# Ciencia del Muestreo

2 ed.

**Mariano Ruiz Espejo**

*A mi familia*

Segunda edición.

Ruiz Espejo, Mariano (2026). *Ciencia del Muestreo 2 ed*. Madrid: Bubok.

# Índice

# Prólogo

Al publicar la primera edición de este libro titulado *Ciencia del Muestreo*, en inglés *Sampling Science*, reviso y amplío el libro con el mismo título, del mismo autor y del publicador Bubok, y proporciono los razonamientos matemáticos sobre los que se basan los métodos inferenciales de uso común en el muestreo y la estimación en poblaciones finitas.

Estos procedimientos son los más objetivos que conozco entre los métodos estadísticos inferenciales, ya que no hay que suponer que la población es de algún modo "sin comprobación posible" como ocurre en la mayor parte de teorías de inferencia estadística y de predicción. Además, la aleatorización y sus propiedades son características que "el investigador controla y no tiene que suponer" a su vez que se comporta de algún modo determinado como ocurría en la inferencia paramétrica clásica, la bayesiana, la no paramétrica, etc. basadas en poblaciones con función de densidad (o en gran parte de modelos de distribución probabilística) y en datos observados. En nuestro libro entendemos que los datos son medidas objetivas y exactas de la realidad física que nos rodea y que podemos observar, cuantificando las observaciones.

El conocimiento objetivo de características poblacionales que deben ser conocidas con cierta precisión, para corregir cualquier deficiencia o atender nuevas necesidades sociales requiere el uso de métodos libres de hipótesis subjetivas o no comprobadas en la práctica. La teoría inferencial en poblaciones finitas que

desarrollamos aporta sólidos conceptos y resultados matemáticos que permiten conocer estas características con métodos objetivos, sencillos, precisos, rápidos y económicos en comparación con la realización de censos que observen toda la población para conocer con perfección todos los datos de interés, como ocurre con la estadística descriptiva que tiene los métodos más objetivos, pero no son de tipo inferencial sino determinístico. Cuando hablamos de una población finita y de una función paramétrica definida, hablamos de realidades que existen en el mundo real, no son operaciones derivadas de suposiciones teóricas como ocurre en la teoría inferencial clásica estadística.

El único camino que garantiza que el diseño muestral no es un mero instrumento de estudio teórico y que se lleva a la práctica es mediante la identificabilidad de las unidades de la población finita, siendo estas unidades accesibles u observables para obtener el dato de ser seleccionada la unidad en la muestra efectiva. Sin estos requisitos el estudio inferencial es exclusivamente teórico sin capacidad para seleccionar muestras de unidades del contexto real al que se trata de aplicar estos conceptos basados en realidades que existen y comprobamos en la práctica de una encuesta o de un estudio por muestreo que busca la información donde está, en el mundo real, no en el ámbito de las meras ideas que no buscan conocer y obtener información para cambiar el mundo con el mejor sentido del bien común o para hacer auténtica una investigación social, sanitaria o política entre otras posibles aplicaciones como también son las de ingeniería, banca, etc.

El manual puede considerarse un libro de referencia en asignaturas de *Teoría de Muestras* en estudios de Estadística en el Grado en Ciencias Matemáticas o similares másteres y asignaturas de doctorado, o en el Grado en Ciencias Estadísticas, en Ciencias Económicas y Empresariales, en Economía, y en Administración y

Dirección de Empresas, así como en Sociología, Ciencias de la Salud, e Ingeniería. También para estudio e investigación.

Requiere conocimientos básicos de Teoría de la Probabilidad, Esperanza Matemática y Varianza de una variable aleatoria discreta, que son también expuestos. Sin duda aportará mayor objetividad a los métodos estadísticos estudiados en Escuelas Técnicas Superiores de Ingeniería, y en Ciencias de la Salud, lo que conllevará una perspectiva más objetiva en sus tradicionales formaciones estadísticas. A modo de ejemplo, y para los lectores que ya manejen con facilidad los conceptos explicados en este libro, les recomiendo la lectura del artículo de Ruiz Espejo (2018f) sobre diseño de experimentos desde una perspectiva objetiva de muestreo de poblaciones finitas.

En este libro también reviso y actualizo otros dos de los que soy autor, editados por Lulu Press, y presento los argumentos y fundamentos que sustentan la mayor o menor objetividad científica y ética entre una selección de métodos de inferencia estadística estudiados tradicionalmente en las universidades de todo el mundo y especialmente en titulaciones en Estadística y en Ciencias Matemáticas, para ser aplicados en el mundo real y práctico con el mayor rigor teórico y aplicado.

Con este fin se presentan resumidos y comentados los fundamentos científicos de la tesis doctoral del autor y que defendió en la Facultad de Ciencias Políticas y Sociología de la Universidad Pontificia de Salamanca, con el título *Observaciones a los Métodos Estadísticos de Investigación del Bienestar Social en el Marco Global* (Madrid, 2003a).

El programa de doctorado al que se adscribió la tesis en aquellos años era *Globalización, Desarrollo y Bienestar Social*. Estaba en boga el estudio de la Globalización como una concepción del mundo y de las relaciones humanas que se fundamentaban en

el uso social de las nuevas tecnologías, como alternativa a una concepción socialista o capitalista de entender la economía y la sociedad, que ya se consideran limitadas o abocadas al fracaso porque no ponen al hombre como fin sino como medio del que servirse los grupos dominantes para alcanzar otros fines generalmente centrados en el bienestar de estos grupos y ralentizando el desarrollo y el bienestar social de los menos cercanos al poder político a los que se considera solo medios para aquellos fines. Sin embargo, hoy podemos decir que la Globalización ha servido hasta ahora también a los grupos políticos y gubernamentales dirigentes que tratan de imponerse con mayor poder e influencia sobre el ciudadano más allá de las fronteras naturales que circunscribían su influencia y gobierno hasta ahora, y llegando a crear serios problemas al desarrollo de regiones y pueblos.

Una alternativa propuesta para superar estas deficiencias del capitalismo, del socialismo y del globalismo es el cristianismo, y su aplicación de la Doctrina Social de la Iglesia Católica a la empresa y a la vida, que busca el bien común.

El cristianismo vinculado a su comunidad o entorno de actividades de las empresas está inspirado en las enseñanzas y el Magisterio de la Iglesia, y pone como fin de dichas actividades al hombre y sus cercanos, que las realiza en continuo camino hacia el conocimiento de Dios, su cercanía, y la trascendencia de nuestra actuación y vida con sentido común y atendiendo las necesidades espirituales y materiales de los que nos rodean, siendo, dando y sirviendo.

En este libro trato de dar los argumentos lógicos que daría un científico y la verdad revelada que aportaría un cristiano para discernir qué tipo de inferencia es objetiva y creíble, y cuáles no lo

son tanto, atendiendo razones y verdades de fe que espero que todos compartan porque no hay sino una buena intención de guiar a la verdad de los fundamentos de sabiduría cristiana y de ciencia, que pueden ser básicamente comunes a la tradición judía pues hay muchas referencias de revelación divina que tanto cristianos como judíos creemos porque compartimos enseñanzas milenarias de revelación de Dios. Las referencias bíblicas a las que me referiré son extraídas de la *Biblia de Jerusalén* (9ª edición, Bilbao, 1999).

La tesis doctoral que da base a esta obra se leyó en una universidad católica, de la Conferencia Episcopal Española, su director, profesor José Ramón Pin Arboledas, y el presidente del tribunal, profesor Francisco José Cano Sevilla, son profesionales universitarios y también políticos de distinta orientación a los que debo honra personal y agradecimiento por haber contribuido a su lectura y reconocimiento.

El capítulo dedicado a la ética y filosofía del muestreo es relativamente breve porque no trato de ser ocioso en divagaciones sino que trato de dar claridad de ideas a los profesionales de la Estadística, tanto universitarios como de la administración, para su trabajo diario, así como orientar posibles futuras aplicaciones de la Estadística en el campo de la salud, ciencia, política, e ingeniería.

También se han incluido las conclusiones del autor en el máster en Bioética, en el trabajo titulado *Investigación Ética y Bioestadística* (2014), realizado en la Universidad Católica San Antonio de Murcia y dirigido por el profesor Jorge López Puga.

Toda inferencia en poblaciones finitas ha de basarse en el marco de la población desde el que las unidades de la población son accesibles al investigador por muestreo. Sin embargo, este marco no siempre está disponible pues implica la colaboración de toda la población en dar datos sensibles de su persona, vivienda, teléfono, etc. y no siempre es posible tener estos censos, lo que

limita el uso de esta ciencia del muestreo. En relación a esto conviene recordar el Libro Primero de las Crónicas 21,17: “Y dijo David a Dios: ‘Yo fui quien mandé hacer el censo del pueblo. Yo fui quien pequé, yo cometí el mal; pero estas ovejas ¿qué han hecho?...’”. En los casos en que estos datos censales, que identifican a las unidades de la población finita, son conocidos y las unidades de la población son accesibles y observables, este libro tiene su pleno interés práctico y objetivo para alcanzar sus fines científicos éticamente. También cuando las unidades no son personas sino objetos especialmente.

Quiero agradecer a los publicadores la oportunidad de editar este libro dirigido a todos los lectores en lengua española. También es justo agradecer a Javier Olivera Ravasi sus 21 artículos titulados “Aprendiendo a pensar: lógica de los sofismas”, que han sido publicados en la publicación digital InfoCatolica.com entre Enero y Febrero de 2015 y que han dado un marco cristiano y filosófico clásico a mis reflexiones, quedando insertado nuestro estudio en la tradición católica y abierta a toda cultura virtuosa y respetuosa con el conocimiento, la sabiduría y la inteligencia cristianos por la tradición fiel al magisterio de la Iglesia.

Debo agradecer a todos los que, presentes o ausentes, han contribuido a que esta publicación sea ofrecida a los lectores interesados en los principios éticos y morales de la ciencia estadística y de sus métodos de inferencia.

Revisamos también los principios, normas y pautas éticas de investigación en seres humanos en sus aspectos bioestadísticos y aportamos métodos y referencias sobre posibles mejoras en este área. Algunos aspectos como voluntarios, consentimiento informado, tratamiento de la no respuesta, y estimación insesgada, son tratados con cierto detalle. Concluimos que incentivando las

condiciones del consentimiento, podríamos aprovechar la información de voluntarios en un segundo intento para inferir con objetividad sobre la función paramétrica de interés. Esto permite extraer conclusiones sobre toda la población de pacientes y no reducirla a la de los primeros voluntarios, pues limitarnos a la población de voluntarios puede determinar un sesgo en las estimaciones sobre la función paramétrica de interés, ya que la población finita de pacientes es más ámplia que aquélla.

Este libro es, por tanto, un compendio resumido de los estudios y las investigaciones del autor.

Agradezco las sugerencias del profesor Guillermo Enrique Ramos, de la Universidad de Morón, Buenos Aires, Argentina.

Mariano Ruiz Espejo

Madrid, Enero de 2026

# Presentación

Es natural y perfectamente lógico que las medidas sean exactas cuando nos afectan en la compra o el consumo de las personas, así como en la retribución por su trabajo. En la práctica muchas veces se presenta la situación de que queremos conocer una magnitud a la que contribuye cada una de las unidades de una población finita pero no tenemos recursos, tiempo o medios para recabar la información exacta de todas las unidades para proceder al cálculo de dicha magnitud.

La inferencia en poblaciones finitas consiste en un procedimiento de muestreo o de selección de unidades de la población para ser observadas o medidas con exactitud, y en un método de estimación que aproveche la información recabada de la muestra a efectos de inferir sobre magnitudes poblacionales que llamamos funciones paramétricas, pues dependen de todos los valores observables fijos en cada una de las unidades de la población finita. La muestra se selecciona de modo probabilístico, mientras que el estimador es una función de los datos muestrales en la recta real a la que pertenece la magnitud que queremos inferir.

Una propiedad de importancia de un método inferencial es su insesgación, que nos indica que en promedio el estimador tiene por esperanza matemática la magnitud que queremos inferir. La medida de dispersión más usada para conocer la variabilidad del estimador insesgado, es su varianza. En general, cuando el

estimador es sesgado, la medida de dispersión usada es su error cuadrático medio.

Antes de poner en práctica un método inferencial, es conveniente estudiar otros métodos alternativos para ser usados en el estudio concreto. Para ello conviene comparar los errores cuadráticos medios de los distintos métodos y ver en cuál de ellos se minimiza la variabilidad, y por tanto será más preciso que los restantes. En un artículo de Ruiz Espejo (1987c), se prueba la no existencia de estimador insesgado uniformemente de mínima varianza (salvo algún caso muy concreto), así como la no existencia de estimador uniformemente de mínimo error cuadrático medio. Sin embargo los métodos más precisos vienen acompañados de mayores costes esperados de uso, lo que les hace no ser los únicos a considerar. A igual coste esperado, sí tiene sentido buscar la mayor precisión o eficiencia. O bien, tiene sentido que a igual precisión, busquemos el método inferencial de menor coste esperado.

Por último, es conveniente estimar sin sesgo para el método inferencial concreto que hemos usado, la varianza o el error cuadrático medio del estimador, con la misma información muestral. Esto nos permitirá dar una estimación insesgada de la magnitud poblacional de interés, y una estimación insesgada del error de muestreo que tiene la estimación anterior.

El libro que está leyendo presenta los argumentos revelados y científicos que orientan en la elección de un método de inferencia estadística para alcanzar objetividad en las conclusiones de sus estudios. Es un diálogo entre la fe y la razón, entre la lógica humana y la revelación divina tal y como se concibe en la cultura judeocristiana española, europea y del mundo que respeta el derecho a la libertad religiosa y a la razón.

Es necesario contar con instrumentos de análisis, ya sea de estadísticas oficiales o de estudios privados, para obtener unas mínimas garantías de conocimiento objetivo y para alcanzar un mayor bienestar, que puede alcanzarse por la interiorización de la responsabilidad personal y por el sentido común. El trabajo se presenta entre dos ciencias humanas: la estadística como ciencia instrumental y la filosofía social como ciencia normativa, ambas iluminadas por la revelación cristiana. La relación entre ambas ciencias es evidente. Sin un contenido social la ciencia carece de contenido moral, pero sin instrumentos de análisis precisos los contenidos morales son inalcanzables en la práctica; la revelación cristiana orienta en la elección moral de estas decisiones. Por ello este trabajo oscila alternativamente entre la reflexión cristiana, la filosófico-social y la necesaria lógica estadística.

La tesis en sociología del autor surgió en un contexto de intentar mejorar el nivel de bienestar social en una economía global, un problema de nuestro tiempo que aún no se ha resuelto y donde ofrecemos una visión y perspectiva cristiana a la resolución del mismo.

El trabajo se puede enmarcar en lo que se denomina literatura social realizable, en el sentido de que intenta marcar un objetivo deseable y posible para el futuro. Para ello se dan métodos estadísticos para alcanzarlo, añadiendo a la revelación y a la ciencia humanística deseable algunos elementos lógicos de factibilidad concretos.

Los principales puntos del debate se sitúan en torno al discernimiento entre algunas metodologías estadísticas que se explican en las universidades, pero que en su mayoría son muy frágiles a la hora de asegurar coherencia y objetividad al describir hechos reales de carácter natural o inferir a partir de ellos, y en

particular hechos de las personas o sus bienes de interés social que en un instante determinado se presentan en las unidades de una población.

Las realidades de carácter social pueden ser y lo son en muchos casos cuantificables. Además las realidades sociales pueden ser observadas individualmente. La realidad social entendida como población estadística compuesta por un número finito de unidades (individuos, empresas, pacientes, etc.), de las que cada una es portadora de características cualitativas o cuantitativas, interesan desde puntos de vista sociales. Visto así tiene sentido discernir qué métodos inferenciales estadísticos son los que aportan objetividad y claridad en los casos prácticos. De otro modo el investigador social quedaría reducido a una posición débil, casi semántica o narrativa, a distancia de su objeto real de estudio si no tuviera en cuenta estas aportaciones estadísticas que reportan instrumentos para el conocimiento y la observación de fenómenos de carácter social.

Nos planteamos qué tipo de metodologías estadísticas basadas en el muestreo son correctas para revelar las realidades cuantitativas acaecidas en un instante o periodo de tiempo determinado, mediante la observación y medida exacta recogida en datos de una parte o muestra de unidades de la población que estudiamos.

La mayor parte de la Estadística universitaria actual no se adapta bien a las condiciones reales cuando interesa conocer hechos relativos a una población finita. Tratar la realidad finita como si no lo fuera refleja actitudes de inercia en los conocimientos estadísticos a costa de no aportar veracidad ni claridad al conocimiento real de los objetos de estudio científico.

Así evitamos que se usurpe la realidad misma por una concepción subjetiva de cómo es y lo que piensa el investigador

estadístico de la realidad objetiva que quiere conocer mediante métodos inferenciales. Es por ello que no aceptamos condiciones distorsionadoras o inasumibles para conocer realidades del mundo natural objetivo. Suponer hipótesis incomprobables o que contradicen elementos básicos de la lógica racional aplicada a su materia de estudio, hacen que sus diseños estadísticos sean inferiores a aquellos diseños estadísticos que no vulneran la regla básica de aceptar la realidad tal cual es para disponer de métodos estadísticos lo más fiables y coherentes con las realidades que se les presenta. Por tanto si queremos "conocer hechos" es preciso no asimilar elementos extraños, innecesarios o distorsionadores para este fin.

Como conclusión llegaremos a que los métodos de muestreo de poblaciones finitas con datos fijos proporcionan los métodos más objetivos y fiables entre los métodos inferenciales más conocidos y tratados en los cursos universitarios. Son por tanto los métodos realmente veraces, útiles, prácticos y lógicamente sólidos. El buen uso de la estadística mejorará la salud, el bienestar, la calidad de vida, la estabilidad social y económica, el desarrollo y la evolución económica de las personas, haciendo un uso cuidadoso de la inferencia estadística objetiva entre otras actuaciones necesarias, lo que aportará información precisa o confiable sobre aspectos sociales en los que actuar con decisiones políticas correctas.

Sin embargo, los métodos de inferencia estadística tradicional necesitan de datos individuales verdaderos para ofrecer conclusiones poblacionales que pueden ser por lo general no verdaderas, sino solo aproximaciones aleatorias (ya que no pueden ser medidas todas las unidades de la población por la limitación de los recursos disponibles), y a veces aproximaciones meramente supuestas a un "valor poblacional verdadero". Es como ocurre en

primer lugar en poblaciones finitas fijadas, y en segundo lugar con otros tipos de inferencia en que podría ser un "valor teórico o no real" y cuya "supuesta exactitud" estimamos.

Al final del libro, pondremos la atención en explicar aspectos éticos de las normas, principios, pautas y consejos ya establecidos en el área de la investigación con seres humanos y destacando los avances bioestadísticos que suponen pasos claros en el tratamiento de los datos para dar luz a las cuestiones de la mejora en la salud y de las mejores terapias posibles en la enfermedad.

En principio no existen límites éticos para el conocimiento de la verdad o en el esfuerzo humano para ello. Pero sí existen límites éticos precisos en cuanto al modo de obrar del hombre que busca dicha verdad, pues no todo lo que es "técnicamente posible" puede considerarse "moralmente admisible". La ciencia y la técnica tienen el límite de que cada persona humana merece respeto por sí misma, y en esto consiste la dignidad y el derecho del ser humano desde el inicio de su vida (cf. Instrucción *Donum Vitae*, I, 1987).

No detallaremos las fórmulas específicas que suponen los avances bioestadísticos, pero sí damos las referencias recientes donde poder encontrarlas y, en ciertos casos de éstas, con las demostraciones matemáticas que justifican sus propiedades objetivas.

De este modo no nos limitamos a una Estadística docente universitaria y tradicional cuyos aspectos mejorables he tratado en otras publicaciones, algunas de las cuales se citan en la Bibliografía, sino a resultados de investigación recientes que no se han impartido hasta la fecha de publicación de este libro en las universidades en que se estudian materias similares y en las que podrían estudiarse nuestras sugerencias.

Mariano Ruiz Espejo

# Conceptos previos

Dos conceptos que son necesarios de antemano para entender el presente libro son los de “Esperanza Matemática” y “Varianza” de una variable estadística, o en general de una variable aleatoria que podemos considerar discreta y con un número finito de posibles valores a tomar.

En concreto, suponemos que la variable estadística o aleatoria $X$ toma los $z$ posibles valores $x_1, x_2, \dots, x_i, \dots, x_z$ con probabilidades respectivas $p_1, p_2, \dots, p_i, \dots, p_z$, verficando además que $p_i \geq 0$ para todo $i = 1, 2, \dots, z$, y que

$$\sum_{i=1}^{z} p_i = 1.$$

La esperanza matemática de la variable aleatoria $X$ se define como

$$E(X) = \sum_{i=1}^{z} x_i p_i.$$

La varianza de la variable aleatoria $X$ se define como

$$V(X) = E(X^2) - [E(X)]^2 = E\{[X - E(X)]^2\} =$$

$$\sum_{i=1}^{z} [x_i - E(X)]^2 p_i.$$

Algunas propiedades del concepto de esperanza matemática son las siguientes, cuyas comprobaciones son relativamente sencillas. Sea $Y$ la variable aleatoria que toma los valores $y_1, y_2, \dots, y_i, \dots, y_z$ con probabilidades respectivas de que ocurran $p_1, p_2, \dots, p_i, \dots, p_z$. Una constante es una variable aleatoria que toma el valor único, la constante, con probabilidad uno, es decir es la misma constante en todos los casos $i$ en que se pueda dar el suceso de probabilidad $p_i$.

Si $c$ es una constante real,

$$E(c) = c.$$

Si $a$ y $b$ son constantes reales,

$$E(aX + b) = aE(X) + b.$$

Si $X$ e $Y$ son variables aleatorias,

$$E(X + Y) = E(X) + E(Y).$$

La variable aleatoria $X + Y$ es la suma de las variables aleatorias $X$ e $Y$, y toma los valores $x_i + y_i$ con probabilidad $p_i$.

Propiedades del concepto de varianza de una variable aleatoria son las siguientes, cuya comprobación es un ejercicio sencillo para el lector.

Si $c$ es una constante real,

$$V(c) = 0.$$

Si $a$ y $b$ son constantes reales,

$$V(aX + b) = a^2V(X).$$

Si $X$ e $Y$ son variables aleatorias,

$$V(X + Y) = V(X) + V(Y) + 2Cov(X, Y).$$

Donde la covarianza de las variables aleatorias $X$ e $Y$ es

$$Cov(X,Y) = E(XY) - E(X)E(Y) =$$

$$E\{[X - E(X)][Y - E(Y)]\},$$

siendo

$$E(XY) = \sum_{i=1}^{z} x_i y_i p_i$$

y

$$E\{[X - E(X)][Y - E(Y)]\} = \sum_{i=1}^{z} [x_i - E(X)][y_i - E(Y)]p_i.$$

También, $XY$ es la variable aleatoria producto de las variables aleatorias $X$ e $Y$, que toma el valor $x_i y_i$ con probabilidad $p_i$.

La varianza de una variable aleatoria $X$ se puede definir, una vez definido el concepto de covarianza, por tanto, como

$$V(X) = Cov(X,X) = E(X^2) - [E(X)]^2 =$$

$$E\{[X - E(X)]^2\}.$$

Una propiedad de la covarianza de variables aleatorias es, por ejemplo, que si $a$, $b$, $c$ y $d$ son constantes reales, y $X$, $Y$, $V$ y $W$ son variables aleatorias discretas tomando un número finito de posibles valores (con probabilidad positiva), entonces

$$Cov(aX + bY, cV + dW) = acCov(X,V) +$$

$$adCov(X,W) + bcCov(Y,V) + bdCov(Y,W).$$

Si tenemos $n$ variables aleatorias discretas $X_1, X_2, \ldots, X_n$ que toman los valores $X_i = x_{i_j}$ con probabilidad

$$p\left(X_i = x_{i_j}\right),$$

éstas variables aleatorias serán independientes si y solo si la probabilidad conjunta es igual al producto de las probabilidades marginales

$$p\left(X_1 = x_{1_j}, \dots, X_n = x_{n_j}\right) = \prod_{i=1}^{n} p\left(X_i = x_{i_j}\right).$$

Es sencillo comprobar ahora que si $X_1, X_2, \dots, X_n$ son variables aleatorias independientes y $f_1, f_2, \dots, f_n$ son $n$ funciones reales de variable real cualesquiera, entonces

$$E\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} E[f_i(X_i)].$$

Aquí, si $X$ es una variable aleatoria y $f$ es una función real de variable real, entonces $f(X)$ es por definición la variable aleatoria que toma los valores $x_j$ con probabilidad

$$p[f(X) = x_j] = \sum_{x_k : f(x_k) = f(x_j)} p(X = x_k).$$

Para demostrar la propiedad basta ver que si $m$ es el número de valores posibles que pueden tomar cualquiera de las $n$ variables aleatorias discretas con probabilidad positiva ($m$ podría ser infinito numerable, pero para nuestro objetivo en este libro basta que con que sea finito), entonces

$$E\left[\prod_{i=1}^{n} f_i(X_i)\right] =$$

$$\sum_{1_j, \dots, n_j = 1}^{m} \left[\prod_{i=1}^{n} f_i\left(x_{i_j}\right)\right] p\left(X_1 = x_{1_j}, \dots, X_n = x_{n_j}\right) =$$

$$\sum_{1_j,\ldots,n_j=1}^{m}\left[\prod_{i=1}^{n} f_i\left(x_{i_j}\right)\right] p\left(X_1 = x_{1_j}\right)\cdots p\left(X_n = x_{n_j}\right) =$$

$$\prod_{i=1}^{n}\left[\sum_{1_j,\ldots,n_j=1}^{m} f_i\left(x_{i_j}\right) p\left(X_1 = x_{1_j}\right)\cdots p\left(X_n = x_{n_j}\right)\right] =$$

$$\prod_{i=1}^{n}\left[\sum_{i_j=1}^{m} f_i\left(x_{i_j}\right) p\left(X_i = x_{i_j}\right)\right] =$$

$$\prod_{i=1}^{n} E[f_i(X_i)].$$

Ya que

$$\sum_{1_j,\ldots,n_j=1}^{m} f_i\left(x_{i_j}\right) p\left(X_1 = x_{1_j}\right)\cdots p\left(X_n = x_{n_j}\right) =$$

$$\sum_{i_j=1}^{m} f_i\left(x_{i_j}\right) p\left(X_i = x_{i_j}\right)\prod_{k\neq i}^{n}\left[\sum_{k_j=1}^{m} p\left(X_k = x_{k_j}\right)\right] =$$

$$E[f_i(X_i)] \times 1^{n-1} = E[f_i(X_i)].$$

Una consecuencia de este resultado es que si dos variables aleatorias $X$ e $Y$ son independientes, su covarianza es nula, ya que

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 0.$$

También se dice entonces que las variables aleatorias $X$ e $Y$ están incorrelacionadas, puesto que el coeficiente de correlación lineal de las variables aleatorias $X$ e $Y$ se define como

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}.$$

Y en el caso de variables aleatorias independientes $X$ e $Y$, entonces $\rho(X,Y) = 0$ , es decir, $X$ e $Y$ son variables aleatorias incorrelacionadas. La propiedad recíproca no siempre es cierta, sino que existen variables aleatorias incorrelacionadas que son dependientes.

# Capítulo 1

# Introducción

El muestreo de poblaciones finitas es un método estadístico que consiste en seleccionar un subconjunto o parte de la población de un número finito de unidades, “subconjunto o parte” que llamamos “muestra”, y con la información adquirida en dicha muestra mediante observación o encuesta de sus unidades, realizar estimaciones o inferencias sobre la población finita entera o magnitudes de ella (como la media poblacional, el total poblacional, la proporción poblacional, el porcentaje poblacional, o la varianza poblacional) y así inferir sobre aspectos importantes de la población finita de los que estamos interesados en conocer.

Al tratar de seleccionar la muestra de la población finita, surge de modo natural la pregunta de cómo seleccionar la muestra en la práctica. La respuesta viene dada por métodos probabilísticos, si queremos tener estimaciones insesgadas, o justas en promedio de las magnitudes o funciones paramétricas poblacionales.

Estudiaremos el modelo de muestreo en el que a cada unidad de la población finita se le asocia un único número real “$y$” desconocido y fijo antes de ser observado, que es el valor de la variable en estudio, también llamada “variable de interés”. Como ejemplo, la variable de interés puede ser el “número de hijos” en una población finita compuesta por todas las “familias de una región administrativa”.

Las unidades de la población están identificadas por un número que las numera a cada una. Esta identificación permite seleccionar la muestra de modo probabilístico, de modo que el estimador basado en la muestra identificada tendrá una distribución probabilística que depende del procedimiento de selección de unidades en la muestra y de los datos observados que se incorporan al estimador. Por ello, la distribución del estimador es algo que el investigador crea y controla al elegir el método de selección de la muestra y del método de estimación, y también depende de los datos fijos de la variable de interés en las unidades de la población finita que el investigador debe respetar al observarlos o al encuestar.

Una población finita (o universo) es una colección o conjunto de unidades numeradas del 1 al $N$, es decir, el conjunto

$$U = \{1, 2, \dots, k, \dots, N\},$$

donde el número entero $N$ se llama "tamaño de la población", y verifica

$$0 < N < \infty.$$

La identificabilidad de las unidades permite acceder a cualquier unidad de la población finita, si dicha unidad es seleccionada en la muestra probabilística o aleatoria concreta. En un caso concreto esta identificabilidad puede ser el listado de nombres y direcciones o teléfonos de las personas que componen la población, o bien la localización con coordenadas GPS de la posición de los árboles si la población son árboles de una plantación. Las unidades de una población finita son identificables si pueden ser numeradas unívocamente de 1 a $N$, y el número de cada unidad es conocido permitiendo la accesibilidad a la unidad por tal número para la observación de su variable de interés.

Cada unidad numerada $k$ tiene un número $y_k$ asociado cuando la característica en estudio es $y$, resultado de la medida exacta y sin error de la variable $y$ en la unidad $k$. De este modo la "observación numerada" será el par $(k, y_k)$.

El vector $\boldsymbol{y} = (y_1, y_2, \dots, y_N)$ es el "vector paramétrico" de la población finita, en el que las unidades $k$ de 1 a $N$ están localizadas por su posición en el parámetro o vector $\boldsymbol{y}$.

El "espacio paramétrico" es el espacio $N$-dimensional donde puede variar el vector paramétrico, puede ser en el caso general $\mathbb{R}^N$ si cualquier valor $y_k$ es un número real, $\mathbb{R}_+^N$ si cualquier valor $y_k$ es un número real positivo, $\{0, 1\}^N$ si $y_k$ puede tomar el valor 0 "si la unidad $k$ no posee cierta cualidad" o el valor 1 "si la unidad $k$ posee cierta cualidad", siendo $k = 1, 2, \dots,$ ó $N$.

Una función real definida sobre el espacio paramétrico se llama "función paramétrica". La inferencia en poblaciones finitas se centra en el diseño de muestreo y en la estimación de una función paramétrica especificada, y a veces teóricamente sobre el propio parámetro $\boldsymbol{y}$. Dos funciones paramétricas de importante relieve son la "media poblacional" que definimos

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k,$$

y la "varianza poblacional" que definimos

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} (y_k - \bar{y})^2,$$

en donde $\bar{y}$ aparece definida anteriormente como media poblacional. Por lo general la inferencia en poblaciones finitas se centra en inferir sobre la media poblacional, mientras que la inferencia sobre la varianza poblacional tiene un interés

suplementario al permitir a veces estimar insesgadamente esta función paramétrica y como consecuencia, también en muchos casos, permite estimar sin sesgo la varianza del estimador insesgado de la media poblacional. Indicamos que tanto la media poblacional como la varianza poblacional son dos valores reales y concretos, verdaderos, que existen objetivamente y por tanto tiene sentido estimar esas cantidades son ciertas, exactas o fijas, aunque desconocidas para los patrocinadores de la encuesta. Esto no ocurre en otros tipos de inferencia, donde solo es posible suponer la existencia de ciertos parámetros si la población fuera como se supone en dichas teorías, pero sin prueba de que esas teorías sean ciertas en las realidades a las que se desean aplicar.

Un caso particular importante de media poblacional se presenta cuando la variable de interés toma exclusivamente valores 0 ó 1, y entonces recibe el nombre de "proporción poblacional". Si llamamos $P$ a la proporción poblacional, entonces la varianza poblacional admite la expresión siguiente

$$\sigma^2 = P - P^2 = P(1 - P).$$

Llamamos "muestra ordenada" a la secuencia

$$\boldsymbol{s} = \left(k_1, k_2, \dots, k_{n(\boldsymbol{s})}\right)$$

tal que $k_i$ es la $i$-ésima unidad de la población finita según el orden de aparición en la muestra ordenada $\boldsymbol{s}$. Recibe el nombre de muestra ordenada porque conserva el orden en que van apareciendo las unidades de la población en la muestra, pudiendo aparecer unidades repetidas en distintos lugares de la muestra ordenada por un procedimiento de muestreo determinado.

El "tamaño muestral", que denotamos $n(\boldsymbol{s})$, es el número de unidades con sus repeticiones aparecidas en la muestra ordenada $\boldsymbol{s}$. Este número llamado tamaño muestral de una muestra ordenada

puede ser mayor que el tamaño poblacional $N$ cuando aparecen unidades repetidas en la muestra.

Notaremos

$$\boldsymbol{S} = \{\boldsymbol{s}: \boldsymbol{s} \text{ es muestra ordenada}\}$$

al conjunto de muestras ordenadas con un procedimiento de muestreo.

Así, por ejemplo, si la población finita es $U = \{1, 2, 3\}$, muestras ordenadas pueden ser $\boldsymbol{s}_1 = (1, 2)$, $\boldsymbol{s}_2 = (3, 1)$, $\boldsymbol{s}_3 = (2, 2)$, ó $\boldsymbol{s}_4 = (1, 3, 2, 1)$.

El "tamaño muestral efectivo" de una secuencia $\boldsymbol{s}$ es el número de componentes distintos que tiene, y se denota $\nu(\boldsymbol{s})$. Así, por ejemplo, $\nu(\boldsymbol{s}_1) = \nu(\boldsymbol{s}_2) = 2$, $\nu(\boldsymbol{s}_3) = 1$, y $\nu(\boldsymbol{s}_4) = 3$. Pero sus tamaños muestrales son $n(\boldsymbol{s}_1) = n(\boldsymbol{s}_2) = n(\boldsymbol{s}_3) = 2$, mientras que $n(\boldsymbol{s}_4) = 4$.

Dada una secuencia o muestra ordenada $\boldsymbol{s}$, podemos construir el conjunto de sus unidades distintas

$$s = \{k: k \text{ es componente de } \boldsymbol{s}\},$$

y entonces, $\nu(\boldsymbol{s}) = \text{card}(s)$, donde hemos denotado por $\text{card}(s)$ al número de unidades o elementos del conjunto $s$. Este número es siempre un número menor o igual que $N$ ya que el conjunto $s$ está contenido en la población finita $U$ cuyo cardinal es $N$, finito.

Llamamos "muestra no ordenada" a todo conjunto $s$ no vacío subconjunto de $U$, es decir que verifica $\phi \neq s \subset U$. Se llama muestra no ordenada porque no influye el orden de selección de las componentes o unidades en el conjunto $s$, así como tampoco influye la multiplicidad de unidades en la muestra. El conjunto de muestras no ordenadas y no vacías lo denotamos por

$$S = \{s: \phi \neq s \subset U\} = \wp(U) - \{\phi\},$$

pues coincide con el conjunto de partes de $U$, excluyendo al conjunto vacío. Si el conjunto $A$ tiene $\mathrm{card}(A) = N$, entonces se demuestra matemáticamente que el cardinal del conjunto de partes de $A$ es

$$\mathrm{card}[\wp(A)] = 2^N,$$

que incluye una unidad al contabilizar el conjunto vacío $\phi$.

Por tanto al excluir como elemento al conjunto vacío dentro de $S$, resulta que $\mathrm{card}(S) = 2^N - 1$, puesto que el conjunto de muestras no ordenadas $S = \wp(U) - \{\phi\}$, y la población finita $U$ tiene cardinal $N$, su "tamaño poblacional" o el número de sus elementos.

Así, por ejemplo, si $U = \{1, 2\}$, el conjunto de muestras no ordenadas será

$$S = \big\{\{1\}, \{2\}, \{1, 2\}\big\}$$

y $\mathrm{card}(S) = 2^2 - 1 = 3$ es el número de muestras no ordenadas no vacías.

El "tamaño muestral efectivo" $\nu(s)$ de una muestra no ordenada $s$ es ahora su número de elementos, es decir

$$\nu(s) = \mathrm{card}(s).$$

Hemos denotado a las muestras por los símbolos $\boldsymbol{s}$ o $s$, respetando la inicial de "sample", que significa "muestra" en inglés.

Llamamos "función de reducción" a la aplicación $r: \boldsymbol{S} \to S$ tal que

$$r(\boldsymbol{s}) = \{k \in U: k \text{ es componente de } \boldsymbol{s}\} = s,$$

es decir, la función de reducción $r$ elimina el orden y la multiplicidad de las unidades de la muestra ordenada $\boldsymbol{s}$, transformándola en una muestra no ordenada $s$.

Por ejemplo, si tenemos un conjunto de muestras ordenadas de tamaño fijo $3$, de una población finita de tamaño $3$, y una muestra ordenada es $\boldsymbol{s} = (1, 1, 2)$, entonces la función de reducción sobre esta muestra es $r(\boldsymbol{s}) = \{1, 2\} = s \in S$. También podemos obtener en este caso la relación de reducción inversa de $s$, que será

$$r^{-1}(s) = \{(1\ ,1\ ,2), (1, 2, 1), (2, 1, 1),$$

$$(1, 2, 2), (2, 1, 2), (2, 2, 1)\} \subset \boldsymbol{S}.$$

Un "diseño muestral" es una función de probabilidad sobre $\boldsymbol{S}$ o $S$. Un "diseño muestral ordenado" es una aplicación o función $p: \boldsymbol{S} \rightarrow [0, 1]$ tal que $p(\boldsymbol{s}) \geq 0$ para toda muestra ordenada $\boldsymbol{s} \in \boldsymbol{S}$, y además

$$\sum_{\boldsymbol{s} \in \boldsymbol{S}} p(\boldsymbol{s}) = 1.$$

Un "diseño muestral no ordenado" es una función $p: S \rightarrow [0, 1]$ tal que $p(s) \geq 0$ para toda muestra no ordenada $s \in S$, y además

$$\sum_{s \in S} p(s) = 1.$$

El diseño muestral puede introducirse a partir de un diseño ordenado $p(\boldsymbol{s})$, y desde éste podemos corresponder con un diseño no ordenado asociado del modo

$$p(s) = \sum_{\boldsymbol{s} \in r^{-1}(s)} p(\boldsymbol{s}),$$

siendo $r^{-1}(s) = \{\boldsymbol{s} \in \boldsymbol{S}: r(\boldsymbol{s}) = s\} \subset \boldsymbol{S}$ el conjunto de muestras ordenadas, $\boldsymbol{s}$, tales que reducidas por la función de reducción $r$ dan lugar a $s$, es decir que $r(\boldsymbol{s}) = s$. También se puede postular como punto de partida un diseño no ordenado. Por ejemplo, si $s = \{1, 2\}$, $r^{-1}(s)$ contiene las siguientes muestras ordenadas: $(1, 2), (2, 1), (1, 1, 2), (1, 2, 1), (1, 2, 2), (2, 1, 1), (2, 1, 2)$, etc. No existe, por tanto, una biyección entre el conjunto de muestras ordenadas y el conjunto de muestras no ordenadas, en general. La función de reducción $r$ no es biyectiva salvo casos triviales, como por ejemplo, con un conjunto de muestras de tamaño fijo menor o igual a 1.

Dado un diseño muestral, se define "probabilidad de inclusión" $\pi_k$ de la unidad $k \in U$ en la muestra aleatoria $\boldsymbol{s}$ o $s$, a

$$\pi_k = \sum_{\boldsymbol{s} \in \boldsymbol{S}_k} p(\boldsymbol{s})$$

o bien

$$\pi_k = \sum_{s \in S_k} p(s),$$

donde $\boldsymbol{S}_k = \{\boldsymbol{s}: k \in \boldsymbol{s}\}$ y $S_k = \{s: k \in s\}$, es decir $\pi_k$ es la suma de las probabilidades de las muestras, ordenadas o no, que tengan como componente la unidad $k \in U$.

La "probabilidad de inclusión de segundo orden" $\pi_{km}$ de las unidades $k$ y $m$ en la muestra es

$$\pi_{km} = \sum_{\boldsymbol{s} \in \boldsymbol{S}_{km}} p(\boldsymbol{s})$$

o bien

$$\pi_{km} = \sum_{s \in S_{km}} p(s),$$

donde ahora $\boldsymbol{S}_{km} = \{\boldsymbol{s}: k, m \in \boldsymbol{s}\}$ y $S_{km} = \{s: k, m \in s\}$. En este caso, se suman las probabilidades de las muestras que tengan como componentes o elementos a las unidades $k \in U$ y $m \in U$.

De modo similar se obtienen las probabilidades de inclusión de órdenes superiores $\pi_{km\ldots z}$.

Un diseño ordenado $p$ se llama "diseño de tamaño fijo" igual a $n$, si el número de componentes de $\boldsymbol{s}$, $n(\boldsymbol{s})$, es constante e igual a $n$ para toda muestra $\boldsymbol{s} \in \boldsymbol{S}$ tal que $p(\boldsymbol{s}) > 0$, y lo denotamos $TF(n)$. Un diseño ordenado (o no ordenado) se llama "diseño de tamaño efectivo fijo" igual a $\nu$, si el tamaño muestral efectivo $\nu(\boldsymbol{s})$ (o $\nu(s)$) es constante e igual a $\nu$ para toda muestra $\boldsymbol{s} \in \boldsymbol{S}$ ($s \in S$) tal que $p(\boldsymbol{s}) > 0$ ($p(s) > 0$), y lo denotaremos diseño $TEF(\nu)$.

En general, el "tamaño muestral efectivo esperado" de un diseño muestral, es

$$\bar{\nu} = \sum_{\boldsymbol{s} \in \boldsymbol{S}} \nu(\boldsymbol{s}) p(\boldsymbol{s})$$

o bien

$$\bar{\nu} = \sum_{s \in S} \nu(s) p(s).$$

El "tamaño muestral esperado de un diseño ordenado" es

$$\bar{n} = \sum_{\boldsymbol{s} \in \boldsymbol{S}} n(\boldsymbol{s}) p(\boldsymbol{s}).$$

**Ejemplo 1.1.** Sea $U = \{1, 2, 3, 4, 5\}$ y tenemos el diseño no ordenado siguiente:

$$p(\{1,2\}) = \frac{1}{3}, p(\{3,4,5\}) = \frac{1}{3}, p(\{3,4\}) = \frac{1}{3}.$$

En este caso los tamaños muestrales efectivos de las muestras son:

$$\nu(\{1,2\}) = 2, \nu(\{3,4,5\}) = 3, \nu(\{3,4\}) = 2.$$

El tamaño muestral efectivo esperado es:

$$\bar{\nu} = 2\frac{1}{3} + 3\frac{1}{3} + 2\frac{1}{3} = \frac{7}{3}.$$

Algunas de sus probabilidades de inclusión son:

$$\pi_1 = \pi_2 = p(\{1,2\}) = \frac{1}{3}.$$

$$\pi_3 = \pi_4 = p(\{3,4,5\}) + p(\{3,4\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

$$\pi_5 = p(\{3,4,5\}) = \frac{1}{3}.$$

$$\pi_{1,2} = \pi_{3,5} = \pi_{4,5} = \frac{1}{3}.$$

$$\pi_{1,3} = \pi_{1,4} = \pi_{1,5} = \pi_{2,3} = \pi_{2,4} = \pi_{2,5} = 0.$$

$$\pi_{3,4} = \frac{2}{3}.$$

**Ejemplo 1.2.** Si tenemos la población $U = \{1,2,3,4,5,6,7\}$ y el diseño muestral ordenado definido por las probabilidades

$$p(1,1,2) = p(3,2,5) = p(4,6,7) = p(6,2,5) = p(7,1,7) = \frac{1}{5}.$$

Ahora los tamaños muestrales efectivos son:

$$\nu(1,1,2) = \nu(7,1,7) = 2,$$

$$\nu(3,2,5) = \nu(4,6,7) = \nu(6,2,5) = 3.$$

El tamaño muestral efectivo esperado es

$$\bar{\nu} = 2\frac{1}{5} + 3\frac{1}{5} + 3\frac{1}{5} + 3\frac{1}{5} + 2\frac{1}{5} = \frac{13}{5}.$$

Algunas probabilidades de inclusión son:

$$\pi_1 = p(1,1,2) + p(7,1,7) = \frac{2}{5},$$

$$\pi_2 = \frac{3}{5}, \pi_3 = \frac{1}{5}, \pi_4 = \frac{1}{5}, \pi_5 = \frac{2}{5}, \pi_6 = \frac{2}{5}, \pi_7 = \frac{2}{5},$$

$$\pi_{1,2} = \frac{1}{5}, \pi_{1,3} = 0, \pi_{1,4} = 0, \pi_{1,5} = 0, \pi_{1,6} = 0, \pi_{1,7} = \frac{1}{5},$$

$$\pi_{2,3} = \frac{1}{5}, \pi_{2,4} = 0, \pi_{2,5} = \frac{2}{5}, \pi_{2,6} = \frac{1}{5}, \pi_{2,7} = 0,$$

$$\pi_{3,4} = 0, \pi_{3,5} = \frac{1}{5}, \pi_{3,6} = \pi_{3,7} = 0,$$

$$\pi_{4,5} = 0, \pi_{4,6} = \pi_{4,7} = \frac{1}{5},$$

$$\pi_{5,6} = \frac{1}{5}, \pi_{5,7} = 0, \pi_{6,7} = \frac{1}{5}.$$

Una vez que la unidad $k$ ha sido seleccionada en una muestra, se procede a su observación y medida para obtener el valor de la variable en estudio o variable de interés de modo exacto, $y_k$, por lo que disponemos del par $(k, y_k)$. El "censo" consiste en conocer el conjunto de todos los pares de este tipo, es decir, conocer el conjunto

$$\{(k, y_k): k \in U\}.$$

Sin embargo, disponer de esta colección de datos puede ser de un trabajo muy complejo y costoso, por lo que la inferencia basada en unos cuantos de ellos seleccionados aleatoriamente, permite conocer aproximadamente mediante una estimación las funciones paramétricas más importantes.

Definimos "dato ordenado" $\boldsymbol{d}$ asociado a la muestra ordenada $\boldsymbol{s}$ a la secuencia

$$\boldsymbol{d} = \big((k, y_k) : k \in \boldsymbol{s}\big).$$

El conjunto de datos ordenados lo denotamos por

$$\boldsymbol{D} = \{\boldsymbol{d} : \boldsymbol{s} \in \boldsymbol{S}\}.$$

El "dato no ordenado" $d$ es el conjunto de pares asociados a la muestra no ordenada $s$, es decir

$$d = \{(k, y_k) : k \in s\}.$$

El conjunto de datos no ordenados lo denotamos por

$$D = \{d : s \in S\}.$$

El concepto de diseño muestral puede entonces extenderse a los datos muestrales, ya que para toda muestra $p(\boldsymbol{d}) = p(\boldsymbol{s})$ y $p(d) = p(s)$, ya que la relación entre dato y muestra es biunívoca y requiere haber observado la variable de interés en las unidades de la muestra, e incorporar dichas observaciones al dato.

Un estimador $t$ es una aplicación del conjunto de datos $\boldsymbol{D}$ o $D$, y que toma valores reales, es decir $t : \boldsymbol{D} \to \mathbb{R}$, o bien $t : D \to \mathbb{R}$. El estimador $t$ es una variable aleatoria discreta que toma un número finito de valores reales $v$ con la probabilidad

$$p\{t = v\} = \sum_{\boldsymbol{d} \in \boldsymbol{D} :\, t(\boldsymbol{d}) = v} p(\boldsymbol{d})$$

si nos referimos a datos ordenados, o bien, si nos referimos a datos no ordenados,

$$p\{t = v\} = \sum_{d \in D:\, t(d)=v} p(d),$$

y cuya esperanza matemática es

$$E(t) = \sum_{\boldsymbol{d} \in \boldsymbol{D}} t(\boldsymbol{d}) p(\boldsymbol{d})$$

o bien

$$E(t) = \sum_{d \in D} t(d) p(d).$$

También podemos sustituir $\boldsymbol{d} \in \boldsymbol{D}$ por $\boldsymbol{s} \in \boldsymbol{S}$, y $d \in D$ por $s \in S$ en los índices de sumación, debido a la correspondencia biunívoca o biyección que hay entre muestras y datos. La "varianza del estimador" puede definirse como

$$V(t) = E\{[t - E(t)]^2\},$$

y el "error cuadrático medio" del estimador $t$ para estimar la función paramétrica $f(\boldsymbol{y})$ es

$$ECM[t; f(\boldsymbol{y})] = E\{[t - f(\boldsymbol{y})]^2\}.$$

Si el estimador $t$ es insesgado para estimar la función paramétrica $f(\boldsymbol{y})$, es decir, si $E(t) = f(\boldsymbol{y})$, entonces

$$ECM(t) = V(t).$$

Pero en general, cuando el estimador sea sesgado (es decir, si su esperanza matemática no coincide con la función paramétrica), no serán iguales el error cuadrático medio y la varianza del estimador, sino que $ECM[t; f(\boldsymbol{y})] = V(t) + \{B[t; f(\boldsymbol{y})]\}^2$ , siendo

$B[t; f(\boldsymbol{y})] = E(t) - f(\boldsymbol{y})$ el sesgo de $t$ para estimar $f(\boldsymbol{y})$, como demostraremos un poco más adelante.

## 1.1 Algunos resultados básicos

Veamos a continuación algunos resultados matemáticos que justifican el uso de la varianza de una variable aleatoria. También veremos propiedades de los datos y de los estimadores.

### Desigualdad de Schwarz

Si $t_1$ y $t_2$ son dos variables aleatorias o estimadores cualesquiera tal que $E(t_1^2)$ y $E(t_2^2)$ existen, entonces:

$$|E(t_1 t_2)| \leq \sqrt{E(t_1^2)E(t_2^2)}.$$

### Demostración

Sea $x$ una variable real, entonces

$$0 \leq E[(t_1 - xt_2)^2] = E(t_1^2) - 2xE(t_1 t_2) + x^2 E(t_2^2).$$

Como tenemos una ecuación de segundo grado que es siempre positiva o cero para todo valor de $x$, tiene a lo sumo una raíz la ecuación, y por tanto su discriminante tiene que ser negativo o cero. Es decir,

$$4[E(t_1 t_2)]^2 - 4E(t_1^2)E(t_2^2) \leq 0.$$

Por lo que podemos concluir que

$$|E(t_1 t_2)| \leq \sqrt{E(t_1^2)E(t_2^2)}.$$

**Teorema 1.1.** Sea $t$ una variable aleatoria cualquiera tal que $E(t^2)$ existe, entonces:

$$E[|t - E(t)|] \leq \sqrt{V(t)}.$$

**Demostración**

Haciendo uso de la desigualdad de Schwarz

$$\{E[|t - E(t)|]\}^2 = |E[|t - E(t)| \cdot 1]|^2 \leq$$

$$E\{[|t - E(t)|]^2\} \cdot E(1^2) = V(t).$$

También se puede demostrar teniendo en cuenta la convexidad de la función $f(x) = x^2$. Así, directamente

$$\{E[|t - E(t)|]\}^2 \leq E\{[t - E(t)]^2\} = V(t).$$

**Corolario 1.1.** Sea $t$ una variable aleatoria cualquiera tal que $E(t^2)$ existe, entonces:

$$E[|t - f(\boldsymbol{y})|] \leq \sqrt{ECM[t; f(\boldsymbol{y})]}.$$

**Demostración**

Sustituyendo en la anterior demostración $E(t)$ por $f(\boldsymbol{y})$.

Por tanto, la desviación típica o raíz cuadrada de la varianza del estimador $t$ acota superiormente la desviación absoluta media del estimador. Esta desviación absoluta media del estimador es su

medida de dispersión más natural y deseable, pues mide el promedio de las desviaciones absolutas del estimador respecto a su esperanza matemática. Pero tiene el inconveniente de que sus propiedades matemáticas la hacen desaconsejable. Sin embargo, esto no ocurre con la varianza del estimador que tiene buenas propiedades para su utilización, y que una vez calculada o bien estimada, su raíz cuadrada o su desviación estándar o típica acota superiormente a la desviación absoluta media del estimador. De aquí la importancia de la función paramétrica "varianza del estimador".

Denotando por "sesgo del estimador" $t$ para estimar la función paramétrica $f(\boldsymbol{y})$, a $B[t; f(\boldsymbol{y})] = E(t) - f(\boldsymbol{y})$, podemos dar otro resultado de interés. Hemos llamado $B$ al sesgo por su inicial de "bias" en inglés.

**Teorema 1.2.** El error cuadrático medio de un estimador $t$ para estimar una función paramétrica $f(\boldsymbol{y})$, es igual a la varianza del estimador más su sesgo al cuadrado, es decir:

$$ECM[t; f(\boldsymbol{y})] = V(t) + \{B[t; f(\boldsymbol{y})]\}^2.$$

**Demostración**

$$ECM[t; f(\boldsymbol{y})] = E\{[t - f(\boldsymbol{y})]^2\} =$$

$$E\{[t - E(t) + E(t) - f(\boldsymbol{y})]^2\} =$$

$$E(\{[t - E(t)] + B[t; f(\boldsymbol{y})]\}^2) =$$

$$E\{[t - E(t)]^2\} + \{B[t; f(\boldsymbol{y})]\}^2 + 2E[t - E(t)]B[t; f(\boldsymbol{y})] =$$

$$V(t) + \{B[t; f(\boldsymbol{y})]\}^2.$$

Por tanto, una propiedad importante para que el estimador sea eficiente es que su desviación cuadrática media sea pequeña, es decir que su error cuadrático medio sea pequeño. Esto puede conseguirse disminuyendo la varianza del estimador, y también disminuyendo el valor absoluto del sesgo del estimador. El sesgo del estimador puede ser cero, mientras que la varianza del estimador tendrá valor positivo en general salvo cuando se realiza un censo de la población o en casos muy particulares que carecen de relevancia para la teoría inferencial en poblaciones finitas.

