Estimación estadística

Vicente Manzano Arrondo – 2012-2014

Estimar qué va a ocurrir respecto a algo (o qué está ocurriendo, o qué ocurrió), a pesar de ser un elemento muy claramente estadístico, está muy enraizado en nuestra cotidianidad. Dentro de ello, además hacemos estimaciones dentro de un intervalo de posibilidades. Por ejemplo: “creo que terminaré la tarea en unos 5-6 días”.

Lo que hacemos en el terreno del análisis de datos es aplicar matizaciones técnicas a este hábito. Vamos a dedicar este documento al concepto de estimación, comenzando con la estimación puntual. Después nos ocuparemos de desarrollar un modelo de estimación por intervalo donde identificaremos los elementos fundamentales, con su significado y símbolo. Y, por último, habrá que desarrollar cómo se calculan esos elementos.

La estimación puntual

Estimar puede tener dos significados interesantes. Significa *querer* e *inferir*. Desde luego, el primer significado es más trascendente. Pero no tiene ningún peso en la estadística, disciplina que no se

ocupa de los asuntos del amor. El segundo significado es el importante aquí. Una estimación estadística es un proceso mediante el que establecemos qué valor debe tener un parámetro según deducciones que realizamos a partir de estadísticos. En otras palabras, estimar es establecer conclusiones sobre características poblacionales a partir de resultados muestrales.

Vamos a ver dos tipos de estimaciones: puntual y por intervalo. La segunda es la más *natural*. Y verás que forma parte habitual de nuestro imaginario como personas sin necesidad de una formación estadística. La primera, la estimación puntual, es la más sencilla y, por ese motivo, vamos a comenzar por ella. Ocurre, además, que la estimación por intervalo surge, poco más o menos, de construir un intervalo de posibles valores alrededor de la estimación puntual.

Una estimación puntual consiste en establecer un valor concreto (es decir, un *punto*) para el parámetro. El valor que escogemos para decir "el parámetro que nos preocupa vale X" es el que suministra un estadístico concreto. Como ese estadístico sirve para hacer esa estimación, en lugar de estadístico suele llamársele *estimador*. Así, por ejemplo, utilizamos el estadístico "media aritmética de la muestra" como estimador del parámetro "media aritmética de la población". Esto significa: si quieres conocer cuál es el valor de la media en la población,

estimaremos que es exactamente el mismo que en la muestra que hemos manejado.

Insesgadez

Del párrafo anterior podemos concluir erróneamente que todo parámetro se infiere a partir de un estadístico que resulta ser la misma fórmula o función pero calculado en la muestra. Si queremos estimar la media poblacional, le asignamos directamente la media de la muestra. Si queremos estimar la proporción poblacional, le asignamos el valor de la proporción en la muestra. Si queremos estimar la varianza poblacional, le asignamos el valor de la varianza de la muestra. Esa norma general tiene excepciones, por lo que es mejor no pensar en ella como norma. De los tres ejemplos, es cierto en los dos primeros casos: estimación puntual de una media o de una proporción; pero no en el tercero: estimación puntual de una varianza. La razón proviene del objetivo de la insesgadez.

Un sesgo es una tendencia constante. En un ejemplo clásico, solemos afirmar que las escopetas de feria están diseñadas para errar, para desviarse. Si esa desviación es fija, es decir, si esa desviación es una tendencia a errar hacia un sentido concreto, entonces hablamos de sesgo. Si no es fija, entonces se trata de una variación aleatoria. Observa la figura 1. El objetivo

es dar al centro de la diana. El área de disparos A muestra una variación aleatoria, pero sin sesgo pues apunta correctamente alrededor del objetivo. El área B muestra un sesgo claro: todos los disparos dan en un mismo punto y ese punto no es el centro de la diana, estamos errando. El área C ejemplifica una mezcla de ambos: existe sesgo y variación aleatoria, puesto que los disparos impactan en un área con cierta dispersión aleatoria pero concentradas en torno a un punto desplazado del objetivo.

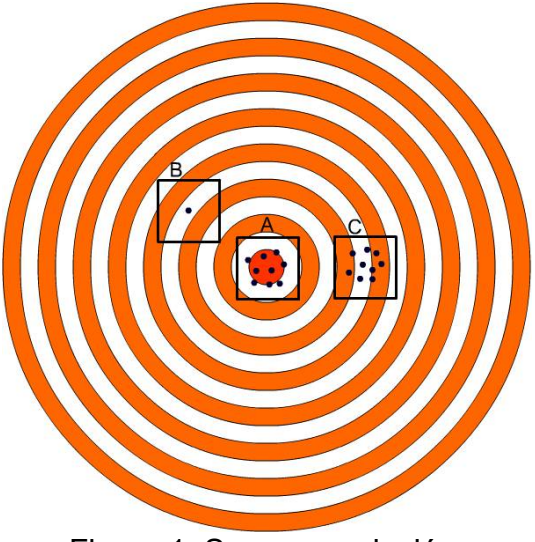

Figura 1. Sesgo y variación.

Los estimadores siempre suministran dispersión aleatoria. Como sabemos del monográfico sobre muestreo, el conjunto de todas las muestras de un mismo diseño que provienen de una misma población suministran valores diferentes. Esta circunstancia indica que existe una variación aleatoria con la que hay que vivir porque es inevitable. Pero todavía sería peor. Es posible que el estimador escogido tenga sesgo, es decir, que no solo esté variando alrededor de un punto, sino que el punto sobre el que varía no es el valor poblacional, verdadero u objetivo de nuestro interés. Esto si es evitable. Así que los estimadores que utilizamos intentamos que sean *insesgados*, es decir, que carezcan de sesgo.

El recurso que utilizamos para ello es el valor esperado, es decir, la media aritmética de la distribución muestral del estimador. Ya lo viste en el monográfico sobre muestreo. El valor esperado es, como dice la expresión, el valor que esperamos. Cabe elegir un estimador tal que el valor esperado coincida con el parámetro. Esto ocurre si utilizamos la media aritmética de la muestra como estimador de la media aritmética de la población, pues $E(\bar{X}) = \mu$. También ocurre con las proporciones, pues $E(p) = \pi$. Pero no ocurre así con la varianza (y, por tanto, tampoco con la desviación tipo) pues $E(S^2) \neq \sigma^2$. Esto ya lo hemos abordado en el monográfico sobre muestreo. Lo que hacemos entonces es escoge otro estimador. En el

muestreo aleatorio simple donde las poblaciones son de gran tamaño, es la cuasivarianza el estadístico escogido como estimador de la varianza poblacional, pues $E(\hat{S}^2) = \sigma^2$, es decir, la cuasivarianza es un estimador insesgado de la varianza poblacional.

Totales

Además de medias, proporciones y variaciones, un parámetro habitual es el total. Llamamos *total* a una frecuencia absoluta calculada en la población. Por ejemplo, podemos tener interés en conocer cuántas personas votarán al partido HH en las próximas elecciones o cuántos cigarrillos van a consumirse en el mes de abril. Para responder, utilizamos un recurso indirecto que parte de una estimación previa, bien sea de una media aritmética o de una proporción. Supongamos que la población que nos interesa cuenta con un millón de habitantes. Hemos trabajado con una muestra de 200. De los que 38 dicen que votarán al partido HH. Esto significa 38/200*100=19%. Una estimación puntual establece que el 19% de la población votará a HH. Como hay un millón de habitantes, entonces, hablamos de 1,000,000*19/100=190,000 personas. Supongamos también que se fuman 50 cigarrillos por término medio cada mes. Si ese es el valor de la media aritmética de la muestra, la estimación puntual afirmará que en la población se fumarán 50 cigarrillos por persona durante

el mes de abril, por término medio. Como hay un millón de habitantes, el mes de abril verá consumidos 50 millones de cigarrillos. Así pues, en la estimación de totales no realizamos un camino alternativo específico sino que ampliamos la estimación realizada previamente, sea de una proporción o de una media.

