con muestras de 30 o más casos ($n \geq 30$). La media y varianza de esta distribución muestral de la media es:
2. $\mu_{\bar{X}}\ [E(\bar{X})] = \mu_X$
3. $\sigma^2_{\bar{X}}\ [VAR(\bar{X})] = \dfrac{\sigma^2_X}{n} \rightarrow \qquad \sigma_{\bar{X}}\ [EE(\bar{X})] = \dfrac{\sigma_X}{\sqrt{n}}$

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, éste será menor: cuanto menor sea la varianza (o desviación típica) de la variable en la población; cuanto mayor sea el tamaño muestral que se considere.

• En nuestro ejemplo de la variable "Nº horas de estudio", la media de la distribución muestral del estadístico media en muestras de n = 10 es (de acuerdo a la revelación recibida):

$$\mu_{\bar{X}} \ [E(\bar{X})] = 5,63$$

Obsérvese, sin embargo, que si se calcula la media de la distribución muestral obtenida con 100 muestras de n = 10 se obtiene:

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

El resultado obtenido no coincide exactamente con el valor de la media de X en la población (μ_X= 5,63) debido que se ha obtenido a partir de una distribución muestral construida con un número finito de muestras y que es, por tanto, una aproximación a la distribución muestral verdadera del estadístico.

Ejercicio 2: Obtener la esperanza de la distribución muestral obtenida con 100 muestra de n = 50. ¿Coincide con el valor *revelado* de la esperanza de la distribución muestral de la media (5,63)?; ¿a qué puede ser debido?; ¿es más o menos próximo al valor verdadero 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 obtención del error estándar de la distribución muestral de la media en muestras de n = 10 y de n = 50 (teniendo en cuenta el valor de σ revelado):

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

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

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 media en la población.

• 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 <u>probabilidad</u> asociada a un rango de valores de media $\rightarrow$ 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:

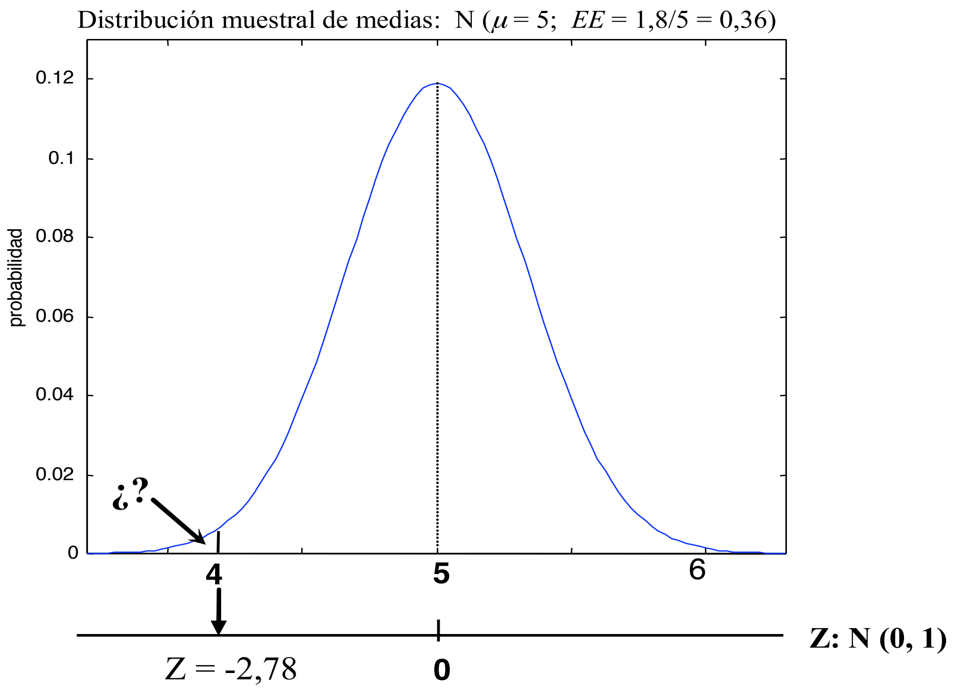
$$\mu_{\bar{X}} = \mu_X = 5 \qquad \text{y} \qquad \sigma_{\bar{X}}\ [EE(\bar{X})] = \frac{\sigma_X}{\sqrt{n}} = \frac{1{,}8}{\sqrt{25}} = 0{,}36$$