**Teorema de la Esperanza Condicionada Promedio**

Si $(u, v)$ es una variable aleatoria que se concreta en un número finito de puntos tal que $p(u_i, v_j) = p_{ij}$ $(i = 1, 2, \dots, N; j = 1, 2, \dots, M)$, entonces

$$E(u) = E[E(u|v)] = E_1 E_2(u).$$

**Demostración**

$$E[E(u|v)] = \sum_{j=1}^{M} p_{.j} E(u|v_j) = \sum_{j=1}^{M} p_{.j} \sum_{i=1}^{N} u_i p_{i|j} =$$

$$\sum_{j=1}^{M} \sum_{i=1}^{N} u_i p_{ij} = \sum_{i=1}^{N} u_i \sum_{j=1}^{M} p_{ij} = \sum_{i=1}^{N} u_i p_{i.} = E(u),$$

donde hemos denotado la probabilidad condicionada

$$p_{i|j} = \frac{p_{ij}}{p_{.j}}.$$

El razonamiento es válido para poblaciones infinitas cuando existen las esperanzas matemáticas.

**Teorema de Madow**

Sea $t$ una variable aleatoria que toma un número finito de valores reales. Entonces

$$V(t) = V_1 E_2(t) + E_1 V_2(t).$$

**Demostración**

Sea $E(t) = T$. Entonces

$$V(t) = E[(t-T)^2] = E_1 E_2[(t-T)^2].$$

Pero

$$E_2[(t-T)^2] = E_2(t^2) - 2TE_2(t) + T^2 =$$

$$[E_2(t)]^2 + V_2(t) - 2TE_2(t) + T^2.$$

Ahora se promedia sobre las concreciones de la primera etapa, y como $T = E_1 E_2(t)$,

$$V(t) = E_1[E_2(t)]^2 - T^2 + E_1[V_2(t)] = V_1[E_2(t)] + E_1[V_2(t)].$$

La fórmula de Madow es generalizable a tres o más etapas, del modo

$$V(t) = V_1E_2E_3(t) + E_1V_2E_3(t) + E_1E_2V_3(t)$$

donde

$$E(t) = E_1E_2E_3(t),$$

etc.

Una generalización del Teorema de Madow es el resultado siguiente, que puede demostrarse de modo similar al ya visto anteriormente. Si $t_1$ y $t_2$ son dos variables aleatorias o dos estimadores, su covarianza incondicional es

$$Cov(t_1, t_2) = Cov_1[E_2(t_1), E_2(t_2)] + E_1[Cov_2(t_1, t_2)],$$

donde $E_2$ es la esperanza condicional, y $Cov_2$ es la covarianza condicional.

El "coeficiente de correlación" entre dos variables aleatorias $X$ e $Y$, y se denota $\rho$, se define como

$$\rho = \rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}.$$

**Desigualdad de Markov**

Si $t$ es una variable aleatoria positiva o cero, existe su esperanza matemática, y $e > 0$ una constante real, entonces:

$$E(t) \geq ep(t \geq e).$$

**Demostración**

$$E(t) = E(t \mid t < e)p(t < e) + E(t \mid t \geq e)p(t \geq e) \geq$$

$$E(t \mid t \geq e)p(t \geq e) \geq ep(t \geq e).$$

**Desigualdad de Chebychev**

Si $t$ es una variable aleatoria cuya varianza existe, y $e > 0$ es una constante real, entonces:

$$p\{|t - E(t)| < e\} \geq 1 - \frac{V(t)}{e^2}.$$

**Demostración**

Haciendo uso de la desigualdad de Markov,

$$p\{[t - E(t)]^2 \geq e^2\} \leq \frac{E\{[t - E(t)]^2\}}{e^2} = \frac{V(t)}{e^2}.$$

Entonces, la probabilidad del suceso complementario es la buscada, que es mayor o igual a 1 menos la cota superior del suceso. Si fuese $e > \sqrt{V(t)}$, la cota inferior es $1 - V(t)/e^2 > 0$.

**Generalización de la desigualdad de Chebychev**

Si $t$ es una variable aleatoria cuya varianza existe, $f(\boldsymbol{y})$ una función paramétrica cualquiera, y $e > 0$ es una constante real, entonces:

$$p\{|t - f(\boldsymbol{y})| < e\} \geq 1 - \frac{ECM[t; f(\boldsymbol{y})]}{e^2}.$$

**Demostración**

La demostración es análoga a la de la desigualdad de Chebychev y sustituyendo ahora $E(t)$ por la función paramétrica general $f(\boldsymbol{y})$. También podemos razonar directamente así:

$$ECM[t; f(\boldsymbol{y})] = E\{[t - f(\boldsymbol{y})]^2\} =$$

$$p\{[t - f(\boldsymbol{y})]^2 < e^2\}E\{[t - f(\boldsymbol{y})]^2 | [t - f(\boldsymbol{y})]^2 < e^2\} +$$

$$p\{[t - f(\boldsymbol{y})]^2 \geq e^2\}E\{[t - f(\boldsymbol{y})]^2 | [t - f(\boldsymbol{y})]^2 \geq e^2\} \geq$$

$$p\{[t - f(\boldsymbol{y})]^2 \geq e^2\}E\{[t - f(\boldsymbol{y})]^2 | [t - f(\boldsymbol{y})]^2 \geq e^2\} \geq$$

$$p[|t - f(\boldsymbol{y})| \geq e] \cdot e^2.$$

Luego,

$$p[|t - f(\boldsymbol{y})| \geq e] \leq \frac{ECM[t; f(\boldsymbol{y})]}{e^2}.$$

O bien,

$$p[|t - f(\boldsymbol{y})| < e] \geq 1 - \frac{ECM[t; f(\boldsymbol{y})]}{e^2}.$$

Además del poder explicativo de la desigualdad de Chebychev en la estimación de una función paramétrica con un estimador insesgado de la misma, en términos de probabilidad de una desviación absoluta máxima, la varianza del estimador es útil para acotar en términos esperados a la desviación absoluta debido al teorema 1.1 consecuencia de la desigualdad de Schwarz. Lo mismo podríamos decir del error cuadrático medio del estimador,

que es útil para acotar en términos esperados a la desviación absoluta del estimador $t$ respecto de la función paramétrica $f(\boldsymbol{y})$, como consecuencia del corolario 1.1.

Una consecuencia de la desigualdad de Chebychev es que si tenemos estimadores independientes $t_1, t_2, \dots, t_u$ con la misma esperanza matemática $T$, podemos decir que el estimador

$$t = \frac{t_1 + t_2 + \cdots + t_u}{u}$$

sigue siendo insesgado para estimar $T$, y además

$$p(|t - T| < e) \geq 1 - \frac{V(t_1) + V(t_2) + \cdots + V(t_u)}{u^2 e^2}.$$

**Teorema 1.3.** Si $t_1, t_2, \dots, t_u$ son $u$ estimadores o estadísticos incorrelacionados con idéntica media $T$ y existen sus varianzas, entonces un estimador insesgado de la varianza del estimador

$$t = \frac{t_1 + t_2 + \cdots + t_u}{u}$$

es el siguiente

$$\hat{V}(t) = \frac{1}{u(u-1)} \sum_{i=1}^{u} (t_i - t)^2 = \frac{1}{u(u-1)} \left( \sum_{i=1}^{u} t_i^2 - ut^2 \right).$$

**Demostración**

$$E\left[\hat{V}(t)\right] = \frac{1}{u(u-1)} \left\{ \sum_{i=1}^{u} [V(t_i) + T^2] - u[V(t) + T^2] \right\} =$$

$$\frac{1}{u(u-1)}\left[\sum_{i=1}^{u} V(t_i) - uV(t)\right] =$$

$$\frac{1}{u(u-1)}[u^2V(t) - uV(t)] = V(t).$$

El estadístico o estimador $t$ es el compendio para estimar sin sesgo $T$, y $\hat{V}(t)$ es el estimador de grupos aleatorios de la varianza de $t$.

**Teorema 1.4.** Sean dos estrategias insesgadas de la misma función paramétrica $f(y_1, y_2, \dots, y_N) = f(\boldsymbol{y})$, que denotamos por $(p_1, t_1)$ y $(p_2, t_2)$, siendo $p_1$ y $p_2$ dos diseños muestrales, y $t_1$ y $t_2$ dos estimadores asociados a sus diseños respectivos. Si la estrategia muestral $(p_1, t_1)$ dispone de un estimador insesgado de su varianza, denotémosle por $\hat{V}(p_1, t_1)$, entonces:

La estrategia insesgada $(p_2, t_2)$ dispone de un estimador insesgado de su varianza cuya expresión es

$$\hat{V}(p_2, t_2) = t_2^2 - t_1^2 + \hat{V}(p_1, t_1).$$

**Demostración**

Para ello, partimos de que por ser estrategias insesgadas

$$E(p_1, t_1) = E(p_2, t_2) = f(\boldsymbol{y}).$$

Además,

$$V(p_1, t_1) = E(p_1, t_1^2) - [f(\boldsymbol{y})]^2.$$

Y

$$V(p_2, t_2) = E(p_2, t_2^2) - [f(\boldsymbol{y})]^2.$$

Por tanto,

$$V(p_2, t_2) = E(p_2, t_2^2) - E(p_1, t_1^2) + V(p_1, t_1).$$

De donde sustituyendo las funciones paramétricas del segundo miembro de esta última ecuación por sus respectivos estimadores insesgados, obtenemos el estimador insesgado $\hat{V}(p_2, t_2)$ de la varianza $V(p_2, t_2)$. En concreto,

$$\hat{V}(p_2, t_2) = t_2^2 - t_1^2 + \hat{V}(p_1, t_1).$$

El estimador $t_1$ depende del dato seleccionado con el diseño muestral $p_1$, el estimador $t_2$ depende del dato seleccionado con el diseño muestral $p_2$, y el estimador $\hat{V}(p_1, t_1)$ depende del dato seleccionado con el diseño muestral $p_1$ y del estimador $t_1$.

Una observación a tener en cuenta es que si los diseños muestrales $p_1$ y $p_2$ no son coincidentes, entonces se tienen que seleccionar dos muestras, independientes o no, cada una de ellas de acuerdo con su diseño muestral. Si fueran coincidentes, solo sería necesario seleccionar una muestra de acuerdo con el diseño muestral común.

Para finalizar este capítulo, veamos el teorema de Rao-Blackwell que garantiza que a efectos inferenciales de la eficiencia de la estimación basta considerar el dato no ordenado, pues es suficiente para conservar la información necesaria para la estimación eficiente. Sin embargo, los diseños muestrales ordenados también se usan en la práctica porque tienen la ventaja de que cuando una unidad de la muestra está repetida, se puede ahorrar el coste en trabajo y económico de obtener el mismo dato para dicha unidad en las veces que se presenta con multiplicidad mayor o igual a 2 en la muestra ordenada.

Dado un estimador $t(\boldsymbol{d})$ definido para datos ordenados, podemos definir el estimador asociado $t^*(d)$ definido sobre el conjunto de datos no ordenados siguiente

$$t^*(d) = E(t \mid d) = \frac{\sum_{\boldsymbol{s}\in r^{-1}(s)} t(\boldsymbol{d})p(\boldsymbol{d})}{p(s)}.$$

Es decir, para el dato $d$ el estimador $t^*$ toma el valor promedio de los valores de $t(\boldsymbol{d})$ siempre que $r(\boldsymbol{d}) = d$. Observar también que

$$p(s) = \sum_{\boldsymbol{s}\in r^{-1}(s)} p(\boldsymbol{s}) = \sum_{\boldsymbol{s}\in r^{-1}(s)} p(\boldsymbol{d}).$$

**Teorema de Rao-Blackwell**

Dado un estimador sobre datos ordenados $t$, el estimador asociado $t^*$ sobre datos no ordenados verifica:

1. $E(t^*) = E(t)$.
2. $V(t^*) \leq V(t)$.
3. $ECM(t^*) \leq ECM(t)$.

**Demostración**

1. Como

$$E(t^*) = E[E(t \mid d)] = E(t).$$

2. Como

$$V(t^*) = E[(t^*)^2] - [E(t^*)]^2,$$

y

$$V(t) = E(t^2) - [E(t)]^2,$$

basta con probar que

$$E[(t^*)^2] = E\{[E(t \mid d)]^2\} \leq E[E(t^2 \mid d)] = E(t^2).$$

3. Cierto porque

$$ECM(t^*) = V(t^*) + [B(t^*)]^2 \leq V(t) + [B(t)]^2 = ECM(t),$$

al ser $V(t^*) \leq V(t)$ y $B(t^*) = B(t)$, ya que $E(t^*) = E(t)$.

## 1.2 Ejercicios resueltos

**Ejercicio 1.1.** Un mismo diseño muestral y estimador insesgado se han empleado en dos ocasiones sucesivas independientes para estimar cierta función paramétrica poblacional. Obtener una estimación mejorada de la misma función paramétrica poblacional, y un estimador insesgado de su varianza.

**Solución.** Sean $t_1$ y $t_2$ las dos estimaciones obtenidas con el mismo diseño muestral y estimador insesgado de $T$. Otro estimador insesgado de la misma función paramétrica $T$ es entonces

$$t = \frac{t_1 + t_2}{2},$$

puesto que $E(t) = [E(t_1) + E(t_2)]/2 = 2T/2 = T$.

Su varianza es menor que cada una de las varianzas de $t_1$ y de $t_2$, pues como

$$V(t_1) = V(t_2) = V,$$

la varianza de $t$ verifica

$$V(t) = V\left(\frac{t_1 + t_2}{2}\right) = \frac{V(t_1) + V(t_2)}{4} = \frac{V}{2} \leq V.$$

Es decir, se ha reducido a la mitad, por lo que el estimador propuesto $t$ es más preciso que $t_1$ y que $t_2$.

Un estimador insesgado de la varianza de $t$ es el estimador insesgado de la varianza para grupos aleatorios que puede escribirse así

$$\hat{V}(t) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (t_i - t)^2$$

donde en nuestro caso, es $n = 2$. Es decir,

$$\hat{V}(t) = \frac{1}{2}[t_1^2 + t_2^2 - 2t(t_1 + t_2) + 2t^2] =$$

$$E(t_1^2) - t^2.$$

Así pues,

$$E\left[\hat{V}(t)\right] = EE(t_1^2) - E(t^2) = V + T^2 - \left(\frac{V}{2} + T^2\right) = \frac{V}{2} = V(t).$$

**Ejercicio 1.2.** El estimador insesgado de la varianza de un estimador $t$, $\hat{V}(t)$, ¿sirve para acotar una medida promedio de la desviación absoluta del estimador respecto a su media?

**Solución.** Una consecuencia de la desigualdad de Schwarz es que si existe la varianza de $t$, como ocurre en la inferencia en poblaciones finitas con observaciones fijas, entonces

$$E[|t - E(t)|] \leq \sqrt{V(t)}.$$

Por tanto,

$$\{E[|t - E(t)|]\}^2 \leq V(t).$$

Así podemos decir que el cuadrado de la desviación absoluta del estimador a su media, está acotado superiormente por la varianza del estimador. Sin embargo, aunque el estimador insesgado de la varianza del estimador estima sin sesgo la varianza del estimador, puede oscilar aleatoriamente y algún valor pequeño de $\hat{V}(t)$ podría no ser cota superior del cuadrado de

$$E[|t - E(t)|],$$

que es esta última fórmula la desviación absoluta media del estimador. En conclusión, el estimador insesgado de la varianza no sirve como cota superior en general. Sin embargo, como este estimador de la varianza $\hat{V}(t)$ suele converger a la varianza $V(t)$ cuando el tamaño muestral es suficientemente grande, tal estimador puede ser una buena aproximación a la cota superior $V(t)$ y valer como cota superior estimada del cuadrado del promedio de la desviación absoluta de $t$.

**Ejercicio 1.3.** Aplicar la desigualdad de Chebychev para estimar por intervalo la media poblacional, conociendo el estimador insesgado $t$ de la media poblacional, y un estimador insesgado de la varianza del estimador, $\hat{V}(t)$, que converge a $V(t)$ para muestras de gran tamaño.

**Solución.** La desigualdad de Chebychev nos garantiza que

$$p\{|t - E(t)| < e\} \geq 1 - \frac{V(t)}{e^2} \approx 1 - \frac{\hat{V}(t)}{e^2}.$$

De esta manera, el intervalo de confianza para la media poblacional es

$$(t - e, t + e),$$

y tiene una probabilidad aproximadamente mayor o igual (tanto más aproximada cuanto la convergencia de $\hat{V}(t)$ sea más rápida) que el valor estimado

$$1 - \frac{\hat{V}(t)}{e^2}.$$

**Ejercicio 1.4.** La media aritmética de diez estimaciones con el mismo diseño muestral y estimador insesgado independiente, es 3. Si la suma de las estimaciones insesgadas de las varianzas de cada uno de los diez estimadores es 1, obtener una cota aproximada del nivel de confianza del intervalo $(1, 5)$ para estimar la esperanza de cada uno de los diez estimadores.

**Solución.** Sea la media aritmética

$$t = \frac{1}{10}\sum_{i=1}^{10} t_i$$

de las diez estimaciones. El intervalo de confianza para la esperanza de $t$ se obtiene con la desigualdad de Chebychev

$$p\{|t - E(t)| < e\} \geq 1 - \frac{V(t)}{e^2} \approx 1 - \frac{\sum_{i=1}^{10} \hat{V}(t_i)}{10^2 e^2} =$$

$$1 - \frac{1}{100e^2}$$

Como el intervalo de confianza es $(1, 5)$ y $t = 3$, resulta que la amplitud del intervalo es $e = 2$, por lo que la cota inferior del nivel de confianza aproximado es

$$1 - \frac{1}{100e^2} = \frac{399}{400}.$$

Luego, es muy probable o prácticamente casi seguro que la media poblacional esté en el intervalo $(1,5)$. En concreto con probabilidad superior aproximadamente a $399/400$.

**Ejercicio 1.5.** Con los datos del Ejercicio anterior, si nos piden contrastar la hipótesis de que la media poblacional es 4, ¿se aceptará la hipótesis al nivel de confianza $399/400$? Y si la hipótesis fuera que la media poblacional es 6, ¿se aceptará la hipótesis al mismo nivel de confianza?

**Solución.** Como el intervalo de confianza $(1,5)$ tiene un nivel de confianza superior o igual aproximadamente a $399/400$, como

$$4 \in (1,5) = R$$

se acepta la hipótesis de que la media poblacional sea 4 a ese nivel de confianza al menos, pues 4 pertenece a la "región de aceptación" $R$. También, como

$$6 \notin (1,5)$$

la hipótesis de que la media poblacional es 6 se rechaza con ese nivel de confianza al menos, aproximadamente. Esto se debe a que 6 no pertenece a la región de aceptación $R$, es decir, 6 pertenece a la "región crítica" o de rechazo $R^c = \mathbb{R} - R$, complementaria a la región de aceptación.

**Ejercicio 1.6.** Sea $t$ un estimador insesgado de la función paramétrica $\alpha$ en una población finita con observaciones fijadas. ¿Qué nivel de confianza mínimo aproximado nos asegura que la función paramétrica se encuentra en el intervalo de confianza $(t-e, t+e)$?

**Solución.** El nivel de confianza mínimo nos viene dado por la desigualdad de Chebychev

$$p\{|t-\alpha|<e\}\geq 1-\frac{V(t)}{e^2}$$

Así el nivel de confianza mínimo para que

$$\alpha\in(t-e,t+e)$$

viene dado por

$$1-\frac{V(t)}{e^2},$$

de donde podemos aproximar esta cota inferior del nivel de confianza por

$$1-\frac{\hat{V}(t)}{(t-\alpha_0)^2}$$

siendo $\hat{V}(t)$ un estimador insesgado de la varianza del estimador $t$, y $\alpha_0$ el valor concreto de la función paramétrica que deseamos contrastar o valor de $\alpha$ en la hipótesis nula a contrastar. Así tenemos un valor aproximado del mínimo nivel de confianza que aceptaría la hipótesis de que la función paramétrica poblacional $\alpha$ tomara el valor $\alpha_0$.

**Ejercicio 1.7.** Demostrar la siguiente relación:

$$2N\sum_{k=1}^{N}(y_k-\bar{y})^2=\sum_{k=1}^{N}\sum_{m\neq k}^{N}(y_k-y_m)^2,$$

donde

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k.$$

**Solución.** Desarrollando por el segundo término de la igualdad pedida,

$$\sum_{k=1}^{N}\sum_{m\neq k}^{N}(y_k - y_m)^2 = \sum_{k=1}^{N}\sum_{m=1}^{N}(y_k - y_m)^2 =$$

$$\sum_{k=1}^{N}\sum_{m=1}^{N} y_k^2 + \sum_{k=1}^{N}\sum_{m=1}^{N} y_m^2 - 2\sum_{k=1}^{N}\sum_{m=1}^{N} y_k y_m =$$

$$2N\sum_{k=1}^{N} y_k^2 - 2N^2\bar{y}^2 = 2N^2\left(\frac{1}{N}\sum_{k=1}^{N} y_k^2 - \bar{y}^2\right) =$$

$$2N^2\frac{1}{N}\sum_{k=1}^{N}(y_k - \bar{y})^2 = 2N\sum_{k=1}^{N}(y_k - \bar{y})^2.$$

# Capítulo 2

# Muestreo aleatorio simple

En este capítulo vamos a estudiar el procedimiento de "muestreo aleatorio simple", al que se le suele añadir la expresión "con reemplazamiento". Le llamamos así porque es el nombre que se da al tipo de muestreo usado en la inferencia estadística tradicional, es decir tomando observaciones independientes e idénticamente distribuidas entre las unidades de la población finita. Para las sucesivas selecciones de las unidades se reincorporan las unidades anteriormente seleccionadas, de aquí la denominación "con reemplazamiento". Se le suele denotar por las siglas *mas*. También se la denota por las siglas *mpir* de "muestreo de probabilidades iguales con reemplazamiento".

## 2.1 Diseño *mas*

El diseño de "muestreo aleatorio simple" o "muestreo aleatorio simple con reemplazamiento" es un diseño muestral ordenado $p$ definido sobre las muestras ordenadas concretadas en las secuencias de tamaño fijo $n$. Es un diseño $TF(n)$ por tanto.

Este diseño muestral puede definirse como el diseño ordenado $p$ sobre el conjunto de muestras ordenadas $\boldsymbol{S}$, de modo que cada secuencia $\boldsymbol{s} \in \boldsymbol{S}$ de tamaño muestral $n(\boldsymbol{s}) = n$ tiene una probabilidad de ser seleccionada $p(\boldsymbol{s}) = 1/N^n$, y para las restantes secuencias $p(\boldsymbol{s}) = 0$.

Una caracterización de este diseño sería reproducible seleccionando una bola de una urna que contiene $N$ bolas, numeradas del 1 al $N$. Una vez seleccionada una bola, se anota su número como la primera componente de la secuencia y seguidamente se reincorpora la bola extraída a la urna, de modo que en la segunda selección se obtenga con igual probabilidad también cualquier unidad de la 1 a la $N$, independientemente del resultado de la primera extracción. Luego se reincorpora la segunda bola seleccionada a la urna de nuevo. Repitiendo este proceso $n$ veces, se selecciona una secuencia de tamaño muestral $n$, con $0 < n$.

Con este diseño muestral ordenado, las distribuciones marginales de la secuencia son iguales e independientes entre sí. Existen en este diseño $N^n$ muestras ordenadas de tamaño fijo $n$. El diseño muestral verifica que

$$\sum_{\boldsymbol{s}\in\boldsymbol{S}} p(\boldsymbol{s}) = N^n \frac{1}{N^n} = 1.$$

Las probabilidades de inclusión en este diseño muestral *mas* son

$$\pi_k = p(k \in \boldsymbol{s}) = 1 - p(k \notin \boldsymbol{s}) = 1 - \left(1 - \frac{1}{N}\right)^n$$

para toda unidad $k$ de la población finita. Observar que si $k_i$ es la unidad $i$-ésima de la secuencia muestral

$$p(k \notin \boldsymbol{s}) = \prod_{i=1}^{n} p(k_i \neq k) = \prod_{i=1}^{n} \frac{N-1}{N} = \left(1 - \frac{1}{N}\right)^n.$$

Las probabilidades de inclusión de segundo orden son

$$\pi_{km} = p(k \text{ y } m \in \boldsymbol{s}) = 1 - p(k \text{ o } m \notin \boldsymbol{s}) =$$

$$1 - p(k \notin \boldsymbol{s}) - p(m \notin \boldsymbol{s}) + p(k \text{ y } m \notin \boldsymbol{s}) =$$

$$1 - \left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n =$$

$$1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n,$$

para todo par de unidades $k$ y $m$ distintas de la población finita.

## 2.2 Estimación de la media poblacional en *mas*

El estimador usual de la media poblacional $\bar{y}$ con este diseño muestral es la media muestral $\bar{y}_{\boldsymbol{s}}$ cuya representación es

$$\bar{y}_{\boldsymbol{s}} = \frac{1}{n}\sum_{i=1}^{n} y_{k_i} = \frac{1}{n}\sum_{k \in \boldsymbol{s}} y_k \,,$$

siendo $k_i$ la $i$-ésima unidad de la secuencia muestral ordenada, es decir cuando la muestra ordenada es $\boldsymbol{s} = (k_1, k_2, \ldots, k_n)$.

Esta media muestral es insesgada para estimar la media poblacional. En efecto, la esperanza matemática de $\bar{y}_{\boldsymbol{s}}$ coincide con $\bar{y}$.

$$E(\bar{y}_{\boldsymbol{s}}) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_{k_i}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(y_{k_i}\right) = \frac{1}{n} n\bar{y} = \bar{y},$$

por distribuirse idénticamente la variable $y_{k_i}$ a la variable $y$ equiprobable en todas las unidades de la población finita, es decir

$$E\left(y_{k_i}\right) = E(y) = y_1 \frac{1}{N} + y_2 \frac{1}{N} + \cdots + y_N \frac{1}{N} = \bar{y}.$$

Por tanto el sesgo de la media muestral con diseño *mas* para estimar la media poblacional es

$$B(\bar{y}_s;\ \bar{y}) = E(\bar{y}_s) - \bar{y} = 0.$$

La varianza de la media muestral $\bar{y}_s$ puede obtenerse así

$$V(\bar{y}_s) = V\left(\frac{1}{n}\sum_{i=1}^{n} y_{k_i}\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} y_{k_i}\right) =$$

$$\frac{1}{n^2}\sum_{i=1}^{n} V(y_{k_i}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

debido a que las variables $y_{k_i}$ son independientes e idénticamente distribuidas a la población con unidades equiprobables.

## 2.3 Estimación de la varianza en *mas*

Un estimador insesgado de la varianza poblacional $\sigma^2$ para el diseño *mas* es la cuasivarianza muestral definida como

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_{k_i} - \bar{y}_s)^2.$$

En efecto,

$$(n-1)s^2 = \sum_{i=1}^{n} (y_{k_i} - \bar{y}_s)^2 = \sum_{i=1}^{n} (y_{k_i} - \bar{y} + \bar{y} - \bar{y}_s)^2 =$$

$$\sum_{i=1}^{n} (y_{k_i} - \bar{y})^2 + n(\bar{y} - \bar{y}_s)^2 + 2(\bar{y} - \bar{y}_s)\sum_{i=1}^{n} (y_{k_i} - \bar{y}) =$$

$$\sum_{i=1}^{n} (y_{k_i} - \bar{y})^2 - n(\bar{y}_s - \bar{y})^2.$$

Por lo que tomando esperanzas matemáticas en el primer y el último miembros, tenemos que

$$(n-1)E(s^2) = \sum_{i=1}^{n} E\left[\left(y_{k_i} - \bar{y}\right)^2\right] - nE[(\bar{y}_s - \bar{y})^2] =$$

$$n\sigma^2 - nV(\bar{y}_s) = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2.$$

De donde deducimos simplificando que $E(s^2) = \sigma^2$, es decir que la cuasivarianza muestral es insesgada en el muestreo aleatorio simple para estimar la varianza poblacional. También podemos escribir que el sesgo $B(s^2;\, \sigma^2) = 0$. Además $s^2$ es estimador óptimo de $\sigma^2$ para distribución libre (Zacks, 1971, p. 150).

Como consecuencia, ya que la media muestral $\bar{y}_s$ es insesgada para estimar la media poblacional $\bar{y}$ y su varianza es $\sigma^2/n$, un estimador insesgado de esta varianza de la media muestral es $s^2/n$. También es usual denotarlo del modo

$$\hat{V}(\bar{y}_s) = \frac{s^2}{n}.$$

Ciertamente,

$$E\left[\hat{V}(\bar{y}_s)\right] = E\left(\frac{s^2}{n}\right) = \frac{1}{n}E(s^2) = \frac{1}{n}\sigma^2 = V(\bar{y}_s).$$

O bien,

$$B\left[\hat{V}(\bar{y}_s); V(\bar{y}_s)\right] = 0.$$

## 2.4 Estimación del total poblacional en *mas*

La función paramétrica “total poblacional” es definida como

$$T = N\bar{y} = \sum_{k=1}^{N} y_k.$$

Un estimador insesgado de $T$, que emplea la información del tamaño poblacional $N$, es $\hat{T} = N\bar{y}_s$. En efecto,

$$E(\hat{T}) = E(N\bar{y}_s) = NE(\bar{y}_s) = N\bar{y} = T.$$

La varianza del estimador $N\bar{y}_s$ es

$$V(\hat{T}) = V(N\bar{y}_s) = N^2 V(\bar{y}_s) = N^2 \frac{\sigma^2}{n}.$$

Un estimador insesgado de esta varianza es $\hat{V}(\hat{T}) = N^2 s^2/n$. En efecto,

$$E[\hat{V}(\hat{T})] = E\left(N^2 \frac{s^2}{n}\right) = \frac{N^2}{n} E(s^2) = \frac{N^2}{n}\sigma^2 = V(\hat{T}).$$

## 2.5 Estimación de la proporción poblacional en *mas*

La "proporción poblacional" $P$ es la función paramétrica "media poblacional" $\bar{y}$ cuando la variable de interés $y$ toma valor 1 ó 0 en cada unidad de la población según posea o no una cualidad respectivamente la unidad. Por ejemplo, tener sexo varón al nacer una persona si la población finita es de seres humanos. La proporción poblacional será

$$P = \frac{1}{N}\sum_{k=1}^{N} y_k = \bar{y},$$

pero ahora

$$y_k = \begin{cases} 1 & \text{si la unidad } k \text{ posee la cualidad} \\ 0 & \text{si la unidad } k \text{ no posee la cualidad} \end{cases}$$

Se trata por tanto de un caso particular de media poblacional que toma valores comprendidos entre 0 y 1.

El estimador insesgado de la proporción poblacional $P$ es la proporción muestral $\hat{P} = \bar{y}_s$, cuya varianza puede expresarse ahora

$$V(\hat{P}) = \frac{\sigma^2}{n} = \frac{PQ}{n},$$

siendo $Q = 1 - P$ la proporción poblacional de unidades que no poseen la cualidad; ya que al tomar $y_k$ valores 1 ó 0, $y_k^2 = y_k$, deducimos que

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{N} y_k^2 - \bar{y}^2 = P - P^2 = P(1-P) = PQ.$$

Del mismo modo, la varianza muestral es $\hat{P}\hat{Q} = (n-1)s^2/n$, siendo $\hat{Q} = 1 - \hat{P}$ la proporción muestral de unidades que no poseen la cualidad. Por tanto, un estimador insesgado de la varianza de la proporción muestral es

$$\hat{V}(\hat{P}) = \frac{\hat{P}\hat{Q}}{n-1}.$$

La estimación del "porcentaje poblacional" es un caso similar al de la "proporción poblacional", ya que el porcentaje es la proporción multiplicada por 100. El estimador insesgado del porcentaje $100P$, es $100\hat{P}$, cuyo estimador insesgado de su varianza es

$$\hat{V}(100\hat{P}) = \frac{10^4\hat{P}\hat{Q}}{n-1}.$$

Su uso es muy práctico, por ejemplo, en la estimación del porcentaje de voto a un partido político.

## 2.6 Tamaño de la muestra con *mas*

La pregunta que nos hacemos es ¿cuál es el tamaño muestral $n$ necesario para alcanzar un "error máximo de muestreo" $e$ con una probabilidad $1-\alpha$? A "$1-\alpha$" se le llama "nivel de confianza" de la estimación. En el muestreo aleatorio simple, la media muestral $\bar{y}_s$ verifica

$$p\{|\bar{y}_s - \bar{y}| < e\} \geq 1 - \frac{V(\bar{y}_s)}{e^2} = 1 - \alpha.$$

De esta manera obligamos a que $1-\alpha$ sea una cota inferior del nivel de confianza verdadero de la estimación. Luego,

$$\alpha = \frac{V(\bar{y}_s)}{e^2} = \frac{\sigma^2}{ne^2}$$

que implica que

$$n = \frac{\sigma^2}{\alpha e^2}$$

es el tamaño muestral que asegura tener un error absoluto máximo de muestreo $e$ con un nivel de confianza mayor o igual a $1-\alpha$. Así, una vez fijados $e$ y $1-\alpha$, el tamaño muestral buscado es una función paramétrica proporcional a la varianza muestral, por lo que es estimable insesgadamente por

$$\hat{n} = \frac{s_0^2}{\alpha e^2},$$

donde $s_0^2$ es la cuasivarianza muestral en una muestra piloto de tamaño $n_0$, previa al estudio con diseño $mas$. En el caso de estimar una proporción poblacional $P$, se verificaría

$$\hat{n} = \frac{n_0 \hat{P}_0 \hat{Q}_0}{(n_0 - 1)\alpha e^2},$$

siendo $\hat{P}_0$ y $\hat{Q}_0$ las proporciones muestrales respectivas de la muestra piloto de tamaño $n_0$.

En el caso de la estimación de un porcentaje $100P$, el tamaño muestral estimado insesgado a partir de la muestra piloto de tamaño $n_0$ verificaría

$$\hat{n} = \frac{n_0 10^4 \hat{P}_0 \hat{Q}_0}{(n_0 - 1)\alpha e^2}.$$

Al tratarse el diseño *mas* de un procedimiento de muestreo cuyas observaciones son idénticamente distribuidas a la población finita e independientes entre sí, podemos utilizar el Teorema Central del Límite y aproximar la distribución del estimador media muestral $\bar{y}_s$ por la distribución normal de la misma media $\bar{y}$ y la misma varianza $\sigma^2/n$. Esto es especialmente práctico cuando el tamaño muestral es grande y nos permitiría obtener tamaños muestrales aproximados al hacer uso de la distribución normal ya tabulada. La idea se formaliza haciendo la aproximación a la distribución normal de parámetros 0 y 1 , estandarizando o tipificando la variable aleatoria media muestral. Como tenemos la distribución aproximada

$$\frac{\sqrt{n}(\bar{y}_s - \bar{y})}{\sigma} \cong N(0,1),$$

podemos buscar en las tablas de la distribución normal el valor $\lambda_\alpha$ tal que

$$p[|N(0,1)| < \lambda_\alpha] = 1 - \alpha.$$

Una vez obtenido el valor tabular $\lambda_\alpha$, tenemos que

$$\left|\frac{\sqrt{n}(\bar{y}_s - \bar{y})}{\sigma}\right| < \lambda_\alpha.$$

Es decir, con probabilidad aproximada $1 - \alpha$

$$|\bar{y}_s - \bar{y}| < \frac{\lambda_\alpha \sigma}{\sqrt{n}} = e,$$

donde $e$ es el error absoluto máximo de muestreo que consideramos. De donde, despejando $n$ tenemos el valor aproximado

$$n = \frac{\lambda_\alpha^2 \sigma^2}{e^2}$$

que puede ser estimado insesgadamente en una muestra piloto previa por

$$\hat{n} = \frac{\lambda_\alpha^2 s_0^2}{e^2},$$

siendo $s_0^2$ la cuasivarianza muestral piloto, que es un estimador insesgado de la varianza poblacional $\sigma^2$.

## 2.7 Ejercicios resueltos

**Ejercicio 2.1.** Disponemos de una población finita de tamaño $N = 5$ y queremos estimar la media poblacional con diseño *mas* de tamaño fijo $n = 3$. Proponer un estimador insesgado de la media

poblacional, estimar ésta y estimar sin sesgo la varianza del estimador propuesto, en estos casos:

a) Si la muestra es $\boldsymbol{s} = (1, 2, 2)$, $y_1 = 4$ e $y_2 = 8$.
b) Si la muestra es $\boldsymbol{s} = (1, 3, 2)$, y además $y_3 = 6$.

**Solución.** El estimador propuesto en ambos casos es la media muestral $\bar{y}_{\boldsymbol{s}}$, que es insesgado para estimar la media poblacional.

a) La media muestral se concreta en este caso así

$$\bar{y}_{(1,2,2)} = \frac{y_1 + y_2 + y_2}{3} = \frac{4 + 8 + 8}{3} = \frac{20}{3}.$$

Una estimación sin sesgo de la varianza de la media muestral viene proporcionada por el estimador

$$\hat{V}(\bar{y}_{\boldsymbol{s}}) = \frac{s^2}{n},$$

donde $n = 3$, y

$$s^2 = \frac{1}{n-1}\sum_{k \in \boldsymbol{s}} (y_k - \bar{y}_{\boldsymbol{s}})^2 = \frac{1}{2}\left(\frac{64}{9} + \frac{16}{9} + \frac{16}{9}\right) = \frac{16}{3},$$

por lo que la estimación insesgada de la varianza de la media muestral es $16/9$.

b) $\bar{y}_{(1,3,2)} = 6$ y la estimación insesgada de su varianza es $4/3$.

**Ejercicio 2.2.** En el problema anterior, ¿qué muestra es más precisa para estimar la media poblacional?

**Solución.** Ambas muestras ordenadas son concreciones del mismo diseño *mas*, por lo que no puede afirmarse que una muestra sea más precisa que otra, ya que la precisión de un estimador se define como la inversa de la varianza del estimador, y esta varianza incluye en

su cálculo a todas las concreciones de las muestras. La precisión no se define para una estimación a no ser que conozcamos también la media poblacional y calculemos su desviación absoluta; en este caso la estimación más precisa sería la que tenga menor desviación absoluta. Como desconocemos la totalidad del parámetro, no podemos calcular la media poblacional y por tanto tampoco la desviación de las dos estimaciones obtenidas a aquélla. En conclusión, no podemos comparar las muestras según su precisión.

Aparentemente la muestra más precisa es la que proporciona una estimación de la varianza del estimador más pequeña, concretamente $4/3 < 16/9$, pero esto no indica necesariamente que las estimaciones de las medias muestrales asociadas guarden un orden de precisión, pues se desconocen los valores $y_4$ e $y_5$ del parámetro y dependiendo de ellos pueden darse un caso u otro o incluso la igualdad en desviación absoluta.

**Ejercicio 2.3.** Tenemos una población de tamaño $N = 1000$ y queremos estimar la media poblacional con un error máximo de muestreo $e = 2$ y con un nivel de confianza mínimo de $0.95$. ¿Qué tamaño muestral es necesario con diseño *mas* para que se verifiquen estas condiciones? Aceptamos que de una muestra piloto, hemos estimado sin sesgo la varianza poblacional por $s_0^2 = 7$.

**Solución.** Aplicando la fórmula obtenida del estimador insesgado del tamaño muestral para cualquier caso,

$$\hat{n} = \frac{s_0^2}{\alpha e^2} = \frac{7}{0.05 \cdot 2^2} = 35 \text{ selecciones.}$$

Admitiendo la hipótesis de normalidad en la distribución del estimador media muestral, el valor de $\alpha = 0.05$ determina un valor

en las tablas de la distribución $N(0,1)$ de $\lambda_{0.05} = 1.96$, por lo que entonces el tamaño muestral estimado sería

$$\hat{n} = \frac{\lambda_\alpha^2 s_0^2}{e^2} = \frac{1.96 \cdot 7}{2^2} \approx 6.7$$

Por lo que bastaría tomar un tamaño muestral de 7 selecciones de unidades. Observar que entonces el Teorema Central del Límite no sería aplicable a la media muestral, pues su tamaño muestral 7 es muy pequeño y el resultado de convergencia es asintótico, cuando $n$ es grande. No obstante el uso de la aproximación por la distribución normal sería razonable si la población finita tuviera una función de cuantía uniforme discreta concentrada de modo cercano a la forma acampanada de aquélla.

En cualquier caso la interpretación es que en el primer caso obtuvimos un tamaño muestral válido para cualquier distribución uniforme discreta de la población finita y asegurando el nivel de confianza, mientras que en el segundo caso se hace una hipótesis aproximativa concreta a la distribución normal que podría fallar, y un nivel de confianza concreto exacto bajo dicha hipótesis falible.

**Ejercicio 2.4.** Se desea estimar la renta total mensual de un colectivo de 200 trabajadores de una planta industrial. A este efecto, se selecciona una muestra aleatoria simple con reemplazamiento de tamaño 20, resultando una media muestral de 1680 euros y una cuasivarianza muestral de 40000 euros al cuadrado. Proponer un estimador insesgado de la renta total, estimarla y estimar sin sesgo su varianza.

**Solución.** En este ejercicio, el tamaño poblacional es $N = 200$, número total de trabajadores de la planta; el tamaño muestral es $n = 20$ selecciones de trabajadores encuestados en la muestra $\boldsymbol{s}$; la renta media muestral es $\bar{y}_{\boldsymbol{s}} = 1680$ euros; la cuasivarianza

muestral es $s^2 = 40000$ euros al cuadrado. La renta total es un total poblacional que puede estimarse por

$$N\bar{y}_s = 200 \cdot 1680 = 336000 \text{ euros.}$$

Un estimador insesgado de la varianza del estimador $N\bar{y}_s$, $V(N\bar{y}_s)$, es el estimador siguiente que se concreta en la estimación que le sigue

$$\hat{V}(N\bar{y}_s) = \frac{N^2}{n}s^2 = \frac{200^2}{20}40000 = 8 \cdot 10^7 \text{ euros al cuadrado.}$$

**Ejercicio 2.5.** Se quiere conocer una estimación insesgada de la varianza del estimador "proporción muestral" y "porcentaje muestral" que ha resultado ser del 4% de productos defectuosos de entre 100 selecciones aleatorias por diseño *mas* de entre los 3546 productos terminados en una fábrica.

**Solución.** En este ejercicio, el tamaño poblacional es $N = 3546$ productos terminados, el tamaño muestral es $n = 100$ selecciones de entre los productos acabados, la proporción muestral es $\hat{P} = 0.04$ y el porcentaje muestral es $100\hat{P} = 4\%$. Los estimadores insesgados de las varianzas y su concreción para la muestra obtenida son

$$\hat{V}(\hat{P}) = \frac{\hat{P}\hat{Q}}{n-1} = \frac{0.04 \cdot 0.96}{99} \approx 0.000388$$

y

$$\hat{V}(100\hat{P}) = 10^4\hat{V}(\hat{P}) \approx 10^4 \cdot 0.000388 = 3.88$$

**Ejercicio 2.6.** Una población finita ("universo") $U$ de tamaño $N$ tiene un subconjunto de unidades (que llamamos "dominio") $D \subset U$ de tamaño $M \leq N$. Tomamos una muestra aleatoria simple de tamaño $n$ de la población finita. Demostrar que la submuestra de tamaño $m$, $0 \leq m \leq n$, de la muestra aleatoria simple de unidades que están en el dominio $D$ constituye una muestra aleatoria simple de la población finita $D$ de $M$ unidades.

**Solución.** Tenemos que demostrar que la submuestra ordenada $\boldsymbol{s}_m$ de tamaño $m$ en $D$ de la muestra aleatoria simple $\boldsymbol{s}_n$ de tamaño fijo $n$ en $U$, tiene una probabilidad de selección

$$p(\boldsymbol{s}_m|\boldsymbol{s}_n, m) = \frac{1}{M^m}.$$

Para ello usamos la definición de la probabilidad condicionada. Si la muestra aleatoria simple $\boldsymbol{s}_n$ ha sido la muestra seleccionada, y $\boldsymbol{s}_m$ es consistente con la muestra anterior $\boldsymbol{s}_n$, es decir, hay tantas unidades de $D$ en $\boldsymbol{s}_n$ como unidades tiene $\boldsymbol{s}_m$,

$$p(\boldsymbol{s}_m|\boldsymbol{s}_n, m) = \frac{1}{M} \cdots \frac{1}{M} = \frac{1}{M^m},$$

siendo la probabilidad $p(\boldsymbol{s}_q|\boldsymbol{s}_n, m) = 0$ en el resto de los casos, es decir, cuando $q$ no sea el número de unidades $m$ de la muestra aleatoria simple ordenada $\boldsymbol{s}_n$ en el dominio $D$. En general el número $m$ es aleatorio. La distribución de probabilidad del valor concreto $m = 0, 1, \ldots, n$, se distribuye binomial de parámetros $n$ y $M/N$. Es decir,

$$p(m|\boldsymbol{s}_n) = \binom{n}{m} \left(\frac{M}{N}\right)^m \left(\frac{N-M}{N}\right)^{n-m}.$$

Por lo que la probabilidad de $\boldsymbol{s}_m$ condicionada a que el tamaño submuestral es $m$ es un diseño muestral ordenado caracterizado por ser una muestra aleatoria simple con reemplazamiento de

tamaño $m$ obtenida del dominio $D$, de tamaño $M$. Así, la distribución exacta de $\boldsymbol{s}_m$ (con $m$ variable) condicionada a una muestra aleatoria simple seleccionada $\boldsymbol{s}_n$ es:

$$p(\boldsymbol{s}_m|\boldsymbol{s}_n) = p(m|\boldsymbol{s}_n)p(\boldsymbol{s}_m|\boldsymbol{s}_n, m).$$

**Ejercicio 2.7.** Obtener el estimador insesgado de la media poblacional $\bar{y}$ de mínima varianza entre los estimadores de la clase

$$t = \sum_{i=1}^{n} t_i y_{k_i},$$

donde $k_i$ es la $i$-ésima unidad seleccionada en la secuencia en la muestra aleatoria simple de tamaño $n$ de una población finita.

**Solución.** La condición que el estimador $t$ debe cumplir para que sea insesgado es que

$$E(t) = \bar{y},$$

es decir que

$$\sum_{i=1}^{n} t_i = 1.$$

Como la varianza de $t$ es

$$V(t) = \sigma^2 \sum_{i=1}^{n} t_i^2,$$

siendo $\sigma^2$ la varianza de la población finita, el problema se reduce a minimizar la función $V(t)$ bajo la restricción de que $E(t) = \bar{y}$. O equivalentemente, a minimizar

$$\sum_{i=1}^{n} t_i^2$$

sujeto a que

$$\sum_{i=1}^{n} t_i = 1.$$

Usando la técnica de los multiplicadores de Lagrange, tenemos el lagrangiano

$$\Lambda = \sum_{i=1}^{n} t_i^2 + \lambda\left(\sum_{i=1}^{n} t_i - 1\right),$$

donde $\lambda$ es el multiplicador de Lagrange. Resolviendo,

$$\frac{\partial \Lambda}{\partial t_i} = 2t_i + \lambda = 0.$$

Por lo que $t_i = c$ (constante). Como la restricción es que $nc = 1$, deducimos que $c = 1/n$. Luego el estimador insesgado de mínima varianza de la media poblacional es la media muestral

$$t = \frac{1}{n}\sum_{i=1}^{n} y_{k_i}.$$

**Ejercicio 2.8.** Proponer un estimador insesgado de la media poblacional de una población finita $U$ de tamaño $N$ que contiene un dominio $D$ de tamaño conocido $M$, en el caso (a) de que la muestra aleatoria simple $\boldsymbol{s}_n$ de tamaño $n \geq 2$ contiene unidades del dominio y fuera del dominio, es decir, contiene $m$ $(1 \leq m \leq n-1)$ unidades de la secuencia de la muestra en $D$; y en el caso (b) de que no sepamos si la muestra aleatoria simple de la población

finita contiene o no unidades del dominio y de fuera del dominio, o bien no sepamos el tamaño del dominio.

**Solución.** En el caso (a), llamando $\boldsymbol{s}_m$ a la submuestra aleatoria simple del dominio cuya media muestral $\bar{y}_{\boldsymbol{s}_m} = \bar{y}_m$ tiene por esperanza matemática $\bar{y}_D = \bar{y}_M$, la media del dominio, y llamando $\boldsymbol{s}_{n-m}$ a la submuestra aleatoria simple fuera del dominio cuya media muestral $\bar{y}_{\boldsymbol{s}_{n-m}} = \bar{y}_{n-m}$ tiene por esperanza matemática $\bar{y}_{U-D} = \bar{y}_{N-M}$, la media fuera del dominio, tenemos que como

$$\bar{y} = \frac{M}{N}\bar{y}_M + \frac{N-M}{N}\bar{y}_{N-M},$$

un estimador insesgado de la media poblacional es directamente

$$\frac{M}{N}\bar{y}_m + \frac{N-M}{N}\bar{y}_{n-m}$$

que recibe el nombre de estimador posestratificado, pues hay dos estratos en los que se divide o clasifica la población finita: dentro del dominio y fuera del dominio.

En el caso (b) la media muestral es un estimador insesgado de la media poblacional en cualquier caso bajo muestreo aleatorio simple de tamaño $n$, que en nuestro caso admite la expresión

$$\frac{m}{n}\bar{y}_m + \frac{n-m}{n}\bar{y}_{n-m}.$$

**Ejercicio 2.9.** Demostrar que con diseño de muestreo aleatorio simple con reemplazamiento de tamaño $n$, la covarianza de dos medias muestrales correspondientes a dos variables definidas sobre la población es igual a la covarianza poblacional de ambas variables dividido por el tamaño muestral $n$.

**Solución.** Sean $y$ y $x$ las variables definidas en las unidades de la población. Denotamos por $\boldsymbol{s}$ a la muestra aleatoria simple de tamaño $n$, y denotamos por $i \in \boldsymbol{s}$ a las selecciones ordenadas por $i = 1, 2, \dots, n$, es decir el orden de aparición en la secuencia muestral. Entonces,

$$Cov(\bar{y}_{\boldsymbol{s}}, \bar{x}_{\boldsymbol{s}}) = E(\bar{y}_{\boldsymbol{s}}\bar{x}_{\boldsymbol{s}}) - \bar{y}\bar{x} =$$

$$E\left[\frac{1}{n^2}\left(\sum_{i \in \boldsymbol{s}} y_{k_i}\right)\left(\sum_{j \in \boldsymbol{s}} x_{k_j}\right)\right] - \bar{y}\bar{x} =$$

$$\frac{1}{n^2}\left[\sum_{i=1}^{n} E\left(y_{k_i}x_{k_i}\right) + \sum_{i=1}^{n}\sum_{j \neq i}^{n} E\left(y_{k_i}x_{k_j}\right)\right] - \bar{y}\bar{x} =$$

$$\frac{1}{n^2}\left[n\alpha_{11} + n(n-1)\alpha_{10}\alpha_{01}\right] - \alpha_{10}\alpha_{01} =$$

$$\frac{1}{n}\left(\alpha_{11} - \alpha_{10}\alpha_{01}\right) = \frac{\mu_{11}}{n},$$

donde hemos denotado los momentos

$$\alpha_{pq} = \frac{1}{N}\sum_{i=1}^{N} y_i^p x_i^q$$

y

$$\mu_{pq} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^p (x_i - \bar{x})^q.$$

**Ejercicio 2.10.** Obtener un estimador insesgado de la covarianza poblacional a partir de una muestra aleatoria simple con

reemplazamiento de tamaño $n$ de dos variables $y$ y $x$ definidas sobre las unidades de la población.

**Solución.** Consideramos la covarianza muestral $m_{11}$ de ambas variables $y$ y $x$. Calculamos su esperanza matemática:

$$E(m_{11}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_{k_i} - \bar{y}_s)(x_{k_i} - \bar{x}_s)\right] =$$

$$\frac{1}{n}E\left(\sum_{i=1}^{n} y_{k_i}x_{k_i} - n\bar{y}_s\bar{x}_s\right) = \frac{1}{n}\left[E\left(\sum_{i=1}^{n} y_{k_i}x_{k_i}\right) - nE(\bar{y}_s\bar{x}_s)\right] =$$

$$\frac{1}{n}\{n\alpha_{11} - n[Cov(\bar{y}_s, \bar{x}_s) + \bar{y}\bar{x}]\} =$$

$$\frac{1}{n}(n\alpha_{11} - \mu_{11} - n\alpha_{10}\alpha_{01}) = \frac{n-1}{n}\mu_{11},$$

puesto que $\mu_{11} = \alpha_{11} - \alpha_{10}\alpha_{01}$. Por tanto, un estimador insesgado de $\mu_{11}$ es:

$$\hat{\mu}_{11} = \frac{n}{n-1}m_{11} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{k_i} - \bar{y}_s)(x_{k_i} - \bar{x}_s).$$

Este estimador recibe también el nombre de cuasicovarianza muestral.

**Ejercicio 2.11.** Proponer un estimador insesgado $\hat{\bar{y}}$ de la media poblacional $\bar{y}$, así como un estimador insesgado $\hat{V}(\hat{\bar{y}})$ de su varianza $V(\hat{\bar{y}})$, cuando se dispone de la media muestral de observaciones *e* con errores de medida (que se aprovechan en el estimador $\hat{\bar{y}}$) obtenida por muestreo aleatorio simple de tamaño $n$, y esta muestra se submuestrea con diseño de muestreo aleatorio

simple con reemplazamiento de tamaño $n'$ en donde se observa el verdadero valor de la variable de interés $y$.