Estimación por intervalo

Las estimaciones puntuales no son una buena opción cuando constituyen el centro del objetivo, aunque solucionan problemas de procedimiento, por lo que son absolutamente necesarias.

Por qué estimar por intervalo

He comenzado prácticamente por el final. Intentemos comprender la afirmación del párrafo anterior. Por un lado, una estimación puntual es una mala opción. Que el parámetro tenga exactamente el valor del estimador es una casualidad de difícil ocurrencia. Queremos estimar el tiempo medio que una persona pasa entre una respiracion y la siguiente cuando duerme. Acotamos la población: nos preocupan los adultos (al menos 18 años de edad) europeos. Demasiados millones como para pensar que podemos abordar a toda la población. Así que seleccionamos una muestra aleatoria simple de 350 habitantes del

continente con 18 o más años. El tiempo medio en la muestra es de 5 segundos. Si hacemos una estimación puntual diremos que el tiempo medio en la población es también de 5 segundos. Imaginemos que somos capaces de conocer el valor real en la población. Es 5,2 segundos. ¿Hemos acertado?

¿Qué significa 5 segundos? En principio, son 5 segundos *exactamente*. Esto lo diferencia de, por ejemplo, 5,0013 o de 4,9987. Sin embargo, la gran mayoría de las personas seguramente aceptarían cualquiera de ambas aproximaciones como un acierto meritorio, pues solo se alejan de 5 en 13 diezmilésimas, una cantidad demasiado pequeña como para penalizar el estudio y afirmar que no acertó. Si nos comportamos de ese modo es que no estamos haciendo una estimación puntual, sino considerando un intervalo alrededor de 5 que marca la desviación admisible o una especie de cuantía máxima de error que nos permite afirmar que realmente se trata de un acierto. Demasiado enrevesado ¿no crees? Si la estimación puntual es utilizar un punto, no podemos estar utilizando un intervalo y seguir hablando de estimación puntual. Así pues, 5 segundos es un error, pues no coincide exactamente con el valor del parámetro, que es 5,2. Le daremos más o menos importancia, pero la estimación no acertó en el valor real. Con poco que pensemos sobre esto, la conclusión

es muy clara: en sentido estricto, las estimaciones puntuales yerran.

Lo que hacemos o deseamos hacer en la práctica son estimaciones por intervalo. Consiste en utilizar el célebre *más o menos*. Diremos, en nuestro ejemplo, que el tiempo medio que una persona dormida ocupa entre dos respiraciones es *más o menos* 5 segundos. No obstante, desde el campo de la estadística, ese *más o menos* es demasiado impreciso. Está incompleto. Es necesario responder a ¿más o menos qué? Hay que manejar alguna precisión. Por ejemplo: más o menos 0,4 segundos. Si la estimación es así, entonces estamos concluyendo con un intervalo: el tiempo medio es de $5 \pm 0{,}4 = \{4{,}6 ; 5{,}4\}$. Acabamos de ver nuestro primer ejemplo de estimación por intervalo.

Dos elementos en la estimación

En una estimación por intervalo podemos observar dos elementos: un centro y un radio o distancia al centro. En el ejemplo, el centro es 5 y el radio es 0,4. El centro es el valor aportado por el estimador. El radio expresa una medida de imprecisión. Cuanto menor es su valor, mayor es la precisión. Así que vamos a llamarlo coherentemente *error de precisión*, utilizando el símbolo e_p. En nuestro ejemplo, el estimador es la media aritmética con valor 5,

mientras que el error de precisión tiene el valor 0,4. Con ambos elementos podemos construir un intervalo. Antes de pasar al *tercer* elemento fundamental de una estimación por intervalo, retomemos la estimación puntual.

He iniciado este apartado afirmando que “Las estimaciones puntuales no son una buena opción cuando constituyen el centro del objetivo, aunque solucionan problemas de procedimiento, por lo que son absolutamente necesarias”. Ya has leído el razonamiento por el que la estimación puntual parece una mala opción. Sin embargo, llegará un momento, dentro de unas páginas, en el que tendremos que calcular el error de precisión. Es algo por lo que hay que pasar comprensiblemente antes de construir el intervalo, ya que este surge de sumar y restar el error de precisión sobre el valor del estimador. En el cálculo del error de precisión veremos que nos hace falta el valor de algún parámetro más. ¿Qué hacemos? Si la estimación por intervalo es la opción razonable, entonces pondremos en marcha un nuevo proceso, anidado en el anterior, donde necesitaremos construir un nuevo intervalo, es decir, calcular un nuevo error de precisión, es decir, encontrar el valor de un nuevo parámetro... y así sucesivamente. Esto debe tener un fin. El fin es la estimación puntual. En pocas palabras:

– cuando la estimación es un objetivo finalista, es decir un fin que deriva de los objetivos de la

investigación, entonces la llevamos a cabo por intervalo, pero

– cuando la estimación es un objetivo instrumental, es decir, una necesidad temporal que surge en el proceso de construcción de un intervalo, entonces la estimación será puntual.

Por ejemplo, para estimar la media de la población mediante un intervalo, el cálculo del error de precisión (como veremos) exige contar con el valor de la desviación tipo de la población. Nuestro objetivo no es encontrar ese valor, pero no tenemos más remedio que acotarlo de algún modo para seguir el proceso que realmente nos interesa. Entonces, para esta segunda necesidad, realizaremos una estimación puntual que, como hemos visto, consistirá en tomar el valor de la cuasidesviación tipo de la muestra.

El tercer elemento

Imagina que hacemos una apuesta. Apuesto contigo a que la siguiente persona que va a pasar por delante nuestra tiene 30 años.

En el contexto en el que nos encontramos, yo de ti aceptaría la apuesta. Acabo de arriesgar una estimación puntual, así que me equivocaré con seguridad. Si esa persona tiene, por ejemplo, 30 años, 9 meses y 17 días, en sentido estricto no son 30 años.

Así que para prevenir estos problemas, utilizaremos una estimación por intervalo. Mejor, me lo pienso. Y pensando no termino de decidir entre dos posibilidades:

A. Esa persona tiene entre 28 y 32 años.
B. Esa persona tiene entre 10 y 50 años.

¿Con cuál de las dos estimaciones por intervalo es más fácil acertar? Es obvio ¿verdad?, con la B. Cuanto más amplio sea el intervalo, es decir, cuanto mayor sea el valor del error de precisión, cabrán más resultados posibles y el acierto será más probable. Ganaré más fácilmente la apuesta si me decido por la versión B que no por la A. He aquí el tercer elemento fundamental de la estimación: la seguridad.

