# Muestreo

Vicente Manzano Arrondo – 2012-2014

## Muestreo cómo y para qué

Existe una multitud de situaciones donde no es posible o no es recomendable estudiar directamente la población y decidimos utilizar una muestra. Imagina, por ejemplo, que para valorar tu estado de salud se decidiera analizar *toda* tu sangre. Muy posiblemente la conclusión sería “¡Qué lástima que se nos fue! ¡Su estado de salud era excelente!”. En muchas ocasiones, abordar la población implica un esfuerzo descomunal, un gasto imposible, un tiempo no disponible o una pérdida de control excesiva. Sea como fuere, la práctica de obtener información sobre *toda* la población estudiando únicamente una parte o *muestra* exige hacer las cosas de tal modo que a la hora de las conclusiones podamos llevarlas a cabo con suficiente credibilidad.

Hay tres criterios para diseñar muestras. Las concreciones en la práctica surgen de la combinación, en alguna medida, de esos criterios:

- Conveniencia. Es lo más habitual, sin lugar a dudas. Se trata de estudiar las unidades que están al alcance, más a la mano o disponibles. Se comenta mucho que la psicología es una

ciencia creada fundamentalmente a partir de muestras de conveniencia de estudiantes universitarios: los investigadores en psicología construyen ciencia a partir de los estudiantes de psicología que asisten a clase. La conveniencia no es un mal criterio cuando lo que estudiamos es común a todos los individuos de la categoría que nos interesa. Por ejemplo, si queremos observar cómo funciona la gravedad, nos da lo mismo tirar desde un balcón una grapadora verde o una cuchara azul. El objeto concreto o su color no son relevantes sino el hecho de ser dos cuerpos con masa. Si vamos a estudiar el reflejo patelar, da igual que trabajemos con estudiantes sevillanos que con ejecutivos gallegos. Son personas y el reflejo viene con el resto de la herencia fisiológica. Pero si estudiamos aspectos relacionados con la opinión, con actitudes o con conocimientos, entre otros muchos ejemplos, el muestreo de conveniencia es una mala decisión.

- Rasgo. Habitualmente intentamos que la muestra responda a una serie de características o rasgos que consideramos relevantes. Por ejemplo, si estudiamos las actitudes de la población frente al sexismo en el lenguaje, es importante tener en cuenta variables relevantes como el sexo de la persona entrevistada o su

nivel de estudios, puesto que son características que suelen estar asociadas con las actitudes sexistas. Una forma de controlar estas variables, de tal forma que entrevistar a la muestra no genere resultados sesgados (es decir, con una tendencia que aleje las conclusiones de lo que ocurre realmente), es provocar que las distribuciones de estas variables sean iguales en la muestra que en la población. Así, por ejemplo, si hay un 51% de mujeres y un 49% de hombres, esos serán también los porcentajes en la muestra.

- Azar. El control mediante rasgo es un hábito saludable pero que complica la vida. Tener esas características bajo control hace que el diseño sea más difícil de llevar a la práctica. Esta norma es tanto más evidente cuanto mayor sea el número de variables que tenemos en cuenta. Una forma de control que suponemos que funciona más o menos con todo es el azar. Ocurre así porque, por definición, el azar no está relacionado con nada. Si lanzamos una moneda al aire es tan probable obtener cara como cruz. Si decimos a un grupo de 100 personas que lancen la moneda, surgirán dos grupos de aproximadamente 50 personas. Dentro de cada uno de ellos observaremos aproximadamente la misma cantidad de mujeres, de personas con

alto nivel de estudios, de gente que calza menos de un 40, de seguidores de Real Betis Fútbol Club, etc. Aplicado este principio al muestreo, lo que hacemos es seleccionar unidades (personas, familias, tiendas, momentos en el tiempo, etc.) al azar. Esto garantiza *aproximadamente* que no existirá sesgo. Es probable (así es el azar) que diversas características no se repartan del mismo modo en la muestra que en la población. Esto ocurre seguro en algunas variables. Lo que deseamos es que este desequilibrio no toque a las variables que tienen efecto sobre nuestro objeto de estudio. Por si las moscas, lo más habitual es realizar un diseño de muestreo donde se controlan expresamente algunos rasgos y el resto se le deja al azar.

Imagina que debes hacer una investigación sobre algunos comportamientos en el interior de los vehículos cuando se encuentran parados ante un semáforo en rojo. Es imposible observar a todos los vehículos que transitan por todos los semáforos durante todo el tiempo. Acotamos la población de estudio a tu ciudad y durante una semana que el equipo ha considerado digna del calificativo “normal”. Así que debes observar el comportamiento de conductores que esperan a que un semáforo se ponga

verde, en tu ciudad, durante una semana concreta. Un muestreo de conveniencia implicaría escoger, por ejemplo, el semáforo que hay junto a tu casa y observar durante una tarde que te venga bien. Un muestreo por rasgo implica 1) identificar las variables que el equipo de investigación considera más relevantes, como puede ser el tipo de vía en la que se encuentra el semáforo o los momentos de una jornada en el que el comportamiento de conducción varía, 2) considerar uno o dos semáforos en cada tipo de vía y uno o dos momentos de observación en cada sección de tiempo, 3) observar en todos esos lugares y momentos previstos. Un muestreo al azar supone meter en un bombo a todos los semáforos de la ciudad y escoger al azar solo algunos; y hacer otro tanto con varias porciones de la jornada. En un muestreo mixto, agrupamos los semáforos por tipos, pongamos cinco o seis, según criterios como tipo de vía, visibilidad, zona de la ciudad, etc. Al azar, escogemos unos pocos semáforos dentro de cada categoría; y hacemos otro tanto con los momentos del día. Los buenos muestreos más habituales suelen ser mixtos.