**Solución.** El estimador insesgado $\hat{\bar{y}}$ de la media poblacional $\bar{y}$, que aprovecha la información con errores de medida en una primera muestra aleatoria simple $\boldsymbol{s}$ de tamaño $n$, es:

$$\hat{\bar{y}} = \bar{e}_{\boldsymbol{s}} + \bar{d}_{\boldsymbol{s}'}$$

Donde $\bar{e}_{\boldsymbol{s}}$ es la media muestral de la muestra de las observaciones con errores de medida $\left(e_{k_1}, e_{k_2}, \dots, e_{k_n}\right)$, y $\bar{d}_{\boldsymbol{s}'}$ es la media muestral de la submuestra aleatoria simple $\boldsymbol{s}'$ de tamaño $n'$ con la variable desviación $d_{k_i} = y_{k_i} - e_{k_i}$ con $i = 1, 2, \dots, n'$. Obviamente el estimador es insesgado, pues

$$E(\hat{\bar{y}}) = E(\bar{e}_{\boldsymbol{s}}) + E\left(\bar{d}_{\boldsymbol{s}'}\right) = \bar{e} + \bar{y} - \bar{e} = \bar{y}.$$

La varianza de $\hat{\bar{y}}$, $V(\hat{\bar{y}})$, se obtiene así:

$$V(\hat{\bar{y}}) = V(\bar{e}_{\boldsymbol{s}}) + V\left(\bar{d}_{\boldsymbol{s}'}\right) + Cov\left(\bar{e}_{\boldsymbol{s}}, \bar{d}_{\boldsymbol{s}'}\right).$$

Ahora bien,

$$V(\bar{e}_{\boldsymbol{s}}) = \sigma_e^2/n,$$

y un estimador insesgado de esta varianza es

$$\hat{V}(\bar{e}_{\boldsymbol{s}}) = s_e^2/n.$$

Aplicando el teorema de Madow,

$$V\left(\bar{d}_{\boldsymbol{s}'}\right) = E_1 V_2\left(\bar{d}_{\boldsymbol{s}'}\right) + V_1 E_2\left(\bar{d}_{\boldsymbol{s}'}\right),$$

donde

$$V_2\left(\bar{d}_{\boldsymbol{s}'}\right) = \frac{\sigma_{d(\boldsymbol{s})}^2}{n'} = \frac{(n-1)s_{d(\boldsymbol{s})}^2}{nn'}$$

y

$$E_1V_2(\bar{d}_{s\prime}) = \frac{(n-1)\sigma_d^2}{nn'},$$

que puede ser estimable insesgadamente por

$$\widehat{E_1V_2}(\bar{d}_{s\prime}) = \frac{s_{d(s\prime)}^2}{n'}.$$

También

$$E_2(\bar{d}_{s\prime}) = \bar{d}_s$$

de donde

$$V_1E_2(\bar{d}_{s\prime}) = \frac{\sigma_d^2}{n},$$

por lo que un estimador insesgado de esta última expresión es

$$\widehat{V_1E_2}(\bar{d}_{s\prime}) = \frac{\widehat{\sigma_d^2}}{n} = \frac{\widehat{s_{d(s)}^2}}{n} = \frac{\widehat{\sigma_{d(s)}^2}}{n-1} = \frac{s_{d(s\prime)}^2}{n-1}.$$

Finalmente, como

$$\bar{e}_s = \frac{1}{n}[(n-n')\bar{e}_{s-s\prime} + n'\bar{e}_{s\prime}],$$

$$Cov(\bar{e}_s, \bar{d}_{s\prime}) = Cov\left(\frac{n'}{n}\bar{e}_{s\prime}, \bar{d}_{s\prime}\right) = \frac{n'}{n}\frac{1}{n'}E[\mu_{11(s)}] = \frac{\mu_{11}}{n-1}.$$

Un estimador insesgado de esta covarianza es:

$$\widehat{Cov}(\bar{e}_s, \bar{d}_{s\prime}) = \frac{\hat{\mu}_{11}}{n-1} = \frac{n\hat{\mu}_{11(s)}}{(n-1)^2} = \frac{nn'm_{11(s\prime)}}{(n-1)^2(n'-1)}.$$

De todo lo cual concluimos que el estimador insesgado de la varianza de $\hat{\bar{y}}$ es:

$$\hat{V}(\hat{\bar{y}}) = \frac{s_e^2}{n} + \frac{(n-1+n')s_{d(\boldsymbol{s}')}^2}{(n-1)n'} + \frac{nn'm_{11(\boldsymbol{s}')}}{(n-1)^2(n'-1)},$$

donde

$$s_e^2 = \frac{1}{n-1}\sum_{i\in\boldsymbol{s}}(e_i - \bar{e}_{\boldsymbol{s}})^2,$$

$$s_{d(\boldsymbol{s}')}^2 = \frac{1}{n'-1}\sum_{i\in\boldsymbol{s}'}\left(d_i - \bar{d}_{\boldsymbol{s}'}\right)^2,$$

y

$$m_{11(\boldsymbol{s}')} = \frac{1}{n'}\sum_{i\in\boldsymbol{s}'}(e_i - \bar{e}_{\boldsymbol{s}'})\left(d_i - \bar{d}_{\boldsymbol{s}'}\right),$$

donde bajo los sumatorios la expresión $i \in \boldsymbol{s}$, o $i \in \boldsymbol{s}'$, indican que el índice $i$ recorre la secuencia completa de la muestra ordenada respectiva.

**Ejercicio 2.12.** Tenemos dos dominios $D_1$ y $D_2$ de una misma población finita de $N$ unidades, de tamaños $NP_1$ y $NP_2$ respectivamente. El dominio $D = D_1 \cap D_2$ tiene un tamaño $NP$. Obtenemos una muestra aleatoria simple con reemplazamiento de tamaño $n$, y observamos las proporciones muestrales $p_1$, $p_2$ y $p$ de los dominios $D_1$ , $D_2$ y $D$ . Obtener la covarianza de las proporciones muestrales $p_1$ y $p_2$, y estimarla sin sesgo a partir de los datos que disponemos. Obtener un estimador de la covarianza de los indicadores de ambos dominios basado en los datos recogidos en la muestra aleatoria simple.

**Solución.** Definimos el indicador del dominio $D$ a la aplicación $I: U \to \{0,1\}$ que a cada unidad $k \in U$ le asigna el valor $I(k) = 1$ si $k \in D$, o bien $I(k) = 0$ si $k \notin D$. Denotando por $I_1$ e $I_2$ a los

indicadores de los dominios $D_1$ y $D_2$ respectivamente, tenemos que si la muestra aleatoria simple es la secuencia $\boldsymbol{s} = (k_1, k_2, \dots, k_n)$, la covarianza pedida es:

$$Cov(p_1, p_2) = E[(p_1 - P_1)p_2] = E(p_1 p_2) - P_1 P_2.$$

Para esto es suficiente ver que:

$$E(p_1 p_2) = \frac{1}{n^2} E\left\{\left[\sum_{i=1}^{n} I_1(k_i)\right]\left[\sum_{j=1}^{n} I_2(k_j)\right]\right\} =$$

$$\frac{1}{n^2}\left\{\sum_{i=1}^{n} E[I(k_i)] + \sum_{i=1}^{n}\sum_{j\neq i}^{n} E\left[I_1(k_i) I_2(k_j)\right]\right\} =$$

$$\frac{1}{n^2}\left\{nP + \sum_{i=1}^{n}\sum_{j\neq i}^{n} E[I_1(k_i)] E\left[I_2(k_j)\right]\right\} =$$

$$\frac{1}{n^2}[nP + n(n-1)P_1 P_2] = \frac{P + (n-1)P_1 P_2}{n}.$$

Luego, de las últimas dos cadenas de igualdades, tenemos que:

$$Cov(p_1, p_2) = \frac{P - P_1 P_2}{n} = \frac{Cov[I_1(k), I_2(k)]}{n} = \frac{\mu_{11}}{n}.$$

Un estimador insesgado de esta covarianza es:

$$\widehat{Cov}(p_1, p_2) = \frac{p - \widehat{P_1 P_2}}{n} = \frac{p - p_1 p_2}{n - 1},$$

donde

$$\widehat{P_1 P_2} = p_1 p_2 - \widehat{Cov}(p_1, p_2) = p_1 p_2 - \frac{p}{n} + \frac{1}{n}\widehat{P_1 P_2},$$

por lo que despejando,

$$\widehat{P_1P_2} = \frac{np_1p_2 - p}{n-1}.$$

De todo ello, tenemos el estimador insesgado de la covarianza poblacional:

$$\hat{\mu}_{11} = \frac{n}{n-1}(p - p_1p_2)$$

que nos indicará aproximadamente la correlación positiva, nula o negativa de los indicadores $I_1$ e $I_2$ de los dominios $D_1$ y $D_2$ en la población finita completa.

**Ejercicio 2.13.** Queremos estimar insesgadamente y con un error absoluto menor que $e$ la media poblacional $\bar{y}$ mediante la media muestral obtenida con diseño de muestreo aleatorio simple con reemplazamiento. Observamos que la cuasivarianza muestral está acotada superiormente por la constante $K$ para valores de $n$ moderados y grandes. ¿Qué tamaño muestral $n$ necesitamos para garantizar un nivel de confianza mayor o igual al $95\%$? Determinar la región de aceptación de un contraste de la hipótesis $H$: $\bar{y} = 7$, al nivel de confianza mayor o igual al $95\%$.

**Solución.** Aplicando la desigualdad de Chebychev, tenemos:

$$p\{|\bar{y}_s - \bar{y}| < e\} \geq 1 - \frac{V(\bar{y}_s)}{e^2} \cong 1 - \frac{s^2}{ne^2} \geq 1 - \frac{K}{ne^2} \geq 0{,}95$$

por lo que el error absoluto máximo $e$ se alcanza con un nivel de confianza mayor o igual a $0{,}95$ cuando, de la última desigualdad de la cadena de ellas,

$$n \geq \frac{K}{0{,}05 \cdot e^2}$$

para que se den las condiciones pedidas; concretamente el valor natural de $n$ inmediatamente superior a la constante $K/(0{,}05 \cdot e^2)$

garantiza un error absoluto en la estimación de la media poblacional con un nivel de confianza superior o igual aproximadamente al 95%.

La región de aceptación $R$ pedida es entonces en la que se acepta la hipótesis $H$ si y solo si

$$7 \in (\bar{y}_s - e, \bar{y}_s + e) = R.$$

**Ejercicio 2.14.** Desarrollar una función derivable dos veces en un entorno de un punto, y aplicarlo a la variable media muestral en el entorno de la media poblacional. Aproximar entonces la esperanza matemática de la función de la media muestral por los dos primeros sumandos no nulos. Aproximar la varianza de la función de la media muestral por el primer término de su expresión. Aplicar esta relación aproximada para evidenciar que la desviación cuadrática de la media muestral a la media poblacional dista de la varianza de la media muestral menos que una cantidad positiva, con una probabilidad aproximadamente 1 cuando el tamaño muestral tiende a infinito en el muestreo aleatorio simple.

**Solución.** Sea la función $f(x)$ derivable dos veces en un entorno de la media poblacional $\bar{y} = E(\bar{y}_s)$. El desarrollo en serie de Taylor de la función en la variable $\bar{y}_s$ en un entorno de $\bar{y}$ es:

$$f(\bar{y}_s) = f(\bar{y}) + (\bar{y}_s - \bar{y}) f'(\bar{y}) + \frac{(\bar{y}_s - \bar{y})^2}{2} f''(\bar{y}) + o[(\bar{y}_s - \bar{y})^2]$$

donde $o(x)$ es un infinitésimo de $x$, es decir, una función de $x$ tal que:

$$\frac{o(x)}{x} \to 0$$

cuando $x \to 0$.

En el desarrollo de Taylor, despreciamos el último sumando infinitesimal, y tomando esperanzas matemáticas en ambos miembros tenemos:

$$E[f(\bar{y}_s)] \cong f(\bar{y}) + \frac{V(\bar{y}_s)}{2} f''(\bar{y}).$$

De ambas aproximaciones, tenemos que:

$$f(\bar{y}_s) - E[f(\bar{y}_s)] \cong (\bar{y}_s - \bar{y}) f'(\bar{y}) +$$

$$\frac{1}{2}[(\bar{y}_s - \bar{y})^2 + V(\bar{y}_s)] f''(\bar{y}),$$

de donde despreciando los términos siguientes del desarrollo aproximado, tenemos

$$\{f(\bar{y}_s) - E[f(\bar{y}_s)]\}^2 \cong (\bar{y}_s - \bar{y})^2 [f'(\bar{y})]^2.$$

Por tanto, tomando esperanzas matemáticas en ambos miembros, tenemos la aproximación

$$V[f(\bar{y}_s)] \cong V(\bar{y}_s)[f'(\bar{y})]^2.$$

En concreto, para la función $f(x) = (x - \bar{y})^2$ tenemos

$$V[(\bar{y}_s - \bar{y})^2] \cong V(\bar{y}_s) \cdot 0^2 = 0.$$

De la desigualdad de Chebychev, tenemos entonces que

$$p\{|(\bar{y}_s - \bar{y})^2 - V(\bar{y}_s)| < e\} \geq 1 - \frac{V[(\bar{y}_s - \bar{y})^2]}{e^2} \cong 1$$

cuando $\bar{y}_s \to \bar{y}$, que es cierto cuando $n \to \infty$.

Este resultado es teórico porque no dispondremos en la práctica de la media poblacional $\bar{y}$ para estimar con los valores $(\bar{y}_s - \bar{y})^2$ a la varianza $V(\bar{y}_s)$. Sin embargo, un estimador insesgado de esta varianza es la cuasivarianza muestral dividida por $n$ como se demuestra en este libro. En Ruiz Espejo *et al.* (2013) se proporciona además un estimador insesgado de la varianza del estimador insesgado de la varianza de la media muestral en muestreo aleatorio simple, como vimos en la sección 2.3 y completamos en ejercicios posteriores.

**Ejercicio 2.15.** Estimar sin sesgo la varianza de una población finita a partir de la cuasivarianza muestral de una submuestra aleatoria simple con reemplazamiento de tamaño fijo $n$, de la muestra ordenada de tamaño fijo $m$ obtenida por muestreo aleatorio simple con reemplazamiento de una población finita de tamaño $N$.

**Solución.** Llamamos $s_1^2$ y $s_2^2$ a las cuasivarianzas muestrales en la primera fase y en la segunda fase respectivamente, es decir el subíndice indica la fase de muestreo a la que se refiere la cuasivarianza muestral. De la primera fase tenemos que

$$E_1(s_1^2) = \sigma^2.$$

De la segunda fase tenemos que

$$E_2\left(\frac{m}{m-1}s_2^2\right) = s_1^2.$$

Por lo tanto,

$$E_1\left[E_2\left(\frac{m}{m-1}s_2^2\right)\right] = \sigma^2.$$

Es decir, el estimador insesgado de la varianza poblacional en muestreo doble o en dos fases de muestreo aleatorio simple con reemplazamiento, es

$$\frac{m}{m-1}s_2^2,$$

donde $m$ es el tamaño fijo de la muestra en la primera fase, y $s_2^2$ es la cuasivarianza muestral en la segunda fase de aleatorización.

**Ejercicio 2.16.** Estimar sin sesgo el momento central poblacional de orden dos con una muestra aleatoria simple con reemplazamiento de tamaño $n$.

**Solución.** Como

$$\sigma^2 = \alpha_2 - \alpha_1^2,$$

tenemos que

$$\hat{\mu}_2 = \widehat{\sigma^2} = \hat{\alpha}_2 - \left[\hat{\alpha}_1^2 - \hat{V}(\hat{\alpha}_1)\right] =$$

$$\hat{\alpha}_2 - \hat{\alpha}_1^2 + \frac{\widehat{\sigma^2}}{n}.$$

Despejando $\widehat{\sigma^2}$ tenemos que

$$\hat{\mu}_2 = \widehat{\sigma^2} = \frac{n}{n-1}(\hat{\alpha}_2 - \hat{\alpha}_1^2) = s^2.$$

Que es la cuasivarianza muestral. Al ser invariante para permutaciones en la muestra ordenada, es también estimador insesgado óptimo para distribución libre (Zacks, 1971, p. 150).

**Ejercicio 2.17.** Estimar sin sesgo el momento central poblacional de orden tres con una muestra aleatoria simple con reemplazamiento de tamaño $n$.

**Solución.** El estimador insesgado óptimo del momento central poblacional de orden tres, para distribución libre, es

$$\hat{\mu}_3 = \frac{n^2}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3) = \frac{n^2 m_3}{n^2 - 3n + 2},$$

donde $\hat{\alpha}_j = (1/n)\sum_{i=1}^{n} y_i^j$ es el momento muestral no central de orden $j$, siendo $y_i^j$ la potencia $j$-ésima del "$i$-ésimo valor observado en la muestra" que hemos denotado $y_i$. En concreto, $\hat{\alpha}_1 = \bar{y}_s$ es la media muestral. También $m_3$ es el momento central muestral de orden tres,

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_s)^3.$$

Básicamente el resultado se basa en que

$$\hat{\mu}_3 = \hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 3\widehat{Cov}(\hat{\alpha}_2, \hat{\alpha}_1) + 2\hat{\alpha}_1^3$$

$$-2\hat{\alpha}_1\hat{V}(\hat{\alpha}_1) - 2\widehat{Cov}(\hat{\alpha}_1^2, \hat{\alpha}_1) + 2\widehat{Cov}\left[\hat{\alpha}_1, \hat{V}(\hat{\alpha}_1)\right].$$

Veamos la prueba de este modo,

$$E(\hat{\alpha}_3) = \alpha_3$$

$$E(\hat{\alpha}_2\hat{\alpha}_1) = \alpha_2\alpha_1 + Cov(\hat{\alpha}_2, \hat{\alpha}_1)$$

$$E(\hat{\alpha}_1^3) = E(\hat{\alpha}_1^2\hat{\alpha}_1) = E(\hat{\alpha}_1^2)E(\hat{\alpha}_1) + Cov(\hat{\alpha}_1^2, \hat{\alpha}_1)$$

$$= [\alpha_1^2 + V(\hat{\alpha}_1)]\alpha_1 + Cov(\hat{\alpha}_1^2, \hat{\alpha}_1)$$

$$= \alpha_1^3 + \alpha_1 V(\hat{\alpha}_1) + Cov(\hat{\alpha}_1^2, \hat{\alpha}_1)$$

Y

$$E\left[\hat{\alpha}_1\hat{V}(\hat{\alpha}_1)\right] = \alpha_1 V(\hat{\alpha}_1) + Cov\left[\hat{\alpha}_1, \hat{V}(\hat{\alpha}_1)\right].$$

Por todo ello, concluimos que

$$E(\hat{\mu}_3) = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3 = \mu_3.$$

Ahora, basándonos en el resultado anterior, si $-\infty < \mu_3 < \infty$, y $\hat{\alpha}_j$ es el momento no central de orden $j$ en la muestra, entonces el estimador

$$\hat{\mu}_3 = \frac{n^2}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3)$$

es insesgado y de mínima varianza para distribución libre (en el sentido explicado por Zacks, 1971, p. 150). La demostración es la siguiente.

$$\widehat{Cov}(\hat{\alpha}_2, \hat{\alpha}_1) = \frac{(\alpha_3 \widehat{-\alpha_2}\alpha_1)}{n}$$

$$= \frac{\hat{\alpha}_3}{n} - \frac{\hat{\alpha}_2\hat{\alpha}_1 - \widehat{Cov}(\hat{\alpha}_2, \hat{\alpha}_1)}{n}.$$

De donde

$$\widehat{Cov}(\hat{\alpha}_2, \hat{\alpha}_1) = \frac{\hat{\alpha}_3 - \hat{\alpha}_2\hat{\alpha}_1}{n-1}.$$

Además,

$$\widehat{Cov}(\hat{\alpha}_1^2, \hat{\alpha}_1) = \widehat{Cov}\left(\hat{\alpha}_2 - \frac{n-1}{n}s^2, \hat{\alpha}_1\right)$$

$$= \widehat{Cov}(\hat{\alpha}_2, \hat{\alpha}_1) - \frac{n-1}{n}\widehat{Cov}(s^2, \hat{\alpha}_1)$$

$$= \frac{\hat{\alpha}_3 - \hat{\alpha}_2\hat{\alpha}_1}{n-1} - \frac{(n-1)\hat{\mu}_3}{n^2}.$$

And

$$\widehat{Cov}\left[\hat{\alpha}_1, \hat{V}(\hat{\alpha}_1)\right] = \widehat{Cov}\left(\hat{\alpha}_1, \frac{s^2}{n}\right) = \frac{\hat{\mu}_3}{n^2}.$$

Sustituyendo estos resultados en el resultado básico inicial tenemos que

$$\hat{\mu}_3 = \hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 3\frac{\hat{\alpha}_3 - \hat{\alpha}_2\hat{\alpha}_1}{n-1} + 2\hat{\alpha}_1^3$$

$$-2\hat{\alpha}_1\frac{\hat{\alpha}_2 - \hat{\alpha}_1^2}{n-1} - 2\left[\frac{\hat{\alpha}_3 - \hat{\alpha}_2\hat{\alpha}_1}{n-1} - \frac{(n-1)\hat{\mu}_3}{n^2}\right] + 2\frac{\hat{\mu}_3}{n^2}$$

$$= \frac{n^2}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3).$$

**Ejercicio 2.18.** Estimar sin sesgo el momento central poblacional de orden cuatro con una muestra aleatoria simple con reemplazamiento de tamaño $n$.

**Solución.** El estimador insesgado óptimo del momento central poblacional de orden cuatro, para distribución libre, es

$$\hat{\mu}_4 = \left(1 - \frac{3}{cn}\right)^{-1}\left[\frac{nm_4}{n-1} + 3\left(\frac{n-1}{n} - \frac{1}{c}\right)s^4\right],$$

donde $m_4$ es el momento central muestral de orden cuatro

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_s)^4,$$

$s^4$ es el cuadrado de la cuasivarianza muestral, y $c$ es la constante

$$c = \frac{n^2 - 2n + 3}{n(n-1)}.$$

Para demostrarlo, tenemos que si $i \neq j$

$$(y_i - y_j)^4 = y_i^4 - 4y_i^3 y_j + 6y_i^2 y_j^2 - 4y_i y_j^3 + y_j^4.$$

Tomando esperanzas en ambos miembros de esta igualdad resulta

$$E\left[(y_i - y_j)^4\right] = 2\alpha_4 - 8\alpha_3\alpha_1 + 6\alpha_2^2.$$

Por otro lado sabemos que

$$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4$$

y que

$$\sigma^4 = \alpha_2^2 - 2\alpha_2\alpha_1^2 + \alpha_1^4.$$

De estas dos últimas fórmulas, tenemos que si $i \neq j$

$$E\left[\frac{1}{2}(y_i - y_j)^4\right] = \mu_4 + 3\sigma^4.$$

Definimos ahora el estadístico

$$t = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^4$$

$$= \frac{n}{n-1}\left[m_4 + 3\left(\frac{n-1}{n}s^2\right)^2\right]$$

$$= \frac{n}{n-1}(\hat{\alpha}_4 - 4\hat{\alpha}_3\hat{\alpha}_1 + 3\hat{\alpha}_2^2).$$

Como además,

$$E(s^4) = V(s^2) + [E(s^2)]^2 =$$

$$\frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\sigma^4 + \sigma^4 = \frac{\mu_4}{n} + c\sigma^4.$$

Implica que

$$\sigma^4 = \frac{1}{c}\left[E(s^4) - \frac{\mu_4}{n}\right].$$

Donde

$$c = \frac{n^2 - 2n + 3}{n(n-1)}.$$

De los anteriores resultados, tenemos que

$$E\left(\frac{n}{n-1}m_4 + 3\frac{n-1}{n}s^4\right) = E(t) = \mu_4 + 3\sigma^4$$

$$= \mu_4 + \frac{3}{c}\left[E(s^4) - \frac{\mu_4}{n}\right] = \mu_4\left(1 - \frac{3}{cn}\right) + \frac{3}{c}E(s^4).$$

O bien, el estimador insesgado de mínima varianza para distribución libre de $\mu_4$ resulta ser

$$\hat{\mu}_4 = \left(1 - \frac{3}{cn}\right)^{-1}\left[\frac{n}{n-1}m_4 + 3\left(\frac{n-1}{n} - \frac{1}{c}\right)s^4\right].$$

**Ejercicio 2.19.** Proponer el valor exacto de la varianza de la cuasivarianza muestral, en muestreo aleatorio simple con reemplazamiento, y un estimador insesgado de la varianza propuesta.

**Solución.** El valor de la varianza pedida es

$$V(s^2) = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}.$$

Demostrar esta fórmula teniendo en cuenta que sabemos que es cierta esta otra

$$V(s^2) = E(s^4) - \sigma^4,$$

pues si $X$ es una variable aleatoria entonces $V(X) = E(X^2) - [E(X)]^2$ y como caso particular si $X = s^2$ es lo que hemos escrito antes, equivale a demostrar que

$$E(s^4) = \frac{\mu_4}{n} + \frac{n^2 - 2n + 3}{n(n-1)}\sigma^4.$$

Veamos pues esta fórmula. Del Ejercicio 1.7 aplicado a una muestra de tamaño fijo $n$, tenemos que la varianza muestral es

$$\frac{n-1}{n}s^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(y_i - y_j)^2.$$

Despejando la cuasivarianza muestral tenemos que

$$s^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(y_i - y_j)^2.$$

Luego,

$$s^4 = \frac{1}{4n^2(n-1)^2}\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n}(y_i - y_j)^2\right]^2 =$$

$$\frac{1}{4n^2(n-1)^2}\left\{\sum_{i=1}^{n}\sum_{j\neq i}^{n}(y_i - y_j)^4 + \right.$$

$$\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{m\neq i,j}^{n}(y_i - y_j)^2\left[(y_i - y_m)^2 + (y_j - y_m)^2\right] +$$

$$\left.\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,j}^{n}\sum_{m\neq i,j,k}^{n}(y_i - y_j)^2(y_k - y_m)^2\right\}.$$

Luego la esperanza matemática de $s^4$ será

$$E(s^4) = \frac{1}{4n^2(n-1)^2}\left\{\sum_{i=1}^{n}\sum_{j\neq i}^{n} E\left[(y_i - y_j)^4\right] + \right.$$

$$\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{m\neq i,j}^{n} 2E\left[(y_i - y_j)^2(y_i - y_m)^2\right] +$$

$$\left.\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k\neq i,j}^{n}\sum_{m\neq i,j,k}^{n} E\left[(y_i - y_j)^2(y_k - y_m)^2\right]\right\}.$$

Ahora, como el número de sumandos del último doble sumatorio es $n(n-1)$ y pueden aparecer los cuadrados en el orden $(i,j)$, $(i,j)$ o bien en el orden $(i,j)$, $(j,i)$, se duplica el número de esperanzas matemáticas

$$E\left[(y_i - y_j)^4\right] = E(y_i^4 - 4y_i^3 y_j + 6y_i^2 y_j^2 - 4y_i y_j^3 + y_j^4)$$

$$= 2\alpha_4 - 8\alpha_3\alpha_1 + 6\alpha_2^2 = 2(\mu_4 + 3\sigma^4),$$

pues $i \neq j$.

En el sumatorio triple, razonando de modo similar tenemos

$$E\left[(y_i - y_j)^2(y_i - y_m)^2\right] = \mu_4 + 3\sigma^4,$$

y el número de sumandos en total es $n(n-1)(n-2)$ pues hemos supuesto que los índices verifican $j \neq i = k \neq m \neq j$. Además debemos multiplicar el número de sumandos por 2, uno para que $k = i$ ó $j$ y $m \neq i,j,k$, y otro para que $m = i$ ó $j$ y $k \neq i,j,m$.

El sumatorio cuádruple tiene $n(n-1)(n-2)(n-3)$ sumandos y cada uno de ellos tiene por esperanza matemática

$$E\left[(y_i - y_j)^2(y_k - y_m)^2\right] = \left\{E\left[(y_i - y_j)^2\right]\right\}^2$$

$$= (2\sigma^2)^2 = 4\sigma^4,$$

donde la primera igualdad se debe a que los factores del primer miembro dentro de la esperanza matemática son independientes al ser los cuatro índices distintos dos a dos.

Por lo que sustituyendo en la esperanza de $s^4$ queda

$$E(s^4) = \frac{1}{4n^2(n-1)^2}[2n(n-1)2(\mu_4 + 3\sigma^4) +$$

$$4n(n-1)(n-2)(\mu_4 + 3\sigma^4) + n(n-1)(n-2)(n-3)4\sigma^4]$$

$$= \frac{\mu_4}{n} + \frac{n^2 - 2n + 3}{n(n-1)}\sigma^4.$$

Que es lo que queríamos demostrar para concluir el resultado de la varianza de la cuasivarianza muestral en muestreo aleatorio simple con reemplazamiento. Lo visto hasta aquí ha sido usado en el Ejercicio 2.17.

Un estimador insesgado de esta varianza $V(s^2)$ es

$$\hat{V}(s^2) = \frac{n-1}{n^2 - 2n + 3}\hat{\mu}_4 - \frac{n-3}{n^2 - 2n + 3}s^4.$$

El estimador insesgado de $\mu_4$, que aparece en el primer sumando del segundo término, ha sido obtenido previamente en el Ejercicio 2.17. Para demostrar el resultado expuesto de $\hat{V}(s^2)$, veamos que

$$\hat{V}(s^2) = \left[\frac{\mu_4}{n} - \frac{\widehat{(n-3)\sigma^4}}{n(n-1)}\right] = \frac{\hat{\mu}_4}{n} - \frac{n-3}{n(n-1)}\widehat{\sigma^4}$$

$$= \frac{\hat{\mu}_4}{n} - \frac{n-3}{n(n-1)}\left[s^4 - \hat{V}(s^2)\right].$$

De donde despejando $\hat{V}(s^2)$ tenemos finalmente el resultado avanzado previamente.

Un ejemplo de aplicación de los resultados anteriores es el cálculo de la varianza del estadístico

$$\bar{y} + ks^2,$$

siendo $\bar{y}$ la media muestral, $k$ una constante real, y $s^2$ la cuasivarianza muestral en el muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$. Dicho estadístico $\bar{y} + ks^2$ es un estimador insesgado de la función paramétrica

$$\alpha_1 + k\sigma^2,$$

siendo $\alpha_1$ la media poblacional, y $\sigma^2$ la varianza poblacional.

De los resultados anteriores tenemos que la varianza del estadístico propuesto es

$$V(\bar{y} + ks^2) =$$

$$V(\bar{y}) + k^2 V(s^2) + 2kCov(\bar{y}, s^2) =$$

$$\frac{\sigma^2}{n} + k^2 \left[\frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}\right] + 2k\frac{\mu_3}{n}.$$

En el último sumando, en el que sustituye la covarianza, hemos usado de un resultado demostrado anteriormente. El valor obtenido de la varianza del estadístico prueba que el estadístico, por ser insesgado, converge en probabilidad a la función paramétrica $\alpha_1 + k\sigma^2$, ya que la varianza obtenida es un infinitésimo de orden $n^{-1}$ y haciendo uso de la desigualdad de Chebychev. Además dicho estadístico es insesgado y óptimo (uniformemente de mínima varianza) para estimar dicha función paramétrica en el modelo de distribución poblacional libre, ya que el estadístico es invariante ante permutaciones en el orden de las observaciones muestrales (Zacks, 1971, p. 150). Además la varianza del estadístico es estimable insesgadamente por el estimador "suma de los

estimadores insesgados óptimos de cada uno de los tres sumandos para distribución poblacional libre" y, como consecuencia, éste es además estimador óptimo o uniformemente de mínima varianza de la función paramétrica $V(\bar{y} + ks^2)$ para distribución poblacional libre.

**Ejercicio 2.20.** Comprobar si se puede seleccionar una muestra aleatoria simple con reemplazamiento de tamaño fijo $n$ de una población finita de tamaño $34.629$ con un generador de números aleatorios independientes, con distribución uniforme en el conjunto $\{0, 1, 2, \dots, 9\}$ de números naturales entre $0$ y $9$. La selección se hace tomando grupos de cinco dígitos sucesivos e identificando las unidades poblacionales de $1$ a $34.629$. Si el primer grupo de cinco dígitos seleccionado está entre los números $00.001$ y $34.629$, se selecciona como primera unidad de la muestra aquella cuyo identificador sea la del indicador de ese grupo; si el grupo seleccionado no estuviese entre tales números, se procede a una nueva selección de cinco dígitos aleatorios sucesivos, y así se repite el proceso hasta seleccionar la primera unidad poblacional de la muestra. Sucesivamente se obtendrían las siguientes unidades de la muestra repitiendo el proceso hasta seleccionar la segunda, tercera y hasta la $n$-ésima.

**Solución.** Es sencillo comprobarlo y ciertamente sí, se puede seleccionar así. En realidad se trata de seleccionar dígitos naturales o del conjunto $\mathbb{N} = \{1, 2, 3, \dots\}$, condicionados a que éstos tienen que estar comprendidos entre 1 y 34.629. La probabilidad de seleccionar un dígito así en cada selección independiente será

$$p\{k|\{1, 2, \dots, 34.629\}\} = \begin{cases} 1/34.629 & \text{si} \quad k \in \mathbb{N} \cap [1, 34.629] \\ 0 & \text{si} \quad k \notin \mathbb{N} \cap [1, 34.629] \end{cases}$$

Por lo que es un procedimiento de selección de probabilidades iguales en el conjunto de números naturales $\{1, 2, \dots, 34.629\}$ y cada selección es independiente de las anteriores. Luego es un diseño de muestreo aleatorio simple con reemplazamiento a partir de una población finita de tamaño $N = 34.629$.

**Ejercicio 2.21.** Obtener la covarianza de los estadísticos cuasivarianza muestral y media muestral en el muestreo aleatorio simple con reemplazamiento de tamaño fijo $n \geq 2$. Y obtener un estimador insesgado de esta covarianza.

**Solución.** Partimos de la propiedad de la varianza de una variable estadística aplicada a una muestra de tamaño fijo $n \geq 2$, y tenemos que la cuasivarianza muestral es

$$s^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (y_i - y_j)^2.$$

Y la media muestral es

$$\bar{y}_s = \frac{1}{n} \sum_{k=1}^{n} y_k.$$

Donde hemos representado por $y_i$ al $i$-ésimo valor que toma la variable de interés $y$ en la muestra aleatoria simple con reemplazamiento de tamaño $n$. Luego,

$$Cov(s^2, \bar{y}_s) =$$

$$\frac{1}{2n^2(n-1)} Cov\left[ \sum_{i=1}^{n} \sum_{j \neq i}^{n} (y_i^2 + y_j^2 - 2y_i y_j), \sum_{k=1}^{n} y_k \right] =$$

$$\frac{1}{2n^2(n-1)} \times$$

$$\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k=1}^{n}\left[Cov\left(y_i^2, y_k\right) + Cov\left(y_j^2, y_k\right) - 2Cov\left(y_i y_j, y_k\right)\right].$$

Ahora, denotando por $\alpha_m$ al momento no central poblacional de orden $m = 1, 2, 3$,

$$\alpha_m = \frac{1}{N}\sum_{i=1}^{N} y_i^m.$$

Si $i = k$ ó $j = k$,

$$Cov\left(y_i^2, y_k\right) = Cov\left(y_j^2, y_k\right) = \alpha_3 - \alpha_2\alpha_1.$$

Si $i \neq k$ y $j \neq k$,

$$Cov\left(y_i^2, y_k\right) = Cov\left(y_j^2, y_k\right) = 0.$$

Similarmente, si $i \neq j = k$ ó $j \neq i = k$,

$$Cov\left(y_i y_j, y_k\right) = \alpha_2\alpha_1 - \alpha_1^3.$$

Y si $i, j \neq k$,

$$Cov\left(y_i y_j, y_k\right) = 0.$$

Por tanto,

$$Cov(s^2, \bar{y}_s) = \frac{1}{2n^2(n-1)} \times$$

$$[2n(n-1)(\alpha_3 - \alpha_2\alpha_1) - 4n(n-1)(\alpha_2\alpha_1 - \alpha_1^3)] =$$

$$\frac{1}{n}(\alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3) = \frac{\mu_3}{n}.$$

El valor 4 que aparece en la primera igualdad de la anterior serie de igualdades se debe a que se ha de duplicar el número de sumandos porque el subíndice $k$ puede ser igual al subíndice $i$ ó bien al subíndice $j$.

Con ello hemos presentado una demostración muy sencilla, de poco más de una página, a este problema clásico de obtener el valor exacto de la covarianza de la cuasivarianza muestral y de la media muestral, en el muestreo aleatorio simple con reemplazamiento de tamaño fijo $n \geq 2$. Como consecuencia, la cuasivarianza muestral y la media muestral estarán incorrelacionadas cuando el momento central poblacional de orden 3 sea nulo. Otra conclusión es que ambos estadísticos en general no son independientes por la misma razón. Pero ya que esta covarianza es un infinitésimo de orden de $n^{-1}$, se puede concluir que asintóticamente ambos estadísticos estarán aproximadamente incorrelacionados pues su covarianza tiende a 0 cuando $n$ tiende a infinito. El teorema de Fisher garantiza que ambos estadísticos son independientes cuando la población de partida es normal, pero este resultado demostrado en este Ejercicio 2.20 nos asegura que en una población finita esto no es cierto en general, es decir, no es posible afirmar lo que el teorema de Fisher afirma para una población normal cuando la población de partida es finita. Y solo se daría la incorrelación de ambos estadísticos cuando la población finita de partida tuviera un coeficiente de asimetría de Fisher $\mu_3/\sigma^3$ igual a cero, donde $\sigma^2 = \mu_2$ es la varianza poblacional.

De este resultado se puede concluir que esta covarianza obtenida es estimable insesgadamente por

$$\widehat{Cov}(s^2, \bar{y}_s) = \frac{nm_3}{n^2 - 3n + 2} = \frac{n}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3).$$

Siendo

$$\hat{\alpha}_m = \frac{1}{n}\sum_{i=1}^{n} y_i^m,$$

el momento no central muestral de orden $m = 1, 2, 3$; y $m_3$ el momento central muestral de orden 3,

$$m_3 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\alpha}_1)^3.$$

Es una consecuencia casi directa del Ejercicio 2.16.

Un ejemplo de aplicación del resultado expuesto anteriormente es la obtención de un estimador insesgado óptimo para distribución libre del parámetro $\alpha_1\mu_2$. En efecto,

$$\alpha_1\mu_2 = E(\bar{y}_s s^2) - Cov(\bar{y}_s, s^2).$$

Luego un estimador insesgado de $\alpha_1\mu_2$ es el estimador

$$\bar{y}_s s^2 - \widehat{Cov}(\bar{y}_s, s^2) =$$

$$\bar{y}_s s^2 - \frac{n}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3) =$$

$$\hat{\alpha}_1 s^2 - \frac{n}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3) =$$

$$\hat{\alpha}_1\left[\frac{n}{n-1}(\hat{\alpha}_2 - \hat{\alpha}_1^2)\right] - \frac{n}{n^2 - 3n + 2}(\hat{\alpha}_3 - 3\hat{\alpha}_2\hat{\alpha}_1 + 2\hat{\alpha}_1^3) =$$

$$\frac{n}{n^2 - 3n + 2}[-\hat{\alpha}_3 + (n+1)\hat{\alpha}_2\hat{\alpha}_1 - n\hat{\alpha}_1^3].$$

Este estimador es además insesgado y óptimo para distribución libre por ser invariante ante permutaciones de las unidades en la muestra ordenada (Zacks, 1971).

**Ejercicio 2.22.** Proponer un estimador insesgado óptimo del producto de dos medias poblacionales de dos variables de interés.

**Solución.** Básicamente partimos del producto de dos medias muestrales, $\bar{y}\bar{x}$, de la función biparamétrica $\bar{Y}\bar{X}$. Entonces, un estimador insesgado de esta función biparamétrica se obtiene de la relación

$$\bar{Y}\bar{X} = E(\bar{y}\bar{x}) - Cov(\bar{y}, \bar{x}) = E(\bar{y}\bar{x}) - \frac{\sigma_{y,x}}{n}.$$

Aquí $n$ es el tamaño muestral. Un estimador insesgado de la covarianza poblacional $\sigma_{y,x}$ en el muestreo aleatorio simple con reemplazamiento es la cuasicovarianza muestral, por lo que el estimador insesgado de $\bar{Y}\bar{X}$ resulta ser

$$\bar{y}\bar{x} - \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}).$$

Aquí $y_i$ y $x_i$ son los valores de la variables de interés $y$ y $x$ para la $i$-ésima observación de la muestra aleatoria simple en la misma unidad poblacional. Finalmente la optimalidad para distribución libre en las variables $y$ y $x$ se obtiene de modo similar al explicado con anterioridad, ya que el estimador insesgado es invariante ante permutaciones en el orden de la muestra ordenada obtenida por muestreo aleatorio simple con reemplazamiento.

# Capítulo 3

# Muestreo irrestricto aleatorio

En este capítulo estudiamos el diseño *mia*, que es un diseño no ordenado llamado "muestreo irrestricto aleatorio" y también es conocido por "muestreo aleatorio simple sin reemplazamiento" con probabilidades iguales. También se le denota por las siglas *mpi* de "muestreo de probabilidades iguales sin reemplazamiento".

## 3.1 Diseño *mia*

Este diseño muestral no ordenado $TF(n)$ y $TEF(n)$ es la distribución de probabilidad definida sobre las posibles muestras o subconjuntos no vacíos de tamaño $n$, $0 < n \leq N$, de la población finita $U$ de tamaño $N \geq 1$. Denotamos por $s$ a una de estas muestras conjunto de tamaño $n$, y el número posible de muestras distintas para el diseño *mia* es

$$\binom{N}{n} = \frac{N!}{(N-n)!\,n!},$$

que coincide con el número de combinaciones de $N$ elementos tomados de $n$ en $n$.

El diseño *mia* recibe también el nombre de "muestreo aleatorio simple sin reemplazamiento con probabilidades iguales" porque si tuviéramos una urna que contuviera $N$ bolas biunívocamente numeradas de la 1 a la $N$, la selección de una

muestra $s$ entre las $\binom{N}{n}$ posibles muestras de tamaño efectivo fijo $n$ se podrá realizar seleccionando una primera bola de la urna y anotamos su número identificador como componente de la muestra no ordenada o conjunto no vacío $s$ ; seguidamente no reincorporaremos a la urna la bola ya seleccionada, con lo cual la urna preparada para la selección de la segunda bola tendrá $N-1$ bolas desde la 1 a la $N$ excluyendo la bola ya seleccionada en la primera selección que queda fuera de la urna, y por tanto su número identificativo no se podrá seleccionar en adelante. La segunda extracción de la urna selecciona una segunda bola que tampoco se reintegra a la urna y por tanto tampoco se repetirá en las siguientes extracciones de la urna. Así actuaríamos hasta seleccionar $n$ unidades $(0 < n \leq N)$ ordenadamente.

De este modo obtenemos una secuencia $(k_1, k_2, \dots, k_n)$ o vector $\boldsymbol{s}$ $n$-dimensional que tendrá como probabilidad de selección, haciendo uso del Teorema de Producto con sucesos dependientes,

$$p(\boldsymbol{s}) = p(k_1) \cdot p(k_2|k_1) \cdots p(k_n|k_1, k_2, \dots, k_{n-1}) =$$

$$\frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-n+1} = \frac{1}{\frac{N!}{(N-n)!}}.$$

Ahora bien, como las muestras conjunto $s$ o no ordenadas de tamaño $n$ están compuestas por tantas muestras vector $\boldsymbol{s}$ como permutaciones de $n$ elementos o unidades distintas, obtenemos que la probabilidad de seleccionar la muestra no ordenada $s$ es

$$p(s) = n!\, p(\boldsymbol{s}) = n! \frac{1}{\frac{N!}{(N-n)!}} = \frac{1}{\binom{N}{n}}.$$

Es decir, la probabilidad total de obtener una muestra conjunto $s$ de tamaño $n$ con diseño *mia* es exactamente $1/\binom{N}{n}$. Por tanto la suma de estas probabilidades al recorrer todas las posibles muestras no ordenadas $s$ de tamaño $n$ es igual a 1 por tratarse de una distribución de probabilidad. En efecto, si

$$S = \{s: \phi \neq s \subset U, n(s) = n\}$$

es el conjunto de $\binom{N}{n}$ muestras no ordenadas $s$ consideradas de tamaño muestral efectivo $n$ con probabilidad positiva $1/\binom{N}{n}$, entonces

$$\sum_{s \in S} p(s) = \binom{N}{n} \frac{1}{\binom{N}{n}} = 1.$$

La probabilidad de inclusión de la unidad $k$, $1 \leq k \leq N$, en la muestra conjunto $s$ con el diseño *mia* utilizando la regla de Laplace para sucesos equiprobables como es el caso, es el cociente entre casos favorables y casos posibles

$$\pi_k = p(k \in s) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!\,(N-n)!}}{\frac{N!}{n!\,(N-n)!}} = \frac{n}{N}.$$

La probabilidad de inclusión de segundo orden para dos unidades distintas $k$ y $m$ de entre 1 y $N$, haciendo uso de la regla de Laplace, es

$$\pi_{km} = p(k, m \in s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!\,(N-n)!}}{\frac{N!}{n!\,(N-n)!}} =$$

$$\frac{n(n-1)}{N(N-1)}.$$

De modo similar se obtienen las probabilidades de inclusión de órdenes superiores.

## 3.2 Estimación de la media poblacional en *mia*

El estimador insesgado usual de la media poblacional $\bar{y}$ con el diseño *mia*, es la media muestral

$$\bar{y}_s = \frac{1}{n}\sum_{k\in s} y_k,$$

siendo $s$ la muestra no ordenada o muestra conjunto de tamaño $n$ seleccionada, e $y_k$ la variable de interés en la unidad $k \in s$. En efecto, veamos que la media muestral $\bar{y}_s$ con el diseño *mia* es insesgado.

$$E(\bar{y}_s) = \sum_{s\in S} \bar{y}_s p(s) = \sum_{s\in S} \frac{1}{n}\left(y_{k_1} + y_{k_2} + \cdots + y_{k_n}\right)\frac{1}{\binom{N}{n}}$$

donde $s = \{k_1, k_2, \ldots, k_n\}$, por lo que sumando tantas veces $y_k$ como muestras $s$ contengan la unidad $k$, es decir tantas como

$$\text{card}\{s: k \in s\} = \binom{N-1}{n-1},$$

tenemos

$$E(\bar{y}_s) = \sum_{k\in U} \frac{1}{n} y_k \text{card}\{s: k \in s\}\frac{1}{\binom{N}{n}} = \sum_{k\in U} y_k \frac{1}{n}\frac{\binom{N-1}{n-1}}{\binom{N}{n}} =$$

$$\sum_{k\in U} y_k \frac{1}{n}\frac{n}{N} = \frac{1}{N}\sum_{k\in U} y_k = \bar{y}.$$

Otra demostración de la insesgación de la media muestral $\bar{y}_s$ con diseño *mia* para estimar la media poblacional $\bar{y}$ es la siguiente. Llamando $k_i$ a la unidad o su identificador de la población finita $U$ que es seleccionada en la muestra $s$ con el orden $i$-ésimo de su secuencia al ser seleccionada $i = 1, 2, \dots, n$

$$E(\bar{y}_s) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_{k_i}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(y_{k_i}\right) = \bar{y},$$

puesto que

$$E\left(y_{k_i}\right) = \sum_{\boldsymbol{s}\in\boldsymbol{S}} y_{k_i} p(\boldsymbol{s}) =$$

$$\sum_{k\in U} y_k \text{card}\{\boldsymbol{s}: k \text{ es la } i-\text{ésima componente de } \boldsymbol{s}\}\frac{1}{\frac{N!}{(N-n)!}} =$$

$$\sum_{k\in U} y_k \frac{(N-1)!}{(N-n)!}\frac{1}{\frac{N!}{(N-n)!}} = \frac{1}{N}\sum_{k\in U} y_k = \bar{y},$$

ya que si la unidad $k = k_i$ es la $i$-ésima unidad fijada de la muestra ordenada y las restantes unidades distintas no están fijadas,

$$\text{card}\{\boldsymbol{s}: k \text{ es la } i-\text{ésima componente de } \boldsymbol{s}\} =$$

$$\text{card}\{\boldsymbol{s}: \boldsymbol{s} = (k_1, k_2, \dots, k_i = k, \dots, k_n)\} =$$

$$(N-1)(N-2)\cdots(N-i+1)\cdot 1\cdot(N-i)\cdots(N-n+1) =$$

$$\frac{(N-1)!}{(N-n)!}.$$

Pues $(N-1)$ es el número de unidades diferentes de $k=k_i$ que pueden ocupar el primer lugar de la secuencia $\boldsymbol{s}$, $(N-2)$ es el número de unidades diferentes de $k_1$ y de $k_i$ que pueden ocupar el segundo lugar de la secuencia $\boldsymbol{s}$, etc. siendo 1 el factor $i$-ésimo por ser $k=k_i$ la única unidad que puede ocupar el lugar $i$-ésimo de la secuencia $\boldsymbol{s}$.

Hemos demostrado entonces que la media muestral no tiene sesgo para estimar la media poblacional con diseño *mia*.

La varianza de la media muestral con diseño *mia* es

$$V(\bar{y}_s)=\frac{N-n}{N}\frac{S^2}{n},$$

donde denotamos $S^2=N\sigma^2/(N-1)$ y se le denomina "cuasivarianza poblacional".

La demostración es la siguiente:

$$V(\bar{y}_s)=E[(\bar{y}_s-\bar{y})^2]=E\left[\left(\frac{1}{n}\sum_{k\in s}y_k-\bar{y}\right)^2\right]=$$

$$E\left\{\left[\frac{1}{n}\sum_{k\in s}(y_k-\bar{y})\right]^2\right\}=$$

$$\frac{1}{n^2}E\left[\sum_{k\in s}(y_k-\bar{y})^2+\sum_{k\neq m\in s}(y_k-\bar{y})(y_m-\bar{y})\right]=$$

$$\frac{1}{n^2}\left\{E\left[\sum_{k\in s}(y_k-\bar{y})^2\right]+E\left[\sum_{k\neq m\in s}(y_k-\bar{y})(y_m-\bar{y})\right]\right\}=$$

$$\frac{1}{n^2}\left[n\sigma^2-\frac{n(n-1)}{N-1}\sigma^2\right]=\frac{N-1-(n-1)}{n(N-1)}\sigma^2=$$

$$\frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{N-n}{N}\frac{S^2}{n}.$$

Veamos ahora que

$$E\left[\sum_{k\in s}(y_k-\bar{y})^2\right] = n\sigma^2$$

y que

$$E\left[\sum_{k\neq m\in s}(y_k-\bar{y})(y_m-\bar{y})\right] = -\frac{n(n-1)}{N-1}\sigma^2.$$

En efecto,

$$E\left[\sum_{k\in s}(y_k-\bar{y})^2\right] = \sum_{s\in S}\left[\sum_{k\in s}(y_k-\bar{y})^2\right]p(s) =$$

$$\sum_{k\in U}(y_k-\bar{y})^2 \text{card}\{s: k\in s\}p(s) =$$

$$\sum_{k\in U}(y_k-\bar{y})^2\binom{N-1}{n-1}\frac{1}{\binom{N}{n}} = N\sigma^2\frac{n}{N} = n\sigma^2.$$

También,

$$E\left[\sum_{k\neq m\in s}(y_k-\bar{y})(y_m-\bar{y})\right] =$$

$$\sum_{s\in S}\left[\sum_{k\neq m\in s}(y_k-\bar{y})(y_m-\bar{y})\right]p(s) =$$

$$\sum_{k\neq m\in U}(y_k-\bar{y})(y_m-\bar{y})\text{card}\{s: k,m\in s\}p(s) =$$

$$-N\sigma^2\binom{N-2}{n-2}\frac{1}{\binom{N}{n}} = -N\sigma^2\frac{n(n-1)}{N(N-1)} = -\frac{n(n-1)}{N-1}\sigma^2.$$

Queda comprobar que

$$\sum_{k\neq m\in U}(y_k-\bar{y})(y_m-\bar{y}) = -N\sigma^2.$$

En efecto, como

$$\sum_{k\in U}(y_k-\bar{y}) = 0,$$

$$0 = \left[\sum_{k\in U}(y_k-\bar{y})\right]^2 =$$

$$\sum_{k\in U}(y_k-\bar{y})^2 + \sum_{k\neq m\in U}(y_k-\bar{y})(y_m-\bar{y}) =$$

$$N\sigma^2 + \sum_{k\neq m\in U}(y_k-\bar{y})(y_m-\bar{y}),$$

lo que concluye la demostración.

## 3.3 Estimación de la varianza en *mia*

Veamos que la "cuasivarianza muestral" $s^2$ es un estimador insesgado de la "cuasivarianza poblacional" $S^2$ con diseño *mia*, y de este modo podremos obtener directamente estimadores insesgados de la "varianza poblacional" y de la "varianza de la media muestral" para este diseño de muestreo irrestricto aleatorio.