Cuanto más seguro quiera estar cuando hago una estimación, es decir, cuanto más difícil quiero que sea la probabilidad de equivocarme ¿qué hago? Una opción que parece clara es incrementar el error de precisión, es decir, aumentar el intervalo.

Así pues, contamos con tres elementos en una estimación por intervalo: el estimador, el error de precisión y la seguridad. El valor del estimador viene determinado por la muestra. No es algo que podamos decidir. Pero ¿y los otros dos? Uno está en función del otro, como hemos razonado. Lo que hacemos es decidir uno y calcular qué valor ha de tener el otro hasta encontrar un equilibrio. La figura 2 expresa esta idea. Los dos elementos se apoyan sobre las

características de la muestra, representadas por el valor del estimador. Conforme aumenta la seguridad disminuye la precisión. Conforme aumenta la precisión disminuye la seguridad. Esto del equilibrio es cosa difícil de explicar. Depende de varios factores, entre los que las consecuencias prácticas deberían constituir el factor protagonista. Pero, como ya veremos, otros criterios menos confesables, como la tradición, el hábito o el miedo a dar explicaciones, han generado otro tipo de soluciones. Abordaremos esto más adelante.

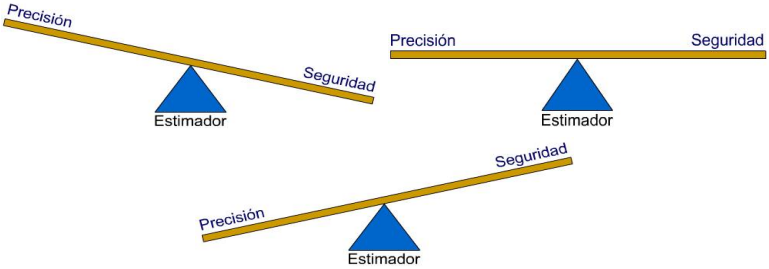

Figura 2. Equilibrio entre precisión y seguridad.

Vayamos por orden

Es posible seguir caminos alternativos, pero el orden más lógico en una estimación por intervalo viene a ser decidir la seguridad y, a partir de ella, calcular el valor que ha de tener el error de precisión,

construyendo acto seguido el intervalo. Si entramos en un esquema más sistemático, he aquí el proceso:

1. Decidir cuáles son los valores deseados para la seguridad y la precisión, siendo conscientes de que valores muy ambiciosos generarán situaciones muy exigentes.
2. Calcular el tamaño que ha de tener la muestra para conseguir esos objetivos iniciales de precisión y seguridad. Este asunto será abordado específicamente en otro monográfico (Tamaño de muestra).
3. Obtener la muestra.
4. Obtener el valor del estadístico utilizado como estimador, que suministra el punto central del intervalo.
5. Calcular el error de precisión a partir de la información de la muestra y de la seguridad deseada.
6. Construir el intervalo.

De todos estos puntos del esquema, el primero está repartido entre este monográfico y el siguiente, sobre el tamaño de la muestra. El segundo punto se cuentra íntegramente en ese otro monográfico. El tercero no nos compete en esta asignatura. El cuarto se encuentra ya abordado en el monográfico Conocer una variable. El quinto será objeto del siguiente

apartado. El sexto ya sabemos cómo hacerlo, si contamos con todo lo anterior. Antes de entrar en ese quinto punto, sobre cómo calcular el error de precisión, vamos a ocuparnos brevemente de reflexionar acerca de cuál debería ser la seguridad.

Qué hacer con la seguridad

Es obvio que queremos tener la máxima seguridad posible. Esto ocurre desde siempre, pero cada vez es más observable. Vivimos un momento histórico muy incierto, con varias dimensiones sujetas a cambios bruscos. Otra de las características de este momento es la intensidad con que se cultiva el miedo, un miedo pronunciado con multitud de objetos: la delincuencia, la inestabilidad laboral, el terrorismo, la gripe de las gallinas, de las vacas o de los cerdos, el tráfico, etc. Así que no es de extrañar que si preguntamos a cualquier persona cuánta seguridad quiere tener en una escala de 0 a 100, responda con 100.

No obstante, es pedir por pedir, además algo imposible. Imagina que decides tirarte con un paracaídas y que las empresas que los fabrican están obligadas a imprimir sobre la mochila que guarda la tela un valor de probabilidad. Es la probabilidad de que el paracaídas se abra cuando has saltado de avión y tiras de la anilla. ¿Qué probabilidad debería ser esa?

Posiblemente digas 1 (en tantos por uno) o 100 (en porcentaje). Vale. Te entiendo. Pero sabes que no es viable. Los paracaídas son inventos humanos. Los humanos somos seres muy entretenidos que, entre otras muchas características, nos dedicamos a la fabricación de productos no perfectos. Todo falla en algún momento y respecto a algún criterio. De vez en cuando un paracaídas no se abre. De vez en cuando, cuando alguien cruza una calle es atropellado, sea por un camión de veinte ruedas o por un niño en monopatín. Si fuera obligatorio que todas las casas en alquiler informaran no solo de las características de la vivienda, de su situación y del precio, sino también de la probabilidad de atropello en las inmediaciones, olvídate de aspirar a una casa cuyo valor de probabilidad de atropello sea 0. Eso no existe. Aún cuando marches en medio de una montaña, dentro de una hermita, nada garantiza que jamás te caerá encima un helicóptero de las fuerzas armadas que perdió el control en medio de unas maniobras. Todo puede ocurrir, aunque sea difícilmente.

La seguridad absoluta no existe. Y sabemos que cuanto más seguridad queramos tener, habremos a su vez que pagar un precio. Ese precio puede ser una baja precisión en las estimaciones, según hemos visto. O puede ser un tamaño de muestra tan grande que no tengamos tiempo, medios humanos ni dinero suficientes como para abordar a todas las unidades de

esa muestra gigante. Esto lo veremos en otro documento. Así que hay que tomar una decisión medianamente razonable, que no es escoger el 100% de seguridad.

En principio, el valor de la seguridad debería estar en íntima relación con las consecuencias de errar. Si afirmo que el paracaídas se abrirá, pero no lo hace, la consecuencia es que me muero. Si eso me parece grave, exigiré mucha seguridad. Imagina que aceptas el oficio de vendedor a domicilio. Vendes vajillas de cristal delicado, con olor a fresa e incrustaciones de pelo de gato común. No sé el precio al que vendes esa preciosidad, ni qué esperanza tienes de vender algo. Pero imagina un valor de probabilidad concreto. Me refiero a la probabilidad de que no vendas una vajilla cuando pulsas el timbre de una puerta. Tu deseo es vender y apuestas por ello. Pero puedes equivocarte. Entonces ¿qué seguridad quieres tener respecto a que vendes una de tus vajillas cuando tocas el timbre de una puerta? ¿100%? ¡Seguro que no! Tal vez aceptes el trabajo si la seguridad es del 10%, es decir, vendes una vajilla cada diez intentos. Eso, sinceramente, sería un exitazo. Observa que las probabilidades manejadas para el caso del paracaídas o de la venta de vajillas son muy diferentes. Lo son porque las consecuencias de equivocarse son también muy diferentes. Habría que situarse en estos términos a la hora de decidir un valor para la seguridad de acertar en una estimación.

No obstante, este discurso sigue sin suministrar un procedimiento para acotar un valor concreto de probabilidad, sino unos principos generales. Por otro lado, muchas ocasiones donde necesitamos hacer una estimación por intervalo no vienen acompañadas por una visibilidad clara de las consecuencias, por lo que estas reflexiones tienen una aplicabilidad al menos difícil.

