

Capítulo 5:

Principios de inferencia

Introducción

¿Qué información proporciona, a un clínico de Barcelona, los resultados obtenidos en un estudio previo realizado en Boston? La evolución de estos casos de Boston se puede conocer perfectamente, sin error. Pero esos casos ya han «evolucionado», no tiene interés predecir una evolución que ya ha sucedido. En cambio, sería muy interesante poder aplicar estos resultados pasados a unos nuevos casos. ¿Cómo hacerlo?

La inferencia estadística, para incorporar al conocimiento teórico la información empírica, define los conceptos de muestra y población. Los valores obtenidos en las muestras permitirán estimar, con un cierto error cuantificable, los parámetros que caracterizan al conjunto de la población. En resumen, la estadística permite cuantificar, tanto la información (o «señal») aportada por los datos, como el error aleatorio (o «ruido») que implica el proceso de generalización.

En este capítulo se exponen los conceptos fundamentales de inferencia estadística. Se introduce la oscilación aleatoria de los valores obtenidos en muestras y cómo la estadística cuantifica esta oscilación. Para ello, se trabaja con el indicador estadístico más usual: el promedio o media muestral. Se estudia cómo oscila, cuánto oscila, alrededor de qué valor oscila y qué forma adopta esta oscilación.

Objetivos

Al terminar este capítulo, un lector que haya realizado los ejercicios:

- Delante de una variable, se planteará la necesidad de resumir y representar los datos.
- Delante de un estudio, identificará la unidad o caso.
- Delante de un estudio, se preguntará por los posibles errores de inferencia.
- Delante de un estudio, se preguntará por la población a la que pueden hacer referencia sus resultados.
- Delante de un estudio, distinguirá entre estadístico, estimador y parámetro.
- Sabrá que la muestra aleatoria permite disponer de estimadores insesgados.
- Conocerá la distribución del estimador media.
- Distinguirá entre desviación típica y error típico.
- Interpretará el error típico como el error esperado al decir que el parámetro toma el valor observado en la muestra.
- Conocerá cómo el incremento del tamaño muestral reduce la incertidumbre de la estimación.

Introducción a la inferencia estadística

¿Qué es la inferencia estadística?

Si, por ejemplo, se desea estimar el tiempo de crecimiento de un cierto tejido, se pueden utilizar dos procedimientos. El primero, teórico, consiste en *deducir* este tiempo a partir de los tiempos de división de sus células. El segundo procedimiento, empírico, consiste en *inducirlo* a partir de un número limitado de casos. Ahora bien, ¿hasta qué punto unas pocas pruebas permiten establecer leyes generales sobre el crecimiento de estos tejidos? O mejor, ¿cuánta información están realmente aportando? En los puntos sucesivos se estudia cómo la inferencia estadística formaliza este proceso, lo que requiere: *a)* definir; *b)* cuantificar, y *c)* acotar los riesgos que conlleva.

Definición



La **inferencia** pretende generalizar la información contenida en unos datos a un cuerpo de conocimiento más amplio.

Historieta



Dos amigos caminan por el Pirineo y, al ver un caballo, uno de ellos comenta: «No sabía que los caballos de La Cerdaña fueran marrones y con las patas anchas». Su amigo, que es lógico, le responde: «Perdona, lo que no sabías es que en La Cerdaña hay, por lo menos, un caballo marrón de patas anchas».

Lectura



Hasta hace relativamente poco tiempo, los filósofos se han quejado de la falta de herramientas técnicas que permitan este salto de las partes al todo (Chalmers, 30; Bunge, 31). Para Hume, la inferencia era simplemente imposible, y para Russell, la inducción seguía siendo un problema de lógica no resuelto. A mediados del siglo pasado, Popper aportó un punto de vista algo más optimista: «Sólo la refutación de una teoría puede ser inferida a partir de datos empíricos y esta inferencia es puramente deductiva». Hoy en día, ya está plenamente aceptado que, en estudios bien diseñados, y ejecutados, la metodología estadística hace posible la inferencia, siempre que se acepten ciertos riesgos.

Respuestas que ofrece la inferencia estadística

Veamos algunas preguntas que pueden ser contestadas con la ayuda de la metodología estadística. El ejemplo más sencillo estudiaría la distribución de una sola variable: ¿cuál es el valor de monóxido de carbono en el aire espirado por fumadores jóvenes? O bien, ¿cuál es la distribución de los valores de homocisteína plasmática en

pacientes con lupus eritematoso? Nótese que si no existiera variabilidad —si la cantidad de monóxido de carbono espirado fuera siempre la misma—, la inferencia sería inmediata: con una observación sería suficiente para conocer el comportamiento de todas ellas. La metodología estadística considera y cuantifica la diversidad entre las unidades.

Ejemplo 5.1



No es necesario hacer un estudio estadístico para conocer la distribución de la variable «número de cerebros» que tiene cada uno de los habitantes de una ciudad. Cada ciudadano tiene un cerebro y sólo uno: así de fácil.

En cambio, sería terriblemente aburrido «decir toda la verdad» sobre la altura de una muestra de 23 pacientes: el primer caso mide 164 cm, el segundo, 173 cm; el tercero 168; ...y el vigésimo tercero, 192.

Ejercicio 5.1



Suponga que, en el ejemplo anterior, por no aburrir, decide hacer un resumen de los datos, ¿qué información le gustaría que este resumen le proporcionara?

Ejercicio 5.2

Proponga otro ejemplo en el que también sea conveniente hacer un resumen estadístico de los datos.

Conocer la distribución de una variable permitirá al equipo clínico realizar de forma científica el diagnóstico, el tratamiento y/o pronóstico de un paciente.

Ejemplo 5.2



Si se conoce cuál es la distribución de la variable tiempo de convalecencia tras cierta enfermedad, el médico de cabecera puede «adelantar» al paciente cuántos días tendrá sus facultades mermadas. Con la media, le dirá al paciente cuál es su valor esperado. Y con la desviación típica, cuál es el error esperado: cuánto cabe esperar que un paciente típico se aleje de esa media.

Población, muestra e individuo

El reto de la inferencia estadística consiste en generalizar un fenómeno observado en unos cuantos casos a todas las observaciones posibles. Para ello, las primeras definiciones que se deben establecer son las de población, muestra y unidad.

Definición

Población: Conjunto de todos los elementos, que cumplen ciertas propiedades comunes, entre los que se desea estudiar un determinado fenómeno.

Muestra: Subconjunto de la población que es estudiado y a partir del cual se sacan conclusiones sobre las características de la población.

Unidad (individuo o caso): Es cada uno de los elementos que componen la muestra y la población.

Población, muestra y unidad se contienen progresivamente, a la manera de las muñecas rusas. La población contiene la muestra y la muestra, a su vez, las unidades. La diferencia es que, conceptualmente, hay un número ilimitado de muestras y de individuos. La población, sin embargo, es única, y representa al conjunto que deseamos conocer.

Ejemplo 5.3

Costa et al. (32). Se invitó a participar en el estudio, de manera consecutiva, a todas las personas que acudieron al Centro de Extracciones del Hospital Clínic i Universitari de Barcelona, desde diferentes servicios, para la realización de una prueba de tolerancia oral a la glucosa (PTOG).

Cilla et al. (33). El grupo de población considerado en el estudio incluyó a mujeres que tuvieron un primer parto después de septiembre de 1989 y un segundo parto entre 2 y 8 años después en la Maternidad del Hospital Nuestra Señora de Aránzazu de San Sebastián (Guipúzcoa).

Por su parte, las unidades no tienen por qué ser «individuos». Pueden ser hospitales, comarcas o visitas clínicas. Es muy importante definir con sumo cuidado estas unidades, ya que se podría llegar a conclusiones diferentes.

Ejemplo 5.4

Cierto facultativo presume de tener un razonable promedio de 7 pacientes por hora. Pero la asociación de usuarios ha preguntado a todos sus pacientes y ha obtenido un promedio de 9. ¡Y pudiera ser que todos digan la verdad, sin trampa!

Pongamos que este profesional tiene 3 horas de visita. En una de ellas ve a las primeras visitas, a razón de 3 por hora. En otra, recibe a las segundas visitas, 6 por hora. Y en la restante hora recibe las demás visitas, 12 por hora. Este facultativo ha definido como unidad del estudio la «hora de visita»: el promedio de 3, 6 y 12 es, efectivamente, 7 pacientes por hora.

Ejemplo 5.4 (Cont.)

$$\sum_{i=1...3} \frac{X_i}{n} = \frac{3+6+12}{3} = 7$$

donde $\sum_{i=1...3}$ indica la suma de los casos 1 a 3

Los usuarios, en cambio, han definido como unidad a cada uno de ellos, de forma que la respuesta «3» la obtienen 3 veces, la respuesta «6», 6 veces y «12», 12 veces. Y el promedio es, efectivamente, 9 pacientes:

$$\sum_{i=1...21} \frac{X_i}{n} = \frac{3+3+3+6+6+6+6+6+6+6+6+12+...+12}{21} = 9$$

Posiblemente, la primera definición represente mejor la pregunta del clínico (¿qué promedio de pacientes visito yo por hora?) y la segunda la del usuario (¿cuánto suele durar mi visita?).

Ambas definiciones son correctas y válidas. Pero no son intercambiables y en cada estudio debe estar muy clara cuál es la unidad que se ha definido. Así, diferentes objetivos requieren diferentes cálculos, todos ellos lícitos y correctos, pero que no deben confundirse: siempre debe quedar bien clara la unidad del estudio.