La cuasivarianza muestral es

$$s^2 = \frac{1}{n-1}\sum_{k \in s}(y_k - \bar{y}_s)^2.$$

Su esperanza matemática es

$$E(s^2) = \frac{1}{n-1}E\left[\sum_{k \in s}(y_k - \bar{y}_s)^2\right] =$$

$$\frac{1}{n-1}\sum_{s \in S}\left[\sum_{k \in s}(y_k - \bar{y}_s)^2\right]p(s),$$

por lo que restando y sumando la media poblacional $\bar{y}$ dentro del primer paréntesis, tenemos

$$[(y_k - \bar{y}) + (\bar{y} - \bar{y}_s)]^2 =$$

$$(y_k - \bar{y})^2 + (\bar{y} - \bar{y}_s)^2 + 2(y_k - \bar{y})(\bar{y} - \bar{y}_s).$$

Ahora,

$$\sum_{s \in S}\left[\sum_{k \in s}(y_k - \bar{y})^2\right]p(s) =$$

$$\sum_{k \in U}(y_k - \bar{y})^2 \text{card}\{s: k \in s\}p(s) = n\sigma^2.$$

$$\sum_{s \in S}\left[\sum_{k \in s}(\bar{y} - \bar{y}_s)^2\right]p(s) = nV(\bar{y}_s).$$

$$\sum_{s \in S}\left[\sum_{k \in s}2(y_k - \bar{y})(\bar{y} - \bar{y}_s)\right]p(s) =$$

$$2\sum_{s \in S}(\bar{y} - \bar{y}_s)\left[\sum_{k \in s}(y_k - \bar{y})\right]p(s) =$$

$$-2n\sum_{s\in S}(\bar{y}_s-\bar{y})^2p(s)=-2nV(\bar{y}_s).$$

Luego, sustituyendo estas expresiones calculadas,

$$E(s^2)=\frac{1}{n-1}[n\sigma^2+nV(\bar{y}_s)-2nV(\bar{y}_s)]=$$

$$\frac{n}{n-1}[\sigma^2-V(\bar{y}_s)]=\frac{\sigma^2}{n-1}\left(n-\frac{N-n}{N-1}\right)=\frac{N\sigma^2}{N-1}=S^2.$$

Por tanto, un estimador insesgado de la varianza poblacional con diseño *mia* es $(N-1)s^2/N$, y un estimador insesgado de la varianza del estimador media muestral es

$$\hat{V}(\bar{y}_s)=\frac{N-n}{N}\frac{s^2}{n}.$$

## 3.4 Estimación del total poblacional en *mia*

El total poblacional $T=N\bar{y}$ es estimado insesgadamente por el estimador $\hat{T}=N\bar{y}_s$ con diseño *mia*, y su varianza es

$$V(\hat{T})=N^2V(\bar{y}_s)=N^2\frac{N-n}{N}\frac{S^2}{n}=N(N-n)\frac{S^2}{n},$$

por lo que un estimador insesgado de la varianza de $\hat{T}$ en *mia* es

$$\hat{V}(\hat{T})=N(N-n)\frac{s^2}{n}.$$

## 3.5 Estimación de la proporción poblacional en *mia*

Como la proporción poblacional $P$ es una media poblacional $\bar{y}$ cuando la variable de interés $y_k$ toma valores 1 ó 0 según la unidad

$k$ posea o no posea una cualidad, el estimador insesgado de $P$ con diseño *mia* es la proporción muestral

$$\hat{P} = \frac{1}{n}\sum_{k \in s} y_k,$$

cuya varianza es

$$V(\hat{P}) = \frac{N-n}{N-1}\frac{PQ}{n},$$

puesto que $\sigma^2 = PQ$, siendo $Q = 1 - P$ la proporción poblacional de unidades que no poseen la cualidad. Un estimador insesgado de $V(\hat{P})$ con este diseño es

$$\hat{V}(\hat{P}) = \frac{N-n}{N}\frac{\hat{P}\hat{Q}}{n-1}.$$

## 3.6 Tamaño de la muestra con *mia*

El problema que vamos a tratar de resolver es el de la determinación del tamaño muestral $n$ necesario para alcanzar un error máximo de muestreo $e$, con una probabilidad mayor o igual a $1 - \alpha$.

Aplicando la desigualdad de Chebychev

$$p\{|\bar{y}_s - \bar{y}| < e\} \geq 1 - \frac{V(\bar{y}_s)}{e^2} = 1 - \alpha$$

ya que no es conocida la distribución del estimador $\bar{y}_s$, y esta desigualdad asegura el resultado independientemente de su distribución, con solo saber que su varianza $V(\bar{y}_s)$ existe. Entonces,

$$\alpha = \frac{V(\bar{y}_s)}{e^2} = \frac{(N-n)\sigma^2}{e^2(N-1)n} = \frac{N\sigma^2}{e^2(N-1)n} - \frac{\sigma^2}{e^2(N-1)},$$

de donde despejando $n$ tenemos

$$n = \frac{\frac{N\sigma^2}{e^2(N-1)}}{\alpha + \frac{\sigma^2}{e^2(N-1)}} = \frac{S^2}{\alpha e^2 + \frac{S^2}{N}},$$

donde $N$ es conocido, $\alpha$ y $e$ vienen determinados en el planteamiento del problema por el nivel de confianza y el error absoluto máximo de muestreo solicitados, y $S^2$ es la cuasivarianza poblacional que puede ser estimada insesgadamente por la cuasivarianza muestral piloto $s_0^2$. Si el tamaño poblacional $N$ es suficientemente grande, es decir $N \to \infty$, entonces el tamaño muestral buscado será $n_\infty = S^2/(\alpha e^2)$, que puede ser estimado sin sesgo por $\hat{n}_\infty = s_0^2/(\alpha e^2)$. Una vez obtenido éste, podemos expresar el tamaño muestral en general como

$$n = \frac{n_\infty}{1 + \frac{n_\infty}{N}} < n_\infty.$$

Obligando a que $n_\infty - n < 1$, obtenemos el primer valor de $n$ a partir del cual no se deben seguir obteniendo unidades muestrales pues $n$ alcanza el valor límite $n_\infty$. En efecto,

$$n_\infty - n = n_\infty - \frac{n_\infty}{1 + \frac{n_\infty}{N}} = n_\infty \left(1 - \frac{N}{N + n_\infty}\right) = \frac{n_\infty^2}{N + n_\infty} < 1,$$

verificándose la desigualdad si y solo si

$$n_\infty^2 < N + n_\infty,$$

que implica

$$n_\infty^2 - n_\infty = n_\infty(n_\infty - 1) < N.$$

Es decir, si $n_\infty = S^2/(\alpha e^2)$ es el primer número natural que verifica que $n_\infty(n_\infty - 1) < N$, entonces tomaremos como tamaño muestral $n = n_\infty$, y en otro caso tomamos la fórmula

$$n = \frac{S^2}{\alpha e^2 + \frac{S^2}{N}},$$

estimando $S^2$ por la cuasivarianza muestral $s_0^2$ en una muestra piloto con diseño *mia*. En la práctica, se sustituye en estos razonamientos el valor teórico $n_\infty$ por su estimación insesgada $\hat{n}_\infty = s_0^2/(\alpha e^2)$, y en su caso estimamos $n$ por

$$\hat{n} = \frac{s_0^2}{\alpha e^2 + \frac{s_0^2}{N}}.$$

En el caso de la estimación del total poblacional $T = N\bar{y}$, el tamaño $n$ de la muestra para un error absoluto máximo $e$ y un nivel de confianza mayor o igual a $1 - \alpha$, se obtiene de la desigualdad de Chebychev:

$$p\{|N\bar{y}_s - N\bar{y}| < e\} \geq 1 - \frac{V(N\bar{y}_s)}{e^2} = 1 - \alpha.$$

De donde

$$\alpha = \frac{V(N\bar{y}_s)}{e^2} = \frac{N(N-n)\frac{S^2}{n}}{e^2} = \frac{\frac{N^2S^2}{n} - NS^2}{e^2},$$

luego despejando $n$

$$n = \frac{N^2S^2}{\alpha e^2 + NS^2}.$$

En el caso particular de la proporción muestral $\hat{P}$ como estimador insesgado de la proporción poblacional $P$, el tamaño $n$ muestral para un error absoluto máximo de muestreo $e$ y un nivel

de confianza mayor o igual a $1-\alpha$, se obtiene de aplicar la desigualdad de Chebychev:

$$p\{|\hat{P}-P|<e\}\geq 1-\frac{V(\hat{P})}{e^2}=1-\alpha,$$

por lo que

$$\alpha=\frac{V(\hat{P})}{e^2}=\frac{\frac{N-n}{N-1}\frac{PQ}{n}}{e^2}=\frac{\frac{NPQ}{(N-1)n}-\frac{PQ}{N-1}}{e^2},$$

de donde despejando $n$ tenemos

$$n=\frac{\frac{NPQ}{N-1}}{\alpha e^2+\frac{PQ}{N-1}}.$$

Cuando el tamaño poblacional $N$ es suficientemente grande, $N\to\infty$, el tamaño muestral límite es $n_\infty=PQ/(\alpha e^2)$. Si ahora dividimos en la fórmula del tamaño muestral tanto numerador como denominador por $\alpha e^2$, tenemos

$$n=\frac{n_\infty\frac{N}{N-1}}{1+\frac{n_\infty}{N-1}}=\frac{Nn_\infty}{N-1+n_\infty}.$$

Si ahora obligamos a que $n_\infty-n<1$, tenemos que

$$n_\infty-n=n_\infty\left(1-\frac{N}{N-1+n_\infty}\right)=n_\infty\frac{n_\infty-1}{N-1+n_\infty}<1$$

o bien,

$$n_\infty(n_\infty-1)<N-1+n_\infty$$

o también, concluimos que si

$$n_\infty(n_\infty-2)<N-1,$$

tomamos como tamaño muestral $n = n_\infty = PQ/(\alpha e^2)$, mientras que si

$$n_\infty(n_\infty - 2) \geq N - 1,$$

tomamos como tamaño muestral el valor obtenido:

$$n = \frac{NPQ}{(N-1)\alpha e^2 + PQ}.$$

En la práctica la varianza $PQ$ ha de ser estimada sin sesgo por

$$\widehat{PQ} = \frac{N-1}{N}\frac{n_0}{n_0 - 1}\hat{P}_0\hat{Q}_0,$$

siendo $n_0$ el tamaño de la muestra piloto, y $\hat{P}_0$ y $\hat{Q}_0$ son las proporciones muestrales respectivas en la muestra piloto. Sustituyendo $PQ$ por su estimador insesgado $\widehat{PQ}$, se obtienen los valores estimados $\hat{n}_\infty$ y $\hat{n}$.

## 3.7 Tamaño muestral con hipótesis de normalidad

Existen críticas para formular la hipótesis de normalidad en poblaciones finitas con diseño *mia*. La idea de estas críticas vienen de que el Teorema Central del Límite no es aplicable porque éste exige que el tamaño muestral tienda a infinito, pero en el diseño *mia* el tamaño muestral $n$ es el efectivo y tiene que ser menor o igual al tamaño poblacional $N$ que es finito. Otros argumentos debidos a la distribución del estimador fueron dados por Plane y Gordon (1982). Básicamente se demuestra que la distribución del estimador cuando $n$ se aproxima a $N$, es la misma que cuando el tamaño muestral es pequeño o próximo a 1, salvo un cambio o transformación lineal. Si la distribución del estimador no es normal cuando $n$ es pequeño, tampoco lo será cuando $n$ tome los valores mayores con diseño *mia*.

No obstante, al estimar una proporción poblacional $P$ con diseño *mia*, en la práctica es usual aproximar la distribución de la proporción muestral $\hat{P}$ por una distribución normal de media $P$ y desviación típica $\sqrt{V(\hat{P})}$. En este caso, con diseño *mia*, el tamaño muestral $n$ para el nivel de confianza $1-\alpha_1 = 0.955$, con $\lambda_{\alpha_1} = 2$, o bien para $1-\alpha_2 = 0.997$, con $\lambda_{\alpha_2} = 3$, se obtiene de la relación

$$p\left\{\frac{|\hat{P}-P|}{\sqrt{V(\hat{P})}} < \lambda_\alpha\right\} = 1-\alpha.$$

Aquí,

$$V(\hat{P}) = \frac{N-n}{N-1}\frac{PQ}{n} \leq \frac{N-n}{4(N-1)n},$$

pues $PQ \leq 1/4$. El intervalo de confianza de la proporción poblacional $P$ es

$$\hat{P} \mp \lambda_\alpha \sqrt{V(\hat{P})}.$$

Pero este intervalo está contenido siempre en el intervalo

$$\hat{P} \mp \lambda_\alpha \sqrt{\frac{N-n}{4(N-1)n}} = \hat{P} \mp \frac{\lambda_\alpha}{2}\sqrt{\frac{N}{(N-1)n} - \frac{1}{N-1}},$$

por lo que fijada la semiamplitud del intervalo de confianza $e$, el intervalo de confianza

$$\hat{P} \mp e$$

se obtiene cuando

$$e = \frac{\lambda_\alpha}{2}\sqrt{\frac{N}{(N-1)n} - \frac{1}{N-1}},$$

de donde despejando $n$ tenemos

$$n = \frac{N}{N-1}\frac{1}{\frac{4e^2}{\lambda_\alpha^2} + \frac{1}{N-1}} = \frac{\lambda_\alpha^2 N}{4e^2(N-1) + \lambda_\alpha^2}.$$

## 3.8 Comparación de precisiones entre *mas* y *mia*

La precisión de un estimador es el inverso de su varianza. Así, para comparar las precisiones de dos estimadores insesgados, bastará comparar sus varianzas. Veamos a continuación que de los resultados obtenidos,

$$V(mas, \bar{y}_s) = \frac{\sigma^2}{n}$$

y

$$V(mia, \bar{y}_s) = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

Luego,

$$V(mia, \bar{y}_s) \leq V(mas, \bar{y}_s).$$

Esto hace pensar que si el muestreo irrestricto aleatorio es más preciso que el muestreo aleatorio simple, deberíamos usar el primero siempre en detrimento del segundo desde un punto de vista de la mejor precisión del estimador para un tamaño muestral $n$ común. Sin embargo, al poder tener unidades repetidas en la muestra ordenada obtenida por *mas*, puede ser más económica en términos esperados la obtención de datos, pues conservando la variable de interés asociada a cada unidad, si ésta se repite

ahorraremos el costo de una nueva observación o encuesta (o sucesivas) de la misma unidad. Observar que el tamaño efectivo esperado del diseño *mia* es $n$, mientras que el tamaño efectivo esperado del diseño *mas* es $\bar{v} \leq n$, lo que se traduce en que tiene este diseño *mas* menor costo esperado que con el diseño *mia*.

## 3.9 Ejercicios resueltos

**Ejercicio 3.1.** Dada una población finita de tamaño $N = 2000$, se toma una muestra de tamaño $n = 20$ con diseño *mia*, de modo que la media muestral es $\bar{y}_s = 537$ y la cuasivarianza muestral es $s^2 = 100$. Se pide una estimación del total poblacional, así como de la varianza del estimador del total poblacional propuesto, utilizando estimadores insesgados.

**Solución.** El estimador del total poblacional es

$$N\bar{y}_s = 2000 \cdot 537 = 1.074.000.$$

Y el estimador insesgado de su varianza es

$$\hat{V}(N\bar{y}_s) = N(N-n)\frac{s^2}{n} = 2000 \cdot 1980 \cdot \frac{100}{20} = 198 \cdot 10^5.$$

**Ejercicio 3.2.** Acotar la varianza de una proporción muestral con diseño *mia* en cualquier caso, independientemente de los posibles valores que pueda tomar la proporción poblacional.

**Solución.** La varianza de la proporción muestral $\hat{P}$, $V(\hat{P})$, es

$$V(\hat{P}) = \frac{N-n}{N-1}\frac{PQ}{n} \leq \frac{N-n}{4(N-1)n},$$

pues definiendo la función $f(P) = P(1-P) = PQ$ , ésta se minimiza en el punto $P = 1/2$, es decir, es un punto crítico de la función $f$, pues $f'(P) = 1 - 2P = 0$ da lugar al punto crítico $P = 1/2$. Al ser $f''(P) = -2 < 0$, el punto crítico es un máximo de $f$. Por esto $PQ = P(1-P) = f(P) \leq f(1/2) = 1/4$.

**Ejercicio 3.3.** Calcular el tamaño muestral necesario para obtener un error máximo de muestreo $e = 10^5$ al nivel de confianza 0.90 para estimar el total de una población finita de tamaño $N = 30.000$. De una muestra piloto, se estima $S^2$ sin sesgo por 50.

**Solución.** Como $1 - \alpha = 0.90$, $\alpha = 0.10$, y entonces

$$n = \frac{N^2 S^2}{\alpha e^2 + N S^2} \approx \frac{9 \cdot 10^8 \cdot 50}{0.10 \cdot 10^{10} + 3 \cdot 10^4 \cdot 50} \approx 45.$$

**Ejercicio 3.4.** Una muestra aleatoria simple sin reemplazamiento ha sido seleccionada de la población compuesta por las familias residentes en cierta provincia, con objeto de estimar el número medio de hijos varones por familia. Se han observado $n = 10$ familias del total de las mismas que eran 39.000. Los datos fueron sintetizados en los estadísticos

$$\sum_{i=1}^{10} y_{k_i} = 19$$

y

$$\sum_{i=1}^{10} y_{k_i}^2 = 71.$$

Estimar insesgadamente la media provincial de hijos varones por familia, y estimar insesgadamente la varianza del primer estimador.

**Solución.** Como estimador de la media provincial de hijos varones por familia, tomamos la media muestral

$$\bar{y}_s = \frac{1}{n}\sum_{i=1}^{10} y_{k_i} = \frac{1}{10}\cdot 19 = 1.9,$$

y como estimador insesgado de su varianza tenemos a

$$\hat{V}(\bar{y}_s) = \frac{N-n}{N}\frac{s^2}{n} \approx \frac{39.000-10}{39.000}\frac{3.877}{10} \approx 0.3876,$$

pues $N = 39.000$, $n = 10$ y

$$s^2 = \frac{\sum_{i=1}^{10}\left(y_{k_i}-\bar{y}_s\right)^2}{n-1} = \frac{\sum_{i=1}^{10} y_{k_i}^2 - \frac{\left(\sum_{i=1}^{10} y_{k_i}\right)^2}{10}}{9} = \frac{71-\frac{19^2}{10}}{9},$$

es decir

$$s^2 \approx 3.877.$$

**Ejercicio 3.5.** Una industria tiene interés en conocer el tiempo semanal que los empleados gastan en ciertas actividades no productivas. Las fichas de control del tiempo de una muestra con diseño *mia* de $n = 70$ empleados muestran que el tiempo promedio dedicado a esas actividades es de 16.45 horas, con una cuasivarianza muestral de $s^2 = 3.01$. La empresa da trabajo a un total de $N = 1.250$ empleados. Estimar el número total de horas-hombre que se pierden por semana en tareas no productivas y dar una estimación de la varianza de tal estimación inicial.

**Solución.** La población consiste en $N = 1.250$ empleados, de los que se selecciona una muestra con diseño *mia* de tamaño $n = 70$ empleados. La cantidad promedio de tiempo que uno de los 70 empleados pierde es de $\bar{y}_s = 16.45$ horas semanales. Luego la estimación del total semanal de horas perdidas por los 1.250 empleados es

$$\hat{T} = N\bar{y}_s = 1250 \cdot 16.45 = 20562.5 \text{ horas.}$$

Un estimador insesgado de la varianza de este estimador $\hat{T}$ es

$$\hat{V}(\hat{T}) = N(N-n)\frac{s^2}{n} = 1250 \cdot 1180 \cdot \frac{3.01}{70} = 63425$$

horas al cuadrado.

**Ejercicio 3.6.** Para estimar la renta familiar disponible al año de una población, en promedio, se sabe que existen un total de 200000 familias y que tras una encuesta piloto, se ha estimado que la cuasivarianza de la renta familiar es $S^2 \approx 2000$. Determinar el tamaño muestral necesario para estimar la renta familiar media poblacional $\bar{y}$ mediante la media muestral $\bar{y}_s$ obtenida por diseño *mia* para alcanzar un error máximo de muestreo $e = 200$ euros con una probabilidad $1 - \alpha = 0.95$.

**Solución.** El tamaño muestral con diseño *mia* directamente es

$$n = \frac{S^2}{\alpha e^2 + \frac{S^2}{N}} \approx \frac{2000}{0.05 \cdot 4 \cdot 10^4 + \frac{2000}{200000}} \approx 1.$$

Luego con el tamaño muestral $n = 1$ se obtiene una estimación de la media muestral con error absoluto máximo $e = 200$ euros y un nivel de confianza mayor o igual al $95\%$, siempre que la

estimación de la cuasivarianza poblacional $\widehat{S^2} = 2000$ sea correcta.

**Ejercicio 3.7.** Una empresa productora de aves para el consumo alimenticio está interesada en estimar la ganancia total de peso de un total de $2000$ aves a lo largo de un mes mediante la alimentación de las aves con una ración. Frente a la alternativa de tener que pesar las $2000$ aves un mes después, se diseña un método de estimación del peso total por el que se pesarán $n$ aves de modo que el error máximo de muestreo sea 3 kg. al nivel de confianza del $90\%$. Usando datos de anteriores estudios similares, se ha estimado la cuasivarianza muestral $s^2 = 40$ gramos al cuadrado. Determinar el tamaño muestral.

**Solución.** El tamaño muestral necesario para estimar el total poblacional es

$$n = \frac{S^2}{\frac{\alpha e^2}{N^2} + \frac{S^2}{N}} \approx \frac{40}{\frac{0.10 \cdot 3000^2}{2000^2} + \frac{40}{2000}} \approx 163.27,$$

por lo que pesando una muestra de $164$ aves podemos estimar dicho peso total con dichos requerimientos. El valor que hemos dado como respuesta es el número entero siguiente al valor aproximado que da la fórmula del tamaño muestral.

**Ejercicio 3.8.** Una muestra aleatoria simple sin reemplazamiento de tamaño $n = 100$ se ha seleccionado para estimar:

a) La fracción de los $300$ estudiantes de un Instituto que asistirán a la Universidad.

b) La fracción de estudiantes que han trabajado a tiempo parcial durante su estancia en el Instituto.

Sean 25 y 30 los totales muestrales de estudiantes que asistirán a la Universidad, y de estudiantes que han trabajado a tiempo parcial durante su estancia en el Instituto. Usando estos datos, estimar la proporción de estudiantes del Instituto que asistirán a la Universidad, y la de estudiantes que ha trabajado a tiempo parcial durante su estancia en el Instituto. Estimar sin sesgo la varianza de estos estimadores de las proporciones de estudiantes del Instituto.

**Solución.** Las proporciones muestrales se obtienen directamente de los datos recogidos,

$$\hat{P}_1 = \frac{25}{100} = 0.25$$

y

$$\hat{P}_2 = \frac{30}{100} = 0.3$$

son los estimadores pedidos, y los estimadores insesgados de sus varianzas son

$$\hat{V}(\hat{P}_1) = \frac{N-n}{N}\frac{\hat{P}_1\hat{Q}_1}{n-1} = \frac{300-100}{300}\frac{0.25 \cdot 0.75}{99} = 0.00126$$

y

$$\hat{V}(\hat{P}_2) = \frac{N-n}{N}\frac{\hat{P}_2\hat{Q}_2}{n-1} = \frac{300-100}{300}\frac{0.3 \cdot 0.7}{99} = 0.00141$$

**Ejercicio 3.9.** Una empresa tiene a su cargo un total de 2000 obreros y el jefe de personal quiere estimar la proporción de obreros que llevan trabajando en la empresa más de 10 años. A tal efecto decide realizar un sondeo entre los obreros, ya que realizar

un censo sería inapropiado debido a la rapidez con la que debe disponer de los datos. Si selecciona una muestra con diseño *mia* para estimar tal proporción, determinar el tamaño muestral $n$, aceptando que la proporción muestral del $50\%$ estima suficientemente bien la proporción poblacional, cuando el error máximo admisible de muestreo es $0.1$ al nivel de confianza del $95\%$.

**Solución.** Si la proporción muestral del $50\%$ estima bien la proporción poblacional, es que aproximadamente $P = Q = 0.5$. Entonces,

$$n = \frac{\frac{NPQ}{N-1}}{\alpha e^2 + \frac{PQ}{N-1}} = \frac{\frac{2000 \cdot 0.5 \cdot 0.5}{1999}}{0.05 \cdot 0.1^2 + \frac{0.5 \cdot 0.5}{1999}} \approx 400.2$$

Por lo tanto, bastará tomar $401$ obreros en la muestra para verificar todos los requisitos, uno más de la parte entera del valor obtenido por la fórmula general. Se toma el tamaño muestral siguiente al número entero que da la fórmula pues de este modo se garantiza que las condiciones del enunciado del ejercicio propuesto se verifican.

**Ejercicio 3.10.** De una población finita se obtienen $m$ estimadores medias muestrales $\bar{y}_i$ $(i = 1, 2, \dots, m)$ independientes, cada uno de ellos con diseño de muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n_i$ $(i = 1, 2, \dots, m)$. Deducir el estimador insesgado lineal en las medias muestrales $\bar{y}_i$ de mínima varianza.

**Solución.** Buscamos el estimador insesgado de mínima varianza de la clase de estimadores lineales insesgados del tipo

$$t = \sum_{i=1}^{m} t_i \, \bar{y}_i.$$

Como este estimador lineal es insesgado, tenemos que

$$\bar{y} = E(t) = \sum_{i=1}^{m} t_i E(\bar{y}_i) = \bar{y} \sum_{i=1}^{m} t_i,$$

por lo que esta condición de insesgación se traduce en que

$$1 = \sum_{i=1}^{m} t_i.$$

La varianza del estimador $t$ es

$$V(t) = \sum_{i=1}^{m} t_i^2 V(\bar{y}_i) = \sum_{i=1}^{m} t_i^2 \frac{N - n_i}{N n_i} \sigma^2.$$

Para obtener los valores concretos de $t_i$ que hacen insesgado al estimador lineal y que es de varianza mínima, hacemos uso del método de los multiplicadores de Lagrange. El lagrangiano será:

$$\Lambda = V(t) + \lambda[E(t) - \bar{y}].$$

O equivalentemente:

$$\Lambda = V(t) + \lambda \left( \sum_{i=1}^{m} t_i - 1 \right).$$

Derivando parcialmente $\Lambda$ con respecto a $t_i$, e igualando a cero, tenemos:

$$\frac{\partial \Lambda}{\partial t_i} = 2 t_i \frac{N - n_i}{N n_i} \sigma^2 + \lambda = 0,$$

de donde

$$t_i = -\lambda \frac{N n_i}{2(N - n_i)\sigma^2} = c \frac{n_i}{N - n_i}.$$

La constante $c$ se determina por la restricción de insesgación, es decir,

$$1 = \sum_{i=1}^{m} t_i = c \sum_{i=1}^{m} \frac{n_i}{N - n_i},$$

de donde

$$c = \frac{1}{\sum_{i=1}^{m} \frac{n_i}{N - n_i}}.$$

Por tanto, el estimador insesgado de mínima varianza del tipo lineal $t = \sum_{i=1}^{m} t_i \bar{y}_i$ tiene por componente $t_i$ a:

$$t_i = \frac{\frac{n_i}{N - n_i}}{\sum_{j=1}^{m} \frac{n_j}{N - n_j}}.$$

El estimador buscado es entonces

$$t = \sum_{i=1}^{m} \frac{\frac{n_i}{N - n_i}}{\sum_{j=1}^{m} \frac{n_j}{N - n_j}} \bar{y}_i.$$

**Ejercicio 3.11.** Indicar un estimador insesgado de la varianza, $\hat{V}(t)$, del estimador insesgado lineal de mínima varianza, $t$, para estimar la media poblacional $\bar{y}$, que hemos obtenido en el ejercicio anterior.

**Solución.** Partimos de que un estimador insesgado de la cuasivarianza poblacional

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

es la cuasivarianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y}_s)^2,$$

donde $y_i$ es el valor observado de la variable $y$ en la unidad $i$ de la muestra conjunto $s$ de $n$ unidades de entre las de la población finita de tamaño $N$.

Denotando por $s_i^2$ a la cuasivarianza muestral obtenida con la $i$-ésima muestra aleatoria simple sin reemplazamiento de tamaño muestral efectivo $n_i$, todas ellas en $i = 1, 2, \dots, m$ selecciones independientes entre sí, llegamos al "estimador insesgado de la varianza" del estimador $t$ lineal insesgado de la media poblacional y de mínima varianza en su clase:

$$\hat{V}(t) = \sum_{i=1}^{m} t_i^2 \hat{V}(\bar{y}_i) = \sum_{i=1}^{m} \left( \frac{\frac{n_i}{N-n_i}}{\sum_{j=1}^{m} \frac{n_j}{N-n_j}} \right)^2 \frac{N-n_i}{(N-1)n_i} s_i^2 =$$

$$\frac{1}{N-1} \frac{\sum_{i=1}^{m} \frac{n_i}{N-n_i} s_i^2}{\left( \sum_{j=1}^{m} \frac{n_j}{N-n_j} \right)^2}.$$

**Ejercicio 3.12.** En algunos casos, de la variable de interés $y$ a observar, sabemos que está acotada inferior y superiormente, es decir $y_k \in [a, b]$ para todo valor o unidad $k \in U$, de la población finita, con $a$ y $b$ constantes reales, $a < b$. Demostrar en estas

condiciones que la media muestral obtenida por muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n$, puede llegar a estimar la media poblacional $\bar{y}$ con un error menor que cualquier cantidad positiva $e > 0$ para cualquier nivel de confianza prefijado $1 - \alpha$, con $0 < \alpha \leq 1$.

**Solución.** Para verlo, empezamos por justificar que si $y_k \in [a, b]$, entonces la varianza poblacional verifica que:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_k - \bar{y})^2 \leq \frac{1}{N}\sum_{i=1}^{N}(b-a)^2 = (b-a)^2.$$

Por lo tanto

$$p\{|\bar{y}_s - \bar{y}| < e\} \geq 1 - \frac{V(\bar{y}_s)}{e^2} = 1 - \frac{(N-n)\sigma^2}{(N-1)ne^2} \geq$$

$$1 - \frac{(N-n)(b-a)^2}{(N-1)ne^2} \geq 1 - \alpha.$$

Esto es cierto si

$$\alpha \geq \frac{N(b-a)^2}{(N-1)e^2n} - \frac{(b-a)^2}{(N-1)e^2}.$$

O bien si

$$n \geq \frac{N(b-a)^2}{\left[\alpha + \frac{(b-a)^2}{(N-1)e^2}\right](N-1)e^2} = \frac{N(b-a)^2}{\alpha e^2(N-1) + (b-a)^2}.$$

Esta última cota inferior del tamaño muestral efectivo prefijado $n$ sabemos que garantiza con seguridad que la media muestral $\bar{y}_s$, con diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n$, estima la media poblacional $\bar{y}$ con un error absoluto de muestreo menor que $e$ a un nivel de confianza mayor o igual a $1 - \alpha$. Por lo

general, a otros tamaños muestrales inferiores es muy posible que también se alcance este nivel de confianza, ya que la acotación $\sigma^2 \leq (b-a)^2$ es muy amplia y la varianza poblacional puede tener cotas superiores más pequeñas que $(b-a)^2$.

**Ejercicio 3.13.** Obtener el tamaño muestral $n$ que haga que la proporción muestral $p$ se desvíe de la proporción poblacional $P$ menos de una cantidad $e > 0$, con un nivel de confianza mayor o igual a $1-\alpha$. La selección de la muestra es con diseño de muestreo irrestricto aleatorio.

**Solución.** Al tomar los valores de la variable de interés ceros y unos, la varianza poblacional $\sigma^2 = PQ = P(1-P) \leq 1/4$ . Entonces, la desigualdad de Chebychev nos dice que

$$p\{|p-P| < e\} \geq 1 - \frac{V(p)}{e^2} \geq 1-\alpha,$$

o bien,

$$V(p) = \frac{(N-n)PQ}{(N-1)n} \leq \frac{N-n}{4(N-1)n} \leq \alpha e^2$$

y esto es cierto cuando

$$\frac{N}{n} - 1 \leq 4(N-1)\alpha e^2,$$

o bien, cuando

$$n \geq \frac{N}{4(N-1)\alpha e^2 + 1}.$$

El valor natural mínimo que satisface la última desigualdad es más pequeña que la que proporcionaba la solución del ejercicio anterior, y es que hemos podido acotar la varianza poblacional por una cota

superior en este ejercicio que es la cuarta parte de la que teníamos en general en el ejercicio anterior.

**Ejercicio 3.14.** Seleccionamos dos muestras independientes de una misma población finita de tamaño $N$, una con diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo $n_1$, y otra con diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n_2$. Obtenemos las dos medias muestrales con ambos diseños, y las denotamos $\bar{y}_{\boldsymbol{s}}$ e $\bar{y}_s$, y definimos la clase de estimadores de la media poblacional por $t = t_1\bar{y}_{\boldsymbol{s}} + t_2\bar{y}_s$. Obtener el estimador insesgado de mínima varianza (óptimo) de esta clase de estimadores, y obtener el estimador insesgado de la varianza del estimador óptimo.

**Solución.** La condición de insesgación para estimar la media poblacional $\bar{y}$ se resume en que

$$\bar{y} = E(t) = E(t_1\bar{y}_{\boldsymbol{s}} + t_2\bar{y}_s) = (t_1 + t_2)\bar{y},$$

o bien

$$1 = t_1 + t_2.$$

La varianza del estimador $t$ resulta ser:

$$V(t) = t_1^2\frac{\sigma^2}{n_1} + t_2^2\frac{N - n_2}{(N-1)n_2}\sigma^2.$$

Para obtener los valores óptimos de $t_1$ y $t_2$, hacemos uso del método de los multiplicadores de Lagrange. El lagrangiano es:

$$\Lambda = V(t) + \lambda(1 - t_1 - t_2).$$

Resolviendo, tenemos el sistema:

$$\frac{\partial \Lambda}{\partial t_1} = 2t_1 \frac{\sigma^2}{n_1} - \lambda = 0$$

$$\frac{\partial \Lambda}{\partial t_2} = 2t_2 \frac{N - n_2}{(N-1)n_2} \sigma^2 - \lambda = 0.$$

Resolviéndolo,

$$t_1 = \frac{\lambda n_1}{2\sigma^2}$$

$$t_2 = \frac{\lambda (N-1) n_2}{2(N - n_2)\sigma^2}.$$

Exigiendo la condición de insesgación, determinamos el valor de $\lambda$ del modo:

$$1 = t_1 + t_2 = \frac{\lambda}{2\sigma^2} \left[ n_1 + \frac{(N-1)n_2}{N - n_2} \right],$$

es decir,

$$\lambda = \frac{2\sigma^2}{n_1 + \frac{(N-1)n_2}{N - n_2}}.$$

Por tanto, los valores óptimos de $t_1$ y de $t_2$ son al sustituir la constante $\lambda$:

$$t_1 = \frac{n_1}{n_1 + \frac{(N-1)n_2}{N - n_2}}$$

y

$$t_2 = \frac{\frac{(N-1)n_2}{N - n_2}}{n_1 + \frac{(N-1)n_2}{N - n_2}}.$$

El estimador insesgado de la varianza del estimador óptimo de la clase es:

$$\hat{V}\left(t_{\acute{o}pt}\right) = t_1^2\hat{V}(\bar{y}_s) + t_2^2\hat{V}(\bar{y}_s) =$$

$$\left[\frac{n_1}{n_1 + \frac{(N-1)n_2}{N-n_2}}\right]^2 \frac{s_1^2}{n_1} + \left[\frac{\frac{(N-1)n_2}{N-n_2}}{n_1 + \frac{(N-1)n_2}{N-n_2}}\right]^2 \frac{N-n_2}{Nn_2} s_2^2,$$

en donde $s_1^2$ es la cuasivarianza muestral con diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo $n_1$, y $s_2^2$ es la cuasivarianza muestral con diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n_2$.

**Ejercicio 3.15.** Demostrar que el estimador

$$v = \frac{N-1}{n(n-1)} \sum_{i<j\in s} \left(y_i - y_j\right)^2$$

es insesgado para estimar la varianza poblacional con diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n$, siendo $N$ el tamaño de la población finita.

**Solución.** Vamos a usar la relación obtenida en el Ejercicio 1.7 de este libro, en concreto, de él tenemos que la varianza poblacional es:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(y_i - y_j\right)^2 = \frac{1}{N}\sum_{i<j}^{N}\left(y_i - y_j\right)^2.$$

Haciendo uso de esta igualdad, veamos que la esperanza matemática del estimador $v$ es exactamente $\sigma^2$:

$$E(v) = \frac{N-1}{n(n-1)} E\left[\sum_{i<j\in s} (y_i - y_j)^2\right] =$$

$$\frac{N-1}{n(n-1)} \sum_{s\in S}\left[\sum_{i<j\in s} (y_i - y_j)^2\right] p(s \in S)$$

donde $S$ es el conjunto de muestras conjunto o no ordenadas obtenidas por muestreo irrestricto aleatorio de tamaño efectivo fijo $n$. Entonces, calculando $p(s \in S) = 1/\text{card}\{s \in S\}$ por la regla de Laplace,

$$E(v) = \frac{N-1}{n(n-1)} \sum_{i<j}^{N} (y_i - y_j)^2 \frac{\text{card}\{s \in S : i < j \in s\}}{\text{card}\{s \in S\}} =$$

$$\frac{N-1}{n(n-1)} \sum_{i<j}^{N} (y_i - y_j)^2 \frac{\binom{N-2}{n-2}}{\binom{N}{n}} =$$

$$\frac{N-1}{n(n-1)} N\sigma^2 \frac{n(n-1)}{N(N-1)} = \sigma^2.$$

En realidad $Nv/(N-1) = s^2$, es la cuasivarianza muestral, que es un estimador insesgado de la cuasivarianza poblacional $S^2$, como ya sabíamos.

**Ejercicio 3.16.** Demostrar que la cuasicovarianza muestral es un estimador insesgado de la cuasicovarianza poblacional en el muestreo irrestricto aleatorio de tamaño efectivo fijo $n$.

**Solución.** Llamamos covarianza muestral al estadístico

$$m_{11} = \frac{1}{n}\sum_{i\in s}(y_i - \bar{y}_s)(x_i - \bar{x}_s).$$

La covarianza poblacional es la función paramétrica

$$\mu_{11} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}).$$

Lo que se nos pide es demostrar que

$$E(s_{yx}) = E\left(\frac{n}{n-1}m_{11}\right) = \frac{N}{N-1}\mu_{11} = S_{yx}.$$

Para ello, bastaría demostrar que

$$E(m_{11}) = \frac{N(n-1)}{(N-1)n}\mu_{11}.$$

Veámoslo.

$$E(m_{11}) = E\left(\frac{1}{n}\sum_{i\in s} y_i x_i\right) - E(\bar{y}_s\bar{x}_s) =$$

$$\alpha_{11} + \alpha_{10}\alpha_{01} - Cov(\bar{y}_s, \bar{x}_s) = \mu_{11} - \frac{N-n}{(N-1)n}\mu_{11} =$$

$$\frac{N(n-1)}{(N-1)n}\mu_{11}.$$

Queda demostrar que

$$Cov\left(\frac{1}{n}\sum_{i\in s} y_i, \frac{1}{n}\sum_{j\in s} x_j\right) = Cov\left(\frac{1}{N}\sum_{i=1}^{N} y_i e_i, \frac{1}{N}\sum_{j=1}^{N} x_j e_j\right) =$$

$$\frac{1}{n^2}\sum_{i=1}^{N} y_i x_i V(e_i) + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i x_j Cov(e_i, e_j).$$

Aquí $e_i$ es una variable aleatoria que toma valor 1 si $i \in s$, y toma valor 0 si $i \notin s$. Entonces, si $\pi_i$ es la probabilidad de inclusión de la unidad $i$ en la muestra,

$$V(e_i) = E(e_i^2) - [E(e_i)]^2 = \pi_i - \pi_i^2 = \frac{n}{N} - \frac{n^2}{N^2} = \frac{(N-n)n}{N^2}.$$

Y si $i \neq j$, denotando por $\pi_{ij}$ a la probabilidad de inclusión de las unidades distintas $i$ y $j$ en la muestra, tenemos

$$Cov(e_i, e_j) = E(e_i e_j) - E(e_i)E(e_j) = \pi_{ij} - \pi_i \pi_j =$$

$$\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{(N-n)n}{N^2(N-1)}.$$

Así, sustituyendo estos resultados en la covarianza de las medias muestrales, tenemos que

$$Cov(\bar{y}_s, \bar{x}_s) =$$

$$\frac{1}{n^2} N\alpha_{11} \frac{(N-n)n}{N^2} + \frac{1}{n^2}\sum_{i=1}^{N} y_i \sum_{j\neq i}^{N} x_j \, Cov(e_i, e_j) =$$

$$\frac{N-n}{Nn}\alpha_{11} + \frac{1}{n^2}\sum_{i=1}^{N} y_i (N\bar{x} - x_i) \frac{-(N-n)n}{N^2(N-1)} =$$

$$\frac{N-n}{Nn}\alpha_{11} - \frac{N-n}{(N-1)n}\alpha_{10}\alpha_{01} + \frac{N-n}{N(N-1)n}\alpha_{11} =$$

$$\frac{N-n}{(N-1)n}(\alpha_{11} - \alpha_{10}\alpha_{01}) = \frac{N-n}{(N-1)n}\mu_{11},$$

donde hemos denotado por

$$\alpha_{km} = \frac{1}{N}\sum_{i=1}^{N} y_i^k x_i^m.$$

**Ejercicio 3.17.** Proponer un estimador insesgado de la covarianza de las medias muestrales obtenidas por muestreo irrestricto aleatorio de tamaño efectivo fijo $n$.

**Solución.** Basándonos en el ejercicio anterior, sabemos que la covarianza de las medias muestrales en muestreo irrestricto aleatorio de tamaño efectivo fijo $n$ es

$$Cov(\bar{y}_s, \bar{x}_s) = \frac{N-n}{(N-1)n}\mu_{11}.$$

Como un estimador insesgado de la covarianza poblacional $\mu_{11}$ es

$$\hat{\mu}_{11} = \frac{(N-1)n}{N(n-1)}m_{11},$$

deducimos que el estimador buscado es

$$\widehat{Cov}(\bar{y}_s, \bar{x}_s) = \frac{N-n}{(N-1)n}\hat{\mu}_{11} = \frac{N-n}{N(n-1)}m_{11},$$

donde hemos denotado la covarianza muestral por

$$m_{11} = \frac{1}{n}\sum_{i\in s}(y_i - \bar{y}_s)(x_j - \bar{x}_s).$$

**Ejercicio 3.18.** De una población finita $U$ de tamaño $N$ se selecciona una muestra irrestricta aleatoria $s$ de tamaño $n$ de dicha población, y posteriormente se selecciona otra muestra irrestricta aleatoria $s'$ de tamaño $n'$, $n + n' \leq N$, de la población resultante

$U - s$. Obtener la covarianza de las medias muestrales de las muestras $s$ y $s'$. Finalmente proponer un estimador insesgado de dicha covarianza, calculable con los datos muestrales.

**Solución.**

$$Cov(\bar{y}_s, \bar{y}_{s'}) = E[Cov(\bar{y}_s, \bar{y}_{s'}|s)] + Cov[E(\bar{y}_s|s), E(\bar{y}_{s'}|s)]$$

Como

$$Cov(\bar{y}_s, \bar{y}_{s'}|s) = \bar{y}_s Cov(1, \bar{y}_{s'}|s) = \bar{y}_s \cdot 0 = 0,$$

deducimos que

$$E[Cov(\bar{y}_s, \bar{y}_{s'}|s)] = E(0) = 0.$$

Por otro lado,

$$E(\bar{y}_s|s) = \bar{y}_s$$

y

$$E(\bar{y}_{s'}|s) = \bar{y}_{U-s} = \frac{N\bar{y} - n\bar{y}_s}{N-n},$$

por lo que

$$Cov[E(\bar{y}_s|s), E(\bar{y}_{s'}|s)] = Cov\left(\bar{y}_s, \frac{N\bar{y} - n\bar{y}_s}{N-n}\right) =$$

$$Cov\left(\bar{y}_s, \frac{N\bar{y}}{N-n}\right) - \frac{n}{N-n} V(\bar{y}_s) =$$

$$0 - \frac{n}{N-n}\frac{N-n}{(N-1)n}\sigma^2 = -\frac{\sigma^2}{N-1},$$

concluyendo que

$$Cov(\bar{y}_s, \bar{y}_{s'}) = -\frac{\sigma^2}{N-1}.$$

Un estimador insesgado de esta covarianza, dada la relación

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

y que la cuasivarianza poblacional es estimada insesgadamente por la cuasivarianza muestral en el diseño de muestreo irrestricto aleatorio, concluimos que

$$\widehat{Cov}(\bar{y}_s, \bar{y}_{s\prime}) = -\frac{\widehat{S^2}}{N} = -\frac{s^2}{N},$$

donde $s^2$ es la cuasivarianza muestral de la muestra irrestricta aleatoria $s$ de tamaño efectivo fijo $n$, de la muestra irrestricta aleatoria $s'$ de tamaño efectivo fijo $n'$, o bien (preferiblemente desde un punto de vista de reducción de la varianza del estimador, y aprovechando toda la información muestral) de la muestra $s \cup s'$ de tamaño efectivo fijo $n + n'$, que es también una muestra irrestricta aleatoria de la población finita.

**Ejercicio 3.19.** Demostrar que un estimador insesgado de la media poblacional $\bar{y}$, es el estimador $t_c$ definido sobre la muestra de tamaño efectivo fijo $s \subset U$ de la población finita de tamaño $N$:

$$t_c = \bar{y}_s + c, \text{ si } 1 \in s \text{ y } N \notin s$$

$$t_c = \bar{y}_s - c, \text{ si } 1 \notin s \text{ y } N \in s$$

$$t_c = \bar{y}_s, \text{ en otro caso,}$$

siendo $\bar{y}_s$ la media muestral con diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n$, y $c$ una constante real.

**Solución.** Podemos clasificar el espacio muestral de muestras de tamaño efectivo fijo $n$, $S$, en cuatro sucesos disjuntos:

$$S = \{1 \in s, N \notin s\} \cup \{1 \notin s, N \in s\} \cup \{1, N \in s\} \cup \{1, N \notin s\}$$

Entonces, llamando a estos sucesos en este orden $S_1$, $S_2$, $S_3$ y $S_4$ respectivamente, tenemos de la esperanza como esperanza de esperanzas condicionadas, sustituyendo las probabilidades de los sucesos obtenidas por la regla de Laplace, y calculando las esperanzas condicionadas por los sucesos enumerados, que

$$E(t_c) = \sum_{i=1}^{4} p(S_i)E(t_c|S_i) =$$

$$\frac{\binom{N-2}{n-1}}{\binom{N}{n}}\left(\frac{y_1}{n} + \frac{n-1}{n}\bar{y}_{U-\{1,N\}} + c\right) +$$

$$\frac{\binom{N-2}{n-1}}{\binom{N}{n}}\left(\frac{y_N}{n} + \frac{n-1}{n}\bar{y}_{U-\{1,N\}} - c\right) +$$

$$\frac{\binom{N-2}{n-2}}{\binom{N}{n}}\left(\frac{y_1 + y_N}{n} + \frac{n-2}{n}\bar{y}_{U-\{1,N\}}\right) +$$

$$\frac{\binom{N-2}{n}}{\binom{N}{n}}\bar{y}_{U-\{1,N\}} =$$

$$\frac{(N-n)n}{N(N-1)}\left(\frac{y_1}{n} + \frac{n-1}{n}\bar{y}_{U-\{1,N\}} + c\right) +$$

$$\frac{(N-n)n}{N(N-1)}\left(\frac{y_N}{n} + \frac{n-1}{n}\bar{y}_{U-\{1,N\}} - c\right) +$$

$$\frac{n(n-1)}{N(N-1)}\left(\frac{y_1 + y_N}{n} + \frac{n-2}{n}\bar{y}_{U-\{1,N\}}\right) +$$

$$\frac{(N-n)(N-n-1)}{N(N-1)}\bar{y}_{U-\{1,N\}} =$$

$$(y_1 + y_N)\left[\frac{N-n}{N(N-1)} + \frac{n-1}{N(N-1)}\right] +$$

$$\bar{y}_{U-\{1,N\}}\left[\frac{2(N-n)(n-1)}{N(N-1)} + \frac{(n-1)(n-2)}{N(N-1)} + \frac{(N-n)(N-n-1)}{N(N-1)}\right] =$$

$$\frac{y_1 + y_N}{N} + \frac{\sum_{i=2}^{N-1} y_i}{N-2}\left[\frac{N^2 - 3N + 2}{N(N-1)}\right] =$$

$$\frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}.$$

Por lo tanto el estimador $t_c$ es insesgado para estimar la media poblacional.

**Ejercicio 3.20.** Comprobar que el coste esperado de seleccionar una muestra aleatoria simple con reemplazamiento de tamaño fijo $n \geq 2$ es menor que el coste esperado de una muestra de tamaño fijo $n$ obtenida por muestreo aleatorio simple sin reemplazamiento.

**Solución.** Si suponemos que el coste por unidad observada es $c > 0$, el coste esperado de una muestra aleatoria simple con reemplazamiento es

$$cE[\nu(\boldsymbol{s})] < cn,$$

ya que el tamaño muestral efectivo, o número de unidades distintas en la muestra aleatoria simple con reemplazamiento $\boldsymbol{s}$ de tamaño fijo $n \geq 2$, es $1 \leq \nu(\boldsymbol{s}) \leq n$, y con $\nu(\boldsymbol{s}) = 1, 2, \ldots, \nu, \ldots, n$, el tamaño muestral efectivo de la muestra $\boldsymbol{s}$, siendo

$$p(\nu) > 0$$

para todo $\nu = 1, 2, \dots, n$.

Por otro lado, tenemos que en el muestreo aleatorio simple sin reemplazamiento de tamaño fijo $n$, el tamaño efectivo fijo de cada muestra con probabilidad positiva es $\nu(s) = n$, constante que al multiplicarla por el coste por unidad nos da el "coste esperado de una muestra aleatoria simple sin reemplazamiento", es decir

$$cn.$$

**Ejercicio 3.21.** Si en el muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$, tomamos como estimador de la media poblacional a la media muestral de las unidades distintas que aparecen en la secuencia muestral, probar que este estimador es menos preciso que la media muestral de las $n$ observaciones en el muestreo aleatorio simple sin reemplazamiento. ¿Qué se puede decir comparando la precisión con la media muestral de las $n$ observaciones en el muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$?

**Solución.** Para ello basta ver que ambos son insesgados para la media poblacional. El caso del muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$ es materia de teoría básica. El otro caso puede verse de este modo,

$$E(\bar{y}_\nu) = EE(\bar{y}_\nu|\nu)$$

Donde $E(\bar{y}_\nu|\nu)$ es la esperanza de la media muestral en muestreo aleatorio simple sin reemplazamiento de tamaño fijo $\nu$, que coincide con la media poblacional. El promedio de medias poblacionales constantes es la media poblacional.

Aplicando el Teorema de Madow, la varianza de la media muestral $\bar{y}_\nu$ es

$$V(\bar{y}_\nu) = EV(\bar{y}_\nu|\nu) = E\left[\frac{N-\nu}{(N-1)\nu}\sigma^2\right] \geq$$

$$\frac{N-n}{(N-1)n}\sigma^2 = V(\bar{y}_n)$$

siendo esta última varianza de la media muestral con diseño de muestreo aleatorio simple sin reemplazamiento de tamaño fijo y efectivo $n$.

Pero también, vamos a ver si es posible probar que

$$V(\bar{y}_\nu) = E\left[\frac{N-\nu}{(N-1)\nu}\sigma^2\right] \leq \frac{\sigma^2}{n} = V(\bar{y}_n)$$

siendo esta última varianza de la media muestral con diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$. Para verlo definimos la función

$$f(\nu) = \frac{1}{n} - \frac{N-\nu}{(N-1)\nu}$$

y comprobar que para $1 \leq \nu \leq n$, $f(\nu) \geq 0$. En efecto,

$$f'(\nu) = \frac{(N-1)\nu - (N-\nu)}{(N-1)^2\nu^2} = \frac{N(\nu-1)}{(N-1)^2\nu^2} \geq 0$$

Por lo que la función $f$ es creciente y la derivada toma valor 0 cuando $\nu = 1$, es decir cuando alcanza el mínimo de $f$. Como

$$f(1) = \frac{1}{n} - 1 \leq 0$$

y

$$f(n) = \frac{1}{n} - \frac{N-n}{(N-1)n} > 0$$

En conclusión, el estimador es más preciso o no que la media aritmética de las $n$ observaciones por muestreo aleatorio simple con reemplazamiento dependiendo del signo positivo o negativo del promedio (de valores positivos y negativos)

$$\sum_{v=1}^{n} f(v)p(v).$$

**Ejercicio 3.22.** Indicar si puede seleccionarse una muestra irrestricta aleatoria de tamaño $5$ de una población finita de tamaño $326$, con un generador de números aleatorios independientes y cada dígito con distribución uniforme en $\{0, 1, 2, \dots, 9\}$.

**Solución.** Sí se puede, bastaría con que numerásemos las unidades de la población finita desde el número $1$ al $326$. Seguidamente seleccionaríamos grupos de tres dígitos aleatorios sucesivos, si el primer grupo está entre $001$ y $326$, la unidad seleccionada es la identificada con ese grupo; si no estuviese entre esas cantidades, se procedería a una nueva selección de tres dígitos aleatorios sucesivos hasta que se seleccionara un identificador de una unidad de la población finita. En la segunda unidad a seleccionar procedemos similarmente, con la particularidad de que si el identificador ya hubiese sido seleccionado en la primera unidad de la muestra, repetiríamos el proceso hasta que fuese un identificador distinto al anterior. Y así sucesivamente hasta seleccionar el quinto grupo de tres dígitos comprendidos entre $001$ y $326$ que no coincidan con el identificador de las unidades ya anteriormente seleccionadas en la muestra irrestricta aleatoria.

**Ejercicio 3.23.** Proponer un estimador insesgado del producto de dos medias poblacionales en muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n$.

**Solución.** De la relación

$$\bar{Y}\bar{X} = E(\bar{y}\bar{x}) - Cov(\bar{y}, \bar{x}) = E(\bar{y}\bar{x}) - \frac{N-n}{Nn} S_{y,x}$$

donde $S_{y,x}$ es la cuasicovarianza poblacional, que es estimable insesgadamente por la cuasicovarianza muestral, resulta como estimador insesgado de $\bar{Y}\bar{X}$ el siguiente

$$\bar{y}\bar{x} - \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}).$$

Donde ahora $y_i$ y $x_i$ son los valores de las variables $y$ y $x$ respectivamente en la $i$-ésima unidad de la muestra aleatoria simple sin reemplazamiento de tamaño efectivo fijo $n$.

**Ejercicio 3.24.** Obtener el tamaño muestral $n$ del muestreo aleatorio simple (diseño muestral *mas*) en función del tamaño muestral efectivo $m$ del muestreo aleatorio simple sin reemplazamiento (diseño muestral *mia*) que conduce a una misma varianza del estimador media muestral, y recíprocamente, es decir, $m$ en función de $n$ en las mismas condiciones.