## Limitaciones

Como todo en esta vida, el muestreo tampoco es la panacea. Las personas solemos perdernos entre los árboles de tal forma que se nos va la visión del bosque.

Un instrumento no es un fin. Cuando se trabaja con una muestra, hemos sustituido a la población, pero no hemos solucionado todos los problemas de representación. Observa la figura 1.

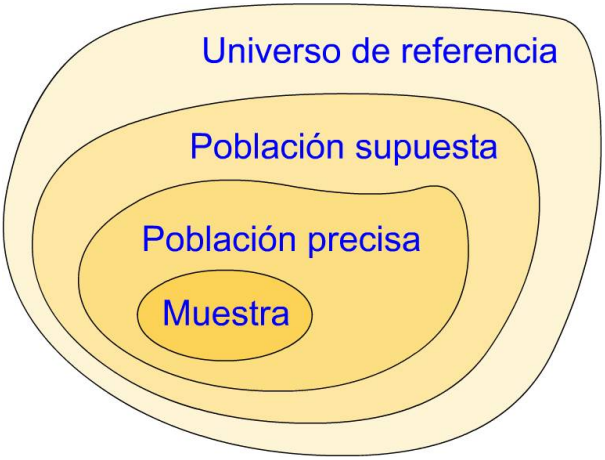


Figura 1. Muestreos anidados.

Imagina que un equipo de investigación quiere saber cómo puede estimular la imaginación de los niños y niñas con edades comprendidas entre 5 y 10 años. Gracias a diversos estudios realizados hasta la fecha, este equipo cuenta con información, conocimientos y herramientas. Diseña algunas intervenciones y algunos procedimientos para valorar el efecto de esas intervenciones. Decide aplicarlo en veinte colegios de enseñanza primaria de distintos

puntos de la Comunidad Autónoma de Andalucía. Entre las mañanas de los días 12 y 15 de noviembre, los entrevistadores pasan por las aulas y recogen la información.

No entremos ahora en la calidad de la recogida de datos. Lo que nos interesa es su fuerza para representar a todos los datos que nos interesan a partir de cómo se ha conseguido la muestra. El foco de interés es el conjunto de niños y niñas de 5 a 10 años. No hay acotación temporal ni espacial. Se entiende que se trata de todo el mundo con esas características explícitas, viva dónde, cómo y cuándo viva. Es el universo de referencia. Estos universos son lógicamente inabarcables. A ellos se les aplica una especie de muestreo de experto basado en una mezcla de conocimiento, confianza y resignación. Tras ese muestreo de experto aterrizamos en la población supuesta: los niños y las niñas de 5 a 10 años que están escolarizados en colegios de enseñanza primaria de la Comunidad Autónoma Andaluza. Fíjate que hemos reducido mucho el universo de referencia. Las conclusiones del estudio se publicarán con ánimo de ser válidas para todos los niños y niñas, aunque la población que se investigó en la práctica no sea exactamente esa. Pues bien, suponer que de esa población de colegios andaluces ha surgido la muestra es mucho suponer. Se trata de la población supuesta, pero no de la precisa. Durante esos días del 12 al 15

de noviembre la población que existía no ha vuelto a existir. En sentido preciso desapareció. Es posible que los aspectos en los que ha variado no sean trascendentes (tal vez la media de longitud de cabello no sea la misma durante esos días que después). O es posible que sí. Lo cierto es que las poblaciones, en sentido estricto, tienen una duración infinitésima, mueren nada más nacer. Nuestra muestra ha surgido de una población que ya no existe pero que, basándonos en otra mezcla de conocimiento, confianza y resignación, suponemos que sigue siendo razonablemente la misma a efectos de las conclusiones de nuestro trabajo.

En sentido estricto, la muestra ha sido obtenida de la población precisa, es decir, de una población que ya no existe y que representa a un extenso ideal que es el universo de referencia. Esto es inevitable. No hay solución. Lo que sí tiene solución es nuestra credulidad, creer que nuestra muestra es una representación indiscutible e infalible del universo de referencia.

Siendo conscientes de estas limitaciones, sigamos.

## Modelos de muestreo

Hay para muchos gustos y posibilidades. En estadística solo podemos utilizar un tipo de muestreo,

el aleatorio. Esto es así porque la inferencia, es decir el camino de vuelta, depende del tipo de muestreo, es decir el camino de ida. Sabemos, por ejemplo, que alguien siente debilidad por los bombones blancos y le pedimos que escoja diez bombones de una caja donde hay de varios tipos, sin que veamos lo que hace. Entonces podemos suponer razonablemente, que los bombones que escogerá serán los blancos. Esta inferencia está basada en el conocimiento que tenemos sobre cómo hace elecciones esa persona.