Ejercicio 5.3

El colegio de odontólogos ha realizado un estudio aleatorio entre los pacientes de sus consultas en la semana anterior. De 1.000 fichas analizadas, 500 habían tenido una visita el año anterior, por lo que concluyen que un 50% de la población acude al dentista cada año. ¿Qué opina? ¿Se puede conocer la frecuencia de visitas al odontólogo en la población general a partir de una muestra obtenida en las consultas?

Ejercicio 5.4

Los centros sanitarios de la Seguridad Social suelen realizar una encuesta de satisfacción a sus usuarios, cuyos resultados suelen ser altamente positivos. ¿Qué le lleva al defensor del pueblo a realizar una encuesta en la población general? (Pista: defina la unidad de ambos estudios y medite sus diferencias.)

Ejercicio 5.5

Para estimar la infección nosocomial, puede hacerse un estudio seleccionando algunos de los pacientes que ingresan o bien seleccionando algunas de las camas ocupadas en el hospital. Asumiendo que los pacientes que están ingresados más tiempo son los que tienen mayor probabilidad de desarrollar esta infección, ¿cuál de los dos estudios dará cifras más altas de infección nosocomial?

Recuerde

Estudie siempre con mucho cuidado cómo se definen las unidades. Dos estudios, para poder ser comparados, requieren la misma definición de las unidades.

Estadísticos, estimadores y parámetros

En el estudio de la información disponible, la inferencia estadística afronta el reto de abarcar un «universo» más amplio que los «pocos casos» disponibles.

Definición

Los indicadores que se calculan en las muestras reciben el nombre de **estadísticos**.

Los indicadores de la población, que estamos interesados en conocer, reciben el nombre de **parámetros**.

Recuerde

Un parámetro hace referencia a un valor de la población, mientras que un estadístico lo hace de la muestra.

Por ejemplo, la media puede representar al «parámetro» media cuando hablamos del centro de gravedad (o **esperanza**) de una distribución poblacional, o al «estadístico» media cuando nos referimos al **promedio** de una serie de valores calculado en una muestra.

Ejemplo 5.5

Supóngase que la probabilidad de que un paciente con anticuerpos del sida tarde, en ciertas condiciones, más de 2 años en desarrollar los primeros síntomas es 0,50. Es decir, expresado en porcentajes, del 50%. Esta probabilidad es un parámetro que resume las expectativas del paciente y que representa una característica intrínseca de la enfermedad. Por otro lado, se han estudiado, en esas condiciones,

Ejemplo 5.5 (Cont.)

una muestra de 25 pacientes de esa población y 15 de ellos han superado los 2 años. Este resultado de $15/25 = 0,60$ (60%) representa la proporción, que es el estadístico o valor observado en la muestra.

Ejercicio 5.6

Proponer un ejemplo similar con la media en lugar de la probabilidad.

El reto de la inferencia estadística es conocer los parámetros, que caracterizan al todo de la población, a partir de los estadísticos, obtenidos en una muestra.

Definición

Cuando un estadístico de una muestra se usa para conocer el valor de un parámetro de la población recibe el nombre de **estimador**.

Nota técnica

Si usted dispone de los datos de toda la población, es decir, si las conclusiones de su estudio se aplicarán únicamente a estos casos y no desea poderlas aplicar a otros datos diferentes, usted no necesita saber qué es la inferencia estadística. Pero tenga cuidado al hablar: no podrá establecer ninguna ley «universal» que vaya más allá de sus propios datos.

Cada muestra es fugaz, en el sentido de ser irrepetible y, en el fondo, irrelevante en sí misma. Una vez terminado el seguimiento de los pacientes de la muestra y cumplidas las responsabilidades sanitarias con ellos, el interés científico se centrará en conocer qué dicen estos casos sobre los pacientes futuros.

Ejemplo 5.6

Las encuestas electorales, a partir de unos pocos miles de entrevistados, intentan conocer la tendencia de unos cuantos millones: el auténtico interés está en lo que votará toda la población. La importancia que tienen los pocos entrevistados es su capacidad para informar sobre la distribución poblacional de esta variable.

Recuerde



Se puede acceder al estadístico observado en la muestra, pero el auténtico objetivo es el parámetro de la población.

Definición



La inferencia estadística es el proceso formal de analizar y cuantificar la información empírica («**evidencia**» o pruebas) que el estimador proporciona del parámetro.

Es tan importante distinguir si se trata de valores muestrales o poblacionales que se les dará diferente símbolo en un caso o en otro. Incluso, la media puede recibir el nombre de esperanza cuando se trata del parámetro poblacional y de promedio cuando es el valor obtenido en la muestra. De la misma forma, el estimador muestral de la probabilidad recibe el nombre de proporción (tabla 5-1).

	Parámetro (θ) (Población)	Estadístico ($\tilde{\theta}$) (Muestra)
Media	$\mu = E(X)$ esperanza	\bar{X} o m promedio
Desviación típica	σ	S
Probabilidad	π probabilidad	p proporción

Tabla 5-1 Símbolos utilizados para distinguir parámetros y estadísticos

Así, una proporción observada en una muestra informa sobre la probabilidad de la población. Pero ¿cuánta información aporta? ¿Se debe «creer» mucho o poco que el valor poblacional se acerca al valor del estadístico observado? La teoría de probabilidad permite valorar la cantidad de información que un estadístico (la proporción muestral P) aporta sobre el valor desconocido del parámetro (la probabilidad π), auténtico objetivo de nuestro estudio.

A continuación se estudia la distribución de los estimadores a lo largo de todas las posibles muestras. Así, se podrá saber cuánto oscila un estadístico de una muestra a otra y se podrá, por lo tanto, proponer medidas que cuantifiquen la cantidad de información que un estadístico, observado en una muestra concreta, proporciona sobre el parámetro de la población. Por simplicidad, veamos esta distribución en el caso de la muestra aleatoria simple (MAS).

Muestra aleatoria

Definición



Muestra aleatoria simple (MAS) es aquella en la que: a) todos los elementos de la población tienen la misma probabilidad de pertenecer a ella, y b) cualquier combinación de n elementos tiene la misma probabilidad de pertenecer a la muestra.

Nota técnica

Todos los elementos de la muestra tienen la misma distribución, ya que vienen de la misma población.

Ejemplo 5.7

Imaginemos: a) la población infinita de todos los posibles pacientes de una enfermedad; b) un procedimiento aleatorio por el que se seleccionan de forma independiente $n = 1.000$ pacientes de esta población.

Contraejemplo 5.8

Una asociación profesional con 25.000 afiliados decide hacer un estudio para conocer qué proporción de ellos han recibido malos tratos en su trabajo. Diseña una muestra aleatoria de 2.000 afiliados a los que les envía un cuestionario, que contestan sólo 500. Se puede saber que los 2.000 representan a los 25.000, pero se desconoce a quién representan estos 500 y, por tanto, qué información aportan sobre el total de la población.

La definición de la población a la que se desea aplicar los resultados puede cambiar la consideración de la muestra.

Contraejemplo 5.9

Supongamos: a) la población finita de los 80 pacientes de una enfermedad determinada de un centro hospitalario; b) un proceso aleatorio de selección de 20 pacientes diferentes. Nótese que, al ser un muestreo sin reemplazamiento, al eliminar un paciente cada vez, la población de los pacientes susceptibles de ser seleccionados va variando, con lo que la variable aleatoria no tiene la misma distribución para cada uno de los elementos de la muestra.

Ejemplo 5.10

En el fondo, el objetivo del estudio del **contraejemplo 5.9** no puede ser conocer cómo se comportan estos 80 pacientes (tema vital para ellos y para el centro que los atiende, pero sin ningún interés para el resto de pacientes). El objetivo del estudio debe ser más ambicioso, de manera que se puedan beneficiar los pacientes de otros centros.

Ejemplo 5.10 (Cont.)

Ahora, por un lado, la situación se simplifica, ya que eliminar un elemento de esta población infinita prácticamente no modifica su distribución. Pero, por otro lado se complica, ya que debe tenerse en cuenta que los casos estudiados (sean 20 o sean 80) no son una muestra aleatoria de la población de todos los pacientes con la misma enfermedad. ¿Hasta qué punto los resultados son extrapolables?

Recuerde

Caso, muestra y población no se definen por separado, de forma aislada. Haga siempre la definición conjunta.

Volvamos a la definición de MAS. También resalta que la información aportada por las diferentes unidades debe ser independiente entre sí. Es decir, el valor obtenido en una observación no aporta información sobre el valor de otras observaciones. Este «no aportar información» debe entenderse como que la distribución de las restantes variables es la misma sea cual sea el valor observado.

Ejemplo 5.11

Sigamos con el **ejemplo 5.7** de pacientes con una enfermedad. Cada uno de los elementos de la muestra aporta exactamente la misma información sobre la población: que cierto paciente tenga un valor elevado no implica que el paciente siguiente ni cualquier otro deba tenerlo ni más alto ni más bajo.

Contraejemplo 5.12

En un estudio multicéntrico, ¿puede creerse que el resultado de un paciente de un centro no aporta información sobre el resultado de otro paciente del mismo centro? O por el contrario, ¿es más razonable pensar que los resultados obtenidos en pacientes de un mismo centro son más similares que los de pacientes de centros diferentes? Si es este último caso, la variable centro es una variable importante que debe ser tomada en cuenta en el análisis posterior.

Recuerde

En una MAS: a) las unidades se escogen al azar; b) todas ellas tienen la misma probabilidad de ser escogidas, y c) todas las posibles combinaciones de elementos tienen la misma probabilidad de configurar la muestra.