La solución nos la dio Fisher, un astrónomo inglés que ideó recursos en estadística, genética (incluso eugenesia), matemáticas y física durante el primer tercio del siglo pasado. Sir Ronald Fisher estuvo pensando en estas cuestiones y se le ocurrió definir una situación concreta: imaginó a una viejecita que decía ser capaz de distinguir si en una taza de té con leche se había volcado antes el té o la leche. Y decía que era capaz de ello con solo probar un poco del líquido. Y el inglés siguió pensando. Dijo que había que acotar cuántos aciertos serían suficientes como para

creer a la viejecita. Si esta mujer fuera capaz de lo que decía, estaba claro que teníamos que aceptar algún error por su parte. Imagina que le damos 10 tazas. Que las acierte todas es algo que puede ocurrir en una ocasión de cada mil veces que sometemos a una persona a esa prueba y resulta que no tiene ni idea, acertando por casualidad. Que acierte casi todas, errando solo en una ocasión es algo que puede pasar en una de cada cien pruebas. Que falle en dos, ocurre en una de cada 25 ocasiones (un 4%). Fisher pensó en 8 tazas donde la mitad se han servido con un orden y la otra mitad con otro y pidió imaginariamente a la mujer que las distinguiera formando dos grupos. Y concluyó que un buen nivel de seguridad era el 95%, es decir, una probabilidad no superior al 5% de que la mujer acertara por casualidad. Y pensó que era una buena cosa ese valor al que había llegado.

Han pasado casi cien años y no hay quien discuta al señor Fisher ni aun después de muerto. No conozco a nadie que investigue sobre habilidad de mujeres mayores distinguiendo el orden en que se han volcado líquidos en un vaso. Pero lo cierto es que el 95% de seguridad se ha extendido tanto que si lo consideras en tus intervalos de estimación, nadie te preguntará nada. Por el contrario, si alguien decide utilizar un 93%, que se prepare a responder a la incómoda pregunta ¿Por qué?

En la práctica se han instalado dos valores para la seguridad. El estándar es 95%. No obstante, cuando alguien quiere mostrar mayor seguridad, más exigencia a la situación, entonces recurre al 99%. Y cualquier otro valor que no sean estos dos es algo muy difícil de encontrar leyendo trabajos de investigación publicados. Hay personas que se lo plantean, lo discuten, se quejan... pero no hay efecto en la práctica. Cohen, un psicólogo estadounidense muy implicado en estas cosas afirma, por ejemplo, que *dios quiere al 0,03 probablemente tanto como al 0,05*. No sé si es creyente, pero es una forma muy ilustrativa de poner en duda un valor que se ha establecido sin discusión.

Cálculo del error de precisión

Esquema simbólico de la estimación

Hemos visto ya que construir un intervalo de estimación tiene un momento de decisión (¿cuál ha de ser la seguridad con que concluimos que el parámetro, por ejemplo φ, se encuentra en el intervalo?) y dos momentos de cálculo. Lo primero que calculamos es el valor de estimador (por ejemplo, *m*). Lo siguiente es el valor del error de precisión (e_p) que se deriva de las características de la muestra y del nivel de seguridad escogido. Llegados a ese punto lo único que nos quedará será construir el intervalo restando y sumando

al estimador el valor del error de precisión. Con los símbolos improvisados en este párrafo, el punto de llegada será:

$$\phi \in \{m - e_p \ , \ m + e_p \}_{seg}$$

La expresión anterior se lee, literalmente, diciendo que afirmamos con una seguridad de valor *seg* que el parámetro φ tiene un valor comprendido entre m - e_p y m + e_p. Si llegados a este punto ya tenemos el valor de *m* y hemos decidido *seg*, entonces solo nos resta calcular e_p.

Elementos para el cálculo

Pues bien, el error de precisión depende de tres elementos. Uno ya lo sabemos: el nivel de seguridad. Cuanto más seguridad queramos tener, el error de precisión será mayor, de tal forma que el intervalo se hará más grande. Vamos a justificarlo más adelante, pero ya adelanto qué recurso vamos a utilizar para representar la seguridad: un valor de distancia estandarizada, Z_{seg}. Conforme mayor sea la seguridad, mayor será el valor de Z_{seg}. Si este elemento fuera el único y dado que a mayor seguridad mayor error de precisión, entonces la fórmula de cálculo sería:

$$e_p = Z_{seg}$$

Pero no ocurre así. La fórmula está incompleta. Faltan elementos. El siguiente que abordamos es la dispersión de la característica que se está midiendo.

Uno de los ejemplos que hemos abordado es el número de cigarrillos consumidos en un mes. En una población donde todo el mundo fuma lo mismo, la estimación tendrá dos características muy apetecibles. Por un lado, será una estimación puntual. No hace falta construir un intervalo alrededor de un valor que sabemos que no varía. Si todo el mundo fuma exactamente 3 cigarrillos, entonces la media es 3 cigarrillos y la varianza es 0. No hay variación, se mida como se mida. Así que cuando obtengamos una muestra, pongamos de 50 personas y encontremos que las cincuenta sin excepción fuman tres cigarrillos al mes, concluiremos que todas las personas de la población fuman exactamente tres cigarrillos al mes, con un error de precisión de valor 0. Es decir, sin intervalo. Es más, no solo se trata de una estimación puntual, sino además de una estimación segura. No hay posibilidad de equivocarse. La seguridad es del 100%.

Ahora imaginemos que ocurre algo bien distinto. En esa población hay mucha gente que no fuma absolutamente nada, mucha gente que fuma una cantidad tal de cigarrillos que resulta increíble, y el resto, que siguen siendo mucha gente, fuma

cantidades muy variadas. Pues bien, una muestra de también cincuenta personas que provienen de esa población arrojará cantidades muy diferentes, bastante diferentes para la variable *número de cigarrillos fumados al mes*. Esta circunstancia favorecerá que, para mantener un nivel de seguridad aceptable, el error de precisión tenga que ser muy elevado, de tal forma que el intervalo de estimación sea suficientemente amplio como para incluir esa variabilidad.

La medida que hemos escogido para indicar la variación de una característica que se expresa en una escala cuantitativa es la desviación tipo. Como hablamos de cómo varía esa característica en la población, se trata entonces de la desviación tipo poblacional. Estamos razonando, pues, que conforme mayor es la desviación tipo poblacional, mayor ha de ser también el valor del error de precisión. Añadiendo esta conclusión a la fórmula, tenemos una segunda versión, más completa:

$$e_p = Z_{seg}\ \sigma$$

El tercer y último elemento es el tamaño de la muestra. ¿Qué influencia crees que tiene en el valor del error de precisión? Vayamos por partes. Primera pregunta: ¿qué efecto tiene el tamaño de la muestra en el parecido entre el estimador y el parámetro? Al aumentar el tamaño de la muestra, si el muestreo sigue