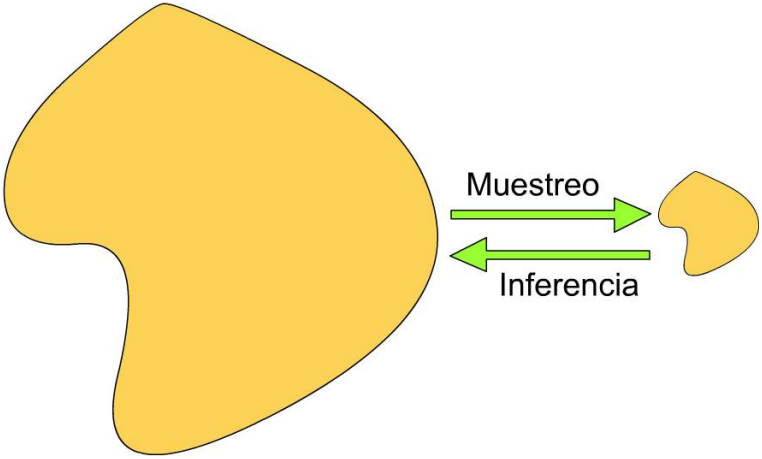


Figura 2. Caminos de ida y vuelta.

En el análisis de los datos suponemos que las muestras han sido aleatorias, es decir, se han generado tirando de probabilidades conocidas. Por este

motivo podemos hacer inferencia estadística, es decir, una inferencia que considera esas probabilidades para hacer deducciones sobre el conjunto amplio (la población precisa, más comúnmente conocida como población) de donde proviene la muestra.

Los modelos que tenemos en cuenta en análisis de datos son aleatorios. Pero de estos también hay muchos.

- El más sencillo o simple es el denominado, asombrosamente, *muestreo aleatorio simple*. No tiene mucho que explicar. Es meter a las *N* unidades en un bombo, moverlo y extraer las *n* unidades que salgan una a una.
- Un muestreo muy habitual es el denominado *de conglomerados*. Ocurre que las poblaciones suelen estar organizadas o agrupadas de algún modo. Buena parte de los niños y niñas de occidente van al colegio. Cada colegio puede ser considerado como un conglomerado, término obtenido de la geología y que designa porciones de terreno con materiales diversos. Un conglomerado es como una población en pequeñito, aunque más bien una población con algún sesgo. Así, por ejemplo, hay colegios privados y otros públicos, de ámbito rural o urbano, mixtos o separados por sexos, no confesionales o religiosos, etc. No podemos suponer que un colegio privado de ciudad y

religioso, por poner un ejemplo concreto, representa a toda la población infantil. En el muestreo de conglomerados, cada conglomerado es como si fuera una unidad en un muestreo aleatorio simple de conglomerados. Lo que se seleccionan son conglomerados enteros, tomando todas las unidades que están en ellos.

- En el muestreo *estratificado* también se consideran agrupaciones o subconjuntos de la población, pero no tienen por qué estar físicamente delimitadas. Los estratos se toman también del lenguaje geológico: son porciones de terreno con materiales homogéneos. El muestreo estratificado es fácil de distinguir del de conglomerados porque en este segundo se seleccionan unos conglomerados enteros y al resto se le ignora, mientras que en el estratificado se consideran *todos* los estratos, pero no se seleccionan enteros (entonces no habría muestreo) sino una porción de cada uno. Por ejemplo, es habitual considerar al sexo como una variable de estratificación: hombres y mujeres. El muestreo estratificado por sexos implica tomar una muestra de hombres y otra de mujeres. De este modo se garantiza que ambos estratos están presentes en la muestra.

No vamos a practicar la exhaustividad. Hay más modelos y, muy especialmente, combinaciones de los anteriores. Pero el objetivo era romper el hielo con este asunto. Lo que nos ocupa a partir de este momento es únicamente el Muestreo Aleatorio Simple (MAS). Es el más habitual, conocido, sencillo, implementado en todo tipo de programas estadísticos y el que sustenta buena parte de los desarrollos en estadística. Si cambiáramos de modelo de muestreo, deberíamos variar las expresiones de cálculo para hacer inferencias.

## Símbolos

Cuando se está haciendo referencia a dos conjuntos de datos (en este caso, muestra y población) es importante utilizar símbolos que nos permitan distinguir entre ellos y evitar confusiones. El recurso al que se acude es expresar con letras latinas todo lo referente a la muestra, mientras que las letras griegas hacen lo mismo con la población. Esta norma, como prácticamente todas, tiene excepciones. El tamaño del conjunto de datos (número de unidades) no recurre a letras latinas/griegas sino minúscula (muestra) o mayúscula (población). La distancia estandarizada se simboliza del mismo modo, con la letra Z, aunque los símbolos que representan a sus elementos de cálculo (valor, media y desviación tipo) sean diferentes. En la

tabla 1 tienes los símbolos que utilizamos en cada caso y que nos van a acompañar en lo sucesivo.

| **Objeto** | **Muestra** *estadísticos* | **Población** *parámetros* |
|---|---|---|
| Tamaño | $n$ | $N$ |
| Media aritmética | $\bar{X}$ | μ (mu o mi) |
| Desviación tipo | $S$ | σ (sigma) |
| Proporción | p | π (pi) |
| Distancia estandarizada | $Z_i = \frac{X_i - \bar{X}}{S}$ | $Z_i = \frac{X_i - \mu}{\sigma}$ |

Tabla 1. Símbolos para muestra y población.

Los índices o recursos suelen denominarse *estadísticos* cuando se refieren a la muestra, mientras que recurrimos al término *parámetros* para hacer referencia a los mismos recursos pero calculados sobre la población. No obstante, es frecuente en los libros de estadística que se hable de estadísticos para referirse a todo tipo de índices, sin distinguir en qué conjunto de datos se calculan. Esto es debido a que la tradición estadística se ha construido sobre el estudio de muestras, por lo que en muchos casos se sobreentiende que todo lo que calculamos son estadísticos.