Papel de la inferencia estadística en los procesos científico y técnico

La capacidad de la estadística para inferir formalmente desde unos pocos datos de la muestra a la totalidad de la población ha permitido un progreso espectacular en todas las ciencias. Hoy en día, se acepta como modelo de razonamiento científico el contenido en el siguiente esquema (cuadro 5-1). Nótese que este modelo integra los razonamientos deductivos con los inductivos. Los primeros son necesarios, por ejemplo, para diseñar la recogida de datos. Los inductivos, por su parte, se requieren para generalizar las observaciones obtenidas en unos cuantos elementos.

1. Descubrir el problema a investigar
2. Documentar y definir el problema o hipótesis
3. Deducir consecuencias contrastables de las hipótesis
4. Diseñar la observación o la experimentación
5. Recoger los datos
6. Analizar los datos mediante inferencia estadística
7. Establecer las conclusiones
8. Integrar las conclusiones en el cuerpo de conocimiento

Cuadro 5-1 Pasos del método científico.

Comentario



Las profesiones científicas se van aproximando progresivamente al esquema de trabajo del método científico (fig. 5-1). Tras un análisis de la situación de partida, el planificador establece una estrategia (hipótesis) para alcanzar unos objetivos. Y se marca unos indicadores numéricos que le permitirán ir verificando (contraste empírico) hasta qué punto ha alcanzado estos objetivos. Se realiza el trabajo de producción (recogida de datos): en un médico, podría ser visitar a los pacientes; en un profesor, impartir la materia. Luego, deben evaluarse los resultados para medir el grado de obtención de los objetivos.

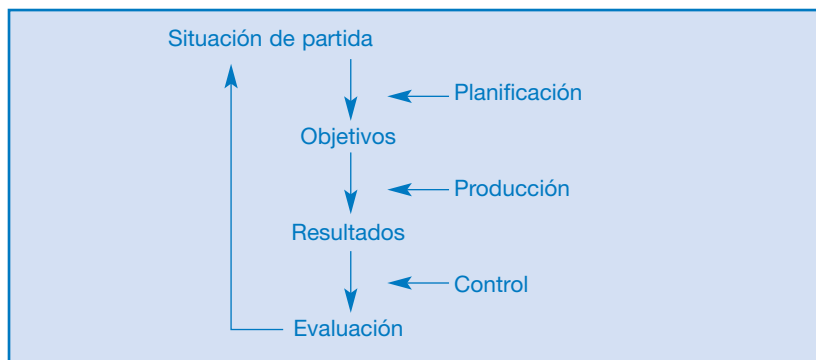


Figura 5-1 Esquema de trabajo del método científico-técnico.

Posibles errores en la inferencia estadística

En los libros de estadística es fácil encontrar definiciones como la siguiente: «Se entiende por **muestra aleatoria** aquella en la que todos los elementos de la población tienen una probabilidad conocida de pertenecer a la muestra. Si esta probabilidad es la misma para cualquier combinación de k elementos, se habla de **muestra aleatoria simple**». A partir de esta definición, la teoría de probabilidad permite cuantificar la información y el ruido aportados por dicha muestra aleatoria. Sin embargo, diferentes hechos hacen irreal la muestra aleatoria. En primer lugar, los individuos tienen derecho a rechazar su participación en el estudio, o incluso a abandonarlo en cualquier momento. En segundo lugar, no se dispone de definiciones operativas de las poblaciones; por ejemplo: no hay ningún listado con todos los pacientes de una determinada enfermedad. Todos estos fenómenos —no aleatorios— pueden provocar distorsiones no aleatorias, llamadas **sesgos**.

Resultados (%)		Previsiones	
		EL PERIÓDICO 23/10/1989 n = 9.524 + 2.000 (%)	LA VANGUARDIA 23/10/1989 n = 3.262 (%)
PSOE	39,6	40,5	41,5
PP	25,8	19,1	25,0
CIU	5,0	4,9	4,5
IU	9,1	10,3	7,5
CDS	7,9	8,5	6,5
Margen		±1	±2

Tabla 5-2 Prospección de voto y resultados electorales de octubre de 1989

Lectura



¿Hasta qué punto debemos creernos las previsiones electorales que se publican en diferentes medios? A continuación, y respecto a las elecciones generales de octubre de 1989, figuran los resultados reales (parámetros poblacionales) junto a las previsiones (estimaciones basadas en muestras) publicadas por El Periódico de Catalunya y por La Vanguardia (tabla 5-2). El segundo diario, que se comprometía con un margen menos ambicioso (2%) ha cumplido. En cambio, el primero ha fallado en dos ocasiones: a pesar de que prometía un margen máximo de 1 punto, para el PP se ha distanciado en 6,7 puntos y para IU en 1,2 puntos. La lectura de las «fichas técnicas» de ambos estudios permite encontrar posibles explicaciones a estas diferencias: que si el tipo de entrevista (personal o telefónica); que si la selección de los casos; que si los días en que se ha realizado la encuesta; que si no se tiene en cuenta la profesión;

Lectura (Cont.)



que si hay 2.000 encuestas adicionales en Cataluña, donde el PP suele estar bajo; etc. Nótese que estas explicaciones se basan en argumentos sociológicos, no estadísticos. Se podría haber argumentado que, por mala suerte, la estimación proporcionada por El Periódico se había alejado del auténtico valor. Éste sería un argumento probabilístico, basado en el error aleatorio debido al azar del muestreo.

La lectura anterior ilustra que en todo procedimiento de muestreo existen dos prototipos de errores: los debidos exclusivamente a las fluctuaciones del azar o **errores aleatorios** y todos los demás, conocidos como **errores sistemáticos** o **sesgos**. La estadística ayuda a cuantificar la magnitud de los primeros. Eliminar o acotar los segundos es una responsabilidad que la estadística comparte con la disciplina objeto del estudio, en nuestro caso, la Medicina. En el ejemplo anterior, han sido razonamientos políticos o sociológicos los que sugerían la existencia de posibles sesgos. En los estudios clínicos, el profesional sanitario debe razonar si las condiciones en las que se ha realizado el estudio le permiten negar la existencia de sesgos.

Recuerde



Al pasar de la muestra a la población, en el proceso inferencial hay dos posibles fuentes de errores: los aleatorios que la Estadística le ayudará a cuantificar; y los sistemáticos, o sesgos, cuya posible existencia debe usted estudiar a la luz de sus conocimientos clínicos.

Poblaciones implicadas en la inferencia estadística

Así pues, un buen estudio debe, en primer lugar, cuantificar la magnitud del error aleatorio. En segundo lugar debe justificar que la magnitud de este error es razonable para los objetivos del estudio. Y en tercer lugar, debe defender la ausencia de sesgos, que podrían haberse producido por desviaciones de la aleatoriedad de la muestra. Si se dan estas condiciones, en epidemiología suele decirse que el estudio es válido. Ahora bien, ¿válido para qué conclusiones, las de los autores del estudio o las de aquellos que desean aplicarlo en una nueva población? Puede ser ilustrativo hacer las siguientes definiciones de poblaciones, progresivamente más amplias.

Definición



Kleinbaum et al. (34) ofrecen las siguientes definiciones:

Población origen de la muestra (*actual population*) o población muestreada: es aquella población imaginaria de la que se hubiera obtenido, por extracción aleatoria pura, la muestra. Excluye, por lo tanto, aquellos casos que se niegan a participar en el estudio, o los que no figuran correctamente en el censo utilizado para generar la selección, o los que fallecen, etc.

Población objetivo o diana (*target population*): es aquella población a la que se desea poder inferir los resultados del estudio.

Poblaciones externas (*external populations*): son cada una de las posibles poblaciones más amplias a las que se puede desear, en algún momento futuro, o por otros investigadores, generalizar los resultados.

Una vez más, se puede recurrir a las muñecas rusas para resaltar que la población externa contiene a la población objetivo y que esta última contiene, a su vez, a la población origen de la muestra (fig. 5-2).

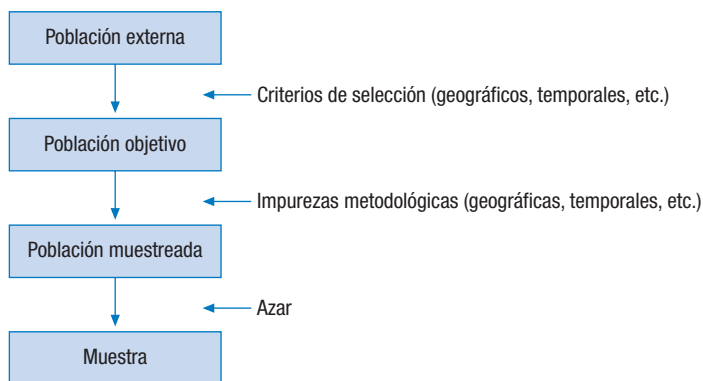


Figura 5-2 Esquema de las poblaciones que intervienen en un estudio.

Ejemplo 5.14



Los ensayos clínicos suelen realizarse con el criterio de inclusión de que la edad debe estar entre 18 y 65 años. Así, los pacientes menores y mayores nunca forman parte de la población objetivo. Quizá sí que formen parte de la externa. En este caso, los argumentos para poder creer que los resultados del estudio también les aplican son ajenos a la inferencia estadística.