siendo el mismo y es bueno, entonces esperaremos que el valor del estimador sea más parecido al parámetro, es decir, nos resultará más raro que la distancia entre el estimador y el parámetro sea elevada. Si hemos entrevistado a una sola persona, es muy probable que su opinión sobre la actuación del ayuntamiento en el asunto de la construcción del puente varíe mucho respecto a entrevistar a otra persona. Y, por tanto, es muy probable que su opinión resulte poco representativa. Pero si hemos preguntado a mil personas mediante un buen cuestionario y la muestra está conseguida mediante un buen procedimiento, entonces esperaremos dar prácticamente en el blanco, en el sentido de que tendremos una sensación bien fundada de que el valor medio de la opinión en la muestra (valor del estimador) será igual o muy parecido al valor medio de la opinión en la población (valor del parámetro). Así pues, conforme el tamaño de la muestra aumenta, el error de precisión ha de disminuir. La fórmula debe considerar, entonces, al tamaño de la muestra dividiendo. Cuando se deduce matemáticamente la expresión, lo que divide no es *n* directamente sino su raíz cuadrada. Por ello, la fórmula definitiva que utilizamos es:

$$e_p = Z_{seg} \frac{\sigma}{\sqrt{n}}$$

Te recuerdo que estamos en un muestreo aleatorio simple aplicado sobre una población de un tamaño muy grande. Si no es así, si la población es pequeña o el muestro tiene otro diseño, entonces la expresión de cálculo sería diferente. No obstante, por diversas razones, incluso cuando se acude a otros modelos de muestreo, sigue siendo habitual utilizar esta expresión.

Seguridad y distribución muestral

Antes de seguir, si no has leído el monográfico sobre muestreo, en el que se habla de la distribución muestral, este es el momento. Si continúas sin haberlo leído es muy posible que te enteres de poco.

Me preocupa conocer durante cuántas horas una persona es capaz de permanecer en un centro comercial abierto 24 horas. He seleccionado una muestra al azar y he preguntado por su capacidad de permanencia en el centro comercial. El resultado es 8 horas por término medio. ¿Y ahora qué? Lo que me interesa no es la muestra, sino la población.

La figura 2 muestra una distribución muestral hipotética. Cada resultado muestral está representado por un bloque o ladrillo. Cada bloque es la media aritmética del número de horas que una muestra aleatoria de personas dice que sería capaz de permanecer en un centro comercial. Muchas muestras

han suministrado el mismo valor, pues sus ladrillos se apilan sobre el mismo punto, formando una columna. En el eje horizontal figura la diferencia entre el valor del estimador en esa muestra y el valor real en la población. Así, por ejemplo, hay 6 muestras en las que se ha obtenido un valor del estimador inferior al parámetro en 5 unidades (la media de esas muestras es 5 horas de permanencia menos que la media de la población), o también hay 11 muestras con valores del estimador que superan al parámetro en 4 unidades.

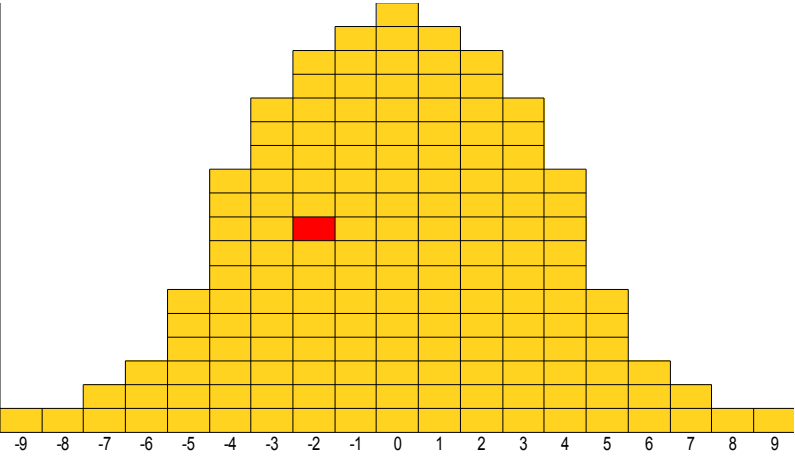

Figura 2. Distribución muestral.

Pues bien, de las 160 muestras que construyen esa distribución muestral, una de ellas es la mía. Tal vez sea la que he marcado con el color rojo. Tal vez

sea alguna de las 159 restantes. Haga lo que haga, no tengo una respuesta precisa a la pregunta ¿dónde está mi muestra? Pero tengo otro tipo de respuesta: puedo hacer una apuesta. Por ejemplo, puedo plantearme cuál es la probabilidad de que mi muestra sea exactamente esa que he marcado en rojo. Si hay 160 posibilidades y he escogido solo una, la probabilidad es muy baja: 1/160 = 0,00625. Es más, esta operación carece de sentido en una distribución muestral más real, puesto que sabemos que el número de muestras es prácticamente infinito por lo que la probabilidad de ocurrencia de una cualquiera de ellas es cero. Una forma de solucionar esto es plantearme la probabilidad de que mi muestra sea una de las que forman un conjunto amplio de resultados muestrales, un intervalo de resultados posibles.

En la figura 3 he marcado un área central alrededor del valor esperado del estimador que, como sabemos, coincide con el parámetro. El área reúne 112 muestras, un 70% de las 160. Se trata de todas las muestras que suministran valores del estimador que se alejan del parámetro en no más de 3 unidades, sea por abajo o por encima. La figura 4 representa una situación similar, pero acotando un área del 95% que afecta a 152 muestras, con una distancia máxima al parámetro de 6 unidades. Observa que, como resulta obvio, cuanto más amplío la superficie de la gráfica, es decir, el porcentaje de muestras posibles consideradas,

también se amplía el intervalo de distancias al parámetro.

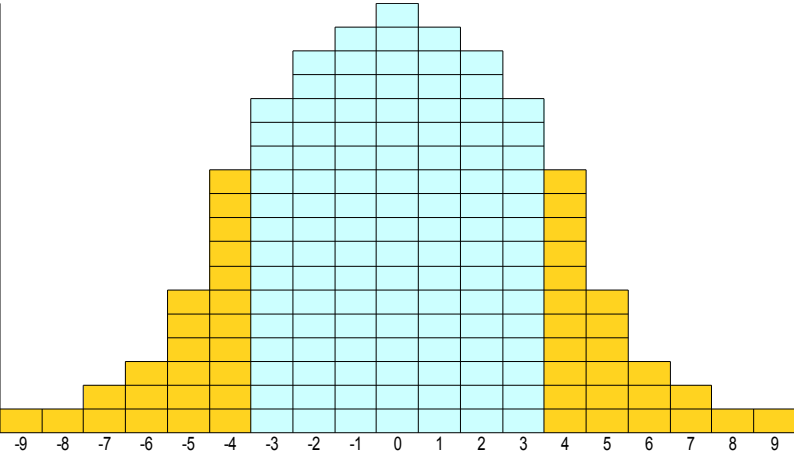

Figura 3. Área centrada del 70%.

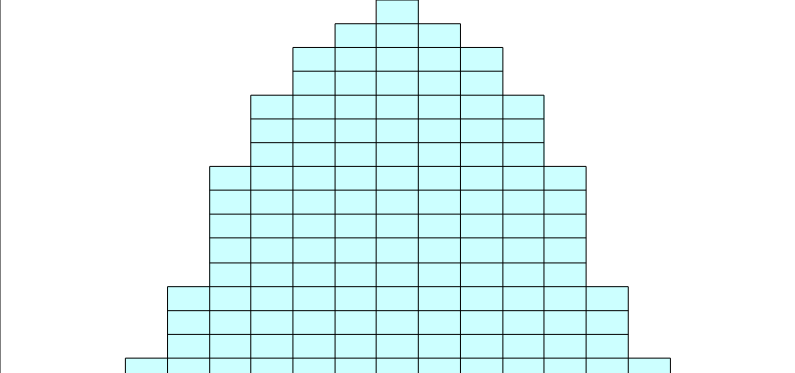

Figura 4. Área centrada del 95%.