Con este nuevo vocabulario podemos decir, por ejemplo, que la inferencia se ocupa, entre otros asuntos, de encontrar el valor de los parámetros a partir de los estadísticos; o que los estadísticos se representan con letras latinas mientras que las griegas se destinan a los parámetros.

**La distribución muestral**

Entre las primeras decisiones que se establecen en relación a las muestras se encuentran el modelo de muestreo y el tamaño de muestra. Ya hemos visto algo sobre los modelos. Ocurre que el tamaño de muestra necesita una serie de conocimientos que todavía no tenemos. La lógica general para decidir el número de unidades que van a ser seleccionadas de la población se basa en decidir antes cómo haremos las inferencias. Como es obvio, todavía no sabemos hacerlas, por lo que tendremos que dejar para más adelante las habilidades implicadas en el calculo del tamaño de una muestra.

El primer paso para adelantar conocimientos necesarios para comprender los procesos de inferencia es saber qué es y cómo se comporta una distribución muestral.

Imagina un aula universitaria con 50 personas. Vamos a extraer una muestra de 10 para preguntarles cómo va la cosa, qué cambiarían, qué mantendrían, etc. respecto al modo en que se está desarrollando la docencia. Seguimos un MAS, por lo que no es necesario considerar ninguna característica en concreto para las selecciones. ¿Cuántas muestras son posibles?

Alguien podría responder impulsivamente con ¡Cinco! Lo siento, pero no es correcto. El razonamiento no es muy malo: 50/10=5. Pero para que dos muestras sean diferentes no es necesario que lo sean *todas* sus unidades, basta con que una de las personas seleccionadas en la primera ocasión no lo sea en la segunda. Es un problema de combinatoria: cuántos subconjuntos de 10 elementos pueden obtenerse de un conjunto de 50. La solución es:

$$\binom{50}{10} = 10272278170$$

¡Elevado! ¿Verdad? Si somos capaces de obtener una muestra cada segundo, necesitamos 326 años sin descanso para completarlas todas. Lo más laborioso sería extraer muestras de 25 personas. En ese caso necesitaríamos algo más de 4 mil siglos. Si te parece, mejor dejamos esa tarea sin hacer.

No es posible extraer todas las muestras posibles desde poblaciones habituales. Por lo general N=50 es un tamaño pequeño. Lo normal es manejar poblaciones mucho más grandes, en las que esos 4 mil siglos necesarios constituyen una cantidad ridícula. No obstante, aunque en la práctica no sea viable obtener todas las muestras, sí que podemos enfrentarnos a deducir qué obtendríamos si nos enfrentáramos a esa tarea. El resultado es muy útil en estadística, como irás viendo.

Vamos a situarnos en que queremos saber algo concreto de una población, por ejemplo cuántos días sus miembros van a cenar fuera de casa cada mes. Las respuestas pueden variar desde 0 hasta 31 para cada persona entrevistada.

Pensemos que obtenemos las k muestras posibles, cada cual con el mismo tamaño n que provienen de una población con tamaño N. No solo nos tomamos el trabajo de conseguir las muestras sino que en cada una de ellas calculamos algo, por ejemplo la media aritmética.Si hemos obtenido k muestras tendremos finalmente k promedios, es decir, k respuestas a la pregunta ¿por término medio cuántas veces va la gente de esta muestra a cenar fuera de casa al mes? Con ese conjunto de k medias podemos hacer lo que hacemos cuando tenemos un conjunto de datos cualesquiera: tabularlos, representarlos gráficamente, calcular estadísticos, etc.

Pues tenemos que el conjunto organizado de las k medias obtenidas de las k muestras recibe el nombre de *distribución muestral de medias*. ¿Puedes adivinar como se llamaría si en lugar de medias aritméticas hubiéramos calculado proporciones? En efecto: *distribución muestral de proporciones*. Lo mismo si fueran medianas, varianzas...

*Un ejemplo con una población completa*

Vamos a ver un ejemplo concreto con una muestra pequeña, de tan solo 6 datos, de donde obtendremos muestras de 4 datos. Estos seis datos (toda la población) son:

3, 7, 7, 9, 4, 5

Antes de abordar la distribución muestral que construiremos con esta población, conozcámosla un poco mejor, calculando la media aritmética, la desviación tipo y, ya puestos, la proporción de números iguales o inferiores a 5.

| | Xi | ci |
|---|---|---|
| | 3 | 1 |
| | 7 | 0 |
| | 7 | 0 |
| | 9 | 0 |
| | 4 | 1 |
| | 5 | 0 |
| Suma = | 35 | 2 |
| Media = | 5,83 | 0,33 |
| Des.tip. = | 2,03 | 0,47 |

Para la distribución muestral de medias es interesante la columna $X_i$, de la que sabemos que la media aritmética de la población es 5,83 y su desviación tipo 2,03. Para la distribución muestral de proporciones, contamos con la otra columna. He llamado $c_i$ a una variable que tiene el valor 1 si se cumple la condición $X_i<5$, y el valor 0 en caso contrario; es decir, el valor $c_i=1$ está indicando que $x_i<5$, por lo que sumar $c_i$ tendremos el número de datos con $x_i$ menor a 5. De esta nueva variable sabemos que su media aritmética es 0,33 y su desviación tipo es 0,47. Observemos este resultado un poco más despacio.