**Solución.** Igualando las varianzas del estimador media muestral con ambos diseños muestrales tenemos que

$$\frac{\sigma^2}{n} = \frac{N-m}{(N-1)m} \sigma^2.$$

Simplificando el factor común $\sigma^2$ y despejando, tenemos que

$$n = \frac{(N-1)m}{N-m} = \frac{N-1}{\frac{N}{m}-1} = \frac{1-\frac{1}{N}}{\frac{1}{m}-\frac{1}{N}}.$$

Obviamente, si $m \to N$, $n \to \infty$.

Recíprocamente, de las igualdades anteriores, concretamente del primer y segundo términos, tenemos que

$$n(N-m) = (N-1)m.$$

De donde despejando $m$ tenemos que

$$nN = m(N-1+n).$$

O bien,

$$m = \frac{nN}{N-1+n}.$$

Obviamente, si $n \to \infty$, $m \to N$.

# Capítulo 4

# Muestreo estratificado

Este tipo de muestreo se presenta cuando la población finita se clasifica en clases o estratos, estimando las funciones paramétricas poblacionales a partir de las estimaciones obtenidas en los estratos.

## 4.1 Diseño estratificado

Si la población finita de tamaño $N$ se clasifica en $L$ estratos o clases de modo que si el tamaño del estrato $h$ $(h = 1, 2, \dots, L)$ es $N_h$, tendremos

$$\sum_{h=1}^{L} N_h = N.$$

El tamaño relativo del estrato $h$-ésimo es $W_h = N_h/N$, de modo que

$$\sum_{h=1}^{L} W_h = 1.$$

Dicha notación viene de “weight” en inglés, que significa “peso”.

Una muestra estratificada se obtiene al seleccionar aleatoriamente $n_h$ unidades en el estrato $h$, con $1 \leq n_h \leq N_h$ si el diseño usado es el *mia* en el estrato $h$-ésimo. En general, $1 \leq n_h$

con diseño *mas* en el estrato $h$-ésimo. Además suponemos que la selección dentro de cada estrato es independiente del resto de estratos, es decir no hay ninguna dependencia entre las unidades seleccionadas en uno y otro estratos cualesquiera. El tamaño de la muestra estratificada es

$$n = \sum_{h=1}^{L} n_h.$$

Si $y$ es la variable de interés o de estudio, notaremos por $y_{hk}$ al valor de la variable de interés en la unidad $k$ del estrato $h$. Entonces la "media del estrato $h$" es

$$\bar{y}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} y_{hk},$$

la "varianza del estrato $h$" es

$$\sigma_h^2 = \frac{1}{N_h} \sum_{k=1}^{N_h} (y_{hk} - \bar{y}_h)^2,$$

el "total del estrato $h$" es

$$T_h = N\bar{y}_h = \sum_{k=1}^{N_h} y_{hk},$$

y la "cuasivarianza del estrato $h$" es

$$S_h^2 = \frac{N_h \sigma_h^2}{N_h - 1}.$$

La media poblacional admite esta nueva expresión en el caso de una población estratificada,

$$\bar{y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{k=1}^{N_h} y_{hk} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h.$$

La varianza poblacional es

$$\sigma^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{k=1}^{N_h} (y_{hk} - \bar{y})^2 = \frac{N-1}{N} S^2,$$

siendo $S^2$ la cuasivarianza poblacional. Finalmente, el total poblacional es

$$T = N\bar{y} = \sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} T_h.$$

La varianza poblacional admite la siguiente descomposición

$$\sigma^2 = \sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h (\bar{y}_h - \bar{y})^2 .$$

Se suele decir que la variabilidad total se descompone en la variabilidad dentro de estratos más la variabilidad entre estratos. En efecto, restando y sumando $\bar{y}_h$ dentro del paréntesis,

$$\sigma^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{k=1}^{N_h} [(y_{hk} - \bar{y}_h) + (\bar{y}_h - \bar{y})]^2 =$$

$$\frac{1}{N}\sum_{h=1}^{L}\sum_{k=1}^{N_h} (y_{hk} - \bar{y}_h)^2 + \frac{1}{N}\sum_{h=1}^{L} N_h (\bar{y}_h - \bar{y})^2 =$$

$$\frac{1}{N}\sum_{h=1}^{L} N_h \sigma_h^2 + \frac{1}{N}\sum_{h=1}^{L} N_h (\bar{y}_h - \bar{y})^2,$$

que es lo que queríamos demostrar. El doble producto se ha anulado porque

$$2\sum_{h=1}^{L}\sum_{k=1}^{N_h}(y_{hk}-\bar{y}_h)(\bar{y}_h-\bar{y})=$$

$$2\sum_{h=1}^{L}(\bar{y}_h-\bar{y})\sum_{k=1}^{N_h}(y_{hk}-\bar{y}_h)=$$

$$2\sum_{h=1}^{L}(\bar{y}_h-\bar{y})(N_h\bar{y}_h-N_h\bar{y}_h)=0.$$

## 4.2 Estimación de la media poblacional

Denotando por $\bar{y}_{s(h)}$ a la media muestral en el estrato $h$, un estimador insesgado de la media poblacional es

$$\bar{y}_{st}=\sum_{h=1}^{L}W_h\bar{y}_{s(h)},$$

donde el subíndice $st$ de $\bar{y}_{st}$ viene de "stratified" que significa "estratificado" en inglés. En efecto, si el diseño *mas* se aplica en cada estrato independientemente

$$E(\bar{y}_{st})=E\left[\sum_{h=1}^{L}W_h\bar{y}_{s(h)}\right]=\sum_{h=1}^{L}W_hE\left[\bar{y}_{s(h)}\right]=\sum_{h=1}^{L}W_h\bar{y}_h=\bar{y}\,.$$

La varianza de este estimador es

$$V(\bar{y}_{st})=\sum_{h=1}^{L}V\left[W_h\bar{y}_{s(h)}\right]=\sum_{h=1}^{L}W_h^2V\left[\bar{y}_{s(h)}\right]=\sum_{h=1}^{L}W_h^2\frac{\sigma_h^2}{n_h},$$

siendo $n_h$ el tamaño muestral en el estrato $h$ $(h = 1, 2, \ldots, L)$ con diseño *mas*. La generalización para diseño *mia* sería directa. Un estimador insesgado de la varianza ya calculada es

$$\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h},$$

siendo $s_h^2$ la cuasivarianza muestral en el estrato $h$, es decir

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left[ y_{hk_i} - \bar{y}_{s(h)} \right]^2$$

y la media muestral en el estrato $h$ es

$$\bar{y}_{s(h)} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hk_i}.$$

## 4.3 Estimación del total poblacional

El estimador usual, en muestreo estratificado con diseño *mas* en cada estrato, del total poblacional $T = N\bar{y}$ es $\hat{T} = N\bar{y}_{st}$ pues

$$E\left(\hat{T}\right) = NE(\bar{y}_{st}) = N\bar{y} = T.$$

La varianza de este estimador es

$$V\left(\hat{T}\right) = N^2 V(\bar{y}_{st}) = N^2 \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} = \sum_{h=1}^{L} N_h^2 \frac{\sigma_h^2}{n_h},$$

que puede estimarse insesgadamente por

$$\hat{V}\left(\hat{T}\right) = \sum_{h=1}^{L} N_h^2 \frac{s_h^2}{n_h},$$

ya que $E(s_h^2) = \sigma_h^2$ con diseño *mas* de tamaño muestral $n_h$. La generalización para diseños *mia* en cada estrato es similar.

## 4.4 Estimación de la proporción poblacional

Como hemos visto, la proporción poblacional y la proporción muestral es una media aritmética de una variable de interés que toma valores cero o uno, y por ello el estimador insesgado de la proporción poblacional $P$ es, en muestreo estratificado con diseño *mas* independiente en cada estrato, el estimador

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{P}_h,$$

donde $\hat{P}_h$ es la proporción muestral en el estrato $h$. Como caso particular de estimador estratificado, tenemos

$$V(\hat{P}) = \sum_{h=1}^{L} W_h^2 \frac{P_h Q_h}{n_h}.$$

Un estimador insesgado de la varianza $V(\hat{P})$ es

$$\hat{V}(\hat{P}) = \sum_{h=1}^{L} W_h^2 \frac{\hat{P}_h \hat{Q}_h}{n_h - 1},$$

pues $s_h^2 = n_h \hat{P}_h \hat{Q}_h / (n_h - 1)$ es la cuasivarianza muestral en el estrato $h$.

La generalización para diseños *mia* en cada estrato es similar.

## 4.5 El problema de asignación muestral

Dado el tamaño muestral $n$, se denomina asignación muestral al reparto de las $n$ selecciones de la muestra en los $L$ estratos, es decir consiste en fijar los tamaños muestrales $n_h$ $(h = 1, 2, \dots, L)$ en cada estrato, de modo que

$$n = \sum_{h=1}^{L} n_h.$$

Algunos tipos de asignación muestral son los siguientes.

**Asignación igual**. Consiste en asignar el mismo tamaño muestral en cada estrato, es decir que $n_1 = \cdots = n_h = \cdots = n_L$. Como

$$n = \sum_{h=1}^{L} n_h = Ln_h,$$

deducimos que la asignación igual es $n_h = n/L$ $(h = 1, 2, \dots, L)$.

**Asignación proporcional**. Consiste en asignar a cada estrato $h$, un tamaño muestral $n_h$ proporcional al tamaño del estrato $N_h$. Entonces, $n_h \propto N_h$ o bien $n_h = cN_h$ donde $c$ es la constante de proporcionalidad, por lo que

$$n = \sum_{h=1}^{L} n_h = \sum_{h=1}^{L} cN_h = cN,$$

de donde $c = n/N$, y por tanto $n_h = nW_h$ $(h = 1, 2, \dots, L)$. La varianza del estimador estratificado de la media poblacional con asignación proporcional es

$$V(\text{prop}, \bar{y}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2 \leq \frac{\sigma^2}{n} = V(mas, \bar{y}_s),$$

debido a la descomposición de la varianza poblacional en variación dentro y entre estratos. Es decir, la asignación proporcional con diseño *mas* dentro de cada estrato proporciona un estimador estratificado más preciso que la media muestral con diseño *mas*.

**Asignación mínima**. Fijado el tamaño muestral $n$, la asignación mínima consiste en asignar a cada estrato un tamaño muestral $n_h$ de modo que la varianza $V(\bar{y}_{st})$ sea mínima. Para calcular los tamaños muestrales utilizamos el método de los multiplicadores de Lagrange con la restricción

$$\sum_{h=1}^{L} n_h = n.$$

El lagrangiano es $L^*$,

$$L^* = V(\bar{y}_{st}) + \lambda\left(\sum_{h=1}^{L} n_h - n\right) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} + \lambda\left(\sum_{h=1}^{L} n_h - n\right),$$

donde $\lambda$ es el multiplicador de Lagrange. Resolviendo,

$$\frac{\partial L^*}{\partial n_h} = -W_h^2 \frac{\sigma_h^2}{n_h^2} + \lambda = 0 \;\; (h = 1, 2, \ldots, L)$$

de donde

$$\sqrt{\lambda} = \frac{W_h \sigma_h}{n_h} = \frac{\sum_{h=1}^{L} W_h \sigma_h}{n},$$

y por esto

$$n_h = n\frac{W_h\sigma_h}{\sum_{h=1}^{L} W_h\sigma_h} \quad (h = 1, 2, \dots, L)$$

es la asignación mínima que consiste en asignar un tamaño muestral en el estrato $h$ proporcional al producto $W_h\sigma_h$ o equivalentemente al producto $N_h\sigma_h$.

La varianza del estimador estratificado de la media poblacional con asignación mínima es ahora, sustituyendo los tamaños muestrales en la fórmula de la varianza

$$V(\text{mín}, \bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h\sigma_h\right)^2 \leq \frac{1}{n}\sum_{h=1}^{L} W_h\sigma_h^2 = V(\text{prop}, \bar{y}_{st}),$$

es decir, con la asignación mínima se mejora la precisión del estimador estratificado de la media poblacional respecto de la asignación proporcional.

Esta misma asignación muestral se hubiera obtenido si minimizáramos el tamaño muestral total

$$n = \sum_{h=1}^{L} n_h$$

sujeto a una varianza prefijada $V = V(\bar{y}_{st})$ como restricción.

**Asignación óptima con costes variables**. Admitiendo que el coste de observación de una unidad muestral del estrato $h$ es $C_h$, y el coste total de la muestra es $C$, podemos minimizar la varianza $V(\bar{y}_{st})$ sujeta a la restricción

$$C = \sum_{h=1}^{L} C_h n_h.$$

En este coste se supone que el coste total depende del número de selecciones de unidades en cada estrato y no del tamaño efectivo de las muestras en los estratos. El mismo resultado se da cuando se minimiza el coste $C$ sujeto a una varianza prefijada $V = V(\bar{y}_{st})$. El lagrangiano será ahora

$$L^* = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} + \lambda \left( \sum_{h=1}^{L} C_h n_h - C \right),$$

y derivando parcialmente $L^*$ respecto a $n_h$ e igualando a cero,

$$\frac{\partial L^*}{\partial n_h} = -\frac{W_h^2 \sigma_h^2}{n_h^2} + \lambda C_h = 0 \ \ (h = 1, 2, \dots, L)$$

de donde

$$\sqrt{\lambda} = \frac{\frac{W_h \sigma_h}{\sqrt{C_h}}}{n_h} = \frac{\sum_{h=1}^{L} \frac{W_h \sigma_h}{\sqrt{C_h}}}{n},$$

y por tanto

$$n_h = n \frac{\frac{W_h \sigma_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{W_h \sigma_h}{\sqrt{C_h}}} \ \ (h = 1, 2, \dots, L),$$

es decir $n_h$ es proporcional a $W_h \sigma_h / \sqrt{C_h}$. En esta asignación se encuentra una solución de compromiso entre el coste y la precisión. El tamaño muestral en el estrato $h$, $n_h$, puede expresarse en función del coste prefijado $C$. En efecto,

$$C = \sum_{h=1}^{L} C_h n_h = n \frac{\sum_{h=1}^{L} W_h \sigma_h \sqrt{C_h}}{\sum_{h=1}^{L} \frac{W_h \sigma_h}{\sqrt{C_h}}},$$

luego

$$n = C\frac{\sum_{h=1}^{L}\frac{W_h\sigma_h}{\sqrt{C_h}}}{\sum_{h=1}^{L}W_h\sigma_h\sqrt{C_h}},$$

de donde sustituyendo, tenemos finalmente

$$n_h = n\frac{\frac{W_h\sigma_h}{\sqrt{C_h}}}{\sum_{h=1}^{L}\frac{W_h\sigma_h}{\sqrt{C_h}}} = C\frac{\frac{W_h\sigma_h}{\sqrt{C_h}}}{\sum_{h=1}^{L}W_h\sigma_h\sqrt{C_h}} \quad (h = 1, 2, \dots, L).$$

La varianza del estimador estratificado de la media poblacional con asignación óptima con costes variables es

$$V(\text{ópt}, \bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L}W_h\sigma_h\sqrt{C_h}\right)\left(\sum_{h=1}^{L}\frac{W_h\sigma_h}{\sqrt{C_h}}\right) =$$

$$\frac{1}{C}\left(\sum_{h=1}^{L}W_h\sigma_h\sqrt{C_h}\right)^2.$$

**Asignación fijada**. Si los tamaños muestrales $n_h$ $(h = 1, 2, \dots, L)$ están prefijados previamente, el tamaño muestral $n$ está también prefijado previamente. Entonces la varianza del estimador de la media poblacional en muestreo estratificado con asignación fijada es la tradicional recogida en este libro.

**Asignación valoral**. Dado el tamaño muestral total $n$, la asignación valoral consiste en distribuir el tamaño muestral $n_h$ en el estrato $h$ de modo que $n_h$ sea proporcional al total del estrato $h$, $N_h\bar{y}_h$. Es decir,

$$\frac{n_1}{N_1\bar{y}_1} = \cdots = \frac{n_h}{N_h\bar{y}_h} = \cdots = \frac{n_L}{N_L\bar{y}_L} = \frac{n}{\sum_{h=1}^{L} N_h\bar{y}_h} = \frac{n}{N\bar{y}},$$

de donde

$$n_h = n\frac{W_h\bar{y}_h}{\bar{y}} \quad (h = 1, 2, \ldots, L).$$

**Asignación** $n_h \propto W_h\sigma_h^2$. Este tipo de asignación produce la misma precisión que la asignación proporcional, y por tanto es más precisa que la media muestral en el diseño *mas*. En efecto, si

$$n_h = n\frac{W_h\sigma_h^2}{\sum_{h=1}^{L} W_h\sigma_h^2} \quad (h = 1, 2, .., L),$$

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h\right)\left(\sum_{h=1}^{L} W_h\sigma_h^2\right) = V(\text{prop}, \bar{y}_{st}),$$

pero esta asignación no tiene utilidad práctica porque requiere el conocimiento adicional de los valores $\sigma_h^2$ además de los de $W_h$, y con éstos últimos ya se consigue la misma precisión con asignación proporcional.

**Asignación especial**. Consiste en realizar un censo en el estrato 2 o de unidades grandes, es decir el tamaño muestral efectivo en el estrato 2 es $n_2 = N_2$, y tomar una muestra de tamaño efectivo fijo $n_1$ con diseño *mas* en el primer estrato de los dos en que se divide la población por un punto $y = \bar{y} + k\sigma$ separador del primero del segundo estrato según los valores que tome la variable de interés en cada unidad. La varianza del estimador es

$$V\left[W_1\bar{y}_{s(1)} + W_2\bar{y}_2\right] = W_1^2\,\frac{N_1 - n_1}{N_1 - 1}\,\frac{\sigma_1^2}{n_1}.$$

Esta varianza es menor que la del muestreo aleatorio simple sin reemplazamiento y la media muestral, cuando y solo cuando

$$V\left(\text{esp}, W_1\bar{y}_{s(1)} + W_2\bar{y}_2\ \right) < \frac{N - n}{N - 1}\,\frac{\sigma^2}{n},$$

o bien

$$N^2(N - N_2 - 1)(n - N_2)\sigma^2 > (N - N_2)^2(N - 1)n\sigma_1^2,$$

pero como

$$\sigma^2 = \frac{N_1}{N}\sigma_1^2 + \frac{N_2}{N}\sigma_2^2 + \frac{N_2}{N_1}(\bar{y}_2 - \bar{y})^2,$$

tenemos

$$(N - N_2)\sigma_1^2 = N\sigma^2 - N_2\sigma_2^2 - \frac{N_2N(\bar{y}_2 - \bar{y})^2}{N - N_2} \leq$$

$$N\sigma^2 - \frac{N_2N(k\sigma)^2}{N - N_2},$$

pues $\bar{y}_2 \geq \bar{y} + k\sigma = y$, y $\sigma_2^2 \geq 0$. De las dos desigualdades últimas, obligando a que

$$N^2(N - N_2 - 1)(n - N_2)\sigma^2 >$$

$$(N - N_2)(N - 1)n\left[N\sigma^2 - \frac{N_2N(k\sigma)^2}{N - N_2}\right]$$

conseguimos una condición suficiente para que el estimador estratificado con asignación especial sea más preciso que la media muestral con diseño *mia* de igual tamaño muestral efectivo $n$, con lo que operando resulta

$$k^2 > \frac{(N-1)(N-N_2)n - N(N-N_2-1)(n-N_2)}{N_2(N-1)n},$$

o equivalentemente al simplificar

$$k^2 > \frac{N}{n} - \frac{NN_2 - n}{n(N-1)},$$

y como $N_2 \leq n$ llegamos a que

$$k^2 > \frac{N}{n} - 1,$$

y al haber supuesto que $k > 0$ queda

$$k > \sqrt{\frac{N}{n} - 1}.$$

Por lo que una cota inferior para que el punto $y$ de estratificación para afijación especial y dos estratos (uno primero de unidades pequeñas que se muestrea, y otro segundo de unidades grandes que se incluyen todas en la muestra a modo de censo en este estrato) proporcione estimaciones más precisas que la media muestral usando diseño *mia* en ambos estimadores, es que el punto de estratificación sea

$$y > \bar{y} + \sigma \times \sqrt{\frac{N}{n} - 1}.$$

Si la variable de interés de las unidades están ordenadas en orden creciente, Glasser (1962) dio la condición suficiente de que el punto de estratificación verifique

$$y > \bar{y} + \sigma \times \sqrt{\frac{N}{n}}$$

para que con diseño *mia* en el primer estrato y asignación especial, se mejore la precisión respecto del diseño *mia* sobre toda la población e igual tamaño muestral efectivo total en ambos casos. Ruiz Espejo (1985) dio la condición suficiente mejorada para este propósito que es compatible con la cota de Glasser.

## 4.6 Estimación de la varianza poblacional

Si el diseño muestral empleado dentro de cada estrato es el *mas*, vamos a obtener un estimador insesgado de la varianza poblacional en el muestreo estratificado. Partimos de la descomposición de la varianza en variación dentro de estratos y entre estratos,

$$\sigma^2 = \sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h (\bar{y}_h - \bar{y})^2 .$$

Para estimar $\sigma^2$, el primer sumando no presenta ningún problema con diseño *mas* en cada estrato puesto que la cuasivarianza muestral $s_h^2$ es un estimador insesgado de $\sigma_h^2$, la varianza del mismo estrato. En cuanto al segundo sumando, sustituyamos $\bar{y}_h$ e $\bar{y}$ por sus estimadores insesgados $\bar{y}_{s(h)}$ e $\bar{y}_{st}$, y calculemos la esperanza matemática:

$$E\left\{\sum_{h=1}^{L} W_h \left[\bar{y}_{s(h)} - \bar{y}_{st}\right]^2\right\} =$$

$$E\left(\sum_{h=1}^{L} W_h \left\{(\bar{y}_h - \bar{y}) + \left[\bar{y}_{s(h)} - \bar{y}_h\right] - (\bar{y}_{st} - \bar{y})\right\}^2\right),$$

sumando y restando $(\bar{y}_h - \bar{y})$. Desarrollando el cuadrado entre llaves y sustituyendo, tenemos que la esperanza anterior es igual a

$$E\left[\sum_{h=1}^{L} W_h(\bar{y}_h - \bar{y})^2\right] + E\left\{\sum_{h=1}^{L} W_h\left[\bar{y}_{s(h)} - \bar{y}_h\right]^2\right\} +$$

$$E\left[\sum_{h=1}^{L} W_h(\bar{y}_{st} - \bar{y})^2\right] - 2E\left\{\sum_{h=1}^{L} W_h\left[\bar{y}_{s(h)} - \bar{y}_h\right](\bar{y}_{st} - \bar{y})\right\} +$$

$$2E\left\{\sum_{h=1}^{L} W_h(\bar{y}_h - \bar{y})\left[\bar{y}_{s(h)} - \bar{y}_h\right]\right\} -$$

$$2E\left[\sum_{h=1}^{L} W_h\,(\bar{y}_h - \bar{y})(\bar{y}_{st} - \bar{y})\right] =$$

$$\sum_{h=1}^{L} W_h(\bar{y}_h - \bar{y})^2 + \sum_{h=1}^{L} W_h \frac{\sigma_h^2}{n_h} + \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} - 2\sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h},$$

porque los dos últimos sumandos se anulan; luego el resultado final es

$$\sum_{h=1}^{L} W_h(\bar{y}_h - \bar{y})^2 + \sum_{h=1}^{L} W_h(1 - W_h)\frac{\sigma_h^2}{n_h},$$

por lo que el segundo sumando de ésta última igualdad es el sesgo de

$$\sum_{h=1}^{L} W_h\left[\bar{y}_{s(h)} - \bar{y}_{st}\right]^2$$

como estimador de

$$\sum_{h=1}^{L} W_h(\bar{y}_h - \bar{y})^2,$$

y en consecuencia, el estimador insesgado de la varianza poblacional $\sigma^2$ en muestreo estratificado con diseño *mas* en cada estrato independientemente, es

$$\widehat{\sigma^2} = \sum_{h=1}^{L} W_h s_h^2 + \sum_{h=1}^{L} W_h \left[\bar{y}_{s(h)} - \bar{y}_{st}\right]^2 - \sum_{h=1}^{L} W_h (1 - W_h) \frac{s_h^2}{n_h}.$$

Una fórmula más compleja se puede dar para estimar la varianza poblacional en muestreo estratificado con diseño *mia* independientemente dentro de cada estrato (Mirás, 1985). Otros estimadores insesgados de la varianza en muestreo estratificado se deben a Ruiz Espejo y Delgado Pineda (2008c).

## 4.7 Posestratificación

A veces se utiliza un diseño no estratificado para seleccionar la muestra, pero una vez seleccionada se decide estratificarla y estimar la media poblacional $\bar{y}$ por una media posestratificada. De este modo, el tamaño muestral en el estrato $h$, $n_h$, es aleatorio antes de seleccionar la muestra, y fijo una vez seleccionada. El estimador posestratificado $\bar{y}_{ps}$ es entonces

$$\bar{y}_{ps} = \sum_{h=1}^{L} W_h \bar{y}_{s(h)},$$

similar al estimador estratificado salvo que la media muestral $\bar{y}_{s(h)}$ tiene ahora un número aleatorio $n_h$ de unidades seleccionadas en el estrato $h$. Ya no son los valores $n_h$ fijados previamente sino que son valores aleatorios que se concretan con la muestra obtenida. Para calcular la esperanza y la varianza del estimador uso de la esperanza y varianza condicionadas al tamaño muestral aleatorio $n_h$. En efecto,

$$E(\bar{y}_{ps}) = E\left[E(\bar{y}_{ps}|n_h)\right] = E\left\{\sum_{h=1}^{L} W_h E\left[\bar{y}_{s(h)}|n_h\right]\right\} =$$

$$E\left(\sum_{h=1}^{L} W_h \bar{y}_h\right) = E(\bar{y}) = \bar{y},$$

por lo que el estimador $\bar{y}_{ps}$ es insesgado para la media poblacional. Su varianza se calcula así,

$$V(\bar{y}_{ps}) = V\left[E(\bar{y}_{ps}|n_h)\right] + E\left[V(\bar{y}_{ps}|n_h)\right],$$

pero el primer sumando es cero porque $E(\bar{y}_{ps}|n_h) = \bar{y}$, y entonces

$$V(\bar{y}_{ps}) = E\left(\sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h}\right) = \sum_{h=1}^{L} W_h^2 \sigma_h^2 E\left(\frac{1}{n_h}\right).$$

Para calcular esta esperanza anterior, sea $\widehat{W}_h = n_h/n$ el tamaño relativo de la muestra en el estrato $h$, mientras que $W_h = N_h/N$ es el tamaño relativo del estrato $h$ en la población. Con diseño *mas*, $\widehat{W}_h$ estima sin sesgo a $W_h$. Podemos escribir

$$n_h = n\widehat{W}_h = n(\widehat{W}_h - W_h + W_h) = nW_h\left(1 + \frac{\widehat{W}_h - W_h}{W_h}\right),$$

luego

$$\frac{1}{n_h} = \frac{1}{nW_h}\frac{1}{1 + \frac{\widehat{W}_h - W_h}{W_h}},$$

pero si

$$\left|\frac{\widehat{W}_h - W_h}{W_h}\right| < 1,$$

algo razonable porque $\widehat{W}_h$ estima sin sesgo a $W_h$ y además converge en probabilidad a dicha función paramétrica, nos permite expresar la aproximación del desarrollo en serie de potencias siguiente

$$\frac{1}{n_h} = \frac{1}{nW_h}\left[1 - \frac{\widehat{W}_h - W_h}{W_h} + \frac{\left(\widehat{W}_h - W_h\right)^2}{W_h^2} - \cdots\right] \approx$$

$$\frac{1}{nW_h}\left[1 - \frac{\widehat{W}_h - W_h}{W_h} + \frac{\left(\widehat{W}_h - W_h\right)^2}{W_h^2}\right]$$

donde hemos tenido en cuenta los tres primeros términos del desarrollo en serie. Tomando esperanzas,

$$E\left(\frac{1}{n_h}\right) \approx \frac{1}{nW_h}\left[1 - 0 + \frac{V\left(\widehat{W}_h\right)}{W_h^2}\right] = \frac{1}{nW_h}\left[1 + \frac{W_h(1 - W_h)}{nW_h^2}\right].$$

Sustituyendo esta relación en la varianza del estimador posestratificado, tenemos

$$V\left(\bar{y}_{ps}\right) = \sum_{h=1}^{L} W_h^2 \sigma_h^2 E\left(\frac{1}{n_h}\right) \approx \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2 \left(1 + \frac{1 - W_h}{nW_h}\right)$$

en términos de su desarrollo en serie de potencias de hasta $n^{-2}$.

Por último, veamos que este método de estimación se basa en que conocemos los tamaños relativos de los estratos $W_h$. En efecto, si utilizáramos el estimador

$$\bar{y}_{ps_0} = \sum_{h=1}^{L} W_{h_0} \bar{y}_{s(h)},$$

con $W_{h_0}$ cualquiera, tendremos que

$$\bar{y}_{ps} = \bar{y}_{ps_0} + \sum_{h=1}^{L} (W_h - W_{h_0})\bar{y}_{s(h)},$$

por lo que

$$B(\bar{y}_{ps_0};\, \bar{y}) = -\sum_{h=1}^{L} (W_h - W_{h_0})\bar{y}_h$$

es el sesgo del posible estimador posestratificado con $W_{h_0}$ cualquiera, que no depende del tamaño muestral $n$. Por lo tanto, no es aconsejable usar la estratificación a posteriori o posestratificación si los tamaños relativos de los estratos $W_h$ $(h = 1, 2, \dots, L)$ no son conocidos en la fase de estimación.

Un estimador insesgado de la varianza del estimador posestratificado puede obtenerse a partir del Teorema 1.4, en concreto, para muestreo aleatorio simple con reemplamiento de tamaño fijo $n$, si denotamos por $\bar{y}_{\boldsymbol{s}}$ a la media muestral y por $s^2$ a la cuasivarianza muestral

$$\hat{V}(\bar{y}_{ps}) = \bar{y}_{ps}^2 - \bar{y}_{\boldsymbol{s}}^2 + \frac{s^2}{n}.$$

## 4.8 Ejercicios resueltos

**Ejercicio 4.1.** En un estudio por muestreo estratificado se decide utilizar asignación especial para el tercer estrato (de unidades grandes) y utilizar asignación igual en los dos primeros estratos de $TF(n_1)$ y de $TEF(n_2)$ respectivamente donde los diseños que se emplean son *mas* con $\bar{y}_{\boldsymbol{s}(1)}$, y *mia* con $\bar{y}_{s(2)}$. Proponer un estimador

insesgado de la media poblacional para este diseño y calcular su varianza.

**Solución.** Un estimador estratificado insesgado de la media poblacional $\bar{y}$ que se propone es

$$\bar{y}_{st} = W_1\bar{y}_{s(1)} + W_2\bar{y}_{s(2)} + W_3\bar{y}_3,$$

siendo $W_h = N_h/N$ el tamaño relativo del estrato $h$ $(h = 1, 2, 3)$. La varianza es

$$V(\bar{y}_{st}) = W_1^2\frac{\sigma_1^2}{n_1} + W_2^2\frac{N_2 - n_2}{N_2 - 1}\frac{\sigma_2^2}{n_2},$$

con $n_1 = n_2 = (n - n_3)/2 = (n - N_3)/2$ , siendo $n_3 = N_3$ el tamaño muestral efectivo en el tercer estrato.

**Ejercicio 4.2.** En las condiciones del ejercicio anterior, proponer un estimador insesgado de la varianza del estimador de $\bar{y}$.

**Solución.** Un estimador insesgado de la varianza del estimador estratificado insesgado de la media poblacional es

$$\hat{V}(\bar{y}_{st}) = W_1^2\frac{s_1^2}{n_1} + W_2^2\frac{N_2 - n_2}{N_2}\frac{s_2^2}{n_2},$$

siendo $s_1^2$ la cuasivarianza muestral obtenida con diseño *mas* de tamaño fijo $n_1$ en el primer estrato, y $s_2^2$ la cuasivarianza muestral obtenida con diseño *mia* de tamaño efectivo fijo $n_2$ en el segundo estrato.

**Ejercicio 4.3.** En una población estratificada en 2 estratos, se ha obtenido que $W_1 = 0.4$, y por una muestra piloto se sabe que aproximadamente $\sigma_1^2 = 100$, $\sigma_2^2 = 81$ y $\sigma^2 = 225$. Suponiendo

que el tamaño poblacional $N$ es suficientemente grande, calcular el tamaño muestral $n$ para que una muestra con asignación mínima proporcione la misma varianza que un diseño *mas* sobre toda la población de tamaño muestral efectivo $n^* = 150$, para estimar la media poblacional.

**Solución.** Como

$$V(\bar{y}_s) = \frac{\sigma^2}{n^*} = \frac{225}{150} = 1.5,$$

y

$$V(\text{mín}, \bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{2} W_h \sigma_h\right)^2 = \frac{1}{n}(0.4 \cdot 10 + 0.6 \cdot 9)^2 = \frac{88.36}{n}.$$

Luego, si

$$1.5 = V(\bar{y}_s) = V(\text{mín}, \bar{y}_{st}) = \frac{88.36}{n},$$

se deduce que

$$n = \frac{88.36}{1.5} \approx 58.91,$$

es decir, el tamaño muestral requerido es $n = 59$.

**Ejercicio 4.4.** Determinar la asignación proporcional en cada estrato, si el tamaño muestral total es $n = 1000$, y hay 5 estratos de tamaños relativos $0.2, 0.3, 0.1, 0.25$ y $0.15$. ¿Cuál es la mayor diferencia absoluta de tamaños muestrales con respecto a la asignación igual?

**Solución.** La asignación proporcional es

$$n_h = nW_h = \begin{cases} 200 \text{ si } h = 1 \\ 300 \text{ si } h = 2 \\ 100 \text{ si } h = 3 \\ 250 \text{ si } h = 4 \\ 150 \text{ si } h = 5 \end{cases}$$

La asignación igual es $n_h = 1000/5 = 200$ $(h = 1, 2, \dots, 5)$. En los estratos 2 y 3 se dan las mayores diferencias absolutas entre ambas asignaciones pues $|300 - 200| = |100 - 200| = 100$, que es mayor que las desviaciones absolutas restantes, que son $0$, $50$ y $50$.

**Ejercicio 4.5.** Para estimar la proporción poblacional $P$ de inclinación de voto a cierto partido político en el conjunto de españoles con derecho a voto, se ha dividido geográficamente a los votantes en dos estratos: litoral y centro, de modo que el tamaño relativo de ambos es $W_1 = W_2$ de un total de $20$ millones de votantes. Se decide usar asignación igual $n_h = 5000$ $(h = 1, 2)$ y resultan, con diseño *mia* en cada estrato, las proporciones muestrales $\hat{P}_1 = 0.35$ y $\hat{P}_2 = 0.28$. Estimar $P$ por muestreo estratificado con asignación igual y estimar insesgadamente la varianza de tal estimador.

**Solución.** El estimador estratificado sin sesgo de la proporción poblacional $P$ es

$$\hat{P} = \sum_{h=1}^{2} W_h \hat{P}_h = (0.5 \cdot 0.35 + 0.5 \cdot 0.28) = 0.315,$$

luego la proporción estimada de voto favorable es del $31.5\%$.

Un estimador insesgado de su varianza es

$$\hat{V}(\hat{P}) = \sum_{h=1}^{2} W_h^2 \frac{N_h - n_h}{N_h} \frac{\hat{P}_h \hat{Q}_h}{n_h - 1} =$$

$$\frac{1}{4} \frac{9995000}{10000000} \frac{1}{4999} (0.35 \cdot 0.65 + 0.28 \cdot 0.72) \approx 0.0000214$$

que representa un error de muestreo estimado muy pequeño, por lo que el estimador $\hat{P}$ es muy preciso.

**Ejercicio 4.6.** En una comarca compuesta por tres pueblos numerados del 1 al 3, se desea conocer la edad media de sus habitantes. Para ello se dispone de un presupuesto de 10000 euros, y se tiene un costo por observación $C_1 = C_2 = 8$ y $C_3 = 12$ euros respectivamente por encuesta. Determinar los tamaños muestrales $n_h$ en cada pueblo, y el tamaño muestral total $n$, si de una encuesta piloto previa se ha estimado que las cuasivarianzas muestrales son $s_1^2 = 30^2$, $s_2^2 = 32^2$ y $s_3^2 = 40^2$, y que se dispone de la información del tamaño total de habitantes en cada pueblo $N_1 = 25000$, $N_2 = 12000$ y $N_3 = 2000$. El objetivo es obtener la máxima precisión a coste fijo.

**Solución.** La asignación $n_h$ debe ser proporcional a $N_h \sigma_h / \sqrt{C_h}$ $(h = 1, 2, 3)$ por tratarse de asignación óptima con costes variables. Por tanto,

$$n_1 \propto \frac{N_1 \sigma_1}{\sqrt{C_1}} = \frac{25000 \cdot 30}{\sqrt{8}} \approx 265165.04$$

$$n_2 \propto \frac{N_2 \sigma_2}{\sqrt{C_2}} = \frac{12000 \cdot 32}{\sqrt{8}} \approx 135764.5$$

$$n_3 \propto \frac{N_3\sigma_3}{\sqrt{C_3}} = \frac{2000 \cdot 40}{\sqrt{12}} \approx 23094.011$$

Por otro lado tenemos que el coste total es

$$10000 = C = \sum_{h=1}^{3} C_h n_h = 8n_1 + 8n_2 + 12n_3 = 3484564.5 \cdot t$$

siendo $t$ la constante de proporcionalidad, de donde

$$t = \frac{10000}{3484564.5} = 0.0028697991,$$

y por tanto,

$$n_1 = t \cdot 265165.04 = 760.97 \approx 761 \text{ habitantes}$$

$$n_2 = t \cdot 135764.5 = 389.60 \approx 390 \text{ habitantes}$$

$$n_3 = t \cdot 23094.011 = 66.27 \approx 66 \text{ habitantes}$$

Es decir, en total entre los tres pueblos de la comarca, se debe encuestar a un total de 1217 habitantes, 761 del primer pueblo, 390 del segundo, y 66 del tercero. Se puede comprobar que el coste total resultante será $761 \cdot 8 + 390 \cdot 8 + 66 \cdot 12 = 10000$.

**Ejercicio 4.7.** Una empresa de publicidad quiere estimar la proporción de hogares en un municipio donde se consume cierto producto. El municipio es dividido en tres estratos de tamaños 155, 62 y 93 hogares respectivamente. Una muestra estratificada de tamaño muestral total 40 hogares se selecciona con asignación proporcional. Estimar la proporción poblacional pedida y dar una estimación insesgada de su varianza, si haciendo uso de diseño *mia* independientemente en cada estrato, el número de hogares en las muestras que consumen el producto son 16 , 2 y 6 respectivamente.

**Solución.** El número total de hogares en el municipio es

$$N = \sum_{h=1}^{3} N_h = (155 + 62 + 93) = 310 \text{ hogares.}$$

La estimación estratificada de la proporción de hogares que consumen el producto es

$$\hat{P} = \sum_{h=1}^{3} \frac{N_h}{N} \hat{P}_h = \frac{155}{310}\frac{16}{20} + \frac{62}{310}\frac{2}{8} + \frac{93}{310}\frac{6}{12} = 0.60$$

Una estimación insesgada de la varianza de este estimador es

$$\hat{V}(\hat{P}) = \sum_{h=1}^{3} \frac{N_h^2}{N^2} \hat{V}(\hat{P}_h),$$

donde

$$\hat{V}(\hat{P}_h) = \frac{N_h - n_h}{N_h} \frac{\hat{P}_h \hat{Q}_h}{n_h - 1} \approx \begin{cases} 0.007 \text{ si } h = 1 \\ 0.024 \text{ si } h = 2 \\ 0.020 \text{ si } h = 3 \end{cases}$$

Es decir, sustituyendo

$$\hat{V}(\hat{P}) \approx 0.0045,$$

que es el valor aproximado del estimador insesgado de la varianza del estimador de la proporción $P$ en muestreo estratificado, es decir, una aproximación por redondeo de la estimación insesgada de la varianza buscada.

**Ejercicio 4.8.** Una población finita $U$ está clasificada en dos dominios o estratos disjuntos que denotamos por su subíndice $h = 1, 2$. Estimamos la diferencia de medias de los estratos, $D = \bar{y}_1 - \bar{y}_2$, por medio del estimador insesgado $d$ diferencia de medias muestrales respectivas obtenidas por diseños de muestreo aleatorio

simple con reemplazamiento independientes de tamaños fijos $n_1$ y $n_2$. Demostrar que el estimador $d$ es insesgado para estimar $D$, obtener su varianza, y minimizarla sujeta a que $n = n_1 + n_2$. Obtener también la asignación muestral óptima que minimiza la varianza de $d$ sujeta a que el coste total $C = c_1 n_1 + c_2 n_2$, así como la varianza de $d$ resultante. Proponer un estimador insesgado de las varianzas resultantes.

**Solución.** El estimador $d = \bar{y}_{\boldsymbol{s}(1)} - \bar{y}_{\boldsymbol{s}(2)}$ diferencia de medias muestrales independientes, es insesgado para estimar $D$ pues $E\left(\bar{y}_{\boldsymbol{s}(h)}\right) = \bar{y}_h$ y la esperanza matemática de una diferencia es la diferencia de las esperanzas, que en nuestro caso es $D = \bar{y}_1 - \bar{y}_2$. Y tiene por varianza

$$V(d) = V\left(\bar{y}_{\boldsymbol{s}(1)}\right) + V\left(\bar{y}_{\boldsymbol{s}(2)}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Para minimizar $V(d)$ sujeto a que $n = n_1 + n_2$, usamos el método de los multiplicadores de Lagrange. El lagrangiano es

$$\Lambda = V(d) + \lambda(n - n_1 - n_2),$$

donde $\lambda$ es el multiplicador de Lagrange, que es solo uno porque hay una sola restricción. Resolviendo, para $h = 1, 2$, tenemos

$$\frac{\partial \Lambda}{\partial n_h} = -\frac{\sigma_h^2}{2n_h^2} - \lambda = 0.$$

Por lo que

$$n_h = \frac{\sigma_h}{\sqrt{-2\lambda}} = n\frac{\sigma_h}{\sum_{h=1}^{2}\sigma_h},$$

despejando $\lambda$ al exigir la restricción sobre la suma de los tamaños muestrales. Al sustituir estos valores de la asignación muestral mínima, obtenemos la varianza mínima del estimador $d$ como

$$V_{\text{mín}}(d) = \frac{1}{n}(\sigma_1 + \sigma_2)^2.$$

De modo similar, la asignación muestral óptima que minimiza la varianza del estimador $d$ sujeto la restricción del coste presupuestado $C = c_1 n_1 + c_2 n_2$ , usando del método de los multiplicadores de Lagrange, nos da como resultado para $h = 1, 2$

$$n_h = C \frac{\sigma_h / \sqrt{c_h}}{\sum_{h=1}^{2} \sigma_h \sqrt{c_h}}.$$

Que sustituidas estas asignaciones muestrales óptimas en la fórmula de la varianza de $d$ nos da como resultado

$$V_{\text{ópt}}(d) = \frac{1}{C}\left(\sigma_1 \sqrt{c_1} + \sigma_2 \sqrt{c_2}\right)^2.$$

Tanto las asignaciones muestrales de mínima varianza y óptima con costes variables dependen de parámetros desconocidos, por lo que dichas asignaciones solo podrían ser estimadas con una muestra piloto. Un estimador insesgado de la varianza de $d$ es

$$\hat{V}(d) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2},$$

siendo $s_h^2$ la cuasivarianza muestral en el estrato $h$ , y $n_h$ la asignación muestral en el estrato $h$ . En los casos de las asignaciones obtenidas no podemos conocer dichos tamaños muestrales en la práctica sin conocer las varianzas de los estratos, por lo que fijados los tamaños muestrales (pudiendo ser tras estimarlos con una muestra piloto), el estimador insesgado de la varianza sería válido.

**Ejercicio 4.9.** Una población finita clasificada en dos dominios o estratos tiene definida sobre sus unidades la variable indicadora de un tercer dominio no necesariamente disjunto de los anteriores. Estimar sin sesgo la proporción desconocida de unidades de la población que están en este tercer dominio, usando muestreo estratificado con asignación proporcional al tamaño de los dos primeros estratos y usando independientemente muestreo aleatorio simple con reemplazamiento en cada uno de los dos primeros estratos. Estimar insesgadamente la varianza de la estimación anterior.

**Solución.** Denotamos por $N$ el tamaño de la población finita, y por $N_h$ el tamaño del estrato $h = 1, 2$. Definimos la variable $y_k = 1$ si la unidad $k$ es del tercer dominio, e $y_k = 0$ si la unidad $k$ no es del tercer dominio. La media de la variable $y$ en el dominio o estrato $h$ es la proporción $P_h$. Como

$$P_3 = \sum_{h=1}^{2} \frac{N_h}{N} P_h,$$

un estimador insesgado de la proporción $P_3$ de unidades de la población finita en el tercer dominio, haciendo uso de técnicas de muestreo estratificado

$$\hat{P}_3 = \sum_{h=1}^{2} \frac{N_h}{N} p_h,$$

donde $p_h$ es la proporción muestral de unidades del estrato $h$ $(= 1, 2)$ que están en el tercer dominio.

La asignación proporcional consiste en que si $n$ es el tamaño muestral total, el reparto de este tamaño muestral en los dos primeros estratos es proporcional al tamaño de los estratos. En concreto, $n_h = nN_h/N$ para $h = 1, 2$.

La varianza de $\hat{P}_3$ es

$$V(\hat{P}_3) = \sum_{h=1}^{2} \frac{N_h^2}{N^2} V(p_h) = \sum_{h=1}^{2} \frac{N_h^2}{N^2} \frac{P_h Q_h}{n_h},$$

por lo que, sustituyendo la asignación proporcional, la varianza admite la expresión

$$V_{\text{prop}}(\hat{P}_3) = \frac{1}{n} \sum_{h=1}^{2} \frac{N_h}{N} P_h Q_h.$$

Y como la varianza $P_h Q_h$ es estimable insesgadamente por el estimador cuasivarianza muestral en el estrato $h$-ésimo en el muestreo aleatorio simple con reemplazamiento, es decir por $n_h p_h q_h/(n_h - 1)$, resulta que el estimador insesgado de la varianza de $\hat{P}_3$ es

$$\hat{V}_{\text{prop}}(\hat{P}_3) = \frac{1}{n} \sum_{h=1}^{2} \frac{N_h}{N} \frac{n_h p_h q_h}{n_h - 1} = \sum_{h=1}^{2} \frac{N_h^2}{N} \frac{p_h q_h}{n(N_h - N/n)} =$$

$$\sum_{h=1}^{2} \frac{N_h^2 p_h q_h}{nNN_h - N^2},$$

donde $q_h = 1 - p_h$ es la proporción muestral en el estrato $h$ de unidades seleccionadas que no pertenecen al tercer dominio.

**Ejercicio 4.10.** Proponer un estimador insesgado de la diferencia $D_i = \bar{y}_i - \bar{y}$ entre la media del dominio o estrato $i$ y la media poblacional. Obtener la varianza de este estimador usando muestreo aleatorio simple con reemplazamiento independientemente en cada estrato. Obtener la asignación

muestral de mínima varianza, y estimar sin sesgo la varianza del estimador de $D_i$.

**Solución.** El estimador natural de la diferencia $D_i = \bar{y}_i - \bar{y}$ es la diferencia

$$d_i = \bar{y}_{\boldsymbol{s}(i)} - \bar{y}_{st} = \bar{y}_{\boldsymbol{s}(i)} - \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_{\boldsymbol{s}(h)} =$$

$$\bar{y}_{\boldsymbol{s}(i)} \left(1 - \frac{N_i}{N}\right) + \sum_{h \neq i}^{L} \frac{N_h}{N} \bar{y}_{\boldsymbol{s}(h)},$$

donde hemos denotado por $\bar{y}_{st}$ al estimador usual de la media poblacional en muestreo estratificado, por $\bar{y}_{\boldsymbol{s}(h)}$ a la media muestral en el estrato $h$, y por $L$ al número de estratos considerado. Este estimador $d_i$ es insesgado de la diferencia $D_i$ pues la esperanza matemática de la diferencia de dos variables aleatorias es la diferencia de las esperanzas matemáticas de dichas variables.

La varianza de este estimador es

$$V(d_i) = \frac{\sigma_i^2}{n_i} \left(1 - \frac{N_i}{N}\right)^2 + \sum_{h \neq i}^{L} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h}.$$

Para obtener la varianza mínima al variar los tamaños muestrales $n_h$ sujetos a la condición

$$n = \sum_{h=1}^{L} n_h,$$

construimos el lagrangiano

$$\Lambda = V(d_i) + \lambda \left( n - \sum_{h=1}^{L} n_h \right).$$

Resolviendo, derivamos parcialmente el lagrangiano con respecto a cada uno de los tamaños muestrales de los estratos, de modo que si $h \neq i$,

$$\frac{\partial \Lambda}{\partial n_h} = -\frac{N_h^2 \sigma_h^2}{2n_h^2 N^2} - \lambda = 0$$

y

$$\frac{\partial \Lambda}{\partial n_i} = -\left(1 - \frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{2n_i^2} - \lambda = 0.$$

De donde si $h \neq i$,

$$\frac{n_h}{N_h \sigma_h} = \frac{n_i}{(N - N_i)\sigma_i} = c,$$

donde $c$ es una constante que se determina imponiendo la restricción

$$n = \sum_{h=1}^{L} n_h = c\left[\sum_{h \neq i}^{L} N_h \sigma_h + (N - N_i)\sigma_i\right],$$

por lo que la asignación muestral de varianza mínima resulta ser para $h \neq i$,

$$n_h = n \frac{N_h \sigma_h}{\sum_{h \neq i}^{L} N_h \sigma_h + (N - N_i)\sigma_i}$$

y

$$n_i = n \frac{(N - N_i)\sigma_i}{\sum_{h \neq i}^{L} N_h \sigma_h + (N - N_i)\sigma_i}.$$

Un estimador insesgado de la varianza del estimador $d_i$ es

$$\hat{V}(d_i) = \left(1 - \frac{N_i}{N}\right)^2 \frac{s_i^2}{n_i} + \sum_{h \neq i}^{L} \frac{N_h^2}{N^2} \frac{s_h^2}{n_h},$$

donde $s_h^2$ es la cuasivarianza muestral en el estrato $h = 1, 2, \ldots, L$, obtenida con muestreo aleatorio simple con reemplazamiento independientemente en cada estrato con el tamaño muestral $n_h$.

**Ejercicio 4.11.** Obtener la asignación muestral óptima con costes variables que minimiza la varianza del estimador diferencia de medias muestrales como estimador insesgado de la diferencia de medias de dos dominios disjuntos, usando muestreo aleatorio simple sin reemplazamiento independientemente en cada dominio.

**Solución.** El estimador de la diferencia de medias de dominios $D = \bar{y}_1 - \bar{y}_2$ es el estimador diferencia de las medias muestrales

$$d = \bar{y}_{s(1)} - \bar{y}_{s(2)},$$

obtenida con diseño de muestreo irrestricto aleatorio de tamaño $n_h$ independiente en cada dominio $h = 1, 2$. Este estimador es insesgado para estimar *D* por ser la esperanza matemática de una diferencia, la diferencia de las esperanzas matemáticas respectivas.

La varianza de $d$ es

$$V(d) = \frac{N_1 - n_1}{N_1 n_1} S_1^2 + \frac{N_2 - n_2}{N_2 n_2} S_2^2,$$

donde $S_h^2 = N_h \sigma_h^2 / (N_h - 1)$ es la cuasivarianza del dominio $h = 1, 2$. Para obtener la asignación óptima con costes variables, el lagrangiano es

$$\Lambda = V(d) + \lambda(C - c_1 n_1 - c_2 n_2),$$

donde $C$ es el presupuesto del procedimiento de muestreo, y $c_h$ es el coste por observación de una unidad en el dominio $h = 1, 2$. Resolviendo, derivamos parcialmente el lagrangiano con respecto a cada una de las variables $n_h$, e igualamos a cero:

$$\frac{\partial \Lambda}{\partial n_h} = -\frac{S_h^2}{2n_h^2} - \lambda c_h = 0,$$

para $h = 1, 2$. De donde,

$$n_h = c\, S_h / \sqrt{c_h}.$$

La constante $c$ se calcula de la restricción

$$C = c_1 n_1 + c_2 n_2 = c\left(S_1\sqrt{c_1} + S_2\sqrt{c_2}\right),$$

por lo que para $h = 1, 2$,

$$n_h = C\,\frac{S_h/\sqrt{c_h}}{S_1\sqrt{c_1} + S_2\sqrt{c_2}}$$

es la asignación muestral pedida, donde $S_h$ es la cuasidesviación típica en el dominio $h$-ésimo.

**Ejercicio 4.12.** Obtener la varianza mínima del estimador usual estratificado

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_{s(h)}.$$

El tamaño de la población finita es $N$ y el tamaño muestral total es $n \leq N$, seleccionando muestras irrestrictas aleatorias de tamaño $n_h$ independientemente en cada estrato.

**Solución.** Minimizando la varianza del estimador usual estratificado sujeto a la restricción

$$n = \sum_{h=1}^{L} n_h,$$

haciendo uso del procedimiento de los multiplicadores de Lagrange, obtenemos que para $h = 1, 2, \ldots, L$,

$$n_h = n \frac{W_h S_h}{\sum_{i=1}^{L} W_i S_i}.$$

Estos tamaños muestrales deben ser aproximados por los números naturales más próximos o que hagan factible el muestreo. Sustituyendo los valores obtenidos de los tamaños muestrales efectivos en cada estrato, en la fórmula de la varianza del estimador usual en muestreo estratificado, que es

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h n_h} S_h^2,$$

obtenemos la varianza mínima pedida,

$$V_{\text{mín}}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2.$$

Este valor obtenido es siempre positivo, aunque aparentemente nos lo haría dudar la relación siguiente,

$$\sum_{h=1}^{L} W_h S_h^2 - \left(\sum_{h=1}^{L} W_h S_h\right)^2 \geq 0,$$

por ser la varianza siempre positiva o cero, duda especialmente aparente pero engañosa para valores de $n$ próximos a $N$.