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 (μ_X) y la desviación típica (σ_X), ¿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 "$1\text{-}\alpha$")

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 ($1\text{-}\alpha = 0{,}90$), 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:

$$\mu_{\bar{X}} = \mu_X = 5 \quad \text{y} \quad \sigma^2_{\bar{X}}\ [VAR(\bar{X})] = \frac{\sigma^2_X}{n} \rightarrow \sigma_{\bar{X}}\ [EE(\bar{X})] = \frac{\sigma_X}{\sqrt{n}} = \frac{1{,}8}{\sqrt{100}} = 0{,}18$$

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:

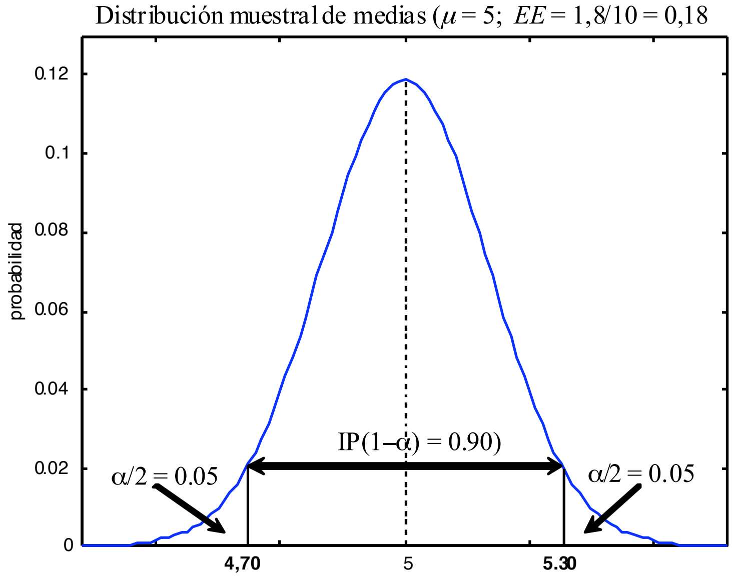
$$z_{0,05} = -1{,}64 \text{ y } z_{0,95} = 1{,}64,$$

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

$$-1{,}64 = \frac{\bar{X} - 5}{0{,}18} \rightarrow \bar{X} = 4{,}70$$

$$1{,}64 = \frac{\bar{X} - 5}{0{,}18} \rightarrow \bar{X} = 5{,}30$$

El proceso ilustrado gráficamente:



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

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

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

Así, para nuestro **ejemplo**:

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

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

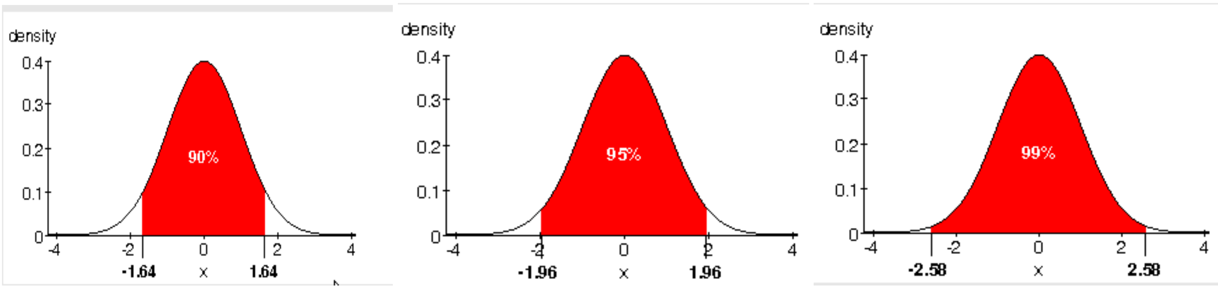
• Como ya se ha señalado, se utiliza la expresión (1-α) o nivel de confianza para hacer referencia a la probabilidad de que el intervalo que obtengamos contenga el valor de interés. En cuanto que probabilidad, $0 \leq (1\text{-}\alpha) \leq 1$, si bien, suele expresarse también como %.