La media es la suma de datos entre el número total de datos. Si los valores son ceros y unos, dado que 0 no incrementa la suma y 1 lo hace una unidad, entonces la suma total es el número de unos. Es más, la media será la proporción de unos, es decir, el número de unos entre el número total de datos. En el

ejemplo, hay 2 datos con el valor 1 y 4 con el valor 0. La suma es 2 y la media será 2/6=0,33. Si calculas la proporción de datos con el valor 1 tendrás el mismo resultado: 2 datos con ese valor respecto a un total de 6 datos, 2/6=0,33. Da lo mismo decir que la media de $c_i$ es 0,33 que afirmar que la proporción de valores inferiores a 5 es 0,33. Esto es muy interesante porque la proporción puede ser considerada como una media aritmética calculada sobre un conjunto de ceros y unos. Gracias a esto, lo que se afirme respecto a las medias vale también para las proporciones.

Hay algo más. La desviación tipo de $c_i$ es 0,47. Es el mismo resultado que obtenemos si calculamos la raíz cuadrada de 0,33*(1-0,33). En otras palabras, cuando estamos trabajando con un listado de ceros y unos, donde llamamos p a la proporción de unos, entonces el cálculo de la desviación tipo se simplifica mucho:

$$S = \sqrt{p\ (1 - p)}$$

*El valor esperado*

Volvamos ahora al asunto de las distribuciones muestrales. Con los N=6 datos pueden obtenerse 15 muestras distintas de n=4 elementos. Por lo tanto, podríamos calcular el valor de 15 medias aritméticas y de 15 proporciones de datos donde se observen

valores inferiores a 5. Es decir, podemos realizar en las 15 ocasiones los mismos movimientos que hemos elaborado para el total de los 6 datos de la población.

Este ejercicio permite observar en la práctica el comportamiento de los conceptos que utilizamos a partir de las distribuciones muestrales. Tal vez te resulte un poco desagradable, aburrido o poco pasional, pero es lo que hay.

Vamos pues a obtener las 15 medias aritméticas y las 15 proporciones de valores inferiores a 5 que provienen de las k=15 muestras posibles de tamaño n=4. En la tabla 2 se encuentra toda esta información. Échale un tiempo de observación. Hay al menos un par de resultados que deberían llamarte la atención.

Un resultado llamativo es que la media de las medias coincide con la media poblacional. Observa que al suma las 15 medias y dividir el resultado entre 15 (media de las medias), obtenemos 5,83 que es el valor de la media aritmética de los 6 datos de la población. También ocurre que la media de las proporciones tiene el mismo valor que la proporción poblacional. Esto no es una casualidad. Es lo que ha de ocurrir necesariamente. Es lo que se demuestra matemáticamente, aunque si me lo permites, no incluyo aquí las demostraciones de estos resultados.

| e. | Elementos (i) | Datos (Xi) | Media | Proporc |
|---|---|---|---|---|
| 1 | 1, 2, 3, 4 | 3, 7, 7, 9, 4, 5 | 6,50 | 1/4 = |
| 2 | 1, 2, 3, 5 | 3, 7, 7, 9, 4, 5 | 5,25 | 2/4 = |
| 3 | 1, 2, 3, 6 | 3, 7, 7, 9, 4, 5 | 5,50 | 1/4 = |
| 4 | 1, 2, 4, 5 | 3, 7, 7, 9, 4, 5 | 5,75 | 2/4 = |
| 5 | 1, 2, 4, 6 | 3, 7, 7, 9, 4, 5 | 6,00 | 1/4 = |
| 6 | 1, 2, 5, 6 | 3, 7, 7, 9, 4, 5 | 4,75 | 2/4 = |
| 7 | 1, 3, 4, 5 | 3, 7, 7, 9, 4, 5 | 5,75 | 2/4 = |
| 8 | 1, 3, 4, 6 | 3, 7, 7, 9, 4, 5 | 6,00 | 1/4 = |
| 9 | 1, 3, 5, 6 | 3, 7, 7, 9, 4, 5 | 4,75 | 2/4 = |
| 10 | 1, 4, 5, 6 | 3, 7, 7, 9, 4, 5 | 5,25 | 2/4 = |
| 11 | 2, 3, 4, 5 | 3, 7, 7, 9, 4, 5 | 6,75 | 1/4 = |
| 12 | 2, 3, 4, 6 | 3, 7, 7, 9, 4, 5 | 7,00 | 0/4 = |
| 13 | 2, 3, 5, 6 | 3, 7, 7, 9, 4, 5 | 5,75 | 1/4 = |
| 14 | 2, 4, 5, 6 | 3, 7, 7, 9, 4, 5 | 6,25 | 1/4 = |
| 15 | 3, 4, 5, 6 | 3, 7, 7, 9, 4, 5 | 6,25 | 1/4 = |
| | | Media= | 5,833 | 0 |
| | | Desviación tipo= | 0,643 | 0 |

Tabla 2.
Distribuciones muestrales de medias y proporciones.

Así pues, 5,83 es tanto la media de los datos de la población como la media de las medias aritméticas calculadas sobre todas las muestras de tamaño n=4 que hemos obtenido de esa población. Y, del mismo modo, 0,33 no solo es la proporción de datos en la población con un valor inferior a 5; es también la media de las proporciones calculadas sobre todas las muestras de tamaño n=4 obtenidas de esa misma población.