**Ejercicio 4.13.** Proponer un estimador insesgado del producto de dos medias de dos dominios disjuntos, calcular su varianza, y obtener un estimador insesgado de la varianza del estimador propuesto.

**Solución.** Seleccionando dos muestras aleatorias simples independientes de tamaños fijos $n_1$ y $n_2$ respectivamente, tenemos como estimador insesgado del producto de medias de dos dominios, $\bar{y}_1\bar{y}_2$, al estimador producto de dos medias muestrales independientes, $\bar{y}_{s(1)}\bar{y}_{s(2)}$. En efecto,

$$E\left(\bar{y}_{s(1)}\bar{y}_{s(2)}\right) = E\left(\bar{y}_{s(1)}\right)E\left(\bar{y}_{s(2)}\right) = \bar{y}_1\bar{y}_2.$$

La varianza del estimador producto de medias muestrales es

$$V\left(\bar{y}_{s(1)}\bar{y}_{s(2)}\right) = E\left(\bar{y}_{s(1)}^2\bar{y}_{s(2)}^2\right) - \left[E\left(\bar{y}_{s(1)}\bar{y}_{s(2)}\right)\right]^2 =$$

$$E\left(\bar{y}_{s(1)}^2\right)E\left(\bar{y}_{s(2)}^2\right) - \bar{y}_1^2\bar{y}_2^2 =$$

$$\left[V\left(\bar{y}_{s(1)}\right) + \bar{y}_1^2\right]\left[V\left(\bar{y}_{s(2)}\right) + \bar{y}_2^2\right] - \bar{y}_1^2\bar{y}_2^2 =$$

$$\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}\bar{y}_2^2 + \frac{\sigma_2^2}{n_2}\bar{y}_1^2.$$

Un estimador insesgado de esta varianza es

$$\hat{V}\left(\bar{y}_{s(1)}\bar{y}_{s(2)}\right) = \frac{s_1^2}{n_1}\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}\left[\bar{y}_{s(2)}^2 + \hat{V}\left(\bar{y}_{s(2)}\right)\right] +$$

$$\frac{s_2^2}{n_2}\left[\bar{y}_{s(1)}^2 + \hat{V}\left(\bar{y}_{s(1)}\right)\right] =$$

$$3\frac{s_1^2}{n_1}\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}\bar{y}_{s(2)}^2 + \frac{s_2^2}{n_2}\bar{y}_{s(1)}^2.$$

**Ejercicio 4.14.** Comparar la relación entre las varianzas obtenidas por muestreo estratificado con asignación proporcional y por muestreo aleatorio simple, usando como diseño básico el muestreo aleatorio simple con reemplazamiento, y también usando como diseño básico el muestreo irrestricto aleatorio.

**Solución.** La asignación proporcional consiste en asignar un tamaño muestral $n_h = nW_h$ proporcional al tamaño relativo del estrato correspondiente.

Con diseño básico de muestreo aleatorio simple (con reemplazamiento), la varianza del estimador estratificado es

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h},$$

donde sustituyendo la asignación proporcional, resulta

$$V_{\text{prop}}(\bar{y}_{st}) = \frac{1}{n} \sum_{h=1}^{L} W_h \sigma_h^2.$$

Como la varianza del estimador media muestral con diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$ es

$$V_{\text{mas}}(\bar{y}_s) = \frac{\sigma^2}{n},$$

entonces la relación pedida será

$$\frac{V_{\text{prop}}(\bar{y}_{st})}{V_{\text{mas}}(\bar{y}_s)} = \frac{\sum_{h=1}^{L} W_h \sigma_h^2}{\sigma^2} \leq 1,$$

relación constante que no depende del tamaño fijo muestral $n$.

En el caso de usar el diseño de muestreo irrectricto aleatorio de tamaño efectivo fijo $n$, la fórmula general de la varianza del estimador usual en muestreo estratificado es

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2,$$

de donde sustituyendo la asignación proporcional, obtenemos

$$V_{\text{prop}}(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} W_h S_h^2.$$

Como sabemos que con diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n$ sobre toda la población finita, la media muestral tiene por varianza

$$V_{\text{mia}}(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2,$$

concluimos que la relación buscada es

$$\frac{V_{\text{prop}}(\bar{y}_{st})}{V_{\text{mia}}(\bar{y}_s)} = \frac{\sum_{h=1}^{L} W_h S_h^2}{S^2},$$

relación que es también constante y no depende del tamaño efectivo fijo muestral.

**Ejercicio 4.15.** Con los datos disponibles de un muestreo estratificado aleatorio con diseño básico de muestreo aleatorio simple con reemplazamiento, proponer un estimador insesgado de la varianza poblacional.

**Solución.** Como la varianza poblacional admite la relación siguiente

$$\sigma^2 = \sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h (\bar{y}_h - \bar{y})^2 =$$

$$\sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h \bar{y}_h^2 - \bar{y}^2,$$

un estimador insesgado de la varianza poblacional es

$$\widehat{\sigma^2} = \sum_{h=1}^{L} W_h s_h^2 + \sum_{h=1}^{L} W_h \left( \bar{y}_{s(h)}^2 - \frac{s_h^2}{n_h} \right) - \bar{y}_{st}^2 + \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h},$$

donde el último sumando es un estimador insesgado de la varianza del estimador usual en muestreo estratificado.

Otro estimador insesgado de la varianza poblacional, ya que

$$\sigma^2 = \sum_{h=1}^{L} W_h \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}^2 - \bar{y}^2,$$

es el estimador

$$\widehat{\sigma^2} = \sum_{h=1}^{L} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}^2 - \bar{y}_{st}^2 + \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h}.$$

**Ejercicio 4.16.** Con una muestra piloto se desea estimar el tamaño muestral deseable $n$ para que la varianza del estimador insesgado usual en muestreo estratificado con asignación proporcional tenga un valor aproximado $v$, usando como diseño básico el muestreo irrestricto aleatorio.

**Solución.** Bastará igualar $v$ a la varianza del estimador estratificado, es decir,

$$v = V_{\text{prop}}(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2,$$

de donde, despejando $n$ y llamándole tamaño muestral exacto para que la varianza del estimador estratificado con asignación proporcional sea $v$, tenemos

$$n = \frac{\sum_{h=1}^{L} W_h S_h^2}{v + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}.$$

Así, el estimador de $n$, que llamamos $\hat{n}$ sería

$$\hat{n} = \frac{\sum_{h=1}^{L} W_h s_h^2}{v + \frac{1}{N}\sum_{h=1}^{L} W_h s_h^2}.$$

En esta aproximación hemos sustituido las cuasivarianzas de los estratos $S_h^2$ por sus estimaciones insesgadas, las cuasivarianzas muestrales en los estratos $s_h^2$ a partir de la muestra piloto.

**Ejercicio 4.17.** Obtener la varianza del estimador usual en muestreo estratificado, con asignación proporcional, para estimar una proporción poblacional. Dar un valor para que el tamaño muestral total asegure una varianza menor o igual a una cantidad constante $v$. Dar un valor exacto para el tamaño muestral si sabemos que la proporción del estrato $h$ verifica que $P_h(1 - P_h) \geq 1/10$.

**Solución.** La varianza del estimador $p_{st}$ usual en muestreo estratificado con asignación proporcional es

$$V_{\text{prop}}(p_{st}) = \frac{N-n}{Nn}\frac{1}{N}\sum_{h=1}^{L} \frac{N_h^2 P_h Q_h}{N_h - 1} \leq v,$$

desigualdad que se verifica cuando

$$n \geq \frac{\frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 P_h Q_h}{N_h - 1}}{v + \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 P_h Q_h}{N_h - 1}}.$$

Pero este valor de $n$, el menor valor de los que verifican la desigualdad, no puede ser conocido ya que depende de las proporciones de los estratos $P_h$ que son desconocidos antes de realizar las observaciones por muestreo.

Si aceptamos que $P_h Q_h \geq 1/10$, sustituyendo la cota inferior obtenida antes, tenemos que como $P_h Q_h \leq 1/4$,

$$n \geq \frac{\frac{1}{4N^2}\sum_{h=1}^{L}\frac{N_h^2}{N_h - 1}}{v + \frac{1}{10N^2}\sum_{h=1}^{L}\frac{N_h^2}{N_h - 1}} \geq \frac{\frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 P_h Q_h}{N_h - 1}}{v + \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 P_h Q_h}{N_h - 1}},$$

que sí sería una cota inferior determinada para la elección del tamaño muestral efectivo fijo $n$ que proporcione un estimador usual en muestreo estratificado de la proporción poblacional con asignación proporcional cuya varianza esta acotada superiormente por la constante $v$. En la práctica, si

$$n \geq \frac{\frac{1}{4N^2}\sum_{h=1}^{L}\frac{N_h^2}{N_h - 1}}{v + \frac{1}{10N^2}\sum_{h=1}^{L}\frac{N_h^2}{N_h - 1}},$$

garantizamos la desigualdad buscada.

**Ejercicio 4.18.** Dar una cota inferior del punto de separación de estratos para que usando muestreo estratificado con dos estratos, el segundo de inclusión segura en la muestra, proporcione

estimadores más precisos que el muestreo aleatorio simple sin reemplazamiento del mismo tamaño muestral sobre toda la población finita. ¿Es útil esta cota inferior directa o indirectamente?

**Solución.** Si la población finita está ordenada en orden creciente por su variable de interés $y_k$ con $k = 1, 2, \ldots, N$, es decir,

$$y_1 \leq y_2 \leq \cdots \leq y_k \leq \cdots \leq y_N$$

el punto de estratificación $y$ que divide la población en dos estratos, el primero formado por las unidades $k$ con

$$y_k \leq y,$$

y el segundo estrato formado por las unidades $k$ con

$$y \leq y_k.$$

Con asignación especial consistente en seleccionar una muestra aleatoria simple sin reemplazamiento de tamaño $n_1$ en el primer estrato, y en seleccionar todo el segundo estrato en la muestra, es decir, $n_2 = N_2$. El tamaño muestral total es $n = n_1 + n_2$. Entonces sabemos que si el punto $y$ de estratificación o de separación de estratos verifica

$$y > \alpha + \sigma \times \sqrt{\frac{N}{n} - 1}$$

siendo $N$ el tamaño poblacional, el estimador estratificado con asignación especial de la media poblacional $\alpha$, concretamente

$$W_1 \bar{y}_{s(1)} + W_2 \bar{y}_2$$

(donde $W_h$ es el tamaño relativo del estrato $h$, $\bar{y}_{s(1)}$ es la media muestral en el primer estrato, e $\bar{y}_2$ la media del segundo estrato),

verifica que es insesgado y más preciso que la media muestral con muestreo aleatorio simple sin reemplazamiento de tamaño muestral común $n$.

Sin embargo el conocimiento de esta cota inferior supone conocer perfectamente tanto la media poblacional $\alpha$, como la desviación estándar poblacional $\sigma$. Si conociéramos $\alpha$, no sería necesario estimarlo. Pero aún desconociendo el verdadero valor de $\alpha$, si disponemos de una muestra piloto previa a la estimación estratificada con asignación especial, podríamos estimar los parámetros $\alpha$ y $\sigma$. Llamemos $\alpha^*$ y $\sigma^*$ a estos estimadores pilotos. Entonces, podemos estimar la cota inferior por la cota piloto inferior

$$\alpha^* + \sigma^* \times \sqrt{\frac{N}{n} - 1}$$

de modo que si aproximadamente el punto de estratificación $y$ es mayor que dicha cota piloto inferior, el estimador estratificado con asignación especial proporcionará estimaciones por lo general más precisas que las proporcionadas por la media muestral con muestreo aleatorio simple sin reemplazamiento del mismo tamaño muestral $n$.

**Ejercicio 4.19.** Desarrollar una teoría de análisis de la varianza en diseños experimentales con una base inferencial objetiva.

**Solución.** En modelos de diseño experimental tradicional se supone que un número infinito de posibles observaciones pueden ser obtenidas de un experimento. Además suele considerarse que estas observaciones pueden ser modeladas estadísticamente e incluir una variable de error que suele estar supuestamente distribuida Normal con algunas condiciones adicionales. La

comprobación práctica de tal distribución de los errores no es posible. Por esto, el uso del diseño experimental tradicional requiere asumir circunstancias que podrían estar lejos de las verdaderas condiciones de trabajo. Algunas consecuencias posibles de tales suposiciones son las conclusiones y resultados inferenciales sin verdadera base lógica sólida.

Algunas aplicaciones de la teoría objetiva desarrollada en este ejercicio son la agricultura natural, industriales, sociales, biomedicina, etc. Con la presente visión tenemos la ventaja de trabajar sin el uso de hipótesis no verificables, algo que no superan los métodos clásicos de diseño de experimentos. Nuestro modelo está basado en hechos, como ocurre con la teoría de muestras de poblaciones finitas de unidades identificadas.

Diseños experimentales de un factor. Partimos del modelo realista siguiente; para $t = 1, 2, \dots, T$ y para cada tratamiento $t$, $i = 1, 2, \dots, N_t$, disponemos de una población finita de tamaño $N_t$. El modelo de un factor es:

$$X_{ti} = A + B_t + \varepsilon_{ti}$$

Donde $T$ es el número de tratamientos, y para cada tratamiento $t$ tenemos un número máximo posible $N_t$ de observaciones diferentes, una por cada unidad de la población en la que se podría experimentar el tratamiento $t$. Los tratamientos (niveles o estratos) son estocásticamente independientes, y para cada tratamiento $t$ realizamos un número finito o tamaño muestral $n_t$ de observaciones o experimentos a partir de la población finita con tamaño o número finito $N_t$ de posibles resultados de los experimentos con el tratamiento común $t$. El valor $X_{ti}$ es la observación fijada de la variable de interés de la población finita o en el estrato de la unidad $i$-ésima para el tratamiento $t$. El valor $A$ es la media común para toda la población finita completa,

considerando todos los tratamientos $t$ y todas las unidades poblacionales $i$ en cada estrato o tratamiento $t$. El valor $B_t$ es el valor medio añadido al valor media común $A$ en el tratamiento $t$. Y $\varepsilon_{ti}$ es el error o desviación de la observación $X_{ti}$ con respecto a la media del tratamiento $t$, es decir, respecto a $A + B_t$. Por ello, se puede definir el error para la unidad $i$ en el tratamiento $t$ como la variable $\varepsilon_{ti} = X_{ti} - A - B_t$.

El número total de unidades experimentales, o tamaño poblacional finito de posibles experimentos observados o de productos en la industria, es

$$N = \sum_{t=1}^{T} \sum_{i=1}^{N_t} 1 = \sum_{t=1}^{T} N_t$$

La media poblacional finita global de las observaciones de la variable de interés es

$$\bar{X} = \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{N_t} X_{ti} = \frac{1}{N} \sum_{t=1}^{T} N_t \bar{X}_t$$

El tamaño muestral de experimentación efectiva para el tratamiento o estrato $t$ es $n_t$, y por tanto el tamaño muestral global de experimentación para los $T$ tratamientos es

$$n = \sum_{t=1}^{T} \sum_{i=1}^{n_t} 1 = \sum_{t=1}^{T} n_t$$

Y el coste total de experimentación es

$$c = \sum_{t=1}^{T} c_t n_t$$

siendo $c_t$ el coste por experimento con el tratamiento $t$.

La media muestral estratificada es

$$\bar{x}_{st} = \frac{1}{N}\sum_{t=1}^{T}\frac{N_t}{n_t}\sum_{i=1}^{n_t} x_{ti}$$

donde $x_{ti} = X_{tj_i}$, siendo el subíndice $j_i$ la $i$-ésima unidad seleccionada en la muestra del estrato o tratamiento $t$.

La media del estrato o tratamiento $t$ es

$$\bar{X}_{t.} = \frac{1}{N_t}\sum_{i=1}^{N_t} X_{ti}$$

La media muestral $t$-ésima, obtenida por observación muestral del tratamiento $t$ en las $n_t$ unidades de la muestra seleccionada de la población de $N_t$ unidades, es

$$\bar{x}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} x_{ti}$$

Sería práctico, aunque no necesario, tomar $n_t$ constante independientemente del tratamiento $t$. En estas condiciones, la descomposición del modelo estudiado de diseños experimentales de un factor será

$$X_{ti} = \bar{X} + (\bar{X}_{t.} - \bar{X}) + (X_{ti} - \bar{X}_{t.})$$

Donde $A = \bar{X}$, $B_t = \bar{X}_{t.} - \bar{X}$ y $\varepsilon_{ti} = X_{ti} - \bar{X}_{t.}$, y entonces tenemos que

$$\sum_{t=1}^{T} N_t B_t = 0$$

ya que

$$\sum_{t=1}^{T} N_t \bar{X}_{t\cdot} = N\bar{X}$$

Y para todo $t = 1, 2, \ldots, T$,

$$\sum_{i=1}^{N_t} \varepsilon_{ti} = 0$$

ya que

$$\sum_{i=1}^{N_t} X_{ti} = N_t \bar{X}_{t\cdot}$$

Los estimadores tradicionales en muestreo estratificado de poblaciones finitas de $A$ y de $B_t$ son respectivamente

$$\hat{A} = \bar{x}_{st}$$

y

$$\hat{B}_t = \bar{x}_t - \bar{x}_{st}$$

La varianza del primero de estos estimadores insesgados es

$$V(\hat{A}) = V(\bar{x}_{st}) = V\left(\frac{1}{N}\sum_{t=1}^{T}\frac{N_t}{n_t}\sum_{i=1}^{n_t} x_{ti}\right)$$

$$= \frac{1}{N^2}\sum_{t=1}^{T} N_t^2 V(\bar{x}_t)$$

Donde $V(\bar{x}_t) = \sigma_t^2/n_t$ con diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo $n_t$ sobre una población finita de tamaño $N_t$. También admite la expresión

$$V(\bar{x}_t) = \frac{N_t - n_t}{N_t - 1}\frac{\sigma_t^2}{n_t}$$

con diseño de muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n_t$ sobre una población finita de tamaño $N_t$. En ambos casos hemos denotado

$$\sigma_t^2 = \frac{1}{N_t}\sum_{i=1}^{N_t}(X_{ti} - \bar{X}_{t\cdot})^2$$

Un estimador insesgado de esta varianza $V(\hat{A})$ es el siguiente

$$\hat{V}(\hat{A}) = \frac{1}{N^2}\sum_{t=1}^{T} N_t^2\,\hat{V}(\bar{x}_t)$$

Donde ahora, $\hat{V}(\bar{x}_t) = s_t^2/n_t$ en el muestreo aleatorio simple con reemplazamiento, y también

$$\hat{V}(\bar{x}_t) = \frac{N_t - n_t}{N_t}\frac{s_t^2}{n_t}$$

en el muestreo aleatorio simple sin reemplazamiento, siendo

$$s_t^2 = \frac{1}{n_t - 1}\sum_{i=1}^{n_t}(x_{ti} - \bar{x}_t)^2$$

la cuasivarianza muestral para el tratamiento $t$.

La estimación insesgada de la función paramétrica $B_t$ es el estimador $\bar{x}_t - \bar{x}_{st}$, y su varianza se obtiene del modo

$$V(\hat{B}_t) = Cov\left(\bar{x}_t - \frac{1}{N}\sum_{t=1}^{T} N_t\bar{x}_t\,, \bar{x}_t - \frac{1}{N}\sum_{t=1}^{T} N_t\bar{x}_t\right)$$

$$= V(\bar{x}_t) - \frac{2N_t}{N} V(\bar{x}_t) + V(\bar{x}_{st})$$

$$= \left(1 - \frac{2N_t}{N}\right) V(\bar{x}_t) + \frac{1}{N^2} \sum_{t=1}^{T} N_t^2 V(\bar{x}_t)$$

Un estimador insesgado de esta varianza se obtiene de este modo,

$$\hat{V}(\hat{B}_t) = \left(1 - \frac{2N_t}{N}\right) \hat{V}(\bar{x}_t) + \frac{1}{N^2} \sum_{t=1}^{T} N_t^2 \hat{V}(\bar{x}_t)$$

$$= \frac{1}{N^2} \left[ (N - N_t)^2 \hat{V}(\bar{x}_t) + \sum_{h \neq t}^{T} N_h^2 \hat{V}(\bar{x}_h) \right]$$

A partir de estos estimadores insesgados es posible obtener intervalos de confianza aproximados para las funciones paramétricas $A$ y $B_t$ haciendo uso de la desigualdad de Chebychev, y consecuentemente es posible contrastar hipótesis nulas relacionadas con dichas funciones paramétricas.

Diseños experimentales de dos factores. De modo similar al caso de diseños experimentales de un factor, el modelo de dos factores es generado por la ecuación

$$X_{tij} = A + F_t + C_i + (FC)_{ti} + \varepsilon_{tij}$$

Donde $t = 1, 2, \dots, T$, siendo $T$ el número de tratamientos del primer factor (factor "fila"), $i = 1, 2, \dots, I$, siendo $I$ el número de tratamientos del segundo factor (factor "columna"), y siendo $j = 1, 2, \dots, N_{ti}$, donde $N_{ti}$ es el número de unidades de la población finita o celda $(ti)$ de la que se se selecciona la muestra con los tratamientos $t$ $e$ $i$ del primer y del segundo factor respectivamente. El valor $A$ viene de "average" (en inglés), que significa "promedio", $F$ viene de "fila" y $C$ de "columna". La población

finita sobre la que se hacen los posibles experimentos tiene un tamaño

$$N = \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{N_{ti}} 1 = \sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti} = \sum_{t=1}^{T} N_{t\cdot} = \sum_{i=1}^{I} N_{\cdot i}$$

Donde hemos denotado, para $t = 1, 2, \dots, T$

$$N_{t\cdot} = \sum_{i=1}^{I} N_{ti}$$

Y para $i = 1, 2, \dots, I$

$$N_{\cdot i} = \sum_{t=1}^{T} N_{ti}$$

Si el tamaño muestral en la celda de los tratamientos $t$ e $i$ es $n_{ti}$, entonces el tamaño muestral total para todos los pares de tratamientos es

$$n = \sum_{t=1}^{T}\sum_{i=1}^{I} n_{ti} = \sum_{t=1}^{T} n_{t\cdot} = \sum_{i=1}^{I} n_{\cdot i}$$

También sería práctico, aunque no necesario, tomar $n_{ti}$ constante independientemente de la celda en que se experimente.

Y el coste total de experimentación será

$$c = \sum_{t=1}^{T}\sum_{i=1}^{I} c_{ti} n_{ti}$$

Siendo $c_{ti}$ el coste por experimentación en la celda $(ti)$, medido en unidades monetarias.

En el diseño experimental de dos factores la media poblacional global es

$$\bar{X} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{N_{ti}} X_{tij} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti}\bar{X}_{ti\cdot}$$

$$= \frac{1}{N}\sum_{t=1}^{T} N_{t\cdot}\bar{X}_{t\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I} N_{\cdot i}\bar{X}_{\cdot i\cdot}$$

Ahora el modelo experimental de dos factores puede descomponerse del siguiente modo más general

$$X_{tij} = \bar{X} + (\bar{X}_{t\cdot\cdot} - \bar{X}) + (\bar{X}_{\cdot i\cdot} - \bar{X})$$

$$+ (\bar{X}_{ti\cdot} - \bar{X}_{t\cdot\cdot} - \bar{X}_{\cdot i\cdot} + \bar{X}) + \left(X_{tij} - \bar{X}_{ti\cdot}\right)$$

El primer sumando representa la función paramétrica promedio general $A$, el segundo representa la función paramétrica del tratamiento $t$ del primer factor, $F_t$, el tercer sumando representa la función paramétrica del tratamiento $i$ del segundo factor, $C_i$, el cuarto sumando representa la función paramétrica interacción de los tratamientos $t$ e $i$ del primer y segundo factor respectivamente, $(FC)_{ti}$, y el quinto sumando representa el error o desviación $\varepsilon_{tij}$.

Un estimador insesgado de $A$ es

$$\hat{A} = \hat{\bar{X}} = \bar{x}_{st} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{I}\frac{N_{ti}}{n_{ti}}\sum_{j=1}^{n_{ti}} x_{tij} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti}\bar{x}_{ti\cdot}$$

Siendo $x_{tij} = X_{tik_j}$ la observación muestral $j$-ésima en la celda $(ti)$. El tamaño muestral total es

$$n = \sum_{t=1}^{T}\sum_{i=1}^{I} n_{ti}$$

Siendo $n_{ti}$ el tamaño muestral en la celda $(ti)$ donde los tratamientos $F_t$ y $C_i$ son experimentados simultáneamente.

La varianza de $\hat{A}$ es

$$V(\hat{A}) = \frac{1}{N^2}\sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti}^2 V(\bar{x}_{ti\cdot})$$

Donde $V(\bar{x}_{ti\cdot}) = \sigma_{ti\cdot}^2/n_{ti}$ en el muestreo aleatorio simple con reemplazamiento de tamaño fijo $n_{ti}$ en la celda $(ti)$, o bien

$$V(\bar{x}_{ti\cdot}) = \frac{N_{ti} - n_{ti}}{N_{ti} - 1}\frac{\sigma_{ti\cdot}^2}{n_{ti}}$$

en el muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n_{ti}$ en la celda $(ti)$.

La expresión de la varianza poblacional de la celda $(ti)$ es

$$\sigma_{ti\cdot}^2 = \frac{1}{N_{ti}}\sum_{j=1}^{N_{ti}}\left(X_{tij} - \bar{X}_{ti\cdot}\right)^2$$

Una estimación insesgada de la varianza de $\hat{A}$ es

$$\hat{V}(\hat{A}) = \frac{1}{N^2}\sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti}^2 \hat{V}(\bar{x}_{ti\cdot})$$

Donde $\hat{V}(\bar{x}_{ti\cdot}) = s_{ti\cdot}^2/n_{ti}$ en muestreo aleatorio simple con reemplazamiento de tamaño fijo $n_{ti}$ en la celda $(ti)$, o bien

$$\hat{V}(\bar{x}_{ti\cdot}) = \frac{N_{ti} - n_{ti}}{N_{ti}}\frac{s_{ti\cdot}^2}{n_{ti}}$$

en el muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n_{ti}$ en la celda $(ti)$. La cuasivarianza muestral es

$$s_{ti.}^2 = \frac{1}{n_{ti}-1}\sum_{j=1}^{n_{ti}}\left(x_{tij} - \bar{x}_{ti.}\right)^2$$

Un estimador insesgado de $F_t$ es

$$\hat{F}_t = \bar{x}_{t..} - \bar{x}_{st}$$

Donde

$$\bar{x}_{t..} = \frac{1}{n_{t.}}\sum_{i=1}^{I}\sum_{j=1}^{n_{ti}} x_{tij} = \frac{1}{n_{t.}}\sum_{i=1}^{I} n_{ti}\bar{x}_{ti.}$$

Y

$$n_{t.} = \sum_{i=1}^{I} n_{ti}$$

Además se puede comprobar que

$$\sum_{t=1}^{T} N_{t.}F_t = \sum_{t=1}^{T} N_{t.}(\bar{X}_{t..} - \bar{X}) = N(\bar{X} - \bar{X}) = 0.$$

La varianza de $\hat{F}_t$ se obtiene como sigue

$$V\left(\hat{F}_t\right) = \frac{1}{N^2}\left[(N - N_{t.})^2 V(\bar{x}_{t..}) + \sum_{k\neq t}^{T} N_{k.}^2 V(\bar{x}_{k..})\right]$$

Donde para $t = 1, 2, \ldots, T$,

$$V(\bar{x}_{t..}) = \sum_{i=1}^{I}\left(\frac{n_{ti}}{n_{t.}}\right)^2 V(\bar{x}_{ti.})$$

También un estimador insesgado de la varianza es

$$\hat{V}(\hat{F}_t) = \frac{1}{N^2}\left[(N - N_{t\cdot})^2 \hat{V}(\bar{x}_{t\cdot\cdot}) + \sum_{k\neq t}^{T} N_{k\cdot}^2 \hat{V}(\bar{x}_{k\cdot\cdot})\right]$$

Siendo para $t = 1, 2, \ldots, T$,

$$\hat{V}(\bar{x}_{t\cdot\cdot}) = \sum_{i=1}^{I} \left(\frac{n_{ti}}{n_{t\cdot}}\right)^2 \hat{V}(\bar{x}_{ti\cdot})$$

Para el segundo factor y el tratamiento $i$, tenemos la estimación insesgada de $C_i$ como

$$\hat{C}_i = \bar{x}_{\cdot i\cdot} - \bar{x}_{st}$$

Donde

$$\bar{x}_{\cdot i\cdot} = \frac{1}{n_{\cdot i}} \sum_{t=1}^{T} \sum_{j=1}^{n_{ti}} x_{tij}$$

Y también

$$\sum_{i=1}^{I} N_{\cdot i} C_i = 0.$$

Similarmente tenemos la varianza

$$V(\hat{C}_i) = \frac{1}{N^2}\left[(N - N_{\cdot i})^2 V(\bar{x}_{\cdot i\cdot}) + \sum_{j=1}^{I} N_{\cdot j}^2 V(\bar{x}_{\cdot j\cdot})\right]$$

Que es estimable insesgadamente por

$$\hat{V}(\hat{C}_i) = \frac{1}{N^2}\left[(N - N_{\cdot i})^2 \hat{V}(\bar{x}_{\cdot i\cdot}) + \sum_{j=1}^{I} N_{\cdot j}^2 \hat{V}(\bar{x}_{\cdot j\cdot})\right]$$

Donde

$$\hat{V}(\bar{x}_{.i.}) = \sum_{t=1}^{T} \left(\frac{n_{ti}}{n_{.i}}\right)^2 \hat{V}(\bar{x}_{ti.})$$

Una estimación insesgada de la interacción $(FC)_{ti}$ es

$$\widehat{(FC)}_{ti} = \bar{x}_{ti.} - \bar{x}_{t..} - \bar{x}_{.i.} + \bar{x}_{st}$$

Y su varianza viene proporcionada por la expresión

$$V\left[\widehat{(FC)}_{ti}\right] =$$

$$V(\bar{x}_{ti.}) - Cov(\bar{x}_{ti.}, \bar{x}_{t..}) - Cov(\bar{x}_{ti.}, \bar{x}_{.i.}) + Cov(\bar{x}_{ti.}, \bar{x}_{st}) +$$

$$V(\bar{x}_{t..}) - Cov(\bar{x}_{t..}, \bar{x}_{ti.}) + Cov(\bar{x}_{t..}, \bar{x}_{.i.}) - Cov(\bar{x}_{t..}, \bar{x}_{st}) +$$

$$V(\bar{x}_{.i.}) - Cov(\bar{x}_{.i.}, \bar{x}_{ti.}) + Cov(\bar{x}_{.i.}, \bar{x}_{t..}) - Cov(\bar{x}_{.i.}, \bar{x}_{st}) +$$

$$V(\bar{x}_{st}) - Cov(\bar{x}_{st}, \bar{x}_{t..}) - Cov(\bar{x}_{st}, \bar{x}_{.i.}) + Cov(\bar{x}_{st}, \bar{x}_{ti.}).$$

Ahora se puede calcular todos los sumandos del segundo miembro de la expresión anterior. Conocemos los valores de las varianzas $V(\bar{x}_{ti.})$, $V(\bar{x}_{t..})$, $V(\bar{x}_{.i.})$ y

$$V(\bar{x}_{st}) = \frac{1}{N^2} \sum_{t=1}^{T} \sum_{i=1}^{I} N_{ti}^2 V(\bar{x}_{ti.}).$$

Y de las covarianzas

$$Cov(\bar{x}_{ti.}, \bar{x}_{t..}) = Cov(\bar{x}_{t..}, \bar{x}_{ti.}) = \frac{N_{ti}}{N_{t.}} V(\bar{x}_{ti.}),$$

$$Cov(\bar{x}_{.i.}, \bar{x}_{ti.}) = Cov(\bar{x}_{ti.}, \bar{x}_{.i.}) = \frac{N_{ti}}{N_{.i}} V(\bar{x}_{ti.}),$$

$$Cov(\bar{x}_{t..}, \bar{x}_{.i.}) = Cov(\bar{x}_{.i.}, \bar{x}_{t..}) = \frac{N_{ti}^2}{N_{t.} N_{.i}} V(\bar{x}_{ti.}),$$

$$Cov(\bar{x}_{ti.}, \bar{x}_{st}) = Cov(\bar{x}_{st}, \bar{x}_{ti.}) = \frac{N_{ti}}{N} V(\bar{x}_{ti.}),$$

$$Cov(\bar{x}_{t\cdot\cdot}, \bar{x}_{st}) = Cov(\bar{x}_{st}, \bar{x}_{t\cdot\cdot}) = \frac{1}{N_{t\cdot}N}\sum_{i=1}^{I} N_{ti}^2 V(\bar{x}_{ti\cdot}),$$

y

$$Cov(\bar{x}_{\cdot i\cdot}, \bar{x}_{st}) = Cov(\bar{x}_{st}, \bar{x}_{\cdot i\cdot}) = \frac{1}{N_{\cdot i}N}\sum_{t=1}^{T} N_{ti}^2 V(\bar{x}_{ti\cdot}).$$

Cada una de las expresiones anteriores puede estimarse sin sesgo de las mismas expresiones sustituyendo $V(\bar{x}_{ti\cdot})$ por su estimación insesgada ya vista anteriormente $\hat{V}(\bar{x}_{ti\cdot})$ . Como consecuencia, es posible estimar sin sesgo la función paramétrica interacción $(FC)_{ti}$ , y estimar sin sesgo la varianza de dicho estimador. También es posible por tanto estimar por intervalo y contrastar hipótesis sobre su valor concreto.

# Capítulo 5

# Muestreo posagrupado

Este tipo de muestreo se presenta cuando queremos tener la precisión del muestreo estratificado o similar, y no disponemos de los tamaños de los estratos pero los podemos estimar en una primera fase. De este modo se puede estimar la media poblacional con un estimador similar al estratificado, pero que incluye estimaciones de los tamaños relativos de los estratos, y pudiendo también estimar sin sesgo su varianza. Una aplicación de este tipo de muestreo es el problema de no respuesta en una encuesta, que queda resuelto a nivel formal con muestreo posagrupado.

## 5.1 Diseño posagrupado

Este diseño consta de dos fases de muestreo.

En la primera fase seleccionamos una muestra de tamaño $m$ con diseño *mas* y observamos el indicador del estrato en cada unidad de la muestra. Clasificamos la muestra seleccionada de tamaño $m$ en $l$ grupos o estratos seleccionados distintos $(1 \leq l \leq m, l \leq L)$. El número $l$ es aleatorio. Sean $m_h$ y $w_h = m_h/m$ respectivamente la frecuencia absoluta y la frecuencia relativa muestrales del grupo o estrato $h$. Tenemos que

$$m = \sum_{h=1}^{l} m_h, \quad 1 = \sum_{h=1}^{l} w_h,$$

y $m_h = w_h = 0$ en los restantes estratos $h = l+1, l+2, \dots, L$. De este modo el valor de $l$ es conocido tras la primera fase.

En la segunda fase, para cada grupo o estrato $h$ seleccionado en la primera fase $(h = 1, 2, \dots, l)$ con $m_h \geq 1$, es decir con alguna unidad seleccionada en el estrato en la primera fase, procedemos a seleccionar con diseño *mas* de tamaño fijo $n_h$ de entre las $N_h$ unidades que contiene su grupo o estrato $h$, tamaño $N_h$ que es conocido antes de la segunda fase de muestreo. Observamos la media muestral $\bar{y}_{s(h)}$ y obtenemos la cuasivarianza muestral $s_h^2$ que requiere un tamaño muestral $n_h \geq 2$.

En estas condiciones, definimos el estimador posagrupado siguiente

$$\bar{y}_{pg} = \sum_{h=1}^{l} w_h \bar{y}_{s(h)},$$

para el cual no son necesarios los tamaños y los marcos de los grupos o estratos $l+1, l+2, \dots, L$, pero para los primeros $l$ grupos o estratos los marcos de trabajo deben ser conocidos. Indicamos que la distribución de los valores

$m_h$ es multinomial de parámetros $m$ y $W_h$.

Veamos que el estimador $\bar{y}_{pg}$ es insesgado para estimar la media poblacional $\bar{y}$. En efecto,

$$E(\bar{y}_{pg}) = E_1\left\{E_2\left[\sum_{h=1}^{L} w_h \bar{y}_{s(h)} | w_h\right]\right\} =$$

$$E_1\left\{\sum_{h=1}^{L} w_h E_2\left[\bar{y}_{s(h)}\right]\right\} = \sum_{h=1}^{L} E_1(w_h)\bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h = \bar{y}.$$

## 5.2 Varianza del estimador posagrupado

La varianza del estimador $\bar{y}_{pg}$ puede calcularse haciendo uso del Teorema de Madow,

$$V(\bar{y}_{pg}) = E_1 V_2(\bar{y}_{pg}) + V_1 E_2(\bar{y}_{pg}),$$

donde

$$V_2(\bar{y}_{pg}) = \sum_{h=1}^{L} w_h^2 V_2[\bar{y}_{s(h)}] = \sum_{h=1}^{L} w_h^2 \frac{\sigma_h^2}{n_h},$$

$$E_1 V_2(\bar{y}_{pg}) = \sum_{h=1}^{L} E_1(w_h^2) \frac{\sigma_h^2}{n_h},$$

y donde

$$E_1(w_h^2) = \frac{E_1(m_h^2)}{m^2} = \frac{V_1(m_h) + [E_1(m_h)]^2}{m^2} =$$

$$\frac{W_h(1 - W_h) + mW_h^2}{m} = \frac{W_h[1 + W_h(m-1)]}{m}.$$

También

$$E_2(\bar{y}_{pg}) = \sum_{h=1}^{L} w_h E_2[\bar{y}_{s(h)}] = \sum_{h=1}^{L} w_h \bar{y}_h,$$

y entonces

$$V_1 E_2(\bar{y}_{pg}) = \sum_{h=1}^{L} \sum_{g=1}^{L} \bar{y}_h \bar{y}_g Cov_1(w_h, w_g) =$$

$$\frac{1}{m}\sum_{h=1}^{L}\bar{y}_h^2 W_h(1-W_h)+\frac{1}{m}\sum_{h=1}^{L}\sum_{g\neq h}^{L}\bar{y}_h\bar{y}_g\left(-W_hW_g\right)=$$

$$\frac{1}{m}\sum_{h=1}^{L}W_h\bar{y}_h^2-\frac{1}{m}\bar{y}^2=\frac{1}{m}\sum_{h=1}^{L}W_h(\bar{y}_h-\bar{y})^2.$$

Sustituyendo ambos sumandos de la descomposición de la varianza total en dos términos, obtenemos la varianza buscada del estimador posagrupado.

Un estimador insesgado de la varianza $V\left(\bar{y}_{pg}\right)$ es el siguiente

$$\hat{V}\left(\bar{y}_{pg}\right)=\sum_{h=1}^{l}w_h^2\frac{s_h^2}{n_h}+$$

$$\frac{1}{m}\sum_{h=1}^{l}\frac{m_h w_h(1-w_h)}{m_h-1}\left[\bar{y}_{s(h)}^2-\frac{s_h^2}{n_h}\right]-$$

$$\frac{1}{m-1}\sum_{h=1}^{l}\sum_{g\neq h}^{l}w_h w_g\bar{y}_{s(h)}\bar{y}_{s(g)},$$

del cual su razonamiento es debido a Ruiz Espejo (1993).

Otros desarrollos semejantes relacionados con diseño *mia* independientemente dentro de cada grupo o estrato han sido obtenidos por Ruiz Espejo y colaboradores (2006).

## 5.3 Estimación insesgada con no respuesta

El problema de la no respuesta surge en las encuestas por muestreo cuando una parte de los encuestados se niegan a facilitar sus respuestas. Si partimos de una muestra con diseño *mas* de tamaño fijo $m$ y parte de la muestra seleccionada se niega a contestar o no colabora con su respuesta, podemos entender que hay dos estratos implícitamente: uno primero de respuesta y otro segundo de no respuesta. Los tamaños de estos estratos son desconocidos. Si en estas condiciones tomáramos como estimador de la media poblacional $\bar{y}$ a la media muestral de las respuestas conseguidas, ésta es la media muestral del estrato de respuestas $\bar{y}_{\boldsymbol{s}(1)}$. Este estimador es sesgado para estimar $\bar{y}$, ya que

$$B\left[\bar{y}_{\boldsymbol{s}(1)};\ \bar{y}\right] = E\left[\bar{y}_{\boldsymbol{s}(1)}\right] - \bar{y} = \bar{y}_1 - (W_1\bar{y}_1 + W_2\bar{y}_2) =$$

$$W_2(\bar{y}_1 - \bar{y}_2),$$

un sesgo desconocido pues depende de tres funciones paramétricas $W_2$, $\bar{y}_1$, e $\bar{y}_2$ que son desconocidas.

La solución a este dilema la proporciona el muestreo posagrupado. En concreto, si en la primera fase obtuvimos los tamaños muestrales aleatorios $m_1 \geq 1$ y $m_2 \geq 1$, de manera que $m = m_1 + m_2$, consideramos que la muestra $\boldsymbol{s}(1)$ es de tamaño fijo $n_1$ que puede ser por ejemplo $n_1 = m_1$, mientras que la muestra del segundo estrato en la segunda fase es una submuestra $\boldsymbol{s}(2)$ de tamaño fijo $n_{(2)}$ de la muestra seleccionada de tamaño ya fijo $n_2 = m_2$ en la primera fase, y que submuestreamos con diseño *mas* de tamaño efectivo prefijado $n_{(2)}$ con $2 \leq n_{(2)}$. Esta submuestra $\boldsymbol{s}(2)$ requiere un mayor cuidado y esmero en la obtención de respuesta, pues ya obtuvimos la no respuesta de dichas unidades en la primera fase. Denotando $w_h = m_h/m$, el estimador insesgado de la media poblacional con no respuesta es

$$\bar{y}_{nr} = w_1\bar{y}_{(1)} + w_2\bar{y}_{(2)},$$

y un estimador insesgado de la varianza de este estimador es algo laborioso de obtener pero técnicamente deducible a partir del muestreo posagrupado y teniendo en cuenta la nueva fase de submuestreo con diseño *mas* que se desarrolla en la muestra de no respuesta en el segundo estrato, también llamado estrato de no respuesta. La fórmula exacta del estimador insesgado de la varianza puede consultarse en el artículo de Ruiz Espejo (2011a):

$$\hat{V}(\bar{y}_{nr}) = \frac{1}{m-1}\left\{\sum_{h=1}^{2} w_h\widehat{\sigma_h^2}\right.$$

$$\left. + \sum_{h=1}^{2} w_h\left[\bar{y}_{(h)}^2 - \hat{V}\left(\bar{y}_{(h)}\right)\right] - \bar{y}_{nr}^2\right\}$$

$$+ \frac{\widehat{\sigma_2^2}}{(m-1)n_{(2)}}(mw_2^2 - w_2).$$

Donde $\widehat{\sigma_1^2} = s_1^2$ y $\widehat{\sigma_2^2} = n_2 s_{(2)}^2/(n_2 - 1)$ siendo $s^2$ la cuasivarianza muestral de las respuestas en cada estrato $h$ que se subindica. Por tanto, $m_1 \geq 2$ y $m_2 \geq 2$. Además, el estimador $\hat{V}\left(\bar{y}_{(1)}\right) = s_1^2/n_1$ y si $n_{(2)}$ es el tamaño muestral en el segundo estrato o número de respuestas de la submuestra

$$\hat{V}\left(\bar{y}_{(2)}\right) = \frac{\widehat{\sigma_2^2}}{n_2} + \frac{s_{(2)}^2}{n_{(2)}} = s_{(2)}^2\left[\frac{1}{n_2 - 1} + \frac{1}{n_{(2)}}\right].$$

Como se puede apreciar, es casi un caso particular de muestreo posagrupado con dos grupos o estratos, en el que ahora se considera una tercera fase de aleatorización por submuestreo en la muestra de no respuesta en segunda fase del segundo estrato.

## 5.4 Ejercicios resueltos

**Ejercicio 5.1.** Una población finita es objeto de muestreo con diseño *mas* de tamaño fijo igual a $10$, donde se aprecian dos estratos cuyos tamaños relativos estimados por las proporciones muestrales son $3/5$ y $2/5$. En una segunda fase se estimaron de cada estrato la media muestral y la cuasivarianza muestral con diseño *mas* dando lugar a los pares de estimaciones $\left(\bar{y}_{s(h)}, s_h^2/n_h\right)$ siguientes para $h = 1, 2$: $(6, 5)$ , $(2, 2)$ . Estimar la media poblacional por muestreo posagrupado y dar un estimador insesgado de su varianza.

**Solución.** El estimador de la media poblacional es

$$\bar{y}_{pg} = \sum_{h=1}^{2} w_h \bar{y}_{s(h)} = \frac{3}{5} \cdot 6 + \frac{2}{5} \cdot 2 = 4.4$$

Y el estimador insesgado de su varianza, aplicando la fórmula general es

$$\hat{V}\left(\bar{y}_{pg}\right) \approx 0.0252.$$

**Ejercicio 5.2.** Para estimar la media poblacional en presencia de no respuesta hemos observado que el $40\%$ de los cien encuestados inicialmente no responden con diseño *mas*. La media muestral de respuestas en esta fase inicial fue de $10$ y la cuasivarianza muestral fue de $60$. Posteriormente se submuestrea con diseño *mas* de tamaño muestral $8$ la muestra de no respuesta en la primera fase dando lugar a una media muestral de $0$ y una cuasivarianza muestral de $16$. Estimar la media poblacional y decir si es posible

proponer un estimador insesgado de la varianza del estimador de la media poblacional.

**Solución.** El estimador de la media poblacional es

$$w_1\bar{y}_{(1)} + w_2\bar{y}_{(2)} = 6,$$

mientras que el estimador insesgado de su varianza es técnicamente posible sustituyendo en la expresión del estimador insesgado de la varianza los valores $w_1 = 0.6$, $w_2 = 0.4$, $m = 100$, $n_1 = 60$, $m_2 = 40$, $n_2 = 8$, $\bar{y}_{(1)} = 10$, $s_1^2 = 60$, $\bar{y}_{(2)} = 0$, y $s_{(2)}^2 = 16$.

**Ejercicio 5.3.** Obtener la esperanza matemática y la varianza del estimador usual en muestreo posagrupado en el caso de dos estratos, grupos o dominios disjuntos, usando en todos los diseños básicos muestreo aleatorio simple con reemplazamiento.

**Solución.** Consideramos una población finita de tamaño $N$, clasificada en dos estratos. El estimador usual en muestreo posagrupado es

$$\bar{y}_{pg} = \sum_{h=1}^{2} w_h \bar{y}_{\boldsymbol{s}(h)},$$

donde $w_h = m_h/m$, donde $m_h$ sigue una distribución binomial de parámetros $m$ y $W_h$, siendo $m$ el tamaño muestral en la primera fase y $W_h = N_h/N$ el tamaño relativo del estrato $h = 1, 2$. Entonces sabemos que $E_1(m_h) = mW_h$ y además $V_1(m_h) = mW_h(1 - W_h) = mW_1W_2$. También $\bar{y}_{\boldsymbol{s}(h)}$ es la media muestral en el estrato $h = 1, 2$, obtenida por muestreo aleatorio simple con reemplazamiento de tamaño fijo $n_h$, muestreos independientes en cada estrato obtenidos en una segunda fase.

La esperanza matemática del estimador $\bar{y}_{pg}$ es la media poblacional, pues

$$E(\bar{y}_{pg}) = E_1E_2(\bar{y}_{pg}|w_h) = E_1\left(\sum_{h=1}^{2} w_h\bar{y}_h\right) = \sum_{h=1}^{2} W_h\bar{y}_h = \bar{y}.$$

La varianza del estimador usual se obtiene mediante el teorema de Madow,

$$V(\bar{y}_{pg}) = E_1V_2(\bar{y}_{pg}|w_h) + V_1E_2(\bar{y}_{pg}|w_h).$$

$$V_2(\bar{y}_{pg}|w_h) = \sum_{h=1}^{2} w_h^2V_2(\bar{y}_{s(h)}) = \sum_{h=1}^{2} w_h^2\frac{\sigma_h^2}{n_h},$$

y

$$E_1V_2(\bar{y}_{pg}|w_h) = \sum_{h=1}^{2} E_1(w_h^2)\frac{\sigma_h^2}{n_h} =$$

$$\frac{1}{m}\left[\sum_{h=1}^{2} W_h\frac{\sigma_h^2}{n_h} + (m-1)\sum_{h=1}^{2} W_h^2\frac{\sigma_h^2}{n_h}\right],$$

pues

$$E_1(w_h^2) = \frac{E_1(m_h^2)}{m^2} = \frac{V_1(m_h) + (mW_h)^2}{m^2} =$$

$$\frac{mW_h(1-W_h) - m^2W_h^2}{m^2} = \frac{W_h[1+(m-1)W_h]}{m}.$$

También,

$$E_2(\bar{y}_{pg}|w_h) = \sum_{h=1}^{2} w_hE_2(\bar{y}_{s(h)}) = \sum_{h=1}^{2} w_h\bar{y}_h,$$

y

$$V_1E_2(\bar{y}_{pg}|w_h) = Cov_1\left(\sum_{h=1}^{2} w_h\bar{y}_h, \sum_{g=1}^{2} w_g\bar{y}_g\right) =$$

$$\sum_{h=1}^{2}\sum_{g=1}^{2} \bar{y}_h\bar{y}_g Cov_1(w_h, w_g) =$$

$$\sum_{h=1}^{2} \bar{y}_h^2 V_1(w_h) + \sum_{h=1}^{2}\sum_{g\neq h}^{2} \bar{y}_h\bar{y}_g Cov_1(w_h, w_g) =$$

$$\sum_{h=1}^{2} \bar{y}_h^2 \frac{mW_h(1-W_h)}{m^2} + 2\bar{y}_1\bar{y}_2 Cov_1(w_1, 1-w_1) =$$

$$\frac{1}{m}\sum_{h=1}^{2} \bar{y}_h^2 W_h\,(1-W_h) - 2\bar{y}_1\bar{y}_2 V_1(w_1) =$$

$$\frac{1}{m} W_1W_2(\bar{y}_1 - \bar{y}_2)^2.$$

Por lo que de todo ello,

$$V(\bar{y}_{pg}) = \frac{1}{m}\left[\sum_{h=1}^{2} W_h\frac{\sigma_h^2}{n_h} + (m-1)\sum_{h=1}^{2} W_h^2\frac{\sigma_h^2}{n_h}\right] +$$

$$\frac{1}{m} W_1W_2(\bar{y}_1 - \bar{y}_2)^2,$$

o bien,

$$V(\bar{y}_{pg}) = E_1\left(\sum_{h=1}^{2} w_h^2\frac{\sigma_h^2}{n_h}\right) + \frac{1}{m} W_1W_2(\bar{y}_1 - \bar{y}_2)^2.$$

**Ejercicio 5.4.** Proponer un estimador insesgado de la varianza del estimador usual en muestreo posagrupado con diseño básico de muestreo aleatorio simple, para dos estratos.

**Solución.** De la última fórmula de la varianza del estimador usual en muestreo posagrupado, sustituyendo los parámetros desconocidos por sus estimaciones insesgadas correspondientes, tenemos como estimador insesgado de la varianza de $\bar{y}_{pg}$ a

$$\hat{V}(\bar{y}_{pg}) = \sum_{h=1}^{2} w_h^2 \frac{s_h^2}{n_h} + \frac{1}{m} \sum_{h=1}^{2} \frac{m_h w_h (1 - w_h)}{m_h - 1} \left( \bar{y}_{s(h)}^2 - \frac{s_h^2}{n_h} \right) - \frac{2}{m-1} w_1 w_2 \bar{y}_{s(1)} \bar{y}_{s(2)},$$

pues la esperanza matemática de este estimador es igual a la varianza de $\bar{y}_{pg}$.

**Ejercicio 5.5.** Obtener la varianza del estimador usual en muestreo posagrupado cuando consideramos dos estratos, y como diseño muestral básico al muestreo irrestricto aleatorio.

**Solución.** El estimador usual en muestreo posagrupado es casi el mismo en apariencia para diseño básico de muestreo irrestricto aleatorio que con muestreo aleatorio simple. Ahora es

$$\bar{y}_{pg} = \sum_{h=1}^{2} w_h \bar{y}_{s(h)},$$

donde $w_h = m_h/m$ es la proporción muestral en la primera fase de unidades de la muestra irrestricta aleatoria que pertenecen al estrato $h = 1, 2$, y ahora $m_h$ se distribuye según geométrica de parámetros $m$, $N$ y $W_h$, siendo $m$ el tamaño muestral en la primera fase, $N$ el tamaño poblacional, y $W_h = N_h/N$ el tamaño relativo del estrato

$h = 1, 2$. Por tanto, $E_1(m_h) = mW_h$, y la varianza es $V_1(m_h) = mW_h(1 - W_h)(N - m)/(N - 1)$. También $\bar{y}_{s(h)}$ es la media muestral independiente en el estrato $h = 1, 2$, obtenidas con muestreo irrestricto aleatorio de tamaño efectivo fijo $n_h$ en una segunda fase.

La justificación de que $\bar{y}_{pg}$ es insesgado para estimar la media poblacional es análoga al caso de muestreo aleatorio simple con reemplazamiento como diseño básico.