Lectura



Dal-Ré et al. (35). «En general, la inclusión de enfermos en ensayos clínicos es un problema más importante de lo que los propios investigadores piensan, y siempre resulta más difícil de lo que en un principio se planeó [...]. En la mayor parte de los ensayos clínicos, muchos de los centros participantes son incapaces de seleccionar el número de enfermos según el calendario acordado; esto se debe a dos razones fundamentales: la selección de los investigadores participantes y las características del protocolo [...]. Las normas de buena práctica clínica de la Unión Europea especifican que los investigadores deben aportar datos de los pacientes que hubieran cumplido los criterios de selección de enfermos señalados en el protocolo, de forma que se pueda asegurar una adecuada tasa de selección.»

Ejercicio 5.7



A partir de un artículo de investigación imaginario o real (p. ej.: «Resultados a los 12 meses de un programa de deshabituación tabáquica en un centro de atención primaria») defina: población externa, población objetivo, población muestreada y muestra (fig. 5-3).

De acuerdo con las definiciones anteriores, se pueden aclarar varios términos de uso habitual.

Definición



Sesgo es toda diferencia entre el valor del parámetro en la población origen de la muestra y el valor del mismo parámetro en la población objetivo.

Error aleatorio es la oscilación de los estadísticos obtenibles en las posibles muestras (siempre centrados en el parámetro de la población origen de la muestra).

En epidemiología (Kleinbaum et al., 34) se dice que un estudio es **preciso** si este error aleatorio es razonablemente pequeño. Se dice que un estudio tiene **validez** si es preciso y no tiene sesgos. Se habla de **validez interna** si no tiene sesgos para con la población objetivo, y de **validez externa** si la ausencia de sesgos abarca también a la población externa.

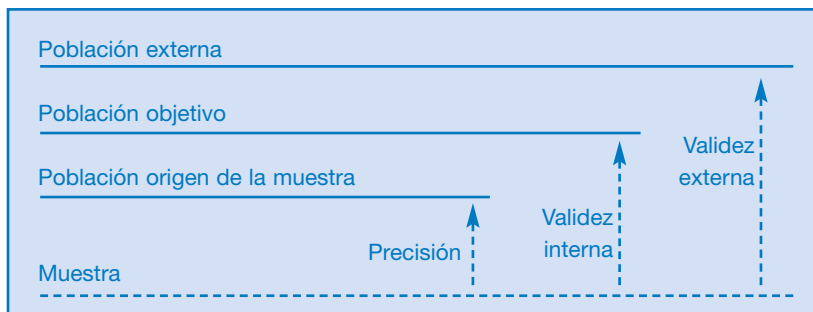


Figura 5-3 Relación entre las propiedades deseables de inferencia y las poblaciones a las que se aplican (Kleinbaum et al. (34)).

Comentario



Si le hacen una pregunta tan vaga como «¿es representativa su muestra?», puede contestar, seguro y con aplomo, que sí. Siempre hay una población origen para la que su muestra será representativa. Las preguntas de interés son: ¿representa su muestra a «cierta» población objetivo? ¿Cómo viene definida ésta? El riesgo de error aparece al redactar las conclusiones donde puede quererse llegar más lejos de lo que permiten los datos. Recuerde que entre los derechos humanos figura el de no participar en una encuesta o en un estudio clínico o epidemiológico, lo que hace inverosímil la muestra aleatoria «pura».

Nota técnica

Seleccionar una muestra «al azar» es un proceso riguroso, muy distinto de seleccionarla «al tuntún».

Ejercicio 5.8

Es bien conocida la tendencia de los autores y los editores de revistas a publicar solamente los resultados que son estadísticamente significativos. ¿Puede esta actitud provocar un sesgo?

Ejercicio 5.9

Los textos médicos suelen estar basados en artículos científicos escritos desde centros de atención terciaria. ¿Puede este hecho provocar un sesgo?

Recuerde

La inferencia estadística sólo cuantifica la magnitud del error aleatorio.

Si un número elevado de casos rehúsa participar en el estudio, el investigador debe dejar clara esta situación y discutir hasta qué punto compromete o invalida las conclusiones.

Estadístico media muestral

Distribución del estadístico media muestral \bar{X}

Recuerde

El cálculo de la media muestral o promedio obtenido en la muestra es bien conocido: $\bar{X} = \sum x_i / n$

Se dijo en el primer capítulo que para usar la media es necesario que la variable esté en escala de intervalo, que exista una unidad de medida que signifique siempre lo mismo.

Historieta

Bienvenido a Valle Sillytont

Fundado en 1.986

Altura 1.191 metros

823 habitantes

TOTAL 4.000

Lectura



Mediavilla et al. (36). La media del colesterol LDL (mmol/l) en los pacientes incluidos es $\bar{X}_{incl} = 3,33$ y la de los pacientes excluidos es $\bar{X}_{excl} = 3,49$.

Detengámonos un segundo: la media muestral ¿tiene distribución? ¿Qué significa esto? La pregunta es si el «estadístico media muestral» es una constante o, por el contrario, se trata de una variable que debe ser caracterizada por su distribución. Es fácil imaginar que la media muestral varía de una muestra a otra, tomando diferentes valores.

Ejemplo 5.14



En el estudio anterior, si se obtienen dos muestras, las medias \bar{X} de colesterol LDL serán algo diferentes, incluso aunque se tratara de casos de la misma población.

Ejemplo 5.15



Seleccionamos al azar $n = 100$ pacientes con hipertensión y calculamos la media muestral o promedio de sus valores del colesterol LDL. Este valor será diferente si obtenemos otra muestra de 100 pacientes y volvemos a calcular su media.

Contraejemplo 5.16



Seleccionamos al azar un paciente y le preguntamos 100 veces por su edad (¡pobre hombre! ¡Qué paciencia! ¿Y qué pensará de nosotros?) y hacemos el promedio. Ahora, cabe esperar que obtengamos el mismo valor si calculamos este promedio en otras 100 preguntas sobre su edad (se asume que el paciente es muy paciente, claro). Al no haber variabilidad en la edad, la media tampoco varía.

Ejemplo 5.17



Al paciente anterior en lugar de preguntarle la edad se le determina 100 veces el colesterol LDL por procedimientos independientes. Ahora, las medias podrían variar por un posible error de medida. Nótese que, en este caso, la variabilidad que se cuantifica hace referencia a la población de las diferentes mediciones en un mismo paciente, no a la población de pacientes. Es decir, cuantifica la variabilidad intra-paciente en lugar de la entre-paciente.

Así, las medias irán variando de una muestra a otra. Si se desea utilizar el estadístico promedio como estimador del parámetro poblacional «esperanza», esta variabilidad inducirá a errores, lo que, por supuesto, nunca es deseable. Ahora bien, ¿se pueden cuantificar estos errores? O, lo que es más importante, ¿se puede limitar su

magnitud? Para responder a estas dos preguntas cruciales, se debe primero contestar a otras más sencillas:

1. ¿Alrededor de qué valor varían? (Es decir, ¿cuál es su centro?)
2. ¿Varían mucho o poco alrededor de este valor? (Es decir, ¿cuál es su dispersión?)
3. ¿Qué forma tiene su distribución?

Nótese que, una vez aceptado que el estadístico promedio o media muestral tiene una cierta distribución, las dos primeras preguntas se reducen a conocer la media poblacional y la varianza o la desviación típica del estadístico. Veámoslas sucesivamente.

Centro de la distribución del estadístico \bar{X}

El centro poblacional del estadístico media muestral \bar{X} recibe el nombre de *esperanza de \bar{X}* y se representa por $E(\bar{X})$. Se sabe que, si la muestra es aleatoria simple, la esperanza de \bar{X} coincide con la esperanza de X que ya denotamos por $E(X)$ o μ .

Fórmula

$$E(\bar{X}) = E(X) = \mu$$



Recuerde



El centro de \bar{X} coincide con el de X . Si vamos obteniendo MAS de 100 pacientes de sus niveles de colesterol LDL, la media de las medias de todas las muestras ($E(\bar{X})$) coincide con la media poblacional de los niveles de colesterol LDL ($E(X) = \mu$).

Ejercicio 5.10



¿Es deseable esta situación? ¿Qué utilidad puede tener este hecho?

En resumen, se sabe que el conjunto de todas las medias de todas las posibles muestras aleatorias se encuentra situado precisamente en el centro de la variable que se está estudiando. Si se utilizara la media de la muestra \bar{X} para conocer la media de la población $\mu = E(X)$, existiría el consuelo de que el conjunto de todas las posibles muestras «apuntan» en la dirección correcta. Los errores podrán ser tanto por exceso como por defecto. Pero dado que el centro de la variable \bar{X} coincide con el objetivo de la estimación, el conjunto de los posibles errores positivos y negativos está equilibrado. Por lo menos, los errores no tienen ningún «favorito».

No sesgo: primera propiedad de un estimador. \bar{X} es insesgado

Definición



Se dice que un **estimador** es **insesgado** si el centro de su distribución a lo largo de todas las posibles muestras coincide con el parámetro que está estimando.

Nota técnica

Demostración de que \bar{X} no tiene sesgo al estimar $E(X)$ en el caso de MAS.

$$E(\bar{X}) = E(\sum x_i/n) = [E(\sum x_i)]/n = [\sum E(x_i)]/n = [\sum E(X)]/n = [n E(X)]/n = E(X)$$

Ejemplo 5.18

Se desea estimar la evolución de los ingresos de los médicos colegiados. Cada año, se pregunta a 30 de ellos, seleccionados al azar, por su salario. Las estimaciones que se obtengan se distribuirán alrededor del auténtico valor. Si se repitiera el muestreo infinitas veces, la mitad de estas muestras tendría valores que inflarían la realidad y la otra mitad valores inferiores. Pero, en general, estaría siendo equilibrado.