Antes de seguir es importante que hagamos una pausa para adquirir vocabulario. En lugar de decir, por ejemplo, “media aritmética de la distribución muestral de medias”, agradecería contar con una referencia algo más breve. La tenemos. Se llama *valor esperado* o también *esperanza matemática* y se simboliza con E(). Así, el valor esperado para la media es $E(\bar{X}) = \mu = 5{,}83$ mientras que el valor esperado para la proporción es $E(p) = \pi = 0{,}33$ .

La expresión “valor esperado” tiene mucho sentido. Recuerda el objetivo de la media aritmética: representar numéricamente a un conjunto de datos. Si tuviéramos que hacer una apuesta sobre qué valor esperar de un grupo de datos medidos según una escala cuantitativa, una buena decisión sería escoger la media aritmética. Como *medida de tendencia central* que es o *valor representativo*, cabe esperar que los datos la ronden, es decir, no se distancien excesivamente de ese valor. Pues bien, la lógica es

todavía más contundente en el caso de las distribuciones muestrales. La media de esa distribución no solo cuenta con el mismo argumento para ser escogida como el valor más representativo, sino que las distribuciones muestrales tienen menos dispersión que la población, puesto que cuando se calcula algo sobre un conjunto de datos, los resultados son más parecidos que cuando se calcula sobre los datos uno a uno. Por otro lado, las distribuciones muestrales suelen mostrar cierta simetría y una coincidencia al menos aproximada entre media, mediana y moda, por lo que el valor más esperado es siempre el mismo se mire como se mire. Así que el nombre parece estar muy bien escogido.

*El error tipo*

La media aritmética no es lo único que vamos a calcular sobre la distribución muestral. Acabamos de recordar que la media aritmética representa numéricamente al conjunto de datos. Ya sabemos de temas anteriores que la media puede ser mejor o peor en su tarea de representación, por lo que procede acompañarla siempre de una medida de bondad de su papel de representación, también llamada índice o estadístico de dispersión, desviación o variabilidad. El estadístico escogido es la desviación tipo. Como sabes, cuanto mayor sea su valor, peor es la media en su

objetivo de representar numéricamente el conjunto de datos. Puede llegar un momento en que lo más característico sea la variación y no el promedio. Así que además de la media, utilizaremos también la desviación tipo para describir una distribución muestral.

La desviación tipo de la distribución muestral se conoce como *error tipo*. Así, hablamos por ejemplo del valor esperado de la media y del error tipo de la media para referirnos, respectivamente, a la media y a la desviación tipo de la distribución muestral de medias. Si el valor esperado es el valor que más esperamos, el error tipo expresa el grado en que nos podemos estar equivocando cuando recurrimos al valor esperado. Como ves, se trata de la misma lógica: un valor que representa y otro valor que expresa cuán bueno es el primero en su cometido de representación.

Lo que nos queda es calcular ese error tipo y conocer el símbolo asociado. Comencemos por esto último. El error tipo utiliza la letra griega sigma (σ), como ocurre en el caso de simbolizar la desviación tipo de la población, pero le añade un subíndice. El subíndice expresa el estadístico que nutre la distribución muestral. Así, el error tipo de la media (es decir, la desviación tipo de la distribución muestral de medias) se representa con el símbolo $\sigma_{\bar{X}}$, mientras que el error tipo de la proporción (la desviación tipo de la distribución muestral de proporciones) es representado con $\sigma_p$.

La expresión de cálculo para el error tipo es muy comprensible y lógica, pero requiere de ti una prueba de fe, pues no voy a demostrarlo matemáticamente. Vamos a adentrarnos en ello por aproximaciones sucesivas.

Observa la siguiente expresión. Es la que se corresponde con el error tipo de la media cuando el muestreo es aleatorio simple (como todo lo que estamos abordando en este documento) y la población tiene un tamaño de tal envergadura que resulta prácticamente infinita:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Esta expresión dice varias cosas. Por un lado afirma que, dado que la muestra tiene varias unidades, el error tipo es siempre de cuantía inferior a la desviación tipo poblacional. En el caso límite, cuando n=1 (imposible menos, como es obvio), el error tipo y la desviación tipo tienen el mismo valor. Es comprensible. Cuando las muestras tienen el tamaño 1, no son muestras, sino que la distribución muestral y la población son la misma cosa. Dicho de otro modo: la población puede ser pensada como una distribución muestral con n=1 (cada individuo es su propia muestra). Esto es ridículo, pero tal vez sea divertido pensarlo así. Seguimos. Otro aspecto muy relevante es que el tamaño de la muestra influye de forma

coherente: cuanto mayor es, más reduce el valor del error tipo. Tiene mucho sentido. Si las muestras van haciéndose cada vez más grandes, las medias se parecerán cada vez más a la media poblacional, por lo que cada vez habrá menos dispersión muestral. De forma sintética: a muestras mayores, errores tipo menores.