Sé que mi muestra es solo una de entre todas las posibles que puedo obtener de la población. Y apuesto a que es una de las del área centrada del 70%. No tengo ninguna información que me haga suponer que deba estar necesariamente ahí. Es una apuesta. Pero hay algo que sé y es que hay una probabilidad del 70% de que gane la apuesta. Ocurre como con cualquier otra situación donde hay juegos de azar: si apuesto a que va a salir un número par al lanzar un dado, como la mitad de sus caras tienen número par, entonces la probabilidad de acertar es la proporción de caras: 0,50 o 50%. Como hay un 70% de muestras en el área marcada de la figura 3, entonces hay una probabilidad del 70% de que mi muestra sea una de

ellas, así como una probabilidad del 30% de que sea una de las que se encuentran fuera de ese área. Dado que ese porcentaje está definido por los estimadores cuyo valor se aleja del parámetro en no más de 3 unidades, entonces afirmo lo siguiente:

> *La distancia que separa lo encontrado en mi muestra respecto a lo que ocurre en la población no es superior a tres unidades. Y hago esta afirmación con una seguridad del 70% (o bien con un riesgo de equivocarme del 30%).*

Observa que acabo de hacer una estimación por intervalo. Si la distancia entre el estimador y el parámetro es no superior a 3 (afirmación que tiene un riesgo del 30% de ser errónea) y en el caso de mi muestra el valor que he encontrado es de 8 horas de permanencia en el centro comercial, entonces, espero (con una seguridad del 70%) que las personas de esa población son capaces de permanecer en un supermercado entre 5 y 11 horas. En términos simbólicos y siguiendo lo que vimos en el apartado anterior:

$$\mu \in \{ \bar{X} \pm e_p \}_{seg} \quad \Rightarrow \quad \mu \in \{ 8 \pm 3 \}_{0,70} = \{ 5 \; ; \; 11 \}_{0,70}$$

Si utilizo el intervalo del 95%, entonces la estimación varía, el intervalo de estimación se hace más grande:

$$\mu \in \{\bar{X} \pm e_p\}_{seg} \quad \Rightarrow \quad \mu \in \{8 \pm 6\}_{0,70} = \{2\ ;\ 14\}_{0,70}$$

Así pues, esa *máxima distancia que cabe esperar entre el estimador y el parámetro* (3 unidades al 70% o 6 al 95%) es otra definición para lo que hemos llamado *error de precisión*. Hemos visto que e_p es lo que restamos y sumamos al estimador para conseguir un intervalo en donde esperamos, con cierta seguridad, que se encuentre el parámetro. Pues aquí lo tenemos. Lo sumamos y lo restamos porque creemos que esa distancia no va a ser superada; una creencia apoyada en una seguridad que en el ejemplo es del 70%.

Para ejemplificar este proceso y ver con claridad su significado he optado por inventar unos datos muestrales y, con ello, una distribución muestral que realmente no existe. Lo que nos queda para cerrar satisfactoriamente este asunto es superar este ejemplo concreto. En él hemos traducido 70% a una distancia de 3 unidades o 95% a una de 6 porque me he inventado la distribución muestral. ¿Qué ocurre si no tengo acceso a la distribución muestral completa? Entonces, para no andarnos con rodeos, tenemos un problema. Pero no hay que lamentarse mucho. Hay

solución, no perfecta pero solución. Es esta: si tenemos indicios de que la distribución muestral tiene una forma concreta y conocida, entonces podemos traducir la seguridad o el error de precisión siguiendo las características de la distribución.

Insisto: lo que necesitamos para traducir una seguridad en un error de precisión es conocer cuál es el modelo de la distribución muestral. En nuestro ejemplo ha sido fácil porque teníamos la distribución completa de las 160 medias aritméticas. Pero cuando esto no ocurra (que es lo que pasa en la práctica), lo que necesito en lugar de todos los resultados muestrales es qué tipo de horma, ley o referencia conocida sigue la distribución muestral a la que pertenece mi muestra. Conociendo esa ley, puedo realizar cálculos en los que paso de distancias o errores de precisión a áreas y viceversa.

Antes de seguir, si no has leído el monográfico sobre curva normal, ahora es el momento, pues vamos a recurrir a ella para terminar este proceso.

La importancia de la curva normal (y su familia)

En el asunto de la curva normal vimos que se trata de una forma muy conocida a la que hacemos referencia como distribución normal, ley normal o modelo normal. Pues bien, sabemos que en determinadas circunstancias las distribuciones

muestrales siguen una ley normal. Si la distribución muestral a la que pertenece mi muestra es normal, entonces puedo hacer lo mismo que hemos hecho en el ejemplo anterior: traducir la seguridad a una distancia o error de precisión.

Como vimos, la curva normal se define por dos únicas características: media y desviación tipo. Por tanto, necesitamos esta informacion para hacer las traducciones.

La normal estandarizada

En nuestro ejemplo, partimos de una muestra con una media aritmética de valor 8. Tal vez la desviación tipo sea S=2. Para traducir una seguridad del 95% (gracias, Fisher) en un error de precisión, necesitamos una fórmula o una tabla que considere esos valores. Para otra ocasión, tal vez con una media de valor 123 y una desviación de 18, necesitaremos otra traducción específica. Esto, como se ve, no es muy operativo. En lugar de ello, tenemos otro recurso: utilizar una única curva normal, la estandarizada.

Del monográfico *Cómo interpretar un caso* conocemos las puntuaciones tipo, típicas, estándar, distancias estandarizadas o Z, expresiones que apuntan al mismo concepto. Sabemos que su media siempre es 0 y su desviación tipo siempre es 1. Luego, si no utilizamos las puntuaciones directas (como el

número de horas de permanencia en el centro comercial o el número de cigarrillos consumidos en un mes) sino que las estandarizamos, nos basta con una única curva normal, la de media 0 y desviación tipo 1, es decir, la normal estandarizada.