Contraejemplo 5.19

No sería correcto extrapolar estos resultados a los enfermeros, cuyos salarios son inferiores. Si se hiciera, se estaría cometiendo un sesgo igual a la diferencia entre las medias de los salarios de ambos colectivos.

Analogía

Sean dos lanzadores con arco que apuntan a sus respectivas dianas (fig. 5-4). El lanzador de la izquierda tiene un sesgo hacia la izquierda y arriba, mientras que el de la derecha está centrado.

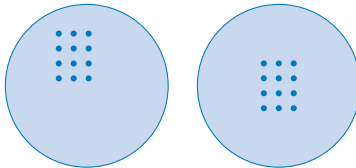


Figura 5-4 El arquero de la izquierda tiene sesgo, pero no el de la derecha.

Ejemplo 5.20

Dos informáticos han diseñado dos experimentos para conocer el rendimiento en tiempo de un nuevo algoritmo que proponen para decodificar el ADN. El primer informático analiza muestras del cromosoma 21, que es más corto, mientras que el segundo selecciona muestras de todos los cromosomas. Las posibles muestras del primero tendrán medias muestrales, cuyo centro, $E(\bar{X})$, estará por debajo de la media poblacional, μ . Las del segundo informático estarán centradas en la auténtica media poblacional (fig. 5-5).

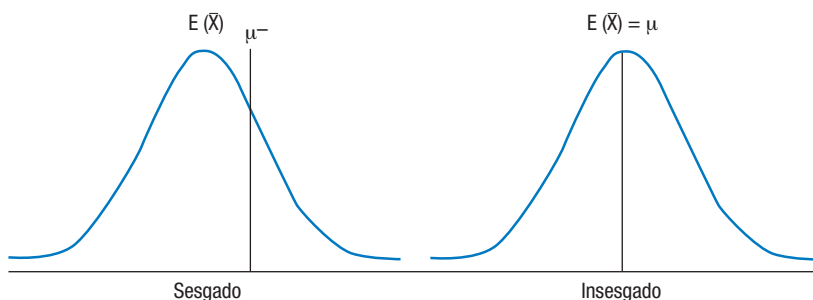


Figura 5-5 Estimador sesgado ($E(\bar{X}) \neq \mu$) e ins sesgado ($E(\bar{X}) = \mu$).

Recuerde



Si la muestra es aleatoria, el promedio muestral es ins sesgado.

Recuerde



Si la muestra NO es aleatoria, la estadística NO puede garantizar ausencia de error sistemático.

En el fondo, se está diciendo que podría ser peor, que podría ser que las estimaciones estuvieran apuntando en dirección incorrecta. Parece un pobre consuelo, ya que las estimaciones, aunque se distribuyan alrededor del parámetro de interés, no acaban de acertar. De hecho, si se ha trabajado con una variable continua (con tantos decimales como se desee), nunca coincidirá el valor estimado con el parámetro de interés: siempre habrá un error que no puede olvidarse.

Por ello, hay que preguntar: ¿es la magnitud de este error tolerable? Ello conduce a la siguiente pregunta: ¿cómo se puede medir o cuantificar este error? A continuación se estudia la información que la varianza de \bar{X} aporta sobre este error.

Dispersión de la distribución del estadístico \bar{X}

Ejercicio 5.11



¿Recuerda el cálculo de la media y de la varianza de una muestra?

REPASO: $S^2 = \sum_{i=1,n} (x_i - \bar{X})^2 / (n - 1)$

$S^2 = [\sum x_i^2 - (\sum x_i)^2 / n] / (n - 1)$ (más eficiente)

Practique ambas fórmulas, a mano y con la calculadora, para el ejemplo sencillo de $n = 5$ alumnos que contestan que en su familia son 1, 2, 3, 4 y 5 hermanos.

Varianza del estadístico media muestral: $V(\bar{X})$. ¿Cuánto vale? ¿Qué significa? Se sabe que la dispersión de \bar{X} es directamente proporcional a la dispersión de X e inversamente proporcional al tamaño n de la muestra.

Fórmula

$$V(\bar{X}) = V(X)/n$$

Nota técnica

Demostración en el caso de MAS.

$$V(\bar{X}) = V(\sum x_i / n) = [V(\sum x_i)] / n^2 = [\sum V(x_i)] / n^2 = [\sum V(x)] / n^2 = [nV(x)] / n^2 = V(x) / n$$

↑ independientes
 ↑ idénticamente distribuidas

Nótese que la relación entre la variabilidad de las medias muestrales, $V(\bar{X})$, y el tamaño muestral, n , es inversa. Ello implica que cuanto mayor sea el tamaño n de la muestra, menor será la variabilidad de las medias muestrales. Lo que coincide con todo lo que ya sabíamos: cuanto mayor es el tamaño de la muestra, mayor credibilidad tienen los resultados.

Lectura

Una sola flor no indica primavera.

Ejemplo 5.21

Supóngase que se están tomando muestras de la altura de niños. La variabilidad de las posibles medias muestrales será mayor si tomamos muestras de tamaño $n = 3$ que si son de tamaño $n = 1.000$.

Ejercicio 5.12

¿Es coherente esta situación?, ¿qué utilidad puede tener?

Contraejemplo 5.22

No se tendrá más información si se mide 1.000 veces al mismo niño. Para que una nueva observación aporte el máximo de información deberá ser lo más independiente posible de la información aportada por las observaciones previas.

Tampoco es una gran sorpresa que cuanto mayor sea la variabilidad de una variable, mayor será el grado de oscilación de la media de una muestra a otra.

Ejemplo 5.23



Supóngase que se están tomando muestras de la altura de niños. La variabilidad de las posibles medias muestrales será mayor si los niños tienen edades comprendidas entre los 5 y los 15 años, que si todos tienen 8 años.

Ejemplo 5.24



Suponga que los ingresos de los titulados de una facultad aumentan con el tiempo que pasa desde que dejan la universidad. Si es así, la dispersión de X será mayor si se estudia el conjunto de todos los titulados, que si se estudia solamente los titulados en un cierto año. En consecuencia, si se obtienen muestras de la población total, las medias muestrales, \bar{X} , fluctuarán más que las muestras de un solo curso.

Recapitulación: Se ha visto que la media muestral \bar{X} oscila de una muestra a otra alrededor del parámetro $\mu = E(X)$ que se desea estimar. ¿Qué significa la varianza de \bar{X} ? Si es el promedio de las distancias (cuadradas) con el centro, representa el error (cuadrado) que se cometería en cada muestra al estimar el parámetro poblacional a partir del valor de \bar{X} obtenido. Así, la varianza es el promedio de todos los posibles errores (cuadrados).

Recuerde



La varianza de \bar{X} proporciona el promedio de los errores al cuadrado.

Por lo tanto, la varianza de \bar{X} informa sobre el promedio del error «cuadrado». Dado que este cuadrado es incómodo a la hora de interpretar los resultados, de la misma manera que se obtenía la desviación típica de una variable mediante la raíz de su varianza, ahora, a través de la raíz de la varianza de \bar{X} , se obtiene el error típico de \bar{X} .

Nota técnica



Por error típico de \bar{X} se entiende el error esperado de la media muestral \bar{X} al estimar el parámetro μ .

FórmulaError típico de \bar{X}

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

Ejercicio 5.13

Suponga ahora que está interesado en conocer el promedio de hermanos de las familias de los alumnos del ejercicio 5.11. Si se considera a esta muestra de $n = 5$ como una muestra aleatoria representativa de todas las familias, ¿qué error cabe esperar que tiene la media observada en la muestra al estimar la media poblacional?

Recuerde

El error típico de la media \bar{X} es la desviación típica de la variable en estudio dividida entre la raíz del número de casos.

Ejemplo 5.25

La siguiente frase «los 100 niños tratados han tenido fiebre durante una media de 3 días; el error típico (o estándar) ha sido de 0,1 día» hace inferencia a los valores de la población: se afirma que, al aplicar este tratamiento en niños de las mismas características, la media de duración de la fiebre es de 3 días y que el error esperado al decir que la media poblacional es de 3 días, es de 0,1 día.

Debe quedar claro que se trata de un «error», por tanto con connotación negativa. Nótese que mientras el término desviación típica no debería tener ninguna connotación, ni positiva ni negativa, ahora el error típico ya deja claro desde el primer momento que se trata de algo negativo, no deseable: el error que cabe esperar que se cometa al estimar el parámetro media poblacional a partir del estimador media muestral.

Ejemplo 5.26

La altura de las mujeres adultas tiene una distribución Normal de media $\mu = 165$ cm, y desviación típica $\sigma = 7$ cm. Que la desviación típica sea de 7 cm no es ni bueno ni malo, simplemente refleja una situación natural. (Nótese que esta diversidad es, para un ecólogo, fuente de riqueza; mientras que para un fabricante de camisas representa un reto que

Ejemplo 5.26 (Cont.)

superar.) En cambio, si para estimar la altura media μ de las mujeres se calcula la media en una muestra de $n = 100$ mujeres, el error típico que conlleva la estimación es:

$$\sigma_{\bar{x}} = \sqrt{V(\bar{X})} = \sigma_x / \sqrt{n} = 7 \text{ cm} / \sqrt{100} = 0,7 \text{ cm}$$

Este valor del error típico dice que la imprecisión de \bar{X} al estimar μ es de 0,7 cm.

Nota

De la misma manera que \bar{X} es el estimador de μ , s es el estimador de σ y $S_{\bar{x}}$ el de $\sigma_{\bar{x}}$. Como en general no se conocerá σ , el error típico que se emplea es $S_{\bar{x}}$.