• 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:

$Z_{(\alpha/2)}$	$Z_{(1-\alpha/2)}$	$(1-\alpha)$	α	$\alpha/2$
-1	1	0,68 [68%]	0,32 [32%]	0,16 [16%]
-1,64	**1,64**	**0,90 [90%]**	**0,10 [10%]**	**0,05 [5%]**
-1,96	**1,96**	**0,95 [95%]**	**0,05 [5%]**	**0,025 [2,5%]**
-2	2	0,954 [95,4%]	0,046 [4,6%]	0,023 [2,3%]
-2,58	**2,58**	**0,99 [99%]**	**0,01 [1%]**	**0,005 [0,5%]**
-3	3	0,9974 [99,74%]	0,0026 [0,26%]	0,0013[0,13%]

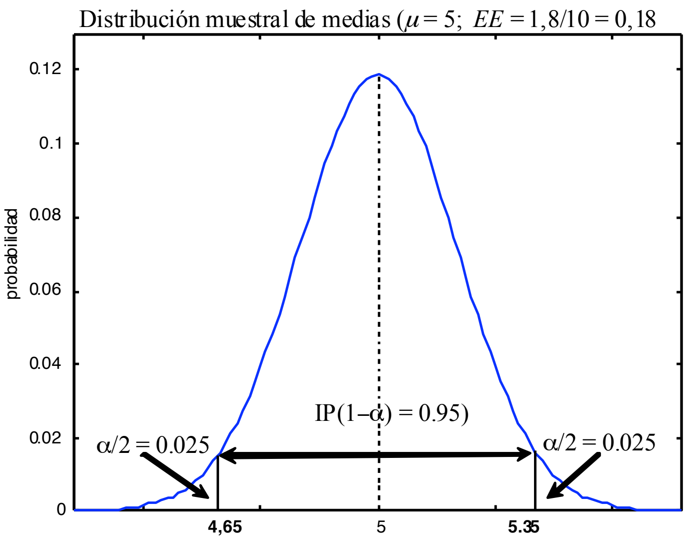
Los valores z correspondientes a los niveles de confianza/riesgo más utilizados en la práctica están subrayados en negrita en la tabla anterior y, a continuación, aparecen representados gráficamente.



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%, se obtiene:

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

Gráficamente:



2.1.2. 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.

2.2. Características de la distribución muestral de la proporción

1. Forma de la distribución: La de la distribución binomial, $B(n, \pi_{Xi})$, donde π_{Xi} 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: $n \cdot \pi_{Xi} \geq 5$ y $n \cdot (1-\pi_{Xi}) \geq 5$
2. $\mu_{P_{Xi}}$ $[E(p_{Xi})] = \pi_{Xi}$
3. $\sigma^2_{P_{Xi}}$ $[VAR(p_{Xi})] = \dfrac{\pi_{Xi} \cdot (1-\pi_{Xi})}{n}$ → $\sigma_{P_{Xi}}$ $[EE(p_{Xi})] = \sqrt{\dfrac{\pi_{Xi} \cdot (1-\pi_{Xi})}{n}}$

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% (π_{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 (π_{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:

$$20 \cdot 0,60 > 5 \text{ y } 20 \cdot 0,40 > 5$$

con parámetros:

$$\mu_{P_{Xi}}\ [E(p_{Xi})] = 0,60$$

$$\sigma_{p_{Xi}}\ [EE(p_{Xi})] = \sqrt{\frac{0,60 \cdot 0,40}{20}} = 0,11$$

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:

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

• Una aplicación fundamental (análoga a la de la distribución muestral de la $\bar{X}$) es que cuando, de acuerdo a la primera propiedad, se pueda considerar que la distribución muestral de la proporción sigue la curva normal, se puede aprovechar la tabla de la distribución normal estándar para contestar a diferentes preguntas de carácter aplicado. En caso contrario, habría que recurrir a la tabla de la distribución binomial. 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 de la UVEG en 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 que 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:

$$\mu_{p_{Xi}}\ [E(p_{Xi})] = 0,68; \qquad \sigma_{p_{Xi}}\ [EE(p_{Xi})] = \sqrt{\frac{0,68 \cdot 0,32}{20}} = 0,104$$

esto es, 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 => $z_{p_{Xi}} = \dfrac{p_{Xi} - \mu_{p_{Xi}}}{\sigma_{p_{Xi}}} = \dfrac{0,50 - 0,68}{0,104} = -1,73$

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:

$$\mu_{p_{Xi}}\ [E(p_{Xi})] = 0,68; \qquad \sigma_{p_{Xi}}\ [EE(p_{Xi})] = \sqrt{\frac{0,68 \cdot 0,32}{120}} = 0,043$$

esto es, N(0,68; 0,043)

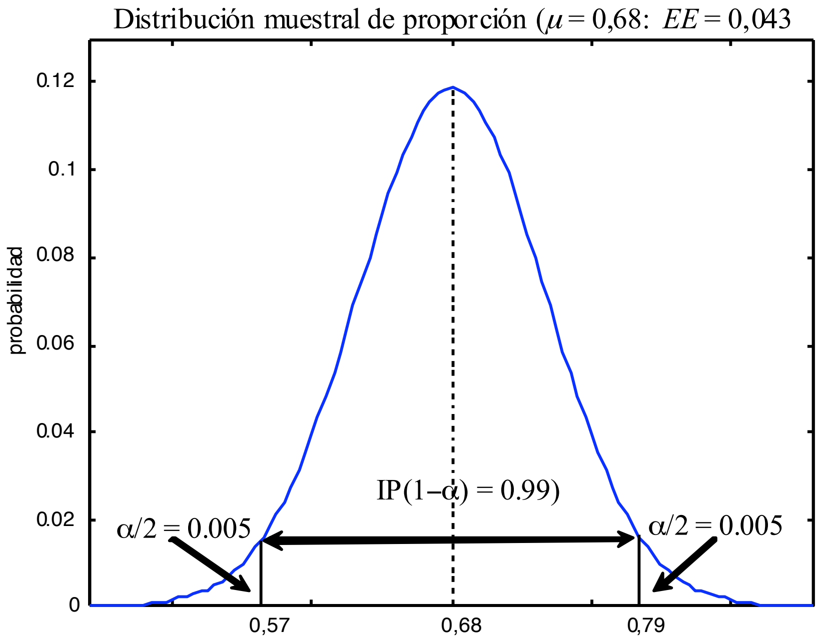
Utilizar la tabla de la curva normal estandarizada implica saber que los valores z que delimitan el intervalo de medias que nos interesa son: $z_{0,005} = -2,58$ y $z_{0,995} = 2,58$

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

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

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

Gráficamente:



Expresión formal de <u>cálculo del *IP*</u> 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:

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

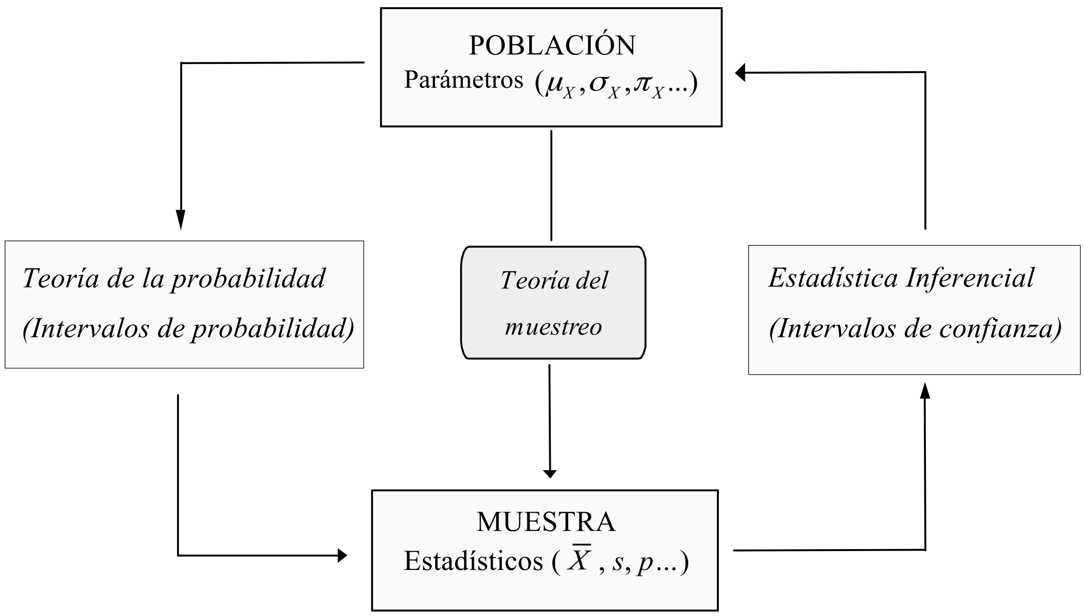
3. Estimación basada en intervalos de confianza

3.1. 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 vió como obtener los *IP* de la media y la proporción.

- La teoría estadística estudia de 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 (nivel de confianza), el verdadero valor del parámetro se encontrará en el intervalo construido.

• La <u>obtención</u> de los dos límites de un *IC* supone sumar y restar al estadístico obtenido en una muestra ($\hat{\theta}$) (estimación puntual del parámetro objeto de interés), un término de error que depende de: (1) el error estándar de la distribución muestral del estadístico en cuestión; (2) el nivel de confianza asumido en la definición del intervalo. Así, la expresión general del *IC* para un determinado parámetro θ es:

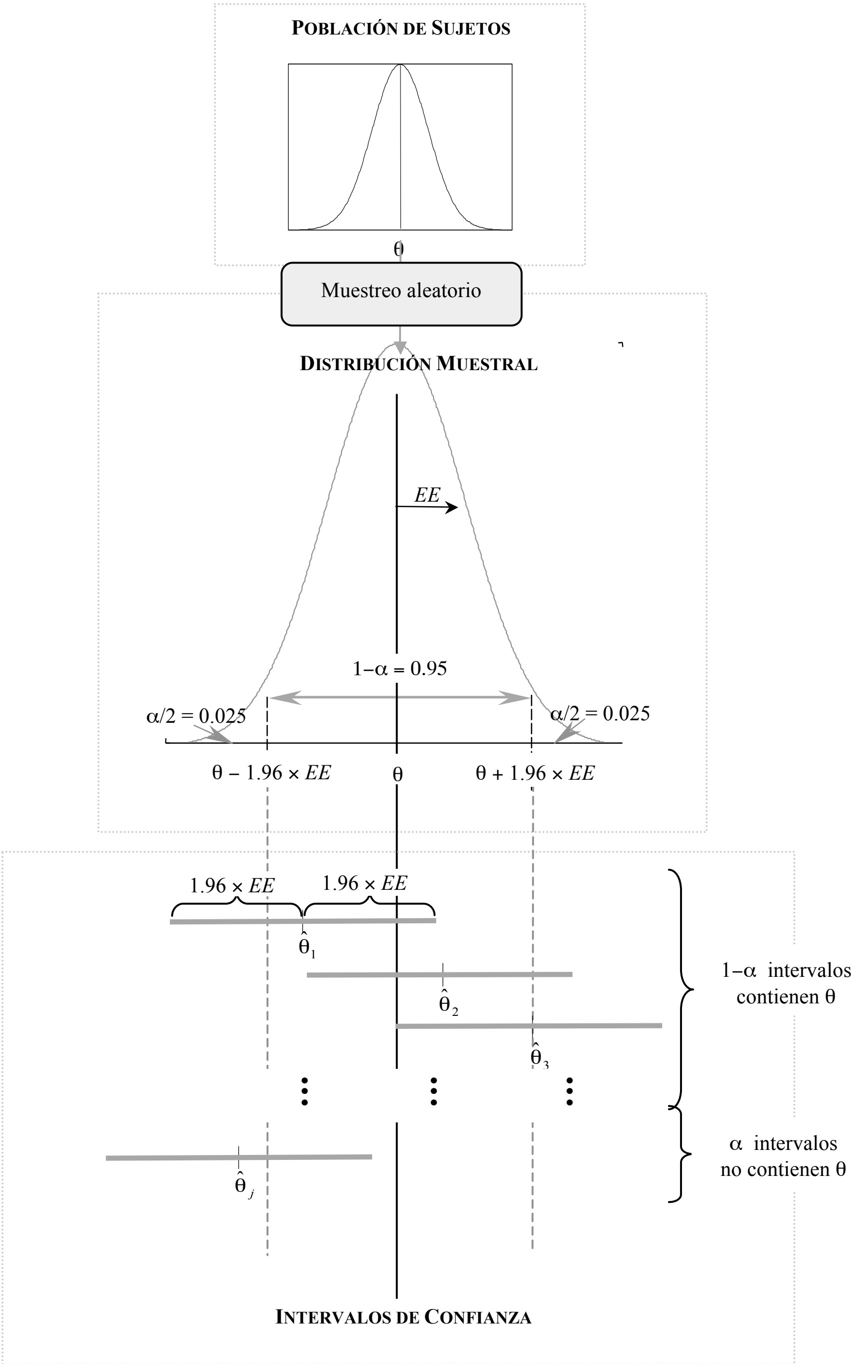
$$IC(1-\alpha)(\theta) = \left[\hat{\theta} + z_{(\alpha/2)} \cdot EE(\hat{\theta})\, ;\, \hat{\theta} + z_{(1-\alpha/2)} \cdot EE(\hat{\theta})\right]$$

Nótese que la expresión para el cálculo de un *IC* es la misma que la utilizada para el cálculo de un *IP* en la sección anterior, a excepción de que se sustituye el valor del parámetro por su estimación puntual en una muestra.

• El <u>nivel de confianza de un *IC*</u> no se ha de interpretar como la probabilidad de que un *IC* concreto contenga el valor del parámetro de interés, sino que la confianza se refiere al porcentaje de éxito del procedimiento de cálculo que se utiliza. Por ejemplo, si creamos un *IC* en que $(1-\alpha)$ es igual a 0,95 (o sea, $\alpha = 0,05$), ello supone que si calculamos un mismo *IC* en distintas muestras, un 95% de los *ICs* contendría el valor del parámetro estimado. Es incorrecto interpretar que un *IC* en concreto tiene una probabilidad de 0,95 de contener el valor del parámetro.

• Siguiendo a Wonnacott y Wonnacott (1991, p. 125-131), la siguiente figura contiene todos los elementos necesarios para la comprensión del mecanismo de construcción del intervalo de confianza de un parámetro θ, siguiendo la distribución muestral del estadístico $\hat{\theta}$ una ley Normal, y asumiendo un riesgo de error del 5% (Nota: esta figura será explicada en clase)

• Dado que el valor que se suma y resta al valor del estadístico obtenido en la muestra para obtener el *IC* es el mismo que el que se utilizaba para calcular el *IP*, la precisión del *IC* depende de los mismos factores que en aquel caso, a saber, del nivel de confianza elegido y del error estándar de la distribución muestral del estadístico.



Construcción de intervalos de confianza de un parámetro en base a la distribución muestral Normal (Losilla y cols., 2005; adaptada de Wonnacott y Wonnacott, 1991, p. 128).