Una vez estandarizada, todo es más sencillo. Del monográfico *La curva normal* sabemos, por ejemplo, que el 95% de las distancias estandarizadas se encuentran entre los valores -1,96 y 1,96. Si queremos hacer una estimación por intervalo con una seguridad del 95%, podremos concluir siempre en términos de distancias estandarizadas afirmando que el error de precisión (estandarizado) es 1,96 o que el parámetro se encontrará entre -1,96 y 1,96 en puntuaciones tipo. Esto es fácil, pero psicológicamente poco recomendable. Es bueno que las conclusiones utilicen la misma escala de medida que la variable original. Así que una vez traducida la seguridad (área centrada en la curva normal) en una puntuación tipo, lo siguiente será una nueva traducción: pasar las puntuación tipo a una puntuación directa. Esto es sencillo, despejando de la expresión de la distancia estandarizada. Como sabemos una distancia estandarizada es la distancia del valor original respecto a la media de su conjunto de datos, expresada en número de desviaciones tipo de su conjunto de datos. Expresada para los contextos de una muestra, una

población y una distribución muestral de medias, respectivamente:

$$Z_i = \frac{X_i - \bar{X}}{S} \quad \Rightarrow \quad Z_i = \frac{X_i - \mu}{\sigma} \quad \Rightarrow \quad Z_i = \frac{\bar{X}_i - \mu}{\sigma_{\bar{X}}}$$

El error de precisión es la máxima distancia que cabe esperar entre el estimador y el parámetro, es decir, $e_p = máx\{\bar{X}_i - \mu\}$. Luego:

$$Z_i = \frac{\bar{X}_i - \mu}{\sigma_{\bar{X}}} = \frac{e_p}{\sigma_{\bar{X}}} \quad \Rightarrow \quad e_p = Z\sigma_{\bar{X}} = Z\frac{\sigma}{\sqrt{n}}$$

La última deducción surge del valor del error tipo de la media (la desviación tipo de la distribución muestral de medias) que conoces del monográfico sobre muestreo, de donde sabemos que el error tipo es la desviación tipo entre la raíz cuadrada del tamaño de la muestra. Observa que hemos llegado al mismo punto que razonamos hace unas pocas páginas: el error tipo está en función directa de la seguridad expresada mediante una distancia estandarizada y la variación de la característica expresada como la desviación tipo de la población, y en función inversa del tamaño de la muestra expresado como su raíz cuadrada. Hemos llegado al mismo punto por dos caminos. En este segundo, además, hemos comprendido de dónde surge

Z_{seg}: es el error de precisión estandarizado que expresa la seguridad de la estimación en la distribución muestral estandarizada y que puede ser traducido a la escala de las puntuaciones directas gracias a que contamos con un modelo de probabilidad para la distribución muestral, la curva normal.

¿Mi distribución muestral es normal?

Del monográfico sobre la curva normal recuerda que De Moivre se dio cuenta de que las distribuciones de algunas variables se aproximaban a una curva normal conforme aumentaba el número de ensayos, es decir, conforme *n* se hace más grande. Así que si las muestras son *suficientemente* grandes, la distribución será al menos operativamente normal. Pero ¿cuánto de grandes? La respuesta depende del estimador.

Se han hecho cálculos y simulaciones para muchas situaciones. Esas operaciones muestran que la distribución muestral de medias es satisfactoriamente normal cuando el tamaño es al menos de 30 unidades ($n \geq 30$). Si la distribución muestral es de varianzas, no hay mucho acuerdo salvo en que las muestras han de ser sensiblemente más grandes (en torno a $n \geq 1000$). Hay otra situación donde podemos suponer que la distribución muestral es normal sin necesidad de atender al tamaño de las muestras: podemos tener la seguridad de que una distribución muestral es normal si

también lo es la distribución poblacional. En otras palabras, si una variable se distribuye según una ley normal en la población, todas las distribuciones muestrales de medias serán también normales. Esto es lógico. Piensa, por ejemplo, en el caso más extremo, muestras de tamaño n=1, y puedes entender que se trata de una distribución muestral normal puesto que coincide con la población y esta lo es.

En el contexto de esta asignatura nos interesa únicamente la estimación por intervalo de medias, de proporciones y de totales. Como los totales los deducimos a partir de medias y proporciones, solo nos quedan estas. Luego ¿cuándo una distribución muestral de proporciones es normal? La respuesta depende del valor de la proporción. Lo lamento. Pero es bastante lógico.

Una proporción tiene valores comprendidos entre 0 y 1. Si la proporción poblacional (el parámetro) es 0,5 entonces está perfectamente centrada. Esto permite a la distribución muestral moverse en 0,5 unidades tanto hacia abajo como hacia arriba. Si el parámetro es 0,3 entonces solo quedan tres décimas de libertad de movimiento para las proporciones muestrales entre el valor esperado y el límite inferior. Es verdad que quedan siete décimas hasta 1, pero son inútiles. Si la distribución muestral de proporciones se dispersa 3 décimas por debajo y 7 por arriba, entonces no es simétrica. Y, como ya sabes, si no es simétrica,

no es normal. Así que conforme más extrema (más cerca de un extremo) sea la proporción poblacional, más difícil se lo pone a la distribución muestral de proporciones para ser normal, pues ha de conseguir mucha menor dispersión de los valores posibles para las proporciones de las muestras. Luego, la condición para que una distribución muestral de proporciones sea normal es una combinación de dos condiciones: un valor no extremo y una muestra grande. En la medida en que una de las dos condiciones se cumpla menos, la otra deberá compensar cumpliéndose más. Por ejemplo, si la proporción es muy extrema, la muestra deberá ser muy grande. Si la muestra es pequeña, la proporción deberá estar muy centrada. La combinación de ambos argumentos se concreta (tras cálculos y simulaciones) en las siguientes dos condiciones:

$$n\pi \geq 5 \; ; \; n(1-\pi) \geq 5$$

Es decir, tanto n por la proporción como n por el complementario de la proporción deben cumplir que suministren un resultado no inferior a 5. Comprueba que si la proporción es centrada ($\pi = 0,5$), basta con una muestra de tamaño n=10 para que se cumpla el supuesto de normalidad, algo muy inferior a la condición sobre la distribución muestral de medias. Pero si la proporción es extrema ($\pi = 0,1$; o bien $\pi = 0,9$, por ejemplo), entonces las exigencias sobre el

tamaño de la muestra son superiores al caso de la distribución muestral de medias (en este caso, n = 50).

Si no se cumplen las condiciones, sea una distribución muestral de medias o de proporciones, no podremos suponer que la distribución muestral es normal, pero tal vez lo sea de otro tipo conocido. Para muestras pequeñas tenemos otra distribución a la que acudiremos, pero que no describiremos: la distribución de Student. No la describiremos porque conociendo la normal, el camino hacia el conocimiento de otras distribuciones es relativamente fácil. Otra posibilidad es adoptar una posición que se llama *conservadora* y que consiste en ponernos en la peor de las situaciones. Hay soluciones para ello, como la *acotación de Chebyshev*. Lo malo es que esa postura conservadora suministra valores muy elevados de tal forma que, para una misma seguridad, tendremos que construir un intervalo de estimación mucho más amplio con Chebyshev que con la curva normal. Pero eso es otra historia. Sigamos con la normalidad.

Un momento, ¿decías σ?

Pues sí. La expresión para calcular e_p necesita el valor de la desviación tipo de la población. Un momento, recapacitemos. No tengo la media aritmética de la población y por eso estoy haciendo la estimación por intervalo. Ahora resulta que para hacer esa

estimación necesito el valor de la desviación tipo poblacional. Si no tengo la media ¿cómo voy a tener la desviación? ¿Qué he de hacer ahora? ¿Tengo que poner en marcha otra estimación por intervalo, esta vez para la desviación tipo poblacional? ¿Este proceso no me exigirá a su vez conocer otro parámetro todavía más rebuscado?

Vale, que no cunda el pánico. Al inicio de este documento ya dije que las estimaciones son de dos tipos: puntuales y por intervalo. Las primeras se utilizan para los procedimientos intermedios y las segundas para los objetivos. Nuestro objetivo es estimar la media aritmética poblacional. Por eso ponemos en marcha una estimación por intervalo. Para ello, en el camino, surge la necesidad de contar con el valor de la desviación tipo de la población. Se trata de un reto procedimental, no de un objetivo de la investigación. Así que la estimación que realizamos en este caso es puntual.