Ejercicio 5.14

Se estima en 4,4 puntos el incremento en la calidad de vida de la semana 0 a la 24 en 43 pacientes. Si la desviación típica observada ha sido de 1,2 puntos, ¿cuánto vale el error típico? Interprete el resultado.

El error *típico* habla del error esperado o promedio, ya que el error exacto que se comete en una muestra concreta permanece desconocido y puede ser más grande o más pequeño.

Nota

Formalmente no se puede interpretar el error típico como el promedio de los errores (es la raíz cuadrada del promedio de los errores cuadrados), pero a nivel práctico, decir que representa el error promedio o esperado es una buena aproximación.

Es muy importante resaltar que este error típico puede hacerse tan pequeño como se quiera: simplemente se trata de ir aumentando el tamaño de la muestra.

¿Desviación típica o error típico?

«¿Y qué debo utilizar, la desviación típica o el error típico?»

Esta pregunta no tiene razón de ser, ya que no son medidas alternativas para un mismo objetivo: la desviación típica es una medida descriptiva de cómo son los casos, mientras que el error típico es una medida del error asociado a un proceso inferencial. Así, se puede usar la desviación típica cuando se describen las condiciones y los casos en los que se ha hecho el estudio (al inicio de «resultados»); y se puede usar el error típico cuando se desea inferir (desde la muestra a la población) el efecto observado. Esta inferencia permitirá a otros científicos utilizar nuestros resultados.

Recuerde

La desviación típica es una medida de dispersión que sirve para describir los datos (¿cómo son mis casos?), mientras que el error típico es una medida del error de estimación al hacer inferencia (¿qué incertidumbre o ruido lleva asociado mi atrevido salto de la muestra a la población?).

Nota técnica

Desviación típica es «*standard deviation*» mientras que error típico es «*standard error*».

Lectura

«Ítem 15 (9): las variables continuas pueden resumirse para cada grupo a través de la media y de la desviación típica. Los errores típicos y los intervalos de confianza no son adecuados para describir la variabilidad; son más bien estadística inferencial que descriptiva.

Ítem 17 (9): resultados y estimación. Para cada variable principal y secundaria, un resumen de resultados para cada grupo y la estimación del tamaño del efecto y su precisión (p. ej., 95% de intervalo de confianza).

Es muy importante resaltar que la relación entre el tamaño de la muestra y el error típico no es inversamente proporcional. Sí que es inversa, pero hay una raíz de por medio. Por tanto, si aumentando el tamaño muestral se quiere disminuir el error típico a la mitad, se deberá multiplicar por cuatro el esfuerzo en recoger datos.

Ejemplo 5.27

El nivel de plaquetas en pacientes de una determinada enfermedad tiene una varianza de 2.500 unidades². Si, para conocer su valor medio, se obtiene una muestra de 25 pacientes, la varianza de la variable promedio \bar{X} en 25 pacientes vale:

$$V(\bar{X}) = V(X) / n = 2.500 \text{ u}^2 / 25 = 100 \text{ u}^2$$

En consecuencia, el error (cuadrado) que cabe esperar de nuestra observación es de 100 unidades (cuadradas). Asimismo, el error típico de la estimación de la media en los 25 pacientes es:

$$\sigma_{\bar{x}} = \sqrt{V(\bar{X})} = \sqrt{100 \text{ u}^2} = 10 \text{ u}$$

Es decir, el valor de la media que obtengamos estará afectado por un error típico de estimación de 10 unidades.

Ejemplo 5.27 (Cont.)

En cambio, si se aumenta la muestra de 25 a 100 casos, la varianza de la media muestral \bar{X} y el correspondiente error típico valen:

$$V(\bar{X}) = V(X) / n = 2.500 \text{ u}^2 / 100 = 25 \text{ u}^2$$

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \sqrt{25 \text{ u}^2} = 5 \text{ u}$$

A pesar de tener una muestra cuatro veces más cara, el ruido de estimación ha bajado a la mitad.

Recuerde

Si desea estimar un parámetro y dispone de un estimador insesgado, el error típico (SE: standard error) de este estimador le informa del error esperado al afirmar que el valor del parámetro poblacional coincide con el valor del estimador obtenido en su estudio.

Estabilidad del conjunto

El hecho de que el error típico de la media se vaya haciendo más pequeño a medida que aumenta el tamaño muestral indica una cierta estabilidad de los grupos que se contraponen a la variabilidad de las unidades.

Recuerde

La variabilidad de los individuos contrasta con la regularidad del conjunto.

Ejemplo 5.28

Pongamos que la probabilidad de nacer varón sea 1/2. El próximo nacimiento de Barcelona tiene esta probabilidad de ser varón. Pero, o bien será niño o bien será niña, pero no será mitad niño y mitad niña. En cambio, podemos tener la tranquilidad de que el próximo año nacerán alrededor de un 50% de niños y un 50% de niñas en Cataluña. No le pediremos a un político que elabore un plan de contingencia por si, por azar, durante unos años sólo nacen bebés de uno de los dos géneros.

Otro ejemplo similar lo tenemos con las quinielas o la lotería. A una persona en concreto o le toca o no le toca. La incertidumbre es absoluta. Pero el que organiza los juegos tiene la confianza de que le tocará a un cierto tanto por ciento. Y esta confianza será tanto mayor cuanto mayor sea el número de personas que apuestan.

Más propiedades de los estimadores

De la misma forma que la esperanza de \bar{X} sirvió para definir el concepto de estimador insesgado, se puede utilizar el error típico de \bar{X} para definir propiedades interesantes de los estimadores.

Definición



Se dice que un estimador es **convergente** si, a medida que crece el tamaño de la muestra, se va acercando progresivamente al valor del parámetro que se desea estimar.

Ejemplo 5.29



\bar{X} es un estimador convergente ya que $V(\bar{X})$ disminuye a medida que aumenta n .

No se insiste más en esta definición que resalta la deseada propiedad de que cuanto mayor sea la muestra, mejor será la estimación.

Definición



Entre dos estimadores insesgados, se dice que es más **eficiente** el que tiene un error típico menor.

Analogía



Imaginemos dos lanzadores con arco, ambos insesgados. El de la izquierda tiene una mayor dispersión alrededor de la diana, por lo que es menos eficiente (fig. 5-5).

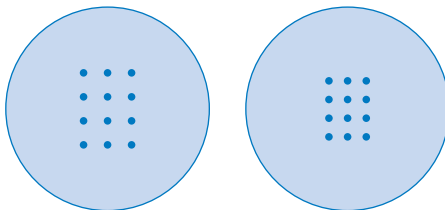


Figura 5-5 Ambos arqueros son insesgados, pero el de la derecha es más eficiente porque su error es más pequeño.

Nota técnica

El término **eficiente** tiene una clara connotación económica, en el sentido de que su relación calidad / coste será más favorable. El estimador más eficiente proporciona más información (tiene menos ruido o error aleatorio) para un mismo tamaño muestral (mismo coste); o también puede obtener la misma cantidad de información con una muestra más pequeña (menor coste).

Ejemplo 5.30

Dos investigadores han diseñado dos experimentos para comparar la biodisponibilidad de dos preparaciones farmacéuticas alternativas. El primero ha obtenido dos muestras de voluntarios, administrando a cada una, una de las dos formulaciones. Luego compara las medias de las dos muestras ($\bar{X}_1 - \bar{X}_2$). El segundo investigador ha recogido la información de las dos preparaciones en un único grupo, y calcula la media de las diferencias (\bar{X}_D), eliminando, de esta forma, la variabilidad debida al voluntario. Los gráficos muestran que, siendo ambos experimentos insesgados, el segundo es más eficiente (fig. 5-6).

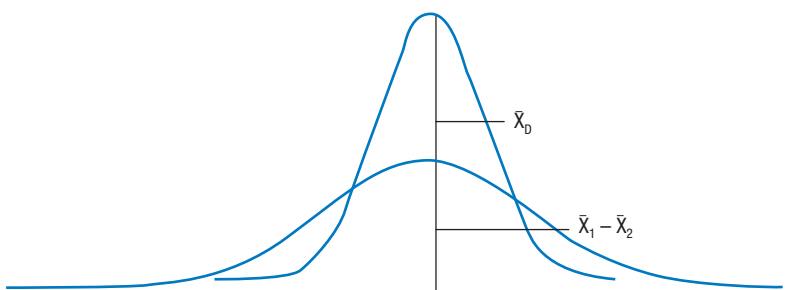


Figura 5-6 Ambos estimadores, $\bar{X}_1 - \bar{X}_2$ y \bar{X}_D , son insesgados pero \bar{X}_D es más eficiente.

Comentario

La estadística permite cuantificar los errores aleatorios. Si conviene que la MAS (Muestra Aleatoria Simple) deje de ser «simple», ningún problema, un profesional de la estadística le ayudará a obtener correctamente el valor del error típico de estimación. Pero si la muestra deja de ser «aleatoria», debe realizarse un detallado estudio sustantivo y estadístico de las consecuencias de las desviaciones de la aleatoriedad. Por estudio sustantivo se entiende el que emplea los conocimientos de la disciplina subyacente, sean clínicos, biológicos, epidemiológicos, etc.

Estimación puntual

Definición



Al valor observado de un estimador en una muestra se le denomina estimación puntual del parámetro.