En el caso de la distribución muestral de proporciones, podemos tirar de la misma expresión recordando que una proporción es un caso particular de media aritmética donde se ha codificado con unos y ceros. Aplicando literalmente la expresión simplificada de cálculo para estas situaciones:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\pi\,(1-\pi)}}{\sqrt{n}} = \sqrt{\frac{\pi\,(1-\pi)}{n}}$$

No obstante, si intentas aplicar estas expresiones a nuestro ejemplo verás que no funciona. Ocurre porque la población que estamos manejando dista mucho de ser infinita, pues N=6. En ese caso, estas fórmulas no nos valen y hemos de acudir a estas otras que, del mismo modo, exigen de ti una cuestión de fe.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Si te fijas, estas expresiones parten de las anteriores, a las que se les ha aplicado un *factor de corrección*, que adapta el cálculo a las situaciones con poblaciones finitas:

$$\sqrt{\frac{N-n}{N-1}}$$

Si las poblaciones son finitas, utilizamos este factor. Si son prácticamente infinitas, lo obviamos, puesto que no tiene efecto en el resultado. Es lógico, pues si N es muy muy grande, le da lo mismo que se le sustraiga 1 o n, pues en ambos casos sigue prácticamente igual. Vamos a aplicar estas fórmulas para obtener el error tipo de medias y proporciones a partir de la desviación tipo de la población. Observaremos entonces que estos valores coinciden con la desviación tipo de las distribuciones muestrales:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2{,}03}{\sqrt{4}}\sqrt{\frac{6-4}{6-1}} = 0{,}64$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0{,}33(1-0{,}33)}{4}}\sqrt{\frac{6-4}{6-1}} = 0$$

*Estimadores insesgados*

Todos los desarrollos de la distribución muestral tienen sentido porque nos sirven para hacer estimaciones, es decir, para inferir qué debe estar ocurriendo en la población a partir de la información que tenemos respecto a la muestra. Como veremos en otra unidad y utilizando el vocabulario con que ya contamos, la estimación consiste en concluir sobre el valor de un parámetro a partir de un estadístico. Para ello, utilizamos el comportamiento que conocemos de las distribuciones muestrales.

En una estimación, queremos conocer el valor de la media poblacional a partir de una muestra. ¿Cómo lo podemos hacer? La versión más simple es lo que llamamos *estimación puntual*. Consiste en afirmar que el parámetro tiene un valor concreto (un punto). ¿Pero cuál? Como sabemos de la distribución muestral que el valor esperado para la media de las muestras es la media de la población, bastaría con tomar el valor esperado y así terminar con el problema. Pero esto no es posible. No tenemos el valor poblacional (es lo que buscamos). No tenemos la distribución muestral (ya sabemos que es imposible conseguirla en la práctica). Así que no podemos contar con el valor esperado. No obstante, podemos hacer algo. Si tenemos que apostar, yo apostaría a que mi muestra ha dado en el clavo de la media poblacional, es decir, tomaría la media de mi

muestra como valor representativo de la media de la población. Es la mejor de las apuestas, según nos dice el comportamiento de la distribución muestral de medias: mi media es la mejor representación del conjunto de datos de la muestra, y la media de las medias (el valor esperado) es la mejor representación para la media poblacional, puesto que coincide plenamente con ella. Así que, de entre todas las posibilidades, escojo la media de mi muestra para apostar que por ahí debe andar la media de la población.

Otra forma de indicar que la media de las medias (el valor esperado de la media) coincide con la media de la población es afirmar que la media aritmética de la muestra es un *estimador insesgado* de la media aritmética de la población. Los estadísticos, es decir, las medidas tomadas en la muestra, pueden servir para muchos cometidos. Cuando utilizamos un estadístico como base para estimar un parámetro (es decir, para inferir cuál será el valor en la población), entonces decimos que ese estadístido es un estimador (como ves, un término bastante literal). Pues bien, si el valor esperado de ese estimador coincide con el parámetro que queremos estimar, entonces decimos que ese estimador no tiene sesgo o bien que es insesgado.

En un muestreo aleatorio simple, la media de la muestra es un estimador insesgado de la media poblacional. También ocurre con la proporción de la

muestra para estimar la proporción de la población. Pero no ocurre así con todo. Por ejemplo, la varianza de la muestra es un estimador *sesgado* de la varianza poblacional. En otras palabras: la media aritmética de la distribución muestral de varianzas (el valor esperado de la varianza) no coincide con la varianza de la población. Es un estimador sesgado porque *apunta* cerca del parámetro pero no da en el blanco. No obstante, no hay que preocuparse (en el caso de que te hayas preocupado, claro). Esto tiene solución. Se demuestra que hay un estadístico cuyo valor esperado coincide con la varianza poblacional. Ese estadístico es la *cuasi*varianza. En otras palabras, el valor esperado de la cuasivarianza coincide con la varianza de la población. Se llama *cuasi*varianza porque es *casi* la varianza. La única diferencia se encuentra en el numerador. Recuerda la expresión de cálculo (o repásala). En la cuasivarianza, el numerador no es *n*, como en la varianza, sino *n*-1. Para indicar que es casi la varianza, se le coloca un sombrero (el acento circunflejo). De este modo, su expresión de cálculo es:

$$\hat{S}^2 = S^2 \frac{n}{n-1} = \frac{\sum \left( X_i - \bar{X} \right)^2}{n-1} \qquad E\left(S^2\right) \neq \sigma^2$$

Coherentemente, la cuasidesviación tipo es la raíz cuadrada de la cuasivarianza. De este modo, la

desviación tipo de la población será la raíz cuadrada del valor esperado de la cuasivarianza. No podemos comprobarlo con los datos de nuestro ejemplo, porque estas afirmaciones son válidas solo para poblaciones de gran tamaño. Si intentas hacer la prueba con nuestra población de N=6 no tendrás el mismo resultado. También hay solución. Si quieres saberlo, continúa por el siguiente párrafo. Si no, salta al siguiente y último apartado (una visión sintética).