3.2. Intervalo de confianza de la media (μ_X)

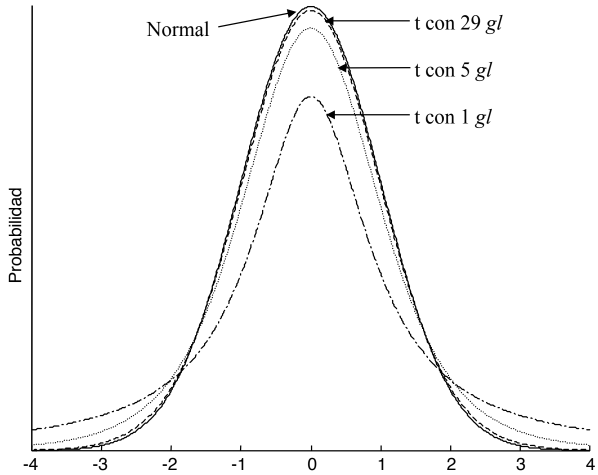
• Dada una muestra de la que se hayan obtenido datos para una variable X y en que se conozca la varianza de esa variable en la población (algo no habitual):

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

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

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

• A medida que se considera un mayor número de grados de libertad en la distribución t de Student, ésta converge con la distribución normal. Las diferencias son ya prácticamente inexistentes para la distribución t con 30 grados de libertad (véase la siguiente figura):



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:

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

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:

$$\sigma_{\bar{X}}\ \left[EE\left(\bar{X}\right)\right]=\frac{s'_X}{\sqrt{n}}=\frac{2,7}{\sqrt{1000}}=0,085$$

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

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

Por tanto, estimamos que la media poblacional se encontrará entre los valores 4,33 y 4,67 con una confianza del 95%.

Si se disminuye el riesgo de error a α=0,01, el *IC* del 99% sería más amplio (menos preciso):

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

Ejercicio 3: 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 ritmo cardiaco promedio 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 si la muestra hubiera sido de 20 sujetos?

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.

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

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

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 **Explorar** permite modificar el nivel de confianza con el que se crea el *IC*.

3.3. 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 *IC* del parámetro de la proporción para una categoría i de esa variable (π_{X_i}) 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 obtenida en la muestra (p_{Xi}).

• La consideración de tamaño grande se basa en el criterio $n\cdot\pi_{Xi}\geq 5$ y $n\cdot(1-\pi_{Xi})\geq 5$, si bien, dado que no se conoce π_{Xi}, se utilizan los límites del *IC* en el que se estima que está π_{Xi}. 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:

$$\sigma_{p_{Xi}}\,[EE(p_{Xi})] = \sqrt{\frac{0,07\cdot 0,93}{200}} = 0,018$$

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

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

Se cumplen los criterios de muestra grande: 0,035·200 = 7 (≥ 5) y 0,105·200 = 21 (≥ 5); y, por otra parte, (1-0,035)= 0,965·200 = 193 (≥ 5) y (1-0,105)= 0,895·200 = 179 (≥ 5)

Ejercicio 4: A la misma muestra del ejercicio 3 (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 con un nivel de confianza del 95%.

Ejemplo con SPSS a partir de los datos obtenidos con el Cuestionario de Vida Académica:
Estimar con una confianza del 95% la proporción de mujeres en la población de estudiantes de APDP 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 ± 1,96·0,029 = [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?

El complementario del *IC* obtenido para las mujeres: $IC(95\%)(\pi_{hombre}) = [1 - 0{,}87;\ 1 - 0{,}76]$

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

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

Referencias:

Losilla, J. M., Navarro, B., Palmer, A., Rodrigo, M. F., y Ato, M. (2005). *Del contraste de hipótesis al modelado estadístico.* Tarrasa: CBS (www.edicionsapeticio.com).

Pardo, A., y San Martín, R. (1998). *Análisis de datos en Psicología II (2ª ed.)* Madrid: Pirámide.

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