Ejemplo 5.31



En una muestra aleatoria de 9 personas, la presión arterial sistólica (PAS) ha tenido una media muestral \bar{X} igual a 120 mmHg y una desviación típica muestral (S) de 12 mmHg. Así, la estimación puntual de la PAS media en esta población ha sido de 120 mmHg. El error típico de esta estimación se puede cifrar en:

$$S_{\bar{x}} = S / \sqrt{n} = 12 \text{ mmHg} / \sqrt{9} = 4 \text{ mmHg}$$

Por lo tanto, hay una señal de 120 mmHg que está afectada por un ruido de 4 mmHg.

Recuerde



El error típico informa del error esperado o promedio, pero el error exacto en una muestra concreta permanece desconocido, pudiendo ser inferior o superior.

Ejercicio 5.15



El descenso de la PAS tras la administración de un fármaco en una muestra de 16 pacientes ha tenido una media de 12 mmHg y una desviación típica de 8 mmHg. Calcule el error típico e interprete los resultados.

Ejercicio 5.16

Si hubiera deseado que el error típico hubiera sido de 1 mmHg, ¿cuántos casos hubiera necesitado (desviación típica de 8 mmHg)?

Forma de la distribución del estadístico \bar{X}

Ya se ha dicho que las posibles medias muestrales \bar{X} se distribuyen alrededor de la media poblacional $\mu = E(X)$ con una distancia promedio que cuantifica el error típico $\sigma_{\bar{x}}$. Ahora bien, ¿qué forma tiene la distribución de \bar{X} ?

Por las leyes de combinatoria y probabilidad, en general la muestra contendrá valores próximos a la media poblacional. También es posible obtener valores alejados, si bien será menos frecuente, ya que, por ejemplo, la probabilidad de obtener una muestra aleatoria que sólo contenga valores muy superiores a la media, y resulte en un valor positivo extremo de \bar{X} , es pequeña. En realidad, cuanto más se aleje una \bar{X}

del valor de la μ , menos probable es observarla. Todos estos razonamientos ilustran que \bar{X} se distribuye con una forma de campana o de montaña como la de la ley normal de Gauss-Laplace.

Recuerde



La media muestral \bar{X} se distribuye siguiendo la ley Normal Gauss-Laplace.

La distribución Normal aparece en variables que sean el resultado de muchos factores o fuerzas que actúen independientemente y con influencias similares. Y eso es precisamente lo que es una media muestral \bar{X} , ya que cada observación de la muestra contribuye con el mismo peso o influencia. Queda por aclarar qué significa «muchos»: ¿cuántos casos se necesitan para que la distribución del promedio de una muestra se acerque a la ley normal?

Ejercicio de Navegación



Entre en la página que se indica a continuación y observe, con la ayuda de la simulación que realiza la aplicación, cómo se comporta la media de una muestra a lo largo de muchas muestras y compruebe cómo se acerca la media muestral \bar{X} a la poblacional μ al crecer n .

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

Nota



Este número puede ser muy inferior si la distribución original de la variable que se está estudiando ya se aproxima a la forma de la distribución normal. De hecho, si X es normal, para $n = 1$, \bar{X} ya es normal.

Nota técnica



El teorema del Límite Central (TLC) establece que si se toman muestras de tamaño n , de una población de media μ y desviación típica σ , a medida que crece n , la distribución de \bar{X} se aproxima a la ley Normal con media μ y desviación típica σ/\sqrt{n} .

Ejercicio 5.17



A partir de lo dicho, ¿la distribución de la variable \bar{X} cambia de forma cuando crece el tamaño de la muestra? ¿Y la de X ?

Ejercicio 5.18

¿Cómo cambia la forma de la distribución de la variable \bar{X} cuando crece el tamaño de las muestras?

Ejemplo 5.32

La edad de los pacientes incluidos en un estudio sigue una distribución uniforme (aplanada, con el mismo número de casos en todas las franjas de edad). Si se toman muestras de tamaño $n = 30$ y se calcula la media muestral \bar{X} de la edad, su distribución también será, a niveles prácticos, normal.

Ejemplo 5.33

La presión arterial sistólica en los adultos sanos tiene una distribución que se asemeja bastante a la ley normal. Si se toman muestras de tamaño $n = 3$ y se calcula la media muestral \bar{X} de la PAS, la distribución de esta media será, a niveles prácticos, normal.

Recuerde

Las condiciones para poder creer que el promedio obtenido en una muestra sigue una distribución normal son, o bien muestra grande ($n \geq 30$) o bien distribución normal de la variable en estudio.

Aplicación práctica de la distribución de la media muestral \bar{X}

Se ha visto que la distribución normal permite construir intervalos que contengan un determinado porcentaje de unidades o casos. Ahora, la variable en estudio es \bar{X} , por lo que, utilizando la distribución normal, se pueden construir intervalos que contengan un deseado porcentaje de las medias que se podrían obtener en todas las posibles muestras.

Ejemplo 5.34

La glucosa en sangre (X) sigue una distribución normal de media 100 y desviación típica 10

$X \rightarrow N(100\text{mg/ml}, 10\text{mg/ml})$

Se desean construir los siguientes intervalos:

- i) el intervalo que contiene el 95% de las unidades de la población;
- ii) el intervalo que contiene el 95% de las posibles \bar{X} de muestras de tamaño $n = 9$; y
- iii) el intervalo que contiene el 95% de las posibles \bar{X} de muestras de tamaño $n = 100$.

Los tres intervalos son similares en el sentido de que deben contener el 95% de las unidades y dejar fuera el 5% ($\alpha = 0,05$).

Ejemplo 5.34 (Cont.)

Y son distintos en el sentido de que se refieren a unidades totalmente diferentes, con distribuciones también diferentes. En el primer intervalo, las unidades son individuos, pacientes o casos, mientras que en el segundo y tercer ejemplo se trata de medias muestrales que se podrían obtener si se repitiera indefinidamente el proceso de tomar muestras de $n = 9$ y $n = 100$ de estos individuos.

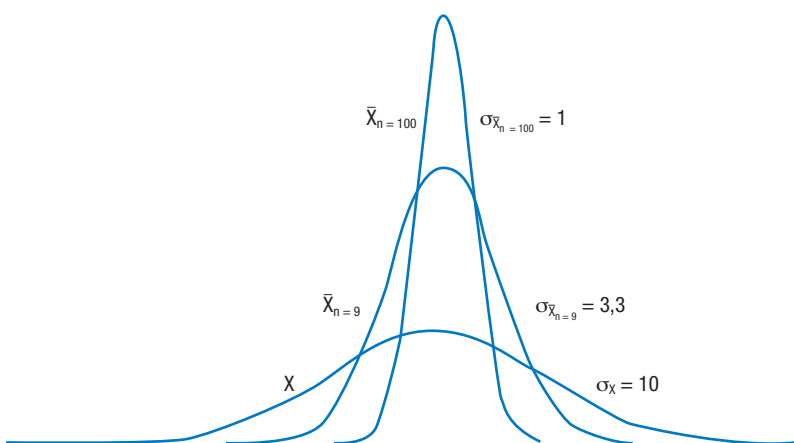


Figura 5-7 Distribución de las variables X , $\bar{X}_{n=9}$ y $\bar{X}_{n=100}$.

Ejemplo 5.34 (Cont.)

Todas estas distribuciones (fig. 5-7) seguirán la ley normal: al ser normal la distribución de la glucosa en los casos, también lo es la distribución de la media \bar{X} , sea cual sea el número de casos. Todas tienen, también, la misma media. Pero cambia la dispersión: para el primer ejemplo, se trata de la desviación típica de la variable original, la glucosa en sangre, 10 mg/ml; mientras que para los restantes dos ejemplos, se trata del error típico, debiéndose dividir la desviación típica por la raíz del número de casos respectivos:

- i) $X \rightarrow N(100 \text{ mg/ml}, 10 \text{ mg/ml})$
- ii) $\bar{X}_{n=9} \rightarrow N(100, 10/\sqrt{n}) = N(100, 10/\sqrt{9}) = N(100 \text{ mg/ml}, 3,33 \text{ mg/ml})$
- iii) $\bar{X}_{n=100} \rightarrow N(100, 10/\sqrt{n}) = N(100, 10/\sqrt{100}) = N(100 \text{ mg/ml}, 1 \text{ mg/ml})$

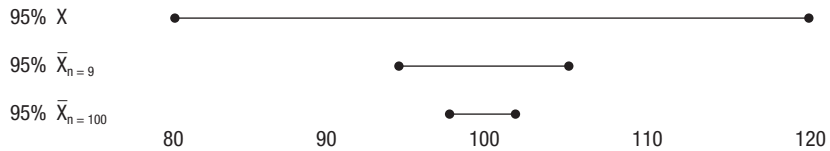


Figura 5-8 Intervalos que contienen el 95% de las observaciones de las variables X , $\bar{X}_{n=9}$ y $\bar{X}_{n=100}$.

Ejemplo 5.34 (Cont.)



Los límites de los intervalos (fig. 5-8) se pueden calcular utilizando las tablas de la normal tipificada: $Z_{0,025} = 1,96$

i) Intervalo que contiene el 95% de las glucemias individuales, X

$$\mu \pm Z_{\alpha/2} \sigma = 100 \pm 1,96 \cdot 10 = 100 \pm 19,6 = [80,4, 119,6]$$

ii) Intervalo que contiene el 95% de las medias ($\bar{X}_{n=9}$) de las infinitas muestras de $n=9$ individuos

$$\mu \pm Z_{\alpha/2} \sigma/\sqrt{n} = 100 \pm 1,96 \cdot 10/3 = 100 \pm 6,53 = [93,47, 106,53]$$

iii) Intervalo que contiene el 95% de las medias ($\bar{X}_{n=100}$) de las infinitas muestras de $n=100$ individuos

$$\mu \pm Z_{\alpha/2} \sigma/\sqrt{n} = 100 \pm 1,96 \cdot 10/10 = 100 \pm 1,96 = [98,04, 101,96]$$

Nota



El ejemplo anterior muestra, una vez más, la mayor variabilidad de las muestras de menor tamaño, ya que la distribución de los casos individuales puede considerarse como la distribución de las muestras de tamaño $n = 1$.