La varianza de este estimador se obtiene por el teorema de Madow,

$$V(\bar{y}_{pg}) = E_1 V_2(\bar{y}_{pg}|w_h) + V_1 E_2(\bar{y}_{pg}|w_h).$$

Desarrollando,

$$V_2(\bar{y}_{pg}|w_h) = \sum_{h=1}^{2} w_h^2 V_2(\bar{y}_{s(h)}) = \sum_{h=1}^{2} w_h^2 \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h},$$

y

$$E_1 V_2(\bar{y}_{pg}|w_h) = \sum_{h=1}^{2} E_1(w_h^2) \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h} =$$

$$\sum_{h=1}^{2} W_h \left[\frac{N - m}{(N - 1)m} + W_h \frac{N(m - 1)}{(N - 1)m}\right] \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h},$$

pues

$$E_1(w_h^2) = V_1(w_h) + [E_1(w_h)]^2 =$$

$$\frac{1}{m} W_h(1 - W_h) \frac{N - m}{N - 1} + W_h^2 =$$

$$W_h\left\{\frac{N-m}{(N-1)m}+W_h\left[1-\frac{N-m}{(N-1)m}\right]\right\}=$$

$$W_h\left[\frac{N-m}{(N-1)m}+W_h\frac{N(m-1)}{(N-1)m}\right].$$

Por otro lado,

$$E_2(\bar{y}_{pg}|w_h)=\sum_{h=1}^{2}w_hE_2(\bar{y}_{s(h)})=\sum_{h=1}^{2}w_h\bar{y}_h,$$

y

$$V_1E_2(\bar{y}_{pg}|w_h)=Cov_1\left(\sum_{h=1}^{2}w_h\bar{y}_h\,,\sum_{g=1}^{2}w_g\bar{y}_g\right)=$$

$$\sum_{h=1}^{2}\bar{y}_h^2V_1(w_h)+2\bar{y}_1\bar{y}_2Cov_1(w_1,1-w_1)=$$

$$\sum_{h=1}^{2}\bar{y}_h^2W_h(1-W_h)\frac{N-m}{(N-1)m}-2\bar{y}_1\bar{y}_2W_1W_2\frac{N-m}{(N-1)m}=$$

$$W_1W_2\frac{N-m}{(N-1)m}(\bar{y}_1-\bar{y}_2)^2.$$

Por tanto, podemos dar una fórmula general exacta de la varianza del estimador usual en muestreo posagrupado sustituyendo los valores obtenidos, o bien, la fórmula

$$V(\bar{y}_{pg})=E_1\left(\sum_{h=1}^{2}w_h^2\frac{N_h-n_h}{N_h-1}\frac{\sigma_h^2}{n_h}\right)+$$

$$W_1W_2\frac{N-m}{(N-1)m}(\bar{y}_1-\bar{y}_2)^2$$

que nos permitirá proponer de modo más sencillo un estimador insesgado de esta varianza.

**Ejercicio 5.6.** Proponer un estimador insesgado de la varianza del estimador usual en muestreo posagrupado con dos estratos y diseño básico de muestreo irrestricto aleatorio.

**Solución.** Por un razonamiento similar al realizado en el Ejercicio 5.4. de este Capítulo, estimamos sin sesgo los parámetros desconocidos en la fórmula de la varianza del estimador usual. Así tenemos,

$$\hat{V}(\bar{y}_{pg}) = \sum_{h=1}^{2} w_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} +$$

$$\frac{N-m}{(N-1)m} \sum_{h=1}^{2} \left( \bar{y}_{s(h)}^2 + \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} \right) \frac{m w_h (1 - w_h)}{m-1} -$$

$$2 \frac{N-m}{(N-1)m} \frac{m w_1 w_2}{m-1} \bar{y}_{s(1)} \bar{y}_{s(2)} =$$

$$\sum_{h=1}^{2} w_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} +$$

$$\frac{N-m}{N-1} \sum_{h=1}^{2} \left( \bar{y}_{s(h)}^2 + \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} \right) \frac{w_h (1 - w_h)}{m-1} -$$

$$2 \frac{N-m}{N-1} \frac{w_1 w_2}{m-1} \bar{y}_{s(1)} \bar{y}_{s(2)}.$$

**Ejercicio 5.7.** Obtener la esperanza matemática y la varianza del estimador usual para el caso de no respuesta en una primera fase del muestreo, con diseño básico de muestreo aleatorio simple.

**Solución.** El estimador usual de la media poblacional con no respuesta es

$$\bar{y}_{nr} = w_1 \bar{y}_{\boldsymbol{s}(1)} + w_2 \bar{y}_{\boldsymbol{s}(2)},$$

donde $w_h = m_h/m$ es la proporción muestral en la primera fase del estrato o dominio de respuesta $(h = 1)$ o bien la proporción muestral en la primera fase del estrato o dominio de no respuesta $(h = 2)$; $\bar{y}_{\boldsymbol{s}(1)}$ es la media muestral de respuestas recogidas en la segunda fase, obtenida por muestreo aleatorio simple con reemplazamiento de tamaño $n_1 = m_1$, en el estrato de respuesta, es decir la media muestral de las respuestas obtenidas en la primera fase; finalmente $\bar{y}_{\boldsymbol{s}(2)}$ es la media muestral obtenida en una submuestra aleatoria simple de tamaño prefijado $n_{(2)}$ en tercera fase de la muestra aleatoria simple de no respuesta de tamaño $n_2 = m_2$ que dio lugar la segunda fase con su no respuesta, y que a su vez es parte, posiblemente con repeticiones de unidades, de la muestra aleatoria simple en el estrato o dominio de no respuesta obtenida en la primera fase.

La esperanza matemática de este estimador es

$$E(\bar{y}_{nr}) = E_1 E_2 E_3 (\bar{y}_{nr}) = E_1 \left[ w_1 E_2 \left( \bar{y}_{\boldsymbol{s}(1)} \right) + w_2 E_2 \left( \bar{y}_{(2)} \right) \right],$$

donde $\bar{y}_{(2)} = E_3 \left( \bar{y}_{\boldsymbol{s}(2)} \right)$ es la media muestral de la muestra en la segunda fase en el estrato de no respuesta. Entonces,

$$E(\bar{y}_{nr}) = E_1 (w_1 \bar{y}_1 + w_2 \bar{y}_2) = E_1 (w_1) \bar{y}_1 + E_1 (w_2) \bar{y}_2 =$$

$$W_1 \bar{y}_1 + W_2 \bar{y}_2 = \bar{y},$$

que es la media poblacional, y por tanto $\bar{y}_{nr}$ es un estimador insesgado.

La varianza del estimador usual $\bar{y}_{nr}$ es, aplicando el teorema de Madow,

$$V(\bar{y}_{nr}) = E_1 V_2(\bar{y}_{nr}) + V_1 E_2(\bar{y}_{nr}).$$

Aplicando esta fórmula por partes,

$$V_2(\bar{y}_{nr}) = V_2(\bar{y}_{nr}|w_h) =$$

$$w_1^2 \frac{\sigma_1^2}{n_1} + w_2^2 \left[E_2 V_3\left(\bar{y}_{\boldsymbol{s}(2)}\right) + V_2 E_3\left(\bar{y}_{\boldsymbol{s}(2)}\right)\right] =$$

$$w_1^2 \frac{\sigma_1^2}{n_1} + w_2^2 \left[E_2\left(\frac{\sigma_{(2)}^2}{n_{(2)}}\right) + V_2\left(\bar{y}_{(2)}\right)\right] =$$

$$w_1^2 \frac{\sigma_1^2}{n_1} + w_2^2 \left[\frac{(n_2 - 1)\sigma_2^2}{n_2 n_{(2)}} + \frac{\sigma_2^2}{n_2}\right] =$$

$$w_1^2 \frac{\sigma_1^2}{n_1} + w_2^2 \frac{\sigma_2^2}{n_2}\left(\frac{n_2 - 1}{n_{(2)}} + 1\right) =$$

$$\frac{w_1 \sigma_1^2}{m} + \frac{w_2 \sigma_2^2}{m}\left(\frac{n_2 - 1}{n_{(2)}} + 1\right) =$$

$$\frac{w_1 \sigma_1^2}{m} + w_2 \sigma_2^2 \left(\frac{w_2}{n_{(2)}} - \frac{1}{m n_{(2)}} + \frac{1}{m}\right).$$

$$E_1 V_2(\bar{y}_{nr}|w_h) =$$

$$\frac{1}{m}\sum_{h=1}^{2} W_h \sigma_h^2 + \frac{\sigma_2^2}{n_{(2)}}\left[E_1(w_2^2) - \frac{W_2}{m}\right] =$$

$$\frac{1}{m}\sum_{h=1}^{2} W_h\sigma_h^2 + \frac{\sigma_2^2}{n_{(2)}}\left(\frac{W_1W_2}{m} + W_2^2 - \frac{W_2}{m}\right) =$$

$$\frac{1}{m}\sum_{h=1}^{2} W_h\sigma_h^2 + \frac{\sigma_2^2}{n_{(2)}}\frac{W_2^2(m-1)}{m},$$

pues

$$E_1(w_2^2) = V_1(w_2) + W_2^2 = \frac{W_1W_2}{m} + W_2^2 =$$

$$\frac{W_2}{m}[1 + W_2(m-1)].$$

Por otro lado,

$$E_2(\bar{y}_{nr}|w_h) = w_1\bar{y}_1 + w_2\bar{y}_2,$$

de donde

$$V_1E_2(\bar{y}_{nr}|w_h) = \frac{W_1W_2}{m}(\bar{y}_1 - \bar{y}_2)^2.$$

Por lo que resumiendo,

$$V(\bar{y}_{nr}) = \frac{\sigma^2}{m} + \frac{(m-1)W_2^2\sigma_2^2}{mn_{(2)}},$$

o incluso también tenemos esta otra fórmula,

$$V(\bar{y}_{nr}) = E_1\left[w_1\frac{\sigma_1^2}{m} + \sigma_2^2\left(\frac{w_2^2}{n_{(2)}} - \frac{w_2}{mn_{(2)}} + \frac{w_2}{m}\right)\right] +$$

$$\frac{W_1W_2}{m}(\bar{y}_1 - \bar{y}_2)^2,$$

que nos permitirá estimarla sin sesgo de modo más sencillo.

**Ejercicio 5.8.** Obtener un estimador insesgado de la varianza del estimador usual con no respuesta, usando como diseño básico el muestreo aleatorio simple.

**Solución.** De la última fórmula de $V(\bar{y}_{nr})$, tenemos que estimando sin sesgo los parámetros desconocidos en dicha fórmula, tenemos el siguiente estimador

$$\hat{V}(\bar{y}_{nr}) = w_1 \frac{s_1^2}{m} + \widehat{\sigma_2^2}\left(\frac{w_2^2}{n_{(2)}} - \frac{w_2}{mn_{(2)}} + \frac{w_2}{m}\right) +$$

$$\frac{w_1 w_2}{m-1}\left[\bar{y}_{s(1)}^2 - \frac{s_1^2}{n_1} + \bar{y}_{s(2)}^2 - \hat{V}\left(\bar{y}_{s(2)}\right) - 2\bar{y}_{s(1)}\bar{y}_{s(2)}\right],$$

donde

$$\widehat{\sigma_2^2} = \widehat{s_2^2} = \frac{n_2}{n_2 - 1}\widehat{\sigma_{(2)}^2} = \frac{n_2}{n_2 - 1}s_{(2)}^2,$$

siendo $s_1^2$ la cuasivarianza muestral de tamaño fijo $n_1$ en el primer estrato o dominio, y $s_{(2)}^2$ la cuasivarianza muestral de tamaño $n_{(2)}$ en el segundo estrato o dominio. Como

$$V\left(\bar{y}_{(2)}\right) = V_1E_2E_3\left(\bar{y}_{(2)}\right) + E_1V_2E_3\left(\bar{y}_{(2)}\right) + E_1E_2V_3\left(\bar{y}_{(2)}\right) =$$

$$V_1(\bar{Y}_2) + E_1V_2(\bar{y}_2) + E_1E_2\left(\frac{\sigma_{(2)}^2}{n_{(2)}}\right) = E_1\left(\frac{\sigma_2^2}{n_2}\right) + E_1E_2\left(\frac{\sigma_{(2)}^2}{n_{(2)}}\right),$$

entonces un estimador insesgado de esta varianza es

$$\hat{V}\left(\bar{y}_{(2)}\right) = \frac{\widehat{\sigma_2^2}}{n_2} + \frac{s_{(2)}^2}{n_{(2)}} = s_{(2)}^2\left[\frac{1}{n_2 - 1} + \frac{1}{n_{(2)}}\right].$$

**Ejercicio 5.9.** Queremos estimar la media poblacional y la varianza de tal estimador ante el problema de no respuesta. ¿Qué tamaño

submuestral debe tomarse de la muestra de no respuesta para estimar sin sesgo la media poblacional? ¿Qué tamaño submuestral debe tomarse de la muestra de no respuesta para estimar sin sesgo la varianza del estimador de la media poblacional?

**Solución.** Para que el estimador insesgado de la media poblacional con no respuesta pueda usarse es necesario que haya respuestas (la necesidad de observar datos del primer estrato, de respuesta, no es indispensable, como en el segundo estrato de no respuesta), pero el tamaño de la submuestra dentro de la muestra de no respuesta debe ser al menos de uno, es decir, el tamaño de la muestra de no respuesta debe ser $n_2 \geq 1$ y el tamaño de la submuestra de la muestra anterior debe ser $n_{(2)} \geq 1$ . Sin embargo, para la estimación insesgada de la varianza de este estimador usual para no respuesta, es necesario un tamaño muestral de no respuesta que sea $n_{(2)} \geq 2$, pues solo así podría obtenerse la estimación de la cuasivarianza muestral en el estrato de no respuesta necesaria para la obtención del estimador insesgado de la varianza del estimador usual de la media poblacional con no respuesta.

# Capítulo 6

# Estimadores indirectos

En este capítulo vamos a ver tres estimadores (de razón, de producto, y de regresión) que además de la información proporcionada por observación de la variable de interés $y$, utilizan otra variable $x$, que llamamos auxiliar, conocida en todas las unidades de la población finita. Esta información permite construir inicialmente estimadores sesgados pero que podrían proporcionar estimaciones con pequeño error cuadrático medio.

## 6.1 Estimador de la razón poblacional

Se define la "razón poblacional" a la función paramétrica

$$R = \frac{N\bar{y}}{N\bar{x}} = \frac{\bar{y}}{\bar{x}}.$$

Es el cociente del total poblacional de la variable de interés entre el total poblacional de la variable auxiliar. Esta función paramétrica $R$ puede estimarse por la "razón muestral"

$$\hat{R} = \frac{\bar{y}_s}{\bar{x}_s} = \frac{n\bar{y}_s}{n\bar{x}_s}.$$

Su sesgo puede obtenerse de este modo,

$$Cov(\hat{R}, \bar{x}_s) = E(\hat{R}\bar{x}_s) - E(\hat{R})E(\bar{x}_s) = E(\bar{y}_s) - E(\hat{R})\bar{x} =$$

$$\bar{y} - E(\hat{R})\bar{x},$$

de donde

$$B(\hat{R}; R) = E(\hat{R}) - R = -\frac{Cov(\hat{R}, \bar{x}_s)}{\bar{x}}.$$

Es la expresión exacta del sesgo de la razón muestral como estimador de la razón poblacional, ya sea con diseño *mas* o *mia*.

El sesgo aproximado se obtiene de que

$$\hat{R} - R = \frac{\bar{y}_s}{\bar{x}_s} - R = \frac{\bar{y}_s - R\bar{x}_s}{\bar{x}_s} = \frac{\bar{y}_s - R\bar{x}_s}{\bar{x}}\frac{\bar{x}}{\bar{x}_s},$$

suponiendo que las variables $y$ y $x$ son positivas, y como además

$$\frac{\bar{x}}{\bar{x}_s} = \frac{\bar{x}}{\bar{x} + \bar{x}_s - \bar{x}} = \frac{1}{1 + \frac{\bar{x}_s - \bar{x}}{\bar{x}}} =$$

$$1 - \frac{\bar{x}_s - \bar{x}}{\bar{x}} + \frac{(\bar{x}_s - \bar{x})^2}{\bar{x}^2} - \cdots$$

siempre y cuando

$$\left|\frac{\bar{x}_s - \bar{x}}{\bar{x}}\right| < 1,$$

con lo que disponemos finalmente del desarrollo en serie

$$\hat{R} - R = \frac{\bar{y}_s - R\bar{x}_s}{\bar{x}}\left[1 - \frac{\bar{x}_s - \bar{x}}{\bar{x}} + \frac{(\bar{x}_s - \bar{x})^2}{\bar{x}^2} - \cdots\right],$$

de donde el sesgo de $\hat{R}$ expresado asintóticamente es

$$B(\hat{R}) = E(\hat{R} - R) =$$

$$-\frac{E[(\bar{y}_s - R\bar{x}_s)(\bar{x}_s - \bar{x})]}{\bar{x}^2} + \frac{E[(\bar{y}_s - R\bar{x}_s)(\bar{x}_s - \bar{x})^2]}{\bar{x}^3} - \cdots$$

pues el primer sumando del desarrollo en serie verifica

$$E\left(\frac{\bar{y}_s - R\bar{x}_s}{\bar{x}}\right) = \frac{1}{\bar{x}}E(\bar{y}_s - R\bar{x}_s) = \frac{1}{\bar{x}}(\bar{y} - \bar{y}) = 0.$$

En concreto, si aproximamos el sesgo de $\hat{R}$ por los dos primeros términos del desarrollo en serie, tenemos que para diseño *mas* o *mia*,

$$B(\hat{R}) \approx -\frac{E[(\bar{y}_s - R\bar{x}_s)(\bar{x}_s - \bar{x})]}{\bar{x}^2} =$$

$$-\frac{E[\bar{y}_s(\bar{x}_s - \bar{x})] - RE[\bar{x}_s(\bar{x}_s - \bar{x})]}{\bar{x}^2} =$$

$$\frac{-Cov(\bar{y}_s, \bar{x}_s) + RV(\bar{x}_s)}{\bar{x}^2}.$$

La varianza aproximada del estimador $\hat{R}$ de $R$ se obtiene considerando el primer término del desarrollo en serie de

$$\hat{R} - R \approx \frac{\bar{y}_s - R\bar{x}_s}{\bar{x}}.$$

Con esta aproximación, $E(\hat{R}) \approx R$ y la varianza aproximada es

$$V(\hat{R}) \approx \frac{V(\bar{y}_s - R\bar{x}_s)}{\bar{x}^2} = \frac{V(\bar{y}_s) + R^2V(\bar{x}_s) - 2RCov(\bar{y}_s, \bar{x}_s)}{\bar{x}^2}.$$

Ahora bien, como $V(\bar{y}_s) = \sigma_y^2/n$ , $V(\bar{x}_s) = \sigma_x^2/n$, y

$$Cov(\bar{y}_s, \bar{x}_s) = \frac{\sigma_{yx}}{n}$$

con diseño *mas*, concluimos que

$$V(\hat{R}) \approx \frac{1}{n\bar{x}^2}(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{yx}).$$

Análogamente, con diseño *mia*,

$$Cov(\bar{y}_s, \bar{x}_s) = \frac{N-n}{Nn} S_{yx}$$

siendo la cuasicovarianza poblacional

$$S_{yx} = \frac{1}{N-1} \sum_{k=1}^{N} (y_k - \bar{y})(x_k - \bar{x}) = S_{xy}$$

(Hansen, Hurwitz y Madow, 1953, p. 97), por lo que

$$V(\hat{R}) \approx \frac{N-n}{Nn\bar{x}^2} \left(S_y^2 + R^2 S_x^2 - 2RS_{yx}\right).$$

## 6.2 Estimador de razón de la media poblacional

Recibe este nombre el estimador de la media poblacional $\bar{y}$ con información auxiliar $x$ el estimador

$$t_R = \hat{R}\bar{x}.$$

Este estimador de razón es sesgado, pero por las propiedades vistas anteriormente, su sesgo es nulo aproximándolo por el primer término del desarrollo en serie, y su varianza aproximada haciendo uso de la misma aproximación es

$$V(t_R) \approx \frac{1}{n} \left(\sigma_y^2 + R^2 \sigma_x^2 - 2R\sigma_{yx}\right)$$

con diseño *mas*, y

$$V(t_R) \approx \frac{N-n}{Nn} \left(S_y^2 + R^2 S_x^2 - 2RS_{yx}\right)$$

con diseño *mia*.

## 6.3 Tamaño muestral del estimador de razón

Para obtener el tamaño muestral $n$ para que el estimador de razón $t_R = \hat{R}\bar{x}$ de la media poblacional $\bar{y}$ difiera de ésta menos que su error máximo absoluto admisible de muestreo $e$ con un cierto nivel de confianza $1-\alpha$, recurrimos a la desigualdad de Chebychev, pues hemos visto que $E(t_R) \approx \bar{y}$. Tenemos que aproximadamente

$$p\{|t_R - \bar{y}| < e\} \geq 1 - \frac{V(t_R)}{e^2} = 1 - \alpha.$$

De donde

$$\alpha e^2 = V(t_R),$$

y sustituyendo sus expresiones aproximadas, obtenemos los tamaños muestrales buscados, ya sea para diseño *mas*

$$n \approx \frac{\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{yx}}{\alpha e^2},$$

o para diseño *mia*

$$n \approx \frac{1}{\frac{1}{N} + \frac{\alpha e^2}{S_y^2 + R^2 S_x^2 - 2RS_{yx}}}.$$

## 6.4 Ganancia en precisión del estimador de razón

Para comparar los métodos inferenciales con diseño *mas* o *mia* entre la media muestral y el estimador de razón, escribimos sus varianzas exactas y aproximadas. Con diseño *mas* tenemos

$$V(\bar{y}_s) = \frac{\sigma_y^2}{n}$$

y

$$V(t_R) \approx \frac{1}{n}\left(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{yx}\right).$$

Luego, si esta aproximación fuera una igualdad,

$$V(t_R) \leq V(\bar{y}_s)$$

si y solo si

$$R^2\sigma_x^2 - 2R\sigma_{yx} \leq 0,$$

o bien, como el coeficiente de correlación de las variables $y$ y $x$ es $\rho = \sigma_{yx}/\left(\sigma_y\sigma_x\right)$ y si $R > 0$ como es habitual cuando las variables de interés y auxiliar son positivas, tenemos si y solo si

$$\rho \geq \frac{R\sigma_x}{2\sigma_y}.$$

Se puede comprobar que esta misma relación debe verificarse en condiciones similares para que con diseño *mia* la precisión del estimador $t_R$ sea mayor o igual a la de la media muestral con idéntico diseño *mia*.

## 6.5 Estimador de razón en el muestreo estratificado

Los dos tipos principales de estimador de razón cuando se usa muestreo estratificado son el estimador separado de razón y el estimador combinado de razón.

El estimador "separado de razón" en muestreo estratificado es aquél que usa del estimador de razón en cada estrato independientemente. Su expresión es

$$t_{SR} = \sum_{h=1}^{L} W_h \hat{R}_h \bar{x}_h,$$

donde $\hat{R}_h \bar{x}_h$ es el estimador independiente de razón en el estrato $h$. Su varianza es

$$V(t_{SR}) = \sum_{h=1}^{L} W_h^2 V\left(\hat{R}_h \bar{x}_h\right),$$

donde

$$V\left(\hat{R}_h \bar{x}_h\right) = \bar{x}_h^2 V\left(\hat{R}_h\right),$$

y con diseño *mas*

$$V\left(\hat{R}_h\right) \approx \frac{1}{n_h}\left(\sigma_{hy}^2 + R_h^2 \sigma_{hx}^2 - 2R_h \sigma_{hyx}\right)$$

y con diseño *mia*

$$V\left(\hat{R}_h\right) \approx \frac{N_h - n_h}{N_h n_h}\left(S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hyx}\right),$$

donde $R_h = \bar{y}_h / \bar{x}_h$,

$$\sigma_{hy}^2 = \frac{N_h - 1}{N_h} S_{hy}^2 = \frac{1}{N_h} \sum_{k=1}^{N_h} (y_{hk} - \bar{y}_h)^2,$$

etc.

El estimador "combinado de razón" en muestreo estratificado es aquél que usa del estimador de la razón al cociente de los estimadores estratificados para la variable $y$ y el de la $x$. Su expresión es

$$t_{CR} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{x}.$$

Debido a que tenemos la aproximación

$$\frac{\bar{y}_{st}}{\bar{x}_{st}} - R \approx \frac{\bar{y}_{st} - R\bar{x}_{st}}{\bar{x}},$$

el estimador $t_{CR}$ es aproximadamente insesgado y su varianza es también aproximadamente

$$V(t_{CR}) \approx V(\bar{y}_{st} - R\bar{x}_{st}) =$$

$$V(\bar{y}_{st}) + R^2 V(\bar{x}_{st}) - 2RCov(\bar{y}_{st}, \bar{x}_{st}),$$

que con diseño *mas* independientemente en cada estrato será igual a

$$\sum_{h=1}^{L} W_h^2 \frac{1}{n_h}\left(\sigma_{hy}^2 + R^2\sigma_{hx}^2 - 2R\sigma_{hyx}\right),$$

mientras que si el diseño fuera *mia* independientemente en cada estrato será igual a

$$\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h n_h}\left(S_{hy}^2 + R^2 S_{hx}^2 - 2RS_{hyx}\right),$$

con $R = \bar{y}/\bar{x}$.

## 6.6 Estimador de producto de la media poblacional

Es un estimador semejante al de razón, pero su uso se limita a cuando existe una relación de proporcionalidad inversa entre la variable de interés y la auxiliar, o bien hay estabilidad en los productos $y_k x_k$ $(k = 1, 2, \ldots, N)$. En estos casos se propone el "estimador de producto" que se define como

$$t_P = \bar{y}_s \frac{\bar{x}_s}{\bar{x}}.$$

Para conocer sus características, lo expresamos del modo

$$t_P = \frac{1}{\bar{x}}(\bar{y}_s - \bar{y} + \bar{y})(\bar{x}_s - \bar{x} + \bar{x}) =$$

$$\bar{y}\left(1 + \frac{\bar{y}_s - \bar{y}}{\bar{y}}\right)\left(1 + \frac{x_s - \bar{x}}{\bar{x}}\right) =$$

$$\bar{y}\left(1 + \frac{\bar{y}_s - \bar{y}}{\bar{y}} + \frac{x_s - \bar{x}}{\bar{x}} + \frac{\bar{y}_s - \bar{y}}{\bar{y}}\frac{x_s - \bar{x}}{\bar{x}}\right),$$

de donde

$$E(t_P) = \bar{y} + \frac{Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}},$$

es decir, el sesgo del estimador de producto $t_P$ para estimar la media poblacional $\bar{y}$ es

$$B(t_P; \bar{y}) = \frac{Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}},$$

que admite distintas expresiones según sea el diseño muestral usado. De la fórmula extendida de $t_P$, tenemos

$$t_P - \bar{y} = \bar{y}\left(\frac{\bar{y}_s - \bar{y}}{\bar{y}} + \frac{x_s - \bar{x}}{\bar{x}} + \frac{\bar{y}_s - \bar{y}}{\bar{y}}\frac{x_s - \bar{x}}{\bar{x}}\right),$$

de donde aproximando por los términos cuadráticos tenemos

$$(t_P - \bar{y})^2 \approx (\bar{y}_s - \bar{y})^2 + 2R(\bar{y}_s - \bar{y})(x_s - \bar{x}) + R^2(x_s - \bar{x})^2,$$

de donde

$$ECM(t_P; \bar{y}) = E[(t_P - \bar{y})^2] \approx$$

$$V(\bar{y}_s) + 2RCov(\bar{y}_s, \bar{x}_s) + R^2V(\bar{x}_s),$$

que toma distintas expresiones según el diseño muestral usado. Haciendo uso de esta aproximación (y considerándola como igualdad), el estimador de producto tiene menor o igual varianza que la media muestral en los diseños *mas* y *mia* siempre y cuando el coeficiente de correlación entre las variables $y$ y $x$ verifica

$$\rho \leq -\frac{R\sigma_x}{2\sigma_y}.$$

Es inmediato proponer estimadores de producto separado y combinado en muestreo estratificado.

## 6.7 Estimador de regresión de la media poblacional

Cuando los puntos $(y_k, x_k)$ con $k = 1, 2, \dots, N$, donde $y$ es la variable de interés y $x$ es la variable auxiliar, están situados sobre una línea recta que pasa por el origen $y_k = ax_k$, el estimador de razón es el más indicado. Si fueran los productos $y_k x_k = a$ los que fueran estables entorno al valor constante $a$, el estimador de producto es el más indicado. Pero si la relación es lineal del tipo

$$y_k = a + bx_k$$

o línea recta que no pasa por el origen $(a \neq 0)$ aunque puede pasar por el origen también $(a = 0)$, entonces el "estimador de regresión lineal" para la media poblacional $\bar{y}$ se obtiene razonando de este modo. Por un lado se dará

$$\bar{y} = a + b\bar{x}$$

e

$$\bar{y}_s = a + b\bar{x}_s,$$

por lo que restando la segunda de la primera igualdad,

$$\bar{y} - \bar{y}_s = b(\bar{x} - \bar{x}_s),$$

de donde se propone como estimador de regresión lineal a

$$\bar{y}_{rl} = \bar{y}_s + b(\bar{x} - \bar{x}_s),$$

siendo $b$ una variable aleatoria. Su sesgo para estimar la media poblacional se obtiene de que

$$E(\bar{y}_{rl}) = \bar{y} + \bar{x}E(b) - E(b\bar{x}_s) = \bar{y} - Cov(b, \bar{x}_s),$$

es decir

$$B(\bar{y}_{rl};\ \bar{y}) = E(\bar{y}_{rl}) - \bar{y} = -Cov(b, \bar{x}_s).$$

Si $b$ fuera constante, el valor de $b$ que minimiza la varianza del estimador de regresión se obtiene de este modo.

$$V(\bar{y}_{rl}) = V(\bar{y}_s) + b^2 V(\bar{x}_s) - 2bCov(\bar{y}_s, \bar{x}_s),$$

que con diseño *mas* admite la expresión

$$V(\bar{y}_{rl}) = \frac{1}{n}\left(\sigma_y^2 + b^2\sigma_x^2 - 2b\sigma_{yx}\right),$$

y con diseño *mia* se expresa del modo

$$V(\bar{y}_{rl}) = \frac{N-n}{Nn}\left(S_y^2 + b^2 S_x^2 - 2bS_{yx}\right).$$

Llamando $f(b) = V(\bar{y}_{rl})$, la función alcanza su mínimo cuando $f'(b) = 0$, o bien

$$b = \frac{\sigma_{yx}}{\sigma_x^2} = \frac{S_{yx}}{S_x^2}.$$

Además es mínimo pues si $n < N$ y $\sigma_x^2 > 0$, $f''(b) > 0$. Para este valor mínimo de $b$ la varianza toma el valor

$$V_{\text{mín}}(\bar{y}_{rl}) = \frac{\sigma_y^2}{n}(1 - \rho^2)$$

en el caso de diseño *mas*, y

$$V_{\text{mín}}(\bar{y}_{rl}) = \frac{N-n}{Nn} S_y^2 (1-\rho^2)$$

en el caso de diseño *mia*. En realidad el valor mínimo de $b$ así obtenido es una función paramétrica que sería desconocida antes de realizarse el muestreo, y estimable después del muestreo, por lo que para aplicar estos resultados tendremos que estimar $b$ por su estimador mínimo-cuadrático (según Cochran, 1977)

$$\hat{b} = \frac{\sum_{k\in s}(y_k - \bar{y}_s)(x_k - \bar{x}_s)}{\sum_{k\in s}(x_k - \bar{x}_s)^2} = \frac{s_{yx}}{s_x^2}.$$

## 6.8 Comparación de precisiones

Una vez obtenidas la varianza exacta para las medias muestrales con diseño *mas* y *mia*, y las correspondientes aproximadas para los estimadores de razón, de producto y de regresión, podemos concluir que el estimador de regresión lineal teórico es más preciso que la media muestral siempre que $\rho \neq 0$, y en el caso $\rho = 0$ las varianzas coinciden (tomando como varianza del estimador de razón a su aproximación). El estimador de regresión lineal teórico es también más preciso que el estimador de razón siempre, y sus varianzas (aproximada en el caso de razón) coinciden cuando

$$R = \rho \frac{\sigma_y}{\sigma_x}.$$

El estimador de regresión es también más preciso que el estimador de producto siempre (si la varianza aproximada de este fuera una igualdad), y sus errores cuadráticos medios aproximados coincidirían cuando

$$R = -\rho \frac{\sigma_y}{\sigma_x}.$$

## 6.9 El estimador de regresión con estratificación

El estimador de regresión separado en el muestreo estratificado es

$$\bar{y}_{rls} = \sum_{h=1}^{L} W_h \bar{y}_{rlh},$$

donde el estimador $\bar{y}_{rlh}$ es el estimador de regresión lineal en el estrato $h$-ésimo. Su uso requiere estimar $L$ valores $\hat{b}_h$, uno por cada estrato. El estimador de regresión combinado es

$$\bar{y}_{rlc} = \bar{y}_{st} + b(\bar{x} - \bar{x}_{st}),$$

el cual requiere estimar un solo valor de $b$.

## 6.10 Ejercicios resueltos

**Ejercicio 6.1.** Se desea estimar la producción de trigo total en cierta comarca. Para ello se toma como unidad de muestreo la parcela dedicada a dicho cultivo, y se conoce como variable auxiliar la superficie de terreno de las parcelas individualmente. Si se supone que la producción de trigo es proporcional a la superficie sembrada en cada parcela o unidad, justificar que el estimador de razón es un indicado estimador para estimar la producción total de trigo en la comarca.

**Solución.** Si existe una relación de proporcionalidad aproximada $y_k = cx_k$ $(k = 1, 2, \dots, N)$, siendo $N$ el número total de parcelas sembradas en la comarca de trigo, $y_k$ la producción de trigo en la

unidad $k$, $x_k$ la superficie sembrada de trigo en la unidad $k$, y $c$ la constante de proporcionalidad, tenemos que el estimador del total de trigo producido de razón es

$$Nt_R = N\bar{y}_s \frac{\bar{x}}{\bar{x}_s},$$

pero como $\bar{y}_s = c\bar{x}_s$ aproximadamente,

$$Nt_R = Nc\bar{x}_s \frac{\bar{x}}{\bar{x}_s} = Nc\bar{x} = N\bar{y},$$

con lo que queda demostrada su adecuación pues

$$c\bar{x} = c\frac{\sum_{k=1}^{N} x_k}{N} = \frac{\sum_{k=1}^{N} cx_k}{N} = \frac{\sum_{k=1}^{N} y_k}{N} = \bar{y},$$

siendo $N\bar{y}$ el total de trigo producido en la comarca.

**Ejercicio 6.2.** Determinar el tamaño muestral $n$ necesario para que el estimador de razón $t_R = (\bar{y}_s\bar{x})/\bar{x}_s$, de la media poblacional $\bar{y}$ de cierta variable de interés $y$, difiera de tal función paramétrica menos que 5 al nivel de confianza del 95%. Además $N = 1000$, y de una muestra piloto se estima que $S_y^2 = 30$, $R = 2$, $S_x^2 = 15$ y $S_{xy} = 3$.

**Solución.**

$$n \approx \frac{1}{\frac{1}{N} + \frac{\alpha e^2}{S_y^2 + R^2 S_x^2 - 2RS_{yx}}} \approx 59.$$

**Ejercicio 6.3.** Para estimar el consumo medio de las familias de un país se ha utilizado el estimador de razón con la variable auxiliar

"renta familiar". Indicar la conveniencia o no de tal estimador para tal objetivo.

**Respuesta.** Razonando como en el ejercicio 6.1, el estimador de razón será adecuado cuando exista una proporcionalidad entre "consumo familiar" y "renta familiar" en tal país. Es decir, el estimador de razón es deseable cuando la dependencia entre el consumo y la renta familiares sea aproximadamente una línea recta que además pase por el origen. Si la dependencia es lineal pero la recta no pasa por el origen, puede utilizarse el estimador de regresión razonando de modo análogo, como se hará en el ejercicio 6.5.

**Ejercicio 6.4.** La experiencia de unos directivos de unos grandes almacenes les hace admitir que las ventas de cierto producto en un día es inversamente proporcional a su precio de venta al público. En esta situación qué estimador propondría, como asesor de la empresa, para la venta media mensual pudiendo conocer las ventas en 5 días diferentes seleccionados con diseño *mia*, y sabiendo los precios de venta de los 25 días que abre al público dichos almacenes en ese mes.

**Respuesta.** Llamamos $y_k$ a las ventas del producto en el día $k$ $(k = 1, 2, \dots, 25)$, y $x_k$ al precio del producto en ese mismo día. Admitimos la relación aproximada $y_k x_k = c$, según nos informan los directivos. Queremos estimar la venta media mensual del producto a lo largo de los 25 días laborables del mes. Tal media poblacional es denotada por $\bar{y}$, y la media muestral de ventas en los 5 días de observación es $\bar{y}_s$. El precio medio mensual del producto es $\bar{x}$, y la media muestral de precios es $\bar{x}_s$. En estas condiciones y ya que se da una relación aproximada entre $y_k$ y $x_k$ de proporcionalidad inversa, un estimador deseable de $\bar{y}$ es el estimador de producto

$$t_P = \bar{y}_s \frac{\bar{x}_s}{\bar{x}} \approx \frac{c}{\bar{x}} \approx \bar{y}.$$

**Ejercicio 6.5.** En una urbanización de $N$ viviendas se dispone de la información auxiliar $x$ número de residentes por vivienda. Se sabe además que se verifica que la superficie en metros cuadrados de las viviendas (variable de interés $y$) mantiene una relación próxima a la lineal, $y_k = a + bx_k \ (k = 1, 2, \dots, N)$. Estudiar la conveniencia del estimador de regresión lineal para estimar la superficie media de las viviendas de dicha urbanización.

**Solución.** Como tenemos la relación lineal entre las variables de interés y la auxiliar, también $\bar{y}_s = a + b\bar{x}_s$ y podemos escribir

$$\bar{y}_{rl} = \bar{y}_s + \frac{\sum_{k\in s}(y_k - \bar{y}_s)(x_k - \bar{x}_s)}{\sum_{k\in s}(x_k - \bar{x}_s)^2}(\bar{x} - \bar{x}_s) =$$

$$a + b\bar{x}_s + \frac{\sum_{k\in s}(a + bx_k - a - b\bar{x}_s)(x_k - \bar{x}_s)}{\sum_{k\in s}(x_k - \bar{x}_s)^2}(\bar{x} - \bar{x}_s) =$$

$$a + b\bar{x}_s + b(\bar{x} - \bar{x}_s) = a + b\bar{x} = \bar{y}.$$

Luego el estimador de regresión lineal es un estimador adecuado si se da la relación aproximada lineal $y = a + bx$ en las unidades de la población finita.

**Ejercicio 6.6.** Si se sabe que el coeficiente de correlación lineal es $\rho_{yx} = 0.4$, determinar la ganancia en precisión del estimador de regresión lineal con respecto a la media muestral, ambas con diseño *mia*.

**Solución.** Como

$$V(\bar{y}_s) = \frac{N-n}{Nn} S_y^2$$

y

$$V(\bar{y}_{rl}) = \frac{N-n}{Nn} S_y^2 (1-\rho^2) = 0.84 V(\bar{y}_s),$$

la ganancia en "precisión" (inversa de la "varianza") será

$$\frac{1}{V(\bar{y}_{rl})} - \frac{1}{V(\bar{y}_s)} = \frac{1}{V(\bar{y}_s)} \left( \frac{1}{0.84} - 1 \right) = 0.19 \frac{1}{V(\bar{y}_s)} > 0,$$

es decir, hay ganancia en precisión positiva del estimador de regresión lineal óptimo teórico respecto del estimador media muestral ambos con diseño *mia*.

**Ejercicio 6.7.** Estimar la media poblacional $\bar{y}$ por el método de regresión lineal sabiendo que se dispone de los siguientes datos. Medias muestrales, de la variable de interés $5$, de la variable auxiliar $3$. Media poblacional de la variable auxiliar $4$. Estimador mínimo-cuadrático del coeficiente de regresión lineal $2$.

**Solución.** El estimador de regresión lineal es

$$\bar{y}_{rl} = \bar{y}_s + \hat{b}(\bar{x} - \bar{x}_s) = 5 + 2(4-3) = 7.$$

**Ejercicio 6.8.** Obtener el sesgo del estimador media de razones,

$$\bar{r}_s = \frac{1}{n} \sum_{k \in s} \frac{y_k}{x_k},$$

como estimador de la razón $R = \bar{y}/\bar{x}$ con muestreo irrestricto aleatorio de tamaño efectivo fijo $n$, de una población finita $U$ de

tamaño $N$ . Obtener estimadores insesgados de la media poblacional $\bar{y}$ basados en el estimador $\bar{r}_s$ y en el sesgo obtenido.

**Solución.** El sesgo de Hartley y Ross de $r_s$ para estimar $R$ es

$$B_{HR}(\bar{r}_s) = E(\bar{r}_s) - R = \frac{1}{N}\sum_{i\in U}\frac{y_i}{x_i} - \frac{1}{N}\sum_{i\in U}\frac{y_i}{\bar{x}} =$$

$$\frac{1}{N}\sum_{i\in U} y_i\left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right) = -\frac{1}{N\bar{x}}\sum_{i\in U}\frac{y_i}{x_i}(x_i - \bar{x}) = -\frac{Cov\left(\frac{y_i}{x_i}, x_i\right)}{\bar{x}}.$$

El estimador insesgado de Hartley y Ross de la media poblacional $\bar{y}$ es entonces

$$t_{HR} = \bar{x}\bar{r}_s - \bar{x}\hat{B}_{HR}(\bar{r}_s) = \bar{x}\bar{r}_s + \widehat{Cov}\left(\frac{y_i}{x_i}, x_i\right) =$$

$$\bar{x}\bar{r}_s + \frac{N-1}{N(n-1)}\sum_{i\in s}\frac{y_i}{x_i}(x_i - \bar{x}_s) = \bar{x}\bar{r}_s + \frac{(N-1)n}{N(n-1)}(\bar{y}_s - \bar{x}_s\bar{r}_s).$$

Otro estimador insesgado de la media poblacional puede obtenerse del estimador insesgado de la covarianza

$$\widehat{Cov}\left(\frac{y_i}{x_i}, x_i\right) = \frac{1}{n}\sum_{i\in s}\frac{y_i}{x_i}(x_i - \bar{x}) = \bar{y}_s - \bar{x}\bar{r}_s,$$

ya que $\bar{x}$ es una magnitud poblacional conocida, que proporciona al estimador insesgado de la media poblacional a la media muestral $\bar{y}_s$.

Otra forma de calcular el sesgo de $\bar{r}_s$ para estimar $R$ es

$$B(\bar{r}_s) = E(\bar{r}_s) - R = E(\bar{r}_s) - \frac{E(\bar{y}_s)}{E(\bar{x}_s)} =$$

$$\frac{E(\bar{r}_s)E(\bar{x}_s) - E(\bar{y}_s)}{\bar{x}} =$$

$$\frac{E(\bar{r}_s)E(\bar{x}_s) - E(\bar{r}_s\bar{x}_s) + E(\bar{r}_s\bar{x}_s) - E(\bar{y}_s)}{\bar{x}} =$$

$$\frac{-Cov(\bar{r}_s, \bar{x}_s) + E(\bar{r}_s\bar{x}_s - \bar{y}_s)}{\bar{x}}.$$

Podemos considerar como estimador insesgado de $E(\bar{r}_s\bar{x}_s - \bar{y}_s)$ al estadístico

$$\bar{r}_s\bar{x}_s - \bar{y}_s = \frac{-t_{HR} + \bar{x}\bar{r}_s}{(N-1)n}N(n-1),$$

siendo $t_{HR}$ el estimador insesgado de la media poblacional propuesto por Hartley y Ross. Como $E(t_{HR}) = E(\bar{y}_s) = \bar{y}$ , igualando los estimadores insesgados obtenidos de los sesgos de Hartley y Ross con el correspondiente a la otra expresión del sesgo de $\bar{r}_s$ podemos construir un estimador insesgado de la covarianza $Cov(\bar{r}_s, \bar{x}_s)$ del modo

$$-\frac{\bar{y}_s}{\bar{x}} + \bar{r}_s = \frac{1}{\bar{x}}\left[-\widehat{Cov}(\bar{r}_s, \bar{x}_s) + \frac{-\bar{y}_s + \bar{x}\bar{r}_s}{(N-1)n}N(n-1)\right],$$

de donde

$$\widehat{Cov}(\bar{r}_s, \bar{x}_s) = \frac{N-n}{(N-1)n}(\bar{y}_s - \bar{x}\bar{r}_s).$$

Sustituyendo este estimador insesgado en la fórmula del estimador insesgado del sesgo de $\bar{r}_s$, tenemos

$$\hat{B}(\bar{r}_s) = \frac{1}{\bar{x}}\left[\frac{N-n}{(N-1)n}(\bar{x}\bar{r}_s - \bar{y}_s) + \bar{x}_s\bar{r}_s - \bar{y}_s\right].$$

Por tanto otro estimador insesgado de la media poblacional es

$$t = \bar{x}\left[\bar{r}_s - \hat{B}(\bar{r}_s)\right] =$$

$$\frac{N(n-1)}{(N-1)n}\bar{x}\bar{r}_s + \frac{N(n+1)-2n}{(N-1)n}\bar{y}_s - \bar{x}_s\bar{r}_s.$$

De este estimador y del de Hartley y Ross, podemos obtener otro también insesgado de la media poblacional que no depende de la media muestral $\bar{y}_s$ en el muestreo irrestricto aleatorio de tamaño efectivo fijo $n$,

$$t^* = \frac{n(N-1)}{N-n}\bar{x}_s\bar{r}_s - \frac{(n-1)N}{N-n}\bar{x}\bar{r}_s.$$

**Ejercicio 6.9.** Definimos el estimador producto del tipo

$$t_P = \frac{\bar{x}_s\bar{y}_s}{m},$$

donde el denominador es la media armónica equiprobable entre las muestras conjunto o no ordenadas de tamaño efectivo fijo $n$ de una población finita de tamaño $N$, y $x$ es una variable auxiliar positiva. Comprobar que el diseño muestral $p$ definido para toda muestra conjunto $s$ de tamaño efectivo fijo $n$,

$$p(s) = \frac{m}{\binom{N}{n}\bar{x}_s} > 0,$$

proporciona un estimador insesgado de la media poblacional $\bar{y}$. Finalmente obtener la expresión de su varianza.

**Solución.** Veamos primero que $p$ es un diseño muestral no ordenado definido sobre el conjunto $S = \{s: s \subset U, \text{card}(s) = n\}$, siendo $U$ el conjunto de $N$ unidades de la población finita. En efecto, para toda muestra conjunto de tamaño muestral efectivo $n$,

$$p(s) = \frac{\binom{N}{n}}{\sum_{s\in S}\frac{1}{\bar{x}_s}}\frac{1}{\binom{N}{n}\bar{x}_s} = \frac{\frac{1}{\bar{x}_s}}{\sum_{s\in S}\frac{1}{\bar{x}_s}} > 0$$

por ser $x$ una variable positiva. Además,

$$\sum_{s\in S} p(s) = \sum_{s\in S}\frac{\frac{1}{\bar{x}_s}}{\sum_{s\in S}\frac{1}{\bar{x}_s}} = \frac{\sum_{s\in S}\frac{1}{\bar{x}_s}}{\sum_{s\in S}\frac{1}{\bar{x}_s}} = 1.$$

La esperanza matemática de $t_P$ es

$$E(t_P) = \sum_{s\in S} t_P(d)p(s) = \sum_{s\in S}\frac{\bar{x}_s\bar{y}_s}{m}\frac{m}{\binom{N}{n}\bar{x}_s} = \frac{1}{\binom{N}{n}}\sum_{s\in S}\bar{y}_s = \bar{y},$$

por ser la media muestral $\bar{y}_s$ de tamaño efectivo fijo $n$ un estadístico insesgado de la media poblacional con diseño de muestreo irrestricto aleatorio.

La varianza de $t_P$ es

$$V(t_P) = E(t_P^2) - [E(t_P)]^2 = \sum_{s\in S}\frac{\bar{x}_s^2\bar{y}_s^2}{m^2}\frac{m}{\binom{N}{n}\bar{x}_s} - \bar{y}^2 =$$

$$\frac{1}{m\binom{N}{n}}\sum_{s\in S}\bar{x}_s\bar{y}_s^2 - \bar{y}^2,$$

que es estimable sin sesgo si y solo si la probabilidad de inclusión de todas las unidades $i$ y $j$ es positiva, lo cual se verifica para cualquier diseño muestral $p$ con $n \geq 2$.

**Ejercicio 6.10.** Proponer un diseño muestral no ordenado para el que el estimador de razón usual de la media poblacional resulte insesgado con una variable auxiliar positiva, y calcular su varianza.

**Solución.** El estimador de razón usual es

$$t_R = \frac{\bar{y}_s}{\bar{x}_s}\bar{x}.$$

Un diseño muestral de tamaño efectivo fijo $n$ que hace de este estimador insesgado para la media poblacional $\bar{y}$ es

$$p(s) = \frac{\bar{x}_s}{\bar{x}\binom{N}{n}} > 0.$$

Además,

$$\sum_{s\in S} p(s) = \frac{1}{\bar{x}\binom{N}{n}} \sum_{s\in S} \bar{x}_s = 1,$$

por ser la media muestral de tamaño efectivo fijo $n$ insesgada para la media poblacional en muestreo irrestricto aleatorio.

Veamos $t_R$ que es insesgado con este diseño muestral. En efecto,

$$E(t_R) = \sum_{s\in S} \frac{\bar{y}_s \bar{x}}{\bar{x}_s} \frac{\bar{x}_s}{\bar{x}\binom{N}{n}} = E(\bar{y}_s) = \bar{y}.$$

La varianza de $t_R$ con este diseño muestral es

$$V(t_R) = E(t_R^2) - [E(t_R)]^2 =$$

$$\sum_{s\in S} \frac{\bar{y}_s^2 \bar{x}^2}{\bar{x}_s^2} \frac{\bar{x}_s}{\bar{x}\binom{N}{n}} - \bar{y}^2 = \frac{\bar{x}}{\binom{N}{n}} \sum_{s\in S} \frac{\bar{y}_s^2}{\bar{x}_s} - \bar{y}^2.$$

**Ejercicio 6.11.** Una estrategia muestral $(p_1, t_1)$ es insesgada para estimar el parámetro poblacional $M_1$. Además $t_2$ es otro estadístico positivo cuya media de sus inversas ponderadas por $p_1$ es $m_2$. Demostrar que el estimador

$$t = \frac{t_1 t_2}{m_2}$$

con el nuevo diseño muestral

$$p(s) = \frac{1}{m_2 t_2} p_1(s)$$

constituye una nueva estrategia $(p, t)$ insesgada para estimar $M_1$. Obtener una expresión de su varianza.

**Solución.** El diseño muestral $p$ lo es porque si $t_2$ es positivo para toda muestra $s \in S$, la media de sus inversas ponderadas

$$m_2 = \sum_{s \in S} \frac{1}{t_2(d)} p_1(s) > 0,$$

donde $d$ es el dato muestral asociado a la muestra $s$. Por tanto, el diseño muestral $p$ verifica las condiciones

$$p(s) = \frac{1}{m_2 t_2(d)} p_1(s) \geq 0$$

y

$$\sum_{s \in S} p(s) = \frac{1}{m_2} \sum_{s \in S} \frac{1}{t_2(d)} p_1(s) = 1.$$

Luego, la esperanza matemática de la estrategia $(p, t)$ es

$$E(p, t) = \sum_{s \in S} \frac{t_1(d) t_2(d)}{m_2} \frac{m_2}{t_2(d)} p_1(s) = M_1.$$

Por lo que $(p, t)$ es una nueva estrategia insesgada de $M_1$.

Su varianza es

$$V(p,t) = \sum_{s \in S} \frac{t_1^2(d) t_2^2(d)}{m_2^2} \frac{m_2}{t_2(d)} p_1(s) - M_1^2 =$$

$$\sum_{s \in S} \frac{t_1^2(d) t_2(d)}{m_2} p_1(s) - M_1^2.$$

**Ejercicio 6.12.** Si $(p, t_i)$ es una estrategia insesgada para estimar $M_i$, con $i = 1, 2$, proponer otra estrategia insesgada de $M_1$ del tipo de razón basada en las primeras, suponiendo que $t_2$ es un estadístico positivo.

**Solución.** La estrategia que proponemos es $(p', t')$, donde

$$t'(d) = \frac{t_1(d)}{t_2(d)} M_2,$$

y el nuevo diseño muestral es

$$p'(s) = \frac{t_2(d)}{M_2} p(s),$$

que es positivo para toda muestra $s$, y verifica que

$$\sum_{s \in S} p'(s) = \frac{M_2}{M_2} = 1.$$

Luego,

$$E(p', t') = \sum_{s \in S} \frac{t_1(d)}{t_2(d)} M_2 \frac{t_2(d)}{M_2} p(s) = M_1.$$

La varianza de esta estrategia es

$$V(p', t') = M_2 \sum_{s \in S} \frac{t_1^2(d)}{t_2(d)} p(s) - M_1^2.$$

**Ejercicio 6.13.** Proponemos como estimador de regresión modificado con diseño de muestreo irrestricto aleatorio a

$$t = \bar{y}_s + \frac{\frac{1}{n}\sum_{i \in s} y_i (x_i - \bar{x})}{\mu_{02}} (\bar{x} - \bar{x}_s),$$

con $\mu_{02}$ la varianza poblacional de la variable auxiliar $x$. Nótese que el término

$$\frac{\frac{1}{n}\sum_{i \in s} y_i (x_i - \bar{x})}{\mu_{02}}$$

es un estimador insesgado del coeficiente $\mu_{11}/\mu_{02}$ que minimiza la varianza del estimador de regresión lineal usual si el coeficiente $b$ fuera una constante. Obtener la esperanza matemática del estimador propuesto $t$ y estimar insesgadamente el sesgo de $t$.