Como también hemos visto al inicio de este documento, escogemos un estimador insesgado de la desviación tipo poblacional: la cuasidesviación tipo de la muestra. Por eso, utilizando solo la información de la muestra y la medida estandarizada de la seguridad decidida, la expresión de cálculo del error de precisión termina siendo:

$$e_p = Z\frac{\sigma}{\sqrt{n}} \quad \simeq \quad Z\frac{\hat{S}}{\sqrt{n}} = Z\frac{S}{\sqrt{n-1}}$$

Surtiéndonos del documento monográfico sobre muestreo, también podemos obtener la expresión de cálculo para la estimación de proporciones:

$$e_p = Z\sqrt{\frac{\pi\,(1-\pi)}{n}} \quad \simeq \quad Z\sqrt{\frac{p\,(1-p)}{n}}$$

Ejemplo

Hemos preguntado a un conjunto de 42 personas, tomadas al azar de la ciudad, cuántas llamadas realizaron la semana pasada desde su teléfono móvil a un fijo. La ciudad cuenta con 130 mil habitantes. El resultado es el siguiente:

7	4	6	2	8	3	5
6	7	3	3	5	2	7
8	7	5	3	3	5	5
4	5	5	7	5	1	3
1	1	2	3	6	7	3
4	3	7	1	3	7	7

Por término medio...

Lo primero que nos interesa es encontrar cuál es el número medio semanal de llamadas a fijos que realizan en la ciudad desde un teléfono móvil. Se trata de estimar la media aritmética poblacional, para lo que necesitamos la media de la muestra (estimador), la desviación tipo, la distancia estandarizada que representa la seguridad y el tamaño de la muestra. Pongamos que, por llevar la contraria a Sir Ronald y porque tampoco se nos ocurre qué problema serio puede derivarse de nuestro posible error, utilizaremos una seguridad del 90%.

Lo primero es observar si podemos suponer que la distribución muestral de medias de la que proviene nuestra muestral es normal. Podemos suponerlo, puesto que $n = 42 \geq 30$.

Lo siguiente es traducir la seguridad a una distancia estandarizada, utilizando la curva normal estandarizada. Según la tabla que tenemos en el monográfico *La curva normal*, un área centrada del 90% se corresponde con una puntuación tipo de valor 1,645.

Acto seguido, nos falta el estimador y la desviación tipo. Lo resolvemos mediante una tabla de frecuencias (donde d2 es la distancia cuadrática a la media):

Tabla de frecuencias			
Xi	fi	Xifi	d2*fi
1	4	4,00	49,00
2	3	6,00	18,75
3	10	30,00	22,50
4	3	12,00	0,75
5	8	40,00	2,00
6	3	18,00	6,75
7	9	63,00	56,25
8	2	16,00	24,50
Suma:	42	189,00	180,50

Media:	189 / 42 =	4,5
Varianza	180,5 / 42 =	4,298
Desv. tipo:	raíz (4,298)=	2,073

Con esta información:

$$e_p = Z\frac{S}{\sqrt{n-1}} = 1{,}645\ \frac{2{,}073}{\sqrt{42-1}} = 0{,}53$$

$$\mu \in \{\bar{X} \pm e_p\}_{seg} \quad \Rightarrow \quad \mu \in \{4{,}5 \pm 0{,}53\}_{0{,}90} = \{\ 3{,}97\ ;$$

Luego, con una seguridad del 90%, afirmamos que la gente de la ciudad realiza semanalmente entre 3,97 y 5,03 llamadas de móvil a fijo por término medio.

Pongamos un porcentaje

¿Qué porcentaje de personas en la ciudad realizan no menos de 5 llamadas de móvil a fijo durante una semana? Utiliza una probabilidad de errar del 5%.

En la muestra, 8+3+9+2=22 personas realizaron no menos de 5 llamadas de ese tipo, lo que significa un 22/42*100 = 52,38% de la muestra. Al consultar la tabla de la curva normal estandarizada, el área centrada del 95% se corresponde con una distancia estandarizada de 1,96. Podemos consultar sin problemas esta tabla, puesto que es asumible que la distribución muestral de proporciones es normal. Utilizando *p* como criterio (pues carecemos del valor de π):

$$np = 42 \cdot 0{,}5238 = 22 \geq 5 \qquad n\,(1-p) = 42 \cdot 0{,}4762$$

Así pues:

$$e_p = Z\sqrt{\frac{p\,(1-p)}{n}} = 1{,}96\sqrt{\frac{52{,}38 \cdot 47{,}62}{42}} = 15$$

$$\pi \in \{\, p \pm e_p \,\}_{seg} \qquad \Rightarrow \qquad \pi \in \{52{,}38\,\% \pm 15\,\%\,\}_{0{,}95} = \{\ 37$$

Luego, con una seguridad del 95% podemos concluir que entre un 37% y un 67% de la gente de la ciudad realiza no menos de 5 llamadas de móvil a fijo en una semana. Observa que he utilizado valores de

porcentaje en lugar de proporciones. Esto es intrascendente. Puedo hacer los cálculos con proporciones y multiplicar finalmente por 100, o arrastrar los porcentajes desde cuando quiera.

En total...

Y, por último, a partir de los resultados anteriores queremos conocer cuántas llamadas de móvil a fijo se hacen en la ciudad y cuánta gente hace no menos de 5 llamadas, ambos semanalmente. Como sabemos que la ciudad alberga a 130 mil habitantes, podemos responder a estas dos inquietudes. No obstante, al tratarse de dos conclusiones realizadas con niveles diferentes de seguridad, no van a ser comparables. Para evitarlo, vamos a repetir los cálculos y, ya puestos, variando también la seguridad. Esta vez vamos a responder a ambas preguntas a través de un riesgo de equivocarnos del 3%.

Al consultar la tabla para un área centrada del 97% de seguridad, el valor estandarizado correspondiente es 2,17. Luego:

Para la media:

$$e_p = Z\frac{S}{\sqrt{n-1}} = 2{,}17\ \frac{2{,}073}{\sqrt{42-1}} = 0{,}70$$

$$\mu \in \{\bar{X} \pm e_p\}_{seg} \quad \Rightarrow \quad \mu \in \{4{,}5 \pm 0{,}7\}_{0{,}97} = \{\ 3{,}8\ ;\ 5{,}$$

Si cada persona realiza entre 3,8 y 5,2 llamadas de móvil a fijo semanales, las 130 mil realizarán entre 3,8*130000= 494 mil y 5,2*130000= 676 mil de este tipo de llamadas a la semana. Respecto a la proporción:

$$e_p = Z\sqrt{\frac{p\ (1-p)}{n}} = 2{,}17\sqrt{\frac{52{,}38 \cdot 47{,}62}{42}} = 16{,}72$$

$$\pi \in \{p \pm e_p\}_{seg} \quad \Rightarrow \quad \pi \in \{52{,}38\,\% \pm 16{,}72\,\%\}_{0{,}97} =$$

Si esto ocurre en una población de 130 mil personas, podemos concluir con una seguridad del 97% que entre 130000*0,3566= 46358 y 130000*0,691=89830 de ellas realizan no menos de cinco llamadas de móvil a fijo a la semana.