**Qué pasa con la estimación de varianzas en una población finita**

En nuestro ejemplo, con poblaciones pequeñas lo que tenemos es que la cuasivarianza de la muestra es un estimador insesgado de la cuasivarianza poblacional. Si queremos conocer la varianza poblacional, entonces partimos del valor esperado de la cuasivarianza muestral y deducimos la varianza a partir de la cuasivarianza, haciendo una deducción inversa a la que hemos hecho dos párrafos más arriba:

$$\sigma^2 = \hat{\sigma}^2 \frac{N-1}{N}$$

La tabla 4 contiene los valores de la varianza y de la cuasivarianza para cada una de las 15 muestras de tamaño n=4 que surgen de nuestra población de

N=6. Al calcular la raíz cuadrada de los valores esperados de la varianza y de la cuasivarianza tenemos, respectivamente, 1,930 y 2,229. Lo primero no es nada. Lo segundo es la cuasidesviación de la población. Para obtener la desviación tipo de la población:

$$\sigma = \hat{\sigma}\sqrt{\frac{N-1}{N}} = 2{,}229\sqrt{5/6} = 2{,}03$$

Observa que es ese, precisamente, el valor de la desviación tipo de la población.

| úm. | Datos (Xi) | Varianza | Cuasivarianz |
|---|---|---|---|
| 1 | 3, 7, 7, 9, 4, 5 | 4,75 | 6,33 |
| 2 | 3, 7, 7, 9, 4, 5 | 3,19 | 4,25 |
| 3 | 3, 7, 7, 9, 4, 5 | 2,75 | 3,67 |
| 4 | 3, 7, 7, 9, 4, 5 | 5,69 | 7,58 |
| 5 | 3, 7, 7, 9, 4, 5 | 5,00 | 6,67 |
| 6 | 3, 7, 7, 9, 4, 5 | 2,19 | 2,92 |
| 7 | 3, 7, 7, 9, 4, 5 | 5,69 | 7,58 |
| 8 | 3, 7, 7, 9, 4, 5 | 5,00 | 6,67 |
| 9 | 3, 7, 7, 9, 4, 5 | 2,19 | 2,92 |
| 10 | 3, 7, 7, 9, 4, 5 | 5,19 | 6,92 |
| 11 | 3, 7, 7, 9, 4, 5 | 3,19 | 4,25 |
| 12 | 3, 7, 7, 9, 4, 5 | 2,00 | 2,67 |
| 13 | 3, 7, 7, 9, 4, 5 | 1,69 | 2,25 |
| 14 | 3, 7, 7, 9, 4, 5 | 3,69 | 4,92 |
| 15 | 3, 7, 7, 9, 4, 5 | 3,69 | 4,92 |
| | Media= | $E(S^2) = 3{,}725$ | $E(\hat{S}^2) = 4{,}96$ |

Tabla 4. Varianzas y cuasivarianzas.

## Una visión sintética

La inferencia estadística es el camino que va de una muestra real a una población supuesta. Lo que nos interesa es la población, pero lo que tenemos es la muestra. ¿Cómo solucionar esta situación? Lo que hacemos en inferencia estadística es utilizar muestras aleatorias, es decir, que se han obtenido mediante procedimientos basados en el azar. Sabemos entonces que la muestra es una de entre muchas posibles. Para hacer inferencia necesitamos valorar ese “muchas posibles” para situar nuestra muestra.

Las “muchas posibles” constituyen lo que llamamos una distribución muestral. Si lo que queremos inferir es una media aritmética, entonces utilizamos la distribución muestral de medias, es decir, el conjunto de todas las mediciones de media aritmética que surgen de obtener todas las muestras posibles de la misma población con las mismas características. Este ejercicio es imposible en la práctica. Lo que hacemos realmente es deducir cómo sería esa distribución caso de construirla. “Cómo sería” implica en la práctica conocer su media aritmética (denominada valor esperado), su desviación tipo (denominada error tipo) y algo sobre su forma o comportamiento. Esto último lo veremos en el siguiente documento.

Una vez conozcamos esas características definitorias de la distribución muestral que nos interesa, entonces estaremos en condiciones de realizar inferencias.

## Glosario

| | |
|---|---|
| Cuasidesviación tipo | Raíz cuadrada de la cuasivarianza. |
| Cuasivarianza | Varianza corregida, donde el denominador no es el número de datos sino el número de datos -1. |
| Desviación tipo | Raíz cuadrada de la varianza. |
| Distribución muestral | Distribución de datos donde cada uno de ellos es el resultado de un mismo estadístico calculado desde cada una de las muestras del mismo diseño obtenidas de la misma población. |
| Error tipo | Desviación tipo de la distribución muestral. |
| Estadístico | Medida cuantitativa de alguna característica de la muestra. |
| Estimador | Estadístico utilizado para inferir el valor de un parámetro. |
| Estimador insesgado | Estadístico cuyo valor esperado coincide con el parámetro que se desea estimar. |
| Muestra | Conjunto de unidades con la que trabajamos y que representa a la población. |
| Parámetro | Medida cuantitativa de alguna característica de la población. |

| | |
|---|---|
| Población | Conjunto de unidades sobre el que deseamos concluir. |
| Valor esperado | Media aritmética de la distribución muestral. |
| Varianza | Medida de dispersión de los datos: media aritmética de las distancias al cuadrado de cada dato respecto a la media del conjunto de dato |