Este ejemplo muestra cómo, utilizando la distribución normal, se puede conocer, a partir de los valores poblacionales de la media [$E(X)=\mu$] y la varianza [$V(X)$], dónde estarán los valores de la media muestral (\bar{X}) en el 95% de las posibles muestras. Puede ser interesante, pero tiene poca utilidad práctica, ya que el problema habitual es, conocidos los estimadores muestrales de la media (\bar{X}) y la varianza (S^2), ¿qué se sabe de la media poblacional $E(X) = \mu$? Esta interesante pregunta se contesta en el siguiente capítulo.

Ejercicio 5.19



Los psicólogos miden la inteligencia por el CI, que sigue, en la población adulta general, una $N(100,15)$. Es decir, $\mu = 100$ y $\sigma = 15$. Si se recolectaran muchas muestras de tamaño $n = 9$ y en cada muestra j se calculara su media \bar{X}_j :

Ejercicio 5.19 (Cont.)

- a) ¿Cómo variarían las medias \bar{X}_j de las muestras?
- b) En una facultad de Medicina se ha recogido una muestra de tamaño $n = 9$ y se ha observado $\bar{X}_j = 104$, ¿se trata de 1) un ejemplo aceptablemente típico; o 2) especialmente afortunado, muy cerca de μ ; o 3) tan raro y alejado que se sospecha que estos alumnos no son de aquella población?
- c) Repetir los dos apartados anteriores, pero con $n = 25$ y $n = 225$.

Soluciones a los ejercicios

5.1 Debe desearse tener información sobre el centro y sobre la dispersión de la distribución, sobre ambos. Usualmente, la comunidad científica enfoca su interés en los valores centrales para poder resumir o representar los casos. Pero hay que hacer el esfuerzo de cuantificar también el grado de dispersión, e incluso conocer la forma de la distribución.

5.2 Cualquier ejemplo es válido. También sería terriblemente aburrido «decir toda la verdad» sobre la carga viral. Nótese, en cambio, que no lo sería sobre el género («53 fueron del género masculino y 47 del femenino»). Y quizá tampoco sobre el número de infecciones oportunistas («2.523 casos no presentaron ninguna; 48 tuvieron una; 7, dos y 1 caso, tres infecciones»).

5.3 En el estudio de los odontólogos, una unidad es una visita a la consulta, mientras que en la población general, una unidad es un habitante. Como hay habitantes que van al dentista más veces que otros, los primeros estarán sobrerrepresentados en un estudio en el que se seleccionen «visitas». Nótese que aquellos que nunca van al dentista tienen una probabilidad nula de ser seleccionados. En definitiva, el estudio de los dentistas ofrecerá cifras más altas que el de la población general.

5.4 Los estudios de satisfacción hospitalaria se basan en las altas hospitalarias, mientras que el estudio del defensor se basó en ciudadanos. Igual que en el ejercicio anterior, aquellos ciudadanos que van menos a los centros públicos tienen una probabilidad menor de ser seleccionados. En resumen, la satisfacción entre las «altas» hospitalarias será mayor que la satisfacción de los habitantes.

5.5 Lo mismo que antes: si los pacientes que desarrollan infecciones nosocomiales permanecen más tiempo en el centro y la selección se hace a partir de las «camas ocupadas», las cifras de infección serán mayores.

5.6 Por ejemplo, en un estudio del perfil lipídico en 41 pacientes con HIV tratados, la media del colesterol total fue de 4,51 mmol/l. Se trata de la media obtenida en la muestra (promedio muestral), y que informa sobre el valor de la media en la población (esperanza poblacional), que es desconocida.

5.7 La muestra serían los casos seleccionados para el estudio; la población muestreada, aquella de la cual se hubiera obtenido por meros mecanismos aleatorios, la muestra; la población objetivo, todos los fumadores de esa región sanitaria; y la población externa, los fumadores de cualquier región sanitaria.

5.8 Se trata del conocido sesgo de publicación: al publicar sólo lo que ha resultado significativo, se da menor oportunidad a difundir resultados «negativos». (Este argumento se complementará en el capítulo 7.)

5.9 No, si se pretende utilizar los resultados para centros de atención terciaria. Pero sí, si se desea utilizarlos en centros donde la severidad de los casos tratados pueda ser menor.

5.10 Si se desea utilizar a la media muestral (\bar{X}) para acercarse a la media poblacional ($E(X)$ o μ) es bueno que la distribución de \bar{X} se disponga alrededor del auténtico valor de μ . Dicho al revés: sería peor que se distribuyera alrededor de cualquier otro valor. Y cuanto más alejado de μ estuviera este otro valor, peor.

5.11 Media $\bar{X} = \sum_i x_i / n = (1+2+3+4+5) / 5 = 3$

Varianza $S^2 = \sum_{i=1,n} (x_i - \bar{X})^2 / (n - 1) =$

$$= [(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2] / 4 = 10/4 = 2,5$$

$$S^2 = [\sum x_i^2 - (\sum x_i)^2 / n] / (n - 1) =$$

$$= [(1^2+2^2+3^2+4^2+5^2) - (1+2+3+4+5)^2/5] / 4 = [55 - 225/5] / 4 = 10/4 = 2,5$$

$$S = \sqrt{S^2} = \sqrt{2,5} \approx 1,58$$

Es decir, la media muestral es 3 hermanos; la varianza muestral, 2,5 hermanos² y la desviación tipo muestral es de aproximadamente 1,6 hermanos. Podemos imaginar que la distancia (o desvío) de una familia típica con la media es de 1,6 hermanos.

5.12 Es coherente: cuantos más casos se tiene, se dispone de más información y, por tanto, hay menos error aleatorio y menos riesgos del muestreo. Es útil (y, por consiguiente, deseable) en el sentido de que un mayor esfuerzo en la recolección de datos se ve recompensado por una menor oscilación de las estimaciones.

5.13 $S_{\bar{X}}^2 = S^2/n = 2,5/5 = 0,5$; $S_{\bar{X}} = \sqrt{S_{\bar{X}}^2} = \sqrt{0,5} \approx 0,707$

Si se afirma que la media de la población es de 3 hermanos, el error esperado al hacer esta afirmación es de 0,7 hermanos.

5.14 $S_{\bar{X}} = \sqrt{S_{\bar{X}}^2} = S/\sqrt{n} = 1,2/\sqrt{43} = 0,18$.

La señal obtenida ha sido 4,4 y el error asociado que lleva esta señal es de 0,18.

5.15 $S_{\bar{X}} = \sqrt{S_{\bar{X}}^2} = S/\sqrt{n} = 8/\sqrt{16} = 2$.

La señal obtenida ha sido de 12 mmHg y el error asociado que lleva esta señal es de 2 mmHg.

5.16 Si se desea que $S_{\bar{X}} = 1 \rightarrow 8/\sqrt{n} = 1 \rightarrow n = 8^2 = 64$.

Una vez más se ve que si se desea reducir a la mitad el ruido de la muestra anterior, hay que multiplicar por 4 el tamaño muestral ($64 = 16 \cdot 4$).

5.17 A medida que crece el tamaño muestral, lo que va cambiando de forma es la distribución de la variable media muestral \bar{X} . La distribución de los valores observados, es decir, lo que se llama la distribución de la variable en estudio es siempre la misma para todos los casos, haya 3, 50 o 1.000. Si no tiene claro que lo que va cambiando es la distribución de la media muestral \bar{X} repita los ejercicios de la última navegación.

5.18 Su centro, la esperanza, no cambia, pero sí que lo hace la dispersión y también puede hacerlo la forma. La dispersión, cuantificada por el error típico, se va haciendo más pequeña a medida que crece el tamaño muestral (la reducción es proporcional a \sqrt{n}). La forma, en el caso de variables que no siguen una distribución normal, se aproxima cada vez más a la de esta distribución (en el caso de variables que siguen la ley Normal, ya tiene esta distribución para cualquier n).

5.19 a) $V(\bar{X}) = V(X) / n = 15^2/9 = 25 \text{ u}^2 \rightarrow S_{\bar{X}} = 5 \text{ u}$

La variabilidad de las medias muestrales es la tercera parte de la variabilidad de la variable

b) $[\bar{X}_i - E(X)]^2 = [104 - 100]^2 = 4^2 \text{ u}^2$ cifra «aceptablemente típica»

ya que su valor esperado era 5^2 u^2

c) Si $n = 25$; $V(\bar{X}) = V(X) / n = 15^2/25 = 9 \text{ u}^2 \rightarrow S_{\bar{X}} = 3 \text{ u}$

$[\bar{X}_i - E(X)]^2 = [104 - 100]^2 = 4^2 \text{ u}^2$ cifra «aceptablemente típica»,

ya que su valor esperado es 3^2 u^2

Si $n = 225$ $V(\bar{X}) = V(X) / n = 15^2/225 = 1^2 \text{ u}^2 \rightarrow S_{\bar{X}} = 1 \text{ u}$

$[\bar{X}_i - E(X)]^2 = [104 - 100]^2 = 4^2 \text{ u}^2$ no es una cifra «aceptablemente típica»,

ya que su valor esperado es 1 u^2 (se sospecha que éstos no son de aquella población).