**Solución.**

$$E(t) = \alpha_{10} + \frac{E\left\{\left[\frac{1}{n}\sum_{i \in s} y_i (x_i - \bar{x})\right](\bar{x} - \bar{x}_s)\right\}}{\mu_{02}} =$$

$$\alpha_{10} - \frac{Cov(\bar{p}_s, \bar{x}_s)}{\mu_{02}}.$$

Aquí hemos denotado por $\bar{p}_s$ a la media muestral de los valores de la variable $p_i = y_i(x_i - \bar{x})$. Por lo que un estimador insesgado del parámetro $Cov(\bar{p}_s, \bar{x}_s)$ es

$$\frac{N-n}{(N-1)n^2}\sum_{i\in s} y_i(x_i-\bar{x})^2,$$

o bien,

$$\frac{N-n}{Nn}s_{px},$$

siendo $s_{px}$ la cuasicovarianza muestral de las variables $p$ y $x$. También podíamos haber procedido así:

$$E\left\{\left[\frac{1}{n}\sum_{i\in s} y_i(x_i-\bar{x})\right](\bar{x}-\bar{x}_s)\right\}=$$

$$E\left[\frac{1}{n}(na_{11}-na_{10}\alpha_{01})(\alpha_{01}-a_{01})\right]=$$

$$\alpha_{01}E(a_{11}-a_{10}\alpha_{01})-E(a_{11}a_{01}-a_{10}a_{01}\alpha_{01})=$$

$$\alpha_{01}\mu_{11}-E(a_{11}a_{01}-a_{10}a_{01}\alpha_{01}).$$

Por lo que, sustituyendo en el numerador de la fracción que explica la esperanza matemática de $t$, tenemos la esperanza buscada.

Un estimador insesgado del sesgo de $t$ es

$$\hat{B}(t)=\frac{\alpha_{01}\hat{\mu}_{11}-(a_{11}a_{01}-a_{10}a_{01}\alpha_{01})}{\mu_{02}},$$

o bien, directamente otra expresión de un estimador insesgado del sesgo de $t$ es

$$\hat{B}(t)=\frac{\frac{1}{n}[\sum_{i\in s} y_i(x_i-\bar{x})](\bar{x}-\bar{x}_s)}{\mu_{02}}=$$

$$\frac{(a_{11}-a_{10}\alpha_{01})(\alpha_{01}-a_{01})}{\mu_{02}}.$$

O bien,

$$\hat{B}(t) = -\frac{\frac{N-n}{Nn} s_{px}}{\mu_{02}}.$$

O también, la más recomendable, al disponer de un mayor grado de libertad en la estimación que la anterior,

$$\hat{B}(t) = -\frac{\frac{N-n}{(N-1)n^2} \sum_{i \in s} y_i (x_i - \bar{x})^2}{\mu_{02}}.$$

Como consecuencia, el estimador $t - \hat{B}(t)$ es insesgado para estimar la media poblacional $\bar{y}$.

**Ejercicio 6.14.** El estimador diferencia, con muestreo irrestricto aleatorio de tamaño efectivo fijo $n$, se define como

$$t_D = \bar{y}_s + \bar{x} - \bar{x}_s.$$

Justificar que este estimador es insesgado para estimar la media poblacional $\bar{y}$. Obtener su varianza, y estimarla sin sesgo.

**Solución.** Es insesgado $t_D$ pues

$$E(t_D) = E(\bar{y}_s) + \bar{x} - E(\bar{x}_s) = \bar{y} - \bar{x} + \bar{x} = \bar{y}.$$

Su varianza es

$$V(t_D) = V(\bar{y}_s) + V(\bar{x}_s) - 2Cov(\bar{y}_s, \bar{x}_s) =$$

$$\frac{N-n}{Nn} S_y^2 + \frac{N-n}{Nn} S_x^2 - 2\frac{N-n}{Nn} S_{yx} =$$

$$\frac{N-n}{Nn} \left(S_y^2 + S_x^2 - 2S_{yx}\right).$$

Un estimador insesgado de la varianza $V(t_D)$ es

$$\hat{V}(t_D) = \frac{N-n}{Nn}\left(s_y^2 + s_x^2 - 2s_{yx}\right),$$

donde aparecen cuasivarianzas muestrales y la cuasicovarianza muestral de las variables $y$ y $x$.

**Ejercicio 6.15.** Disponemos de dos estimadores de la media poblacional $\bar{y}$, el estimador diferencia $t_D = \bar{y}_s + \bar{x} - \bar{x}_s$, y el estimador suma que definimos $t_S = \bar{y}_s + \bar{x}_s - \bar{x}$. En ellos, $x$ es una variable auxiliar. Obtener sus varianzas y compararlas. ¿Podemos saber con los datos de la muestra cuál estimador es mejor en precisión en un caso práctico?

**Solución.** Ambos estimadores $t_D$ y $t_S$ son insesgados para estimar la media poblacional $\bar{y}$. El estimador diferencia es más preciso que el estimador suma si y solo si

$$V(t_D) < V(t_S).$$

El concepto de "precisión de un estimador insesgado" es la inversa de su varianza. La condición anterior es equivalente en muestreo aleatorio simple de tamaño fijo $n$ a que

$$Cov(\bar{y}_s, \bar{x}_s) = \frac{\mu_{11}}{n} > 0,$$

que equivale a que

$$\mu_{11} > 0.$$

En ellas hemos denotado por $\mu_{11}$ a la covarianza poblacional de las variables $y$ y $x$, parámetro desconocido que puede ser estimado insesgadamente, en muestreo aleatorio simple de tamaño fijo $n$, por el estimador

$$\hat{\mu}_{11} = \frac{1}{n}\sum_{i \in \boldsymbol{s}} y_i(x_i - \bar{x}).$$

Si este estimador fuese positivo, como estima insesgadamente a la covarianza poblacional (y por ser una media muestral, su varianza en muestreo aleatorio simple es una constante por $n^{-1}$), indicará que aproximadamente dicha covarianza poblacional es positiva también, y por tanto el estimador diferencia es mejor, desde el punto de vista de su precisión aproximada, que el estimador suma. Otro estimador insesgado de la covarianza poblacional en muestreo aleatorio simple es

$$\frac{n}{n-1} m_{11\boldsymbol{s}} = \frac{1}{n-1}\sum_{i \in \boldsymbol{s}} y_i(x_i - \bar{x}_{\boldsymbol{s}}),$$

que es la cuasicovarianza muestral en muestreo aleatorio simple de tamaño $n$, que tiene un grado de libertad menos que el anterior estimador insesgado.

En el razonamiento anterior podemos cambiar el sentido de todas las desigualdades para concluir cuando conviene elegir el estimador suma como mejor que el estimador diferencia. Finalmente, si hacemos de todas ellas igualdades, nos indicarán cuando ambos estimadores son equivalentes, y si $\hat{\mu}_{11} \approx 0$ de la muestra inferimos que ambos estimadores serían aproximadamente equivalentes o tendrían similar precisión.

Por último, si en lugar de muestreo aleatorio simple hubiéramos usado el muestreo irrestricto aleatorio de tamaño efectivo fijo $n$, la muestra, $s$, sería ahora un conjunto. Los razonamientos serían semejantes, teniendo en cuenta que ahora

$$Cov(\bar{y}_s, \bar{x}_s) = \frac{N-n}{Nn} S_{yx} = \frac{N-n}{(N-1)n}\mu_{11},$$

por lo que las comparaciones y desigualdades para elegir entre los estimadores diferencia y suma (sustituyendo la muestra ordenada $\boldsymbol{s}$ por la muestra conjunto $s$) siguen siendo válidas en el muestreo irrestricto aleatorio de tamaño efectivo fijo $n$. El estimador insesgado de la covarianza poblacional con el máximo grado de libertad es ahora

$$\hat{\mu}_{11} = \frac{1}{n}\sum_{i \in s} y_i(x_i - \bar{x}),$$

que es una media muestral. Otro estimador insesgado en muestreo irrestricto aleatorio, de la covarianza poblacional $\mu_{11}$, con un grado de libertad menos que el anterior, y recomendado cuando se desconoce $\bar{x}$, es

$$\frac{N-1}{N}\frac{n}{n-1}m_{11s} = \frac{N-1}{N(n-1)}\sum_{i \in s} y_i(x_i - \bar{x}_s).$$

**Ejercicio 6.16.** Los estimadores tradicionales en muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$ con información auxiliar, es decir el estimador de razón, el de regresión lineal y el de producto, ¿proporcionan estimaciones insesgadas de la media poblacional? Describir consecuencias de la respuesta anterior cuando los estimadores se aplican a cantidades económicas.

**Solución.** Como se estudia en la parte teórica de estimadores indirectos, los estimadores de razón, de regresión lineal y de producto son aproximadamente insesgados. Es decir, son insesgados para estimar la media poblacional bajo hipótesis que por lo general no se dan en la práctica, como una relación de proporcionalidad directa, una relación lineal, o una relación inversa

entre la variable a observar de interés y la variable auxiliar. De hecho, si estas relaciones se dieran, un solo dato daría estimaciones exactas en el caso de los estimadores de razón y de producto, mientras que con dos datos diferentes tendríamos estimaciones exactas en el caso del estimador de regresión.

La consecuencia es que estos tres estimadores son sesgados en condiciones generales, lo que se traduce en que estas "estimaciones" con datos económicos no son "justas en promedio" o son sesgados para el parámetro media poblacional a estimar.

**Ejercicio 6.17.** Corregir los estadísticos "producto de medias" y "media de productos" para conseguir con ellos estimadores insesgados de la media poblacional en muestreo irrestricto aleatorio de tamaño fijo $n$.

**Solución.** El estimador "producto de medias" está proporcionado por la fórmula

$$t_p = \frac{\bar{y}_s \bar{x}_s}{\bar{x}}.$$

Para muestreo irrestricto aleatorio o muestreo aleatorio simple sin reemplazamiento de tamaño muestral fijo $n$. El sesgo del estimador "producto de medias" $t_p$ es

$$B(t_p) = \frac{N-n}{(N-1)nA_{01}} M_{11},$$

donde $A_{01}$ es la media poblacional $\bar{x}$, y $M_{11}$ es la covarianza poblacional de las variables $y$ de interés y $x$ auxiliar. Un estimador insesgado de este sesgo es

$$\widehat{B_1}(t_p) = \frac{N-n}{(N-1)nA_{01}} \widehat{(M_{11})} = \frac{N-n}{N(n-1)A_{01}} m_{11},$$

donde $m_{11}$ es la covarianza muestral de las variables $y$ y $x$. Otro estimador insesgado del sesgo de $t_p$ es

$$\widehat{B_2}(t_p) = \frac{N-n}{(N-1)nA_{01}}(a_{11} - a_{10}A_{01}),$$

donde $a_{11}$ es la media muestral de los productos $y_i x_i$, o momento muestral con respecto al origen de órdenes 1 y 1. Entonces, dos estimadores "producto de medias" corregidos insesgados para la media poblacional $\bar{y}$ en muestreo irrestricto aleatorio de tamaño muestral fijo $n$ son

$$t_{up} = t_p - \widehat{B_1}(t_p) = \frac{a_{10}a_{01}}{A_{01}} - \frac{N-n}{N(n-1)A_{01}}m_{11}$$

y

$$t_{up}^* = t_p - \widehat{B_2}(t_p) = \frac{a_{10}a_{01}}{A_{01}} - \frac{N-n}{(N-1)nA_{01}}(a_{11} - a_{10}A_{01}).$$

Otro estimador tipo producto, basado en el estadístico "media de productos", es

$$t_{mp} = \frac{\frac{1}{n}\sum_{i\in s} y_i x_i}{\bar{x}} = \frac{a_{11}}{A_{01}},$$

el cual tiene el sesgo, para diseño de muestreo irrestricto aleatorio de tamaño muestral fijo $n$,

$$B(t_{mp}) = E(t_{mp}) - \bar{y} = \frac{A_{11}}{A_{01}} - A_{10} = \frac{M_{11}}{A_{01}}.$$

Este sesgo es estimable insesgadamente por

$$\widehat{B_1}(t_{mp}) = \frac{\widehat{(M_{11})}}{A_{01}} = \frac{(N-1)n}{N(n-1)A_{01}}m_{11}$$

y por

$$\widehat{B_2}(t_{mp}) = \frac{a_{11} - a_{10}A_{01}}{A_{01}} = \frac{a_{11}}{A_{01}} - a_{10}.$$

Así, se pueden elaborar dos estimadores corregidos insesgados basados en el estadístico "media de productos", concretamente

$$t_{ump} = t_{mp} - \widehat{B_1}(t_{mp}) = \frac{a_{11}}{A_{01}} - \frac{(N-1)n}{N(n-1)A_{01}} m_{11},$$

el cual es un estimador media de productos corregido insesgado, y

$$t^*_{ump} = t_{mp} - \widehat{B_2}(t_{mp}) = a_{10}.$$

Este último estimador corresponde directamente a la media muestral en el muestreo irrestricto aleatorio de tamaño muestral fijo $n$, que no depende del estadístico media de productos.

**Ejercicio 6.18.** Proponer un estimador insesgado de la media poblacional que aproveche el estadístico producto de medias muestrales, en muestreo irrestricto aleatorio.

**Solución.** El estimador producto $t_P$ fue propuesto por Murthy (1964) con diseño de muestreo aleatorio simple sin reemplazamiento de tamaño muestral efectivo $n$. El estimador producto, de la media poblacional $\bar{y} = (1/N)\sum_{i=1}^{N} y_i$, siendo $N$ el tamaño poblacional, puede expresarse del modo

$$t_P = \frac{\bar{y}_s \bar{x}_s}{\bar{x}}.$$

Aquí $\bar{y}_s = (1/n)\sum_{i\in s} y_i$ es la media muestral de la variable de interés $y$, $\bar{x}_s = (1/n)\sum_{i\in s} x_i$ es la media muestral de la variable auxiliar $x$ positiva en todas las unidades de la población finita, el subíndice $s$ es la muestra no ordenada o conjunto de tamaño

muestral efectivo o cardinal $n$, y $\bar{x} = (1/N)\sum_{i=1}^{N} x_i$ es la media poblacional de la variable auxiliar.

El objeto de este ejercicio es indicar que aunque el uso de información auxiliar en la estimación de funciones paramétricas es realizada en los artículos de investigación sobre el tema con mucha frecuencia con aproximaciones del sesgo y del error cuadrático medio, sin embargo en algunos casos, como los que presentamos, estos sesgos pueden corregirse. Además damos orientaciones útiles para el uso eficiente de estos procedimientos mediante métodos de estimación insesgada del valor óptimo del que depende el estimador insesgado de mínima varianza de una clase de dichos estimadores de la media poblacional.

Proponemos varias clases de estimadores producto corregidos insesgados. Unas clases de ellas son basadas en un estimador producto generalizado corregido insesgado en base al estimador producto generalizado, que era sesgado para la media poblacional y cuyo sesgo fue identificado con exactitud, debido a Ruiz Espejo (1991). Otras clases de estimadores producto corregidos insesgados han sido introducidas en este libro, que nos permitirán seleccionar un estimador óptimo teórico que depende de una función paramétrica desconocida antes del muestreo, pero esta función paramétrica puede ser estimable insesgadamente a partir de la misma muestra e información disponibles; lo que nos permite obtener estimadores aproximadamente óptimos. El estimador insesgado teóricamente óptimo tiene una varianza menor o igual a la “varianza del estimador producto corregido insesgado” debido al autor (2016c).

El estimador producto generalizado fue propuesto por Ruiz Espejo (1991) y su expresión es la siguiente

$$t_{PG} = \frac{\bar{y}_s[\bar{x}_s + k(\bar{x} - \bar{x}_s)]}{\bar{x}}.$$

Aquí $k$ es una constante. Este estimador general $t_{PG}$ es en general sesgado, y generaliza a la media muestral insesgada $\bar{y}_s$ (para $k = 1$) y al estimador producto sesgado $t_P$ (para $k = 0$).

Para obtener el estimador insesgado a partir del estimador $t_{PG}$ obtenemos la esperanza matemática de éste.

$$E(t_{PG}) = \frac{E(\bar{y}_s\bar{x}_s)}{\bar{x}} - k\frac{Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}}$$

$$= \frac{Cov(\bar{y}_s, \bar{x}_s) + \bar{y}\bar{x}}{\bar{x}} - k\frac{Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}} = \bar{y} + (1-k)\frac{Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}}.$$

De aquí, el sesgo de este estimador producto generalizado $t_{PG}$ es

$$B(t_{PG}) = E(t_{PG}) - \bar{y} = (1-k)\frac{Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}}.$$

Este sesgo puede ser estimado insesgadamente por

$$\hat{B}(t_{PG}) = (1-k)\frac{\widehat{Cov}(\bar{y}_s, \bar{x}_s)}{\bar{x}} = (1-k)\frac{N-n}{Nn\bar{x}}\hat{S}_{y,x}.$$

Aquí hemos denotado por $\widehat{Cov}(\bar{y}_s, \bar{x}_s)$ a un estimador insesgado de la covarianza

$$Cov(\bar{y}_s, \bar{x}_s) = \frac{N-n}{Nn}S_{y,x}.$$

Siendo $S_{y,x}$ la cuasicovarianza poblacional de las variables $y$ y $x$, es decir

$$S_{y,x} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}).$$

Esta cuasicovarianza poblacional $S_{y,x}$ puede ser estimada insesgadamente a partir de cualquiera de uno de estos dos estadísticos, con $n-1$ grados de libertad

$$s_{y,x} = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)(x_i - \bar{x}_s).$$

O bien, con $n$ grados de libertad

$$s'_{y,x} = \frac{N}{(N-1)n}\sum_{i\in s} y_i(x_i - \bar{x}).$$

Por tanto, un estimador insesgado de la cuasicovarianza poblacional $S_{y,x}$ puede ser $\hat{S}_{y,x} = s_{y,x}$ o bien $\hat{S}_{y,x} = s'_{y,x}$. Como consecuencia, una clase (al variar la constante $k$) de estimadores producto generalizado insesgados será

$$t_{PGu} = t_{PG} - \hat{B}(t_{PG}) = t_{PG} + (1-k)\frac{N-n}{Nn\bar{x}}s_{y,x}.$$

Y otra clase relacionada de estimadores producto generalizado insesgado es

$$t'_{PGu} = t_{PG} + (1-k)\frac{N-n}{Nn\bar{x}}s'_{y,x}.$$

De este modo queda corregido el sesgo del estimador producto generalizado. Sin embargo el valor exacto de $k$ que daría lugar a un estimador insesgado óptimo de cada clase de estimadores no es fácil de obtener ni de estimar. Por esto vamos a proponer otra clase de estimadores insesgados basada en el estimador producto generalizándolo de modo más sencillo y tratable estadísticamente.

Proponemos la siguiente clase de estimadores producto generalizado

$$t_{PG\prime} = \frac{\bar{y}_s \bar{x}_s + k(\bar{x} - \bar{x}_s)}{\bar{x}}.$$

Donde aquí, $k$ es una constante. Este estimador tiene por esperanza matemática

$$E(t_{PG\prime}) = \frac{E(\bar{y}_s \bar{x}_s)}{\bar{x}} = \frac{\bar{y}\bar{x} + Cov(\bar{y}_s, \bar{x}_s)}{\bar{x}} = \bar{y} + \frac{N-n}{Nn\bar{x}} S_{y,x}.$$

Por lo que un estimador insesgado, para todo valor constante posible de $k$, basado en el estimador $t_{PG\prime}$ es

$$t_{PG\prime u} = t_{PG\prime} - \frac{N-n}{Nn\bar{x}} s'_{y,x}.$$

La varianza del estimador producto generalizado insesgado $t_{PG\prime u}$ se obtiene a partir de

$$V(t_{PG\prime}) = \frac{1}{\bar{x}^2}[V(\bar{y}_s \bar{x}_s) + k^2 V(\bar{x}_s) - 2kCov(\bar{y}_s \bar{x}_s, \bar{x}_s)].$$

$$V\left(\frac{N-n}{Nn\bar{x}} s'_{y,x}\right) =$$

$$\left(\frac{N-n}{Nn\bar{x}}\right)^2 \left(\frac{N}{N-1}\right)^2 \frac{N-n}{(N-1)n}\left[A_{2;\, y(x-\bar{x})} - A^2_{1;\, y(x-\bar{x})}\right].$$

Y

$$Cov\left(t_{PG\prime}, \frac{N-n}{Nn\bar{x}} s'_{y,x}\right) =$$

$$\frac{N-n}{Nn\bar{x}^2}\left[Cov\left(\bar{y}_s \bar{x}_s, s'_{y,x}\right) - kCov\left(\bar{x}_s, s'_{y,x}\right)\right].$$

Hemos denotado por *A* con subíndice al momento poblacional de orden que se subindica, para la variable que se describe a continuación en el subíndice tras el orden del momento no central y del punto y coma ";". Por esto, si queremos optimizar el valor de $k$ que minimice la varianza $V(t_{PG\prime u})$, debemos derivar esta

varianza con respecto a $k$, e igualar a cero. De aquí se obtiene que multiplicando ambos miembros por $\bar{x}^2$, resulta la ecuación

$$2kV(\bar{x}_s) = 2Cov(\bar{y}_s\bar{x}_s, \bar{x}_s) - 2\frac{N-n}{Nn}Cov\left(\bar{x}_s, s'_{y,x}\right).$$

Por lo que el valor óptimo de $k$ que hace mínima la varianza $V(t_{PG\prime u})$ será

$$k_{\acute{o}pt} = \frac{Cov(\bar{y}_s\bar{x}_s, \bar{x}_s) - \frac{N-n}{Nn}Cov\left(\bar{x}_s, s'_{y,x}\right)}{V(\bar{x}_s)}.$$

Este valor es una función paramétrica que no es conocida antes de realizar el muestreo, por lo que su valor es teórico con miras a determinar el estimador producto generalizado corregido insesgado óptimo. Sin embargo, dicho valor óptimo puede ser estimado insesgadamente antes de elegir el estimador insesgado $t_{PG\prime u}$ concreto con el que estimar la media poblacional $\bar{y}$. Sustituyendo en la fórmula de $k_{\acute{o}pt}$ las covarianzas del numerador por sus estimadores insesgados respectivos, se obtiene un estimador insesgado de $k_{\acute{o}pt}$, concretamente

$$\hat{k}_{\acute{o}pt} = \frac{\widehat{Cov}(\bar{y}_s\bar{x}_s, \bar{x}_s) - \frac{N-n}{Nn}\widehat{Cov}\left(\bar{x}_s, s'_{y,x}\right)}{V(\bar{x}_s)}.$$

El denominador es una constante conocida, que no depende de los valores de la variable de interés $y$, sino que solo depende de la variable auxiliar $x$ definida y conocida para todas las unidades de la población finita de tamaño $N$ y del tamaño muestral $n$, ambos tamaños conocidos. Los estimadores insesgados concretos de dichas covarianzas pueden obtenerse del modo siguiente

$$\widehat{Cov}(\bar{y}_s\bar{x}_s, \bar{x}_s) = \bar{y}_s\bar{x}_s^2 - \hat{E}(\bar{y}_s\bar{x}_s)\bar{x} = \bar{y}_s\bar{x}_s^2 - \bar{y}_s\bar{x}_s\bar{x} =$$

$$\bar{y}_s\bar{x}_s(\bar{x}_s - \bar{x}).$$

Y

$$\widehat{Cov}\left(\bar{x}_s, s'_{y,x}\right) = \frac{N-n}{Nn}\hat{S}_{x,y(x-\bar{x})} = \frac{N-n}{Nn}s'_{x,y(x-\bar{x})} =$$

$$\frac{N-n}{(N-1)n^2}\sum_{i\in s} y_i(x_i-\bar{x})^2.$$

Finalmente, hacemos notar que el valor óptimo de $k = k_{ópt}$ es un mínimo global, porque la derivada segunda de la varianza $V(t_{PG\prime u})$ con respecto a $k^2$ , da un valor positivo independientemente de $k$ (salvo en el caso particular en que la variable auxiliar $x$ fuera una constante positiva en todas las unidades de la población finita, un caso trivial en el que $t_{PG\prime u} = t_{PG\prime} = t_P = \bar{y}_s$). Concretamente

$$\frac{d^2V(t_{PG\prime u})}{dk^2} = \frac{2V(\bar{x}_s)}{\bar{x}^2} > 0.$$

Como consecuencia, la varianza $V(t_{PG\prime u})$ alcanza el mínimo global para una variedad infinita de valores reales de $k$ entre los que se encuentra $k = 0$, en cuyo caso tendríamos el "estimador producto corregido insesgado" debido al autor. Luego con el valor óptimo de $k = k_{ópt}$ se obtendría un estimador $t_{PG\prime u}$ que mejora la precisión del último estimador indicado. En la práctica, sustituyendo $k_{ópt}$ por su estimador insesgado $\hat{k}_{ópt}$ los resultados obtenidos con el estimador $t_{PG\prime u}$ son aproximadamente óptimos ya que su varianza estará próxima a la varianza mínima global.

Se puede estimar insesgadamente la varianza del estimador $t_{PG\prime u}$. Para ello sabemos que

$$V(t_{PG'u}) = V(t_{PG'}) + V\left(-\frac{N-n}{Nn\bar{x}}s'_{y,x}\right)$$

$$-2Cov\left(t_{PG'}, \frac{N-n}{Nn\bar{x}}s'_{y,x}\right).$$

Por lo que sustituyendo en el segundo miembro las dos varianzas y la covarianza por sus respectivos estimadores insesgados, obtenemos el estimador insesgado de la varianza $V(t_{PG\prime u})$, y que denotamos por la notación habitual $\hat{V}(t_{PG\prime u})$. Los tres sumandos del segundo miembro de la última fórmula son estimables insesgadamente. Veámoslo detenidamente.

Primer sumando:

$$V(t_{PG\prime}) = \frac{1}{\bar{x}^2}[V(\bar{y}_s\bar{x}_s) + k^2V(\bar{x}_s) - 2kCov(\bar{y}_s\bar{x}_s, \bar{x}_s)].$$

De aquí, el estimador insesgado del primer miembro es

$$\hat{V}(t_{PG\prime}) = \frac{1}{\bar{x}^2}\left[\hat{V}(\bar{y}_s\bar{x}_s) + k^2V(\bar{x}_s) - 2k\widehat{Cov}(\bar{y}_s\bar{x}_s, \bar{x}_s)\right].$$

Donde

$$\hat{V}(\bar{y}_s\bar{x}_s) = \bar{y}_s^2\bar{x}_s^2 - \{[E(\widehat{\bar{y}_s\bar{x}_s})]^2\}.$$

Siendo

$$\{[E(\widehat{\bar{y}_s\bar{x}_s})]^2\} = \left[\left(\bar{y}\bar{x} + \widehat{\frac{N-n}{Nn}S_{y,x}}\right)^2\right] =$$

$$\left[\bar{y}^2\bar{x}^2 + \left(\frac{N-n}{Nn}\right)^2\widehat{S_{y,x}^2} + 2\frac{N-n}{Nn}\bar{y}\bar{x}S_{y,x}\right] =$$

$$\left[\bar{y}_s^2 - \hat{V}(\bar{y}_s)\right]\bar{x}^2 + \left(\frac{N-n}{Nn}\right)^2\left(\frac{N}{N-1}\right)^2\left[\left(A_{1,1} - \widehat{A_{1,0}A_{0,1}}\right)^2\right] +$$

$$2\bar{x}\frac{N-n}{Nn}(\widehat{\bar{y}S_{y,x}}).$$

Ahora,

$$\hat{V}(\bar{y}_s) = \frac{N-n}{Nn} s_y^2 = \frac{N-n}{Nn(n-1)} \sum_{i \in s} (y_i - \bar{y}_s)^2.$$

$$\left[\widehat{(A_{1,1} - A_{1,0}A_{0,1})^2}\right] = \left[\widehat{(A_{1,1})^2}\right] + \widehat{(\bar{y}^2)}\bar{x}^2 - 2\widehat{(A_{1,1}A_{1,0})}\bar{x}.$$

Donde

$$\left[\widehat{(A_{1,1})^2}\right] = a_{1;\, y^2x^2} - \frac{N-n}{Nn} s_{yx}^2 =$$

$$\frac{1}{n} \sum_{i \in s} y_i^2 x_i^2 - \frac{N-n}{Nn(n-1)} \sum_{i \in s} (y_i x_i - a_{1;yx})^2.$$

Siendo

$$a_{1;yx} = a_{1,1} = \frac{1}{n} \sum_{i \in s} y_i x_i.$$

También

$$\widehat{(\bar{y}^2)} = \bar{y}_s^2 - \hat{V}(\bar{y}_s) = \bar{y}_s^2 - \frac{N-n}{Nn} s_y^2.$$

Y

$$\widehat{(A_{1,1}A_{1,0})} = [\widehat{E(yx)E(y)}] = \hat{E}(y^2x) - \widehat{Cov}(yx, y) =$$

$$a_{2,1} - \frac{N-1}{N} \hat{S}_{yx,y} =$$

$$\frac{1}{n} \sum_{i \in s} y_i^2 x_i - \frac{N-1}{N(n-1)} \sum_{i \in s} (y_i x_i - a_{1,1})(y_i - a_{1,0}).$$

Finalmente,

$$\widehat{(\bar{y}S_{y,x})} = \bar{y}_s s'_{y,x} - \widehat{Cov}(\bar{y}_s, s'_{y,x}).$$

Donde

$$s'_{y,x} = \frac{N}{(N-1)n}\sum_{i\in s} y_i(x_i - \bar{x}).$$

Por lo que

$$\widehat{Cov}(\bar{y}_s, s'_{y,x}) = \frac{N-n}{Nn}\hat{S}_{y(x-\bar{x}),y} = \frac{N-n}{Nn}s_{y(x-\bar{x}),y} =$$

$$\frac{N-n}{Nn(n-1)}\sum_{i\in s} y_i(x_i - \bar{x})(y_i - \bar{y}_s).$$

Por otro lado,

$$V(\bar{x}_s) = \frac{N-n}{(N-1)n}\left(A_{0,2} - A_{0,1}^2\right).$$

Para terminar, ya hemos visto que

$$\widehat{Cov}(\bar{y}_s\bar{x}_s, \bar{x}_s) = \bar{y}_s\bar{x}_s(\bar{x}_s - \bar{x}).$$

Segundo sumando:

$$\hat{V}\left(\frac{N-n}{Nn\bar{x}}s'_{y,x}\right) =$$

$$\left(\frac{N-n}{Nn\bar{x}}\right)^2\left(\frac{N}{N-1}\right)^2\frac{N-n}{(N-1)n}\left[A_{2;\,\widehat{y(x-\bar{x})}} - A_{1;\,y(x-\bar{x})}^2\right] =$$

$$\left(\frac{N-n}{Nn\bar{x}}\right)^2\left(\frac{N}{N-1}\right)^2\frac{N-n}{Nn}s_{y(x-\bar{x})}^2.$$

Donde

$$s_{y(x-\bar{x})}^2 = \frac{1}{n-1}\sum_{i\in s}\left[y_i(x_i - \bar{x}) - a_{1;y(x-\bar{x})}\right]^2.$$

Con

$$a_{1;y(x-\bar{x})} = \frac{1}{n}\sum_{i\in s} y_i(x_i - \bar{x}).$$

Tercer sumando:

$$-2Cov\left(t_{PG'}, \frac{N-n}{Nn\bar{x}} s'_{y,x}\right) =$$

$$-2\frac{N-n}{Nn\bar{x}^2}\left[Cov\left(\bar{y}_s\bar{x}_s, s'_{y,x}\right) - kCov\left(\bar{x}_s, s'_{y,x}\right)\right].$$

Que es estimable insesgadamente por

$$-2\frac{N-n}{Nn\bar{x}^2}\left[\widehat{Cov}\left(\bar{y}_s\bar{x}_s, s'_{y,x}\right) - k\widehat{Cov}\left(\bar{x}_s, s'_{y,x}\right)\right].$$

Donde

$$\widehat{Cov}\left(\bar{y}_s\bar{x}_s, s'_{y,x}\right) = \bar{y}_s\bar{x}_s s'_{y,x} - \left[E(\bar{y}_s\bar{x}_s)\widehat{E\left(s'_{y,x}\right)}\right].$$

Como

$$E(\bar{y}_s\bar{x}_s) = Cov(\bar{y}_s, \bar{x}_s) + E(\bar{y}_s)\bar{x} = \frac{N-n}{Nn} S_{y,x} + E(\bar{y}_s)\bar{x}.$$

Y

$$E\left(s'_{y,x}\right) = S_{y,x} = \frac{1}{N-1}\sum_{i=1}^{N} y_i(x_i - \bar{x}).$$

Entonces

$$\left[E(\bar{y}_s\bar{x}_s)\widehat{E\left(s'_{y,x}\right)}\right] = \frac{N-n}{Nn}\widehat{\left(S^2_{y,x}\right)} + \bar{x}\left[E(\bar{y}_s)\widehat{E\left(s'_{y,x}\right)}\right].$$

Donde

$$\widehat{\left(S^2_{y,x}\right)} = \left(s'_{y,x}\right)^2 - \hat{V}\left(s'_{y,x}\right) = \left(s'_{y,x}\right)^2 - \frac{N-n}{(N-1)n}\left[\widehat{S^2_{y(x-\bar{x})}}\right] =$$

$$\left(s'_{y,x}\right)^2 - \frac{N-n}{(N-1)n} s^2_{y(x-\bar{x})}.$$

Siendo

$$s^2_{y(x-\bar{x})} = \frac{1}{n-1}\sum_{i\in s}\left[y_i(x_i-\bar{x}) - a_{1;y(x-\bar{x})}\right]^2.$$

Y

$$\left[E(\bar{y}_s)\widehat{E(s'_{y,x})}\right] = \left(\widehat{\bar{y}S_{y,x}}\right) = \bar{y}_s s'_{y,x} - \widehat{Cov}\left(\bar{y}_s, s'_{y,x}\right) =$$

$$\bar{y}_s s'_{y,x} - \frac{N-n}{Nn(n-1)}\sum_{i\in s} y_i(x_i-\bar{x})(y_i-\bar{y}_s).$$

Que ha sido calculado con anterioridad. También,

$$\widehat{Cov}\left(\bar{x}_s, s'_{y,x}\right) = \frac{N-n}{Nn}\hat{S}_{x,y(x-\bar{x})} = \frac{N-n}{Nn} s'_{x,y(x-\bar{x})} =$$

$$\frac{N-n}{(N-1)n^2}\sum_{i\in s} y_i(x_i-\bar{x})^2.$$

Por todo ello concluimos que el estimador producto generalizado corregido insesgado propuesto admite un estimador insesgado de su varianza con la información auxiliar disponible y con el diseño de muestreo irrestricto aleatorio de tamaño efectivo fijo $n$. Además, este estimador insesgado de $V(t_{PG\prime u})$ puede calcularse en cada caso concreto mediante las estimaciones proporcionadas en esta sección, que aunque puedan ser laboriosas, resultan realizables cuidadosamente en la práctica.

**Ejercicio 6.19.** Proponer un estimador insesgado de la media poblacional con información de dos variables auxiliares, en muestreo irrestricto aleatorio.

**Solución.** Un estimador insesgado general $t_u$ que aprovecha toda la información auxiliar disponible, concretamente el estimador

$$t_u = \bar{y}_s + \sum_{i=1}^{2} k_i\left(\bar{x}_i - \bar{x}_{i,s}\right).$$

Donde los dos valores $k_i$ son constantes conocidas para todo $i = 1, 2$; $\bar{x}_i$ es la media poblacional de la variable auxiliar $i$-ésima; y $\bar{x}_{i,s}$ es la media muestral de la variable auxiliar $i$-ésima para la misma muestra aleatoria simple sin reemplazamiento $s$, de tamaño $n$, seleccionada. Así, tenemos

$$\bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{i,k}.$$

Y

$$\bar{x}_{i,s} = \frac{1}{n}\sum_{k\in s} x_{i,k}.$$

Siendo $x_{i,k}$ el valor de la variable auxiliar $i$-ésima en la unidad $k$ de la población finita, es decir, con uno de los valores posibles de $k = 1, 2, \ldots, N$. Sabemos que la esperanza matemática de la media muestral coincide con la media poblacional de la misma variable. Por tanto, $E(\bar{y}_s) = \bar{y}$, y también para todo $i = 1, 2$, tenemos que $E\left(\bar{x}_{i,s}\right) = \bar{x}_i$, haciendo uso de las propiedades del diseño de muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n$.

Ya que $\bar{x}_{i,s}$ es una media muestral, es un estimador insesgado de la media poblacional $\bar{x}_i$, por lo que tomado la esperanza matemática de $t_u$ tenemos

$$E(t_u) = E\left[\bar{y}_s + \sum_{i=1}^{2} k_i\left(\bar{x}_i - \bar{x}_{i,s}\right)\right] =$$

$$E(\bar{y}_s) + \sum_{i=1}^{2} k_i\left[\bar{x}_i - E\left(\bar{x}_{i,s}\right)\right] = \bar{y}.$$

Debido a las propiedades de la esperanza matemática, ya que para todos los valores posibles de $i = 1, 2$, tanto $k_i$ como $\bar{x}_i$ son constantes. En resumen, el estimador general $t_u$ es insesgado para estimar la media poblacional de interés, con muestreo irrestricto aleatorio.

Haciendo uso de las propiedades de la varianza de una variable aleatoria, tenemos que

$$V(t_u) = V\left[\bar{y}_s + \sum_{i=1}^{2} k_i\left(\bar{x}_i - \bar{x}_{i,s}\right)\right]$$

$$= V(\bar{y}_s) + \sum_{i=1}^{2} k_i^2 V\left(\bar{x}_{i,s}\right) - 2\sum_{i=1}^{2} k_i Cov\left(\bar{y}_s, \bar{x}_{i,s}\right)$$

$$+ \sum_{i=1}^{2}\sum_{j\neq i}^{2} k_i k_j Cov\left(\bar{x}_{i,s}, \bar{x}_{j,s}\right).$$

Aquí, en el último miembro, todo son constantes conocidas antes de proceder al muestreo y a la fase de estimación, salvo las funciones paramétricas $V(\bar{y}_s)$ y $Cov\left(\bar{y}_s, \bar{x}_{i,s}\right)$, con $i = 1, 2$. Por esto, la varianza del estimador general $t_u$ puede ser estimada sin sesgo del modo

$$\hat{V}(t_u) = \hat{V}(\bar{y}_s) + \sum_{i=1}^{2} k_i^2 V\left(\bar{x}_{i,s}\right) - 2\sum_{i=1}^{2} k_i \widehat{Cov}\left(\bar{y}_s, \bar{x}_{i,s}\right)$$

$$+\sum_{i=1}^{2}\sum_{j\neq i}^{2} k_i k_j Cov\left(\bar{x}_{i,s}, \bar{x}_{j,s}\right).$$

Donde $\hat{V}(\bar{y}_s)$ y $\widehat{Cov}\left(\bar{y}_s, \bar{x}_{i,s}\right)$ son los estimadores insesgados respectivos uno a uno de las funciones paramétricas $V(\bar{y}_s)$ y $Cov\left(\bar{y}_s, \bar{x}_{i,s}\right)$, de modo similar a como explico en el artículo reciente de Ruiz Espejo *et al.* (2013). A continuación vamos a obtener dichos estimadores insesgados en el muestreo irrestricto aleatorio de tamaño muestral efectivo $n$.

$$\hat{V}(\bar{y}_s) = \frac{N-n}{Nn}\widehat{\left(S_y^2\right)} = \frac{N-n}{Nn}s_y^2 = \frac{N-n}{Nn(n-1)}\sum_{k\in s}(y_k - \bar{y}_s)^2.$$

Y

$$\widehat{Cov}\left(\bar{y}_s, \bar{x}_{i,s}\right) = \frac{N-n}{Nn}\widehat{\left(S_{y,x_\iota}\right)} = \frac{N-n}{Nn}s'_{y,x_i} =$$

$$\frac{N-n}{(N-1)n^2}\sum_{k\in s} y_k\left(x_{i,k} - \bar{x}_i\right).$$

Hasta aquí hemos supuesto que los valores constantes $k_i$ estaban fijados de antemano y eran conocidos para concretar el estimador insesgado $t_u$. Sin embargo, es posible estudiar qué valores concretos de $k_i$ minimizan la varianza del estimador general insesgado bivariante $t_u$. Para ello, derivamos parcialmente la expresión de la varianza $V(t_u)$ con respecto a $k_i$, e igualándolas a cero obtenemos un sistema de 2 ecuaciones lineales con 2 incógnitas (que son las constantes óptimas $k_i = k_{i,\acute{o}pt}$). En efecto, el sistema de ecuaciones lineales es el siguiente

$$\begin{cases} \dfrac{\partial V(t_u)}{\partial k_i} = 0 \\ i = 1, 2 \end{cases}$$

Que resulta ser entonces

$$\begin{cases} k_i V(\bar{x}_{i,s}) + \sum_{j \neq i}^{2} k_j Cov(\bar{x}_{i,s}, \bar{x}_{j,s}) = Cov(\bar{y}_s, \bar{x}_{i,s}) \\ i = 1, 2 \end{cases}$$

También se puede comprobar que

$$\frac{\partial^2 V(t_u)}{\partial k_i^2} = 2V(\bar{x}_{i,s}).$$

Que es una constante positiva, salvo que la variable auxiliar $i$-ésima sea constante en todas las unidades de la población finita, en cuyo caso el término correspondiente a dicha variable auxiliar se anula en la fórmula del estimador $t_u$, por lo que su expresión se reduciría a una estimación basada en una variable auxiliar al eliminar aquélla en la que la variable auxiliar no aportara una información con alguna variabilidad.

Para $i \neq j$, tenemos que

$$\frac{\partial^2 V(t_u)}{\partial k_i \partial k_j} = 2Cov(\bar{x}_{i,s}, \bar{x}_{j,s}).$$

Finalmente, las derivadas parciales de orden tres se anulan en todos los casos, por lo cual concluimos que se obtiene un mínimo global de la función real bidimensional para ciertos valores $k_i = k_{i,\acute{o}pt}$ que son óptimos y calculables teóricamente en cada caso concreto. Salvo casos triviales, los valores críticos son los óptimos que minimizan la varianza del estimador $t_u$, ya que los menores principales de la matriz de covarianzas son positivos. Excluimos el caso trivial en que exista un coeficiente de correlación $1$ ó $-1$ entre las medias muestrales de las dos variables auxiliares. Veamos a continuación la solución óptima teórica en el caso de disponer de

dos variables auxiliares con un coeficiente de correlación absoluto menor que 1.

En el caso en que el número de variables auxiliares es 2, tenemos que la solución concreta del sistema de ecuaciones lineales viene dada por estas fórmulas.

$$k_{1,\acute{o}pt} = \frac{V(\bar{x}_{2,s})Cov(\bar{y}_s, \bar{x}_{1,s}) - Cov(\bar{y}_s, \bar{x}_{2,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}.$$

$$k_{2,\acute{o}pt} = \frac{V(\bar{x}_{1,s})Cov(\bar{y}_s, \bar{x}_{2,s}) - Cov(\bar{y}_s, \bar{x}_{1,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}.$$

Que son constantes óptimas desconocidas, pues son funciones paramétricas que dependen de todos los valores de la variable de interés en las unidades de la población finita. Con estas constantes, si las conociéramos antes de realizar el muestreo y de observar en la muestra seleccionada la variable de interés, el estimador insesgado de regresión bivariante sería

$$t_u = \bar{y}_s + \sum_{i=1}^{2} k_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s}).$$

Y alcanzaría su varianza el valor mínimo global con $(k_{1,\acute{o}pt}, k_{2,\acute{o}pt})$ entre todos los posibles valores del plano real para $(k_1, k_2)$. Pero la realidad es que no conocemos estas constantes óptimas teóricas en un estudio concreto, por lo que cabe estimarlas sin sesgo sustituyendo, en el numerador de la expresión de cada una de dichas constantes óptimas, las funciones paramétricas $Cov(\bar{y}_s, \bar{x}_{i,s})$ por sus estimadores insesgados (al variar $i = 1, 2$) que obtenemos a continuación.

$$\widehat{Cov}(\bar{y}_s, \bar{x}_{i,s}) = \frac{N-n}{Nn} s'_{y,x_i} = \frac{N-n}{(N-1)n^2} \sum_{k \in s} y_k (x_{i,k} - \bar{x}_i).$$

De ese modo, ya que los demás términos de $k_{i,\acute{o}pt}$ son constantes conocidas de antemano, obtenemos los valores óptimos estimados sin sesgo siguientes

$$\hat{k}_{1,\acute{o}pt} = \frac{V(\bar{x}_{2,s})\widehat{Cov}(\bar{y}_s, \bar{x}_{1,s}) - \widehat{Cov}(\bar{y}_s, \bar{x}_{2,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}.$$

$$\hat{k}_{2,\acute{o}pt} = \frac{V(\bar{x}_{1,s})\widehat{Cov}(\bar{y}_s, \bar{x}_{2,s}) - \widehat{Cov}(\bar{y}_s, \bar{x}_{1,s})Cov(\bar{x}_{1,s}, \bar{x}_{2,s})}{V(\bar{x}_{1,s})V(\bar{x}_{2,s}) - \left[Cov(\bar{x}_{1,s}, \bar{x}_{2,s})\right]^2}.$$

Por todo ello, parece indicado partir del estimador

$$t' = \bar{y}_s + \sum_{i=1}^{2} \hat{k}_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s}).$$

Este estimador es similar al que hemos estudiado como bivariante insesgado $t_u$ al sustituir los valores $k_i$ por los valores que estiman sus valores óptimos, es decir, por $\hat{k}_{i,\acute{o}pt}$. Pero como estos últimos estimadores no son constantes sino variables aleatorias, tienen un efecto en $t'$ que lo hacen sesgado en general para estimar la media poblacional $\bar{y}$.

El estimador bivariante óptimo teórico es

$$t_u = \bar{y}_s + \sum_{i=1}^{2} k_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s}).$$

Tiene una varianza

$$V_{\acute{o}pt}(t_u) = V(\bar{y}_s) + \sum_{i=1}^{2} k_{i,\acute{o}pt}^2 V(\bar{x}_{i,s})$$

$$-2\sum_{i=1}^{2} k_{i,\acute{o}pt} Cov\left(\bar{y}_s, \bar{x}_{i,s}\right) + 2\, k_{1,\acute{o}pt} k_{2,\acute{o}pt} Cov\left(\bar{x}_{1,s}, \bar{x}_{2,s}\right).$$

Por lo que esta varianza óptima teórica $V_{\acute{o}pt}(t_u)$ puede ser estimada sin sesgo a partir de las estimaciones insesgadas siguientes.

$$\hat{V}(\bar{y}_s) = \frac{N-n}{Nn} s_y^2 = \frac{N-n}{Nn(n-1)} \sum_{k \in s} (y_k - \bar{y}_s)^2.$$

También

$$\left\{\left[\widehat{Cov\left(\bar{y}_s, \bar{x}_{\iota,s}\right)}\right]^2\right\} = \left(\frac{N-n}{Nn}\right)^2 \left(\widehat{S^2_{y,x_\iota}}\right)$$

$$= \left(\frac{N-n}{Nn}\right)^2 \left[\left(s'_{y,x_i}\right)^2 - \hat{V}\left(s'_{y,x_i}\right)\right].$$

Donde

$$s'_{y,x_i} = \frac{N}{(N-1)n} \sum_{k \in s} y_k \left(x_{i,k} - \bar{x}_i\right).$$

Y

$$\hat{V}\left(s'_{y,x_i}\right) = \frac{N^2}{(N-1)^2} \hat{V}\left[\frac{1}{n} \sum_{k \in s} y_k \left(x_{i,k} - \bar{x}_i\right)\right]$$

$$= \frac{N^2}{(N-1)^2} \frac{N-n}{Nn} \left[\widehat{S^2_{y(x_\iota - \bar{x}_\iota)}}\right] = \frac{N(N-n)}{(N-1)^2 n} s^2_{y(x_i - \bar{x}_i)}$$

$$= \frac{N(N-n)}{(N-1)^2 n(n-1)} \sum_{k \in s} \left[y_k \left(x_{i,k} - \bar{x}_i\right) - a_{1;y(x_i - \bar{x}_i)}\right]^2.$$

Siendo

$$a_{1;y(x_i-\bar{x}_i)} = \frac{1}{n}\sum_{k\in s} y_k\left(x_{i,k}-\bar{x}_i\right).$$

Y también

$$\left[Cov\left(\bar{y}_s,\bar{x}_{1,s}\right)\widehat{Cov}\left(\bar{y}_s,\bar{x}_{2,s}\right)\right] = \left(\frac{N-n}{Nn}\right)^2\left(S_{y,x_1}\widehat{S_{y,x_2}}\right)$$

$$= \left(\frac{N-n}{Nn}\right)^2\left[s'_{y,x_1}s'_{y,x_2} - \widehat{Cov}\left(s'_{y,x_1},s'_{y,x_2}\right)\right].$$

Donde

$$\widehat{Cov}\left(s'_{y,x_1},s'_{y,x_2}\right) = \frac{N^2}{(N-1)^2}\widehat{Cov}\left[a_{1;y(x_1-\bar{x}_1)},a_{1;y(x_2-\bar{x}_2)}\right]$$

$$= \frac{N^2}{(N-1)^2}\frac{N-n}{Nn}\left[\widehat{S_{y(x_1-\bar{x}_1),y(x_2-\bar{x}_2)}}\right]$$

$$= \frac{N(N-n)}{(N-1)^2 n}s_{y(x_1-\bar{x}_1),y(x_2-\bar{x}_2)}$$

$$= \frac{N(N-n)}{(N-1)^2 n(n-1)}$$

$$\times\sum_{k\in s}\left[y_k\left(x_{1,k}-\bar{x}_1\right)-a_{1;y(x_1-\bar{x}_1)}\right]\left[y_k\left(x_{2,k}-\bar{x}_2\right)-a_{1;y(x_2-\bar{x}_2)}\right].$$

El resto de la demostración es un ejercicio algebraico relativamente asequible.

El estimador que hemos estudiado anteriormente no es posible llevarlo a la práctica pues, aunque tiene muy buenas propiedades teóricas, depende de funciones paramétricas que son desconocidas y que deben ser estimadas sin sesgo. Así si sustituimos los valores óptimos $k_{i,\acute{o}pt}$ por sus estimadores

insesgados $\hat{k}_{i,\acute{o}pt}$ , el estimador resultante $t'$ es sesgado, concretamente

$$t' = \bar{y}_s + \sum_{i=1}^{2} \hat{k}_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s}).$$

Sin embargo, se puede corregir para que sea insesgado, del modo siguiente

$$t'_u = \bar{y}_s + \sum_{i=1}^{2} \hat{k}_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s}) - \sum_{i=1}^{2} \widehat{Cov}(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s}).$$

Aquí $\widehat{Cov}(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s})$ es un estimador insesgado de la covarianza $Cov(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s})$, que más adelante pasaremos a concretar cómo obtenerlo para que sea útil en la práctica. Para demostrar que $t'_u$ es insesgado nos basamos en que $\widehat{Cov}(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s})$ es un estimador insesgado de la esperanza matemática de $\hat{k}_{i,\acute{o}pt}(\bar{x}_i - \bar{x}_{i,s})$. En concreto se puede ver que

$$\{E[\hat{k}_{\imath,\acute{o}pt}\widehat{(\bar{x}_\imath - \bar{x}_{\imath,s})}]\} =$$

$$[E(\hat{k}_{\imath,\acute{o}pt})\widehat{E(\bar{x}_\imath - \bar{x}_{\imath,s})}] + \widehat{Cov}(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s}) =$$

$$[E(\widehat{\hat{k}_{\imath,\acute{o}pt}}) \times 0] + \widehat{Cov}(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s}) = \widehat{Cov}(\hat{k}_{i,\acute{o}pt}, \bar{x}_{i,s}).$$

Para calcular este último estimador, es un ejercicio asequible pero cuidadoso en el caso bivariante a partir de los estimadores insesgados necesarios siguientes.

$$\widehat{Cov}[\widehat{Cov}(\bar{y}_s, \bar{x}_{i,s}), \bar{x}_{i,s}] = \widehat{Cov}\left(\frac{N-n}{Nn}\hat{S}_{y,x_i}, \bar{x}_{i,s}\right) =$$

$$\widehat{Cov}\left[\frac{N-n}{(N-1)n}a_{1;y(x_i-\bar{x}_i)}, \bar{x}_{i,s}\right] = \frac{N-n}{(N-1)n}\frac{N-n}{Nn}\left[\widehat{S^2_{y(x_\imath-\bar{x}_\imath)^2}}\right]$$

$$= \frac{(N-n)^2}{N(N-1)n^2} s^2_{y(x_i-\bar{x}_i)^2}$$

$$= \frac{(N-n)^2}{(N-1)^2 n^2 (n-1)} \sum_{k \in s} \left[ y_k \left( x_{i,k} - \bar{x}_i \right)^2 - a_{1;y(x_i-\bar{x}_i)^2} \right]^2.$$

Y de modo similar, en el caso bivariante,

$$\widehat{Cov}\left[\widehat{Cov}\left(\bar{y}_s, \bar{x}_{2,s}\right), \bar{x}_{1,s}\right] = \frac{(N-n)^2}{N(N-1)n^2} \hat{S}_{y(x_2-\bar{x}_2),x_1}$$

$$= \frac{(N-n)^2}{N(N-1)n^2} s'_{y(x_2-\bar{x}_2),x_1}$$

$$= \frac{(N-n)^2}{N(N-1)n^3} \sum_{k \in s} y_k \left( x_{2,k} - \bar{x}_2 \right) \left( x_{1,k} - \bar{x}_1 \right).$$

Etc.

**Ejercicio 6.20.** Ajustar un modelo de regresión lineal multivariante por el método de mínimo error cuadrático medio en una población finita, y estimarlo de modo insesgado a partir de muestras aleatorias simples sin reemplazamiento.

**Solución.** El modelo a ajustar es

$$y = k_0 + k_1 x_1 + k_2 x_2 + \cdots + k_m x_m + e.$$

Teniendo en cuenta que, en este caso general, hay $m$ variables explicativas o auxiliares, que son las que hemos denotado por $x_1, x_2, \ldots, x_m$. Los valores $k_0$, $k_1$, $k_2$, …, $k_m$ son las constantes que determinan el modelo de regresión lineal multivariante óptimo. La variable $y$ es la variable explicada o de interés. El error

cuadrático total poblacional, proporcional al error cuadrático medio poblacional, en este caso es

$$\phi = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left( y_i - k_0 - \sum_{r=1}^{m} k_r x_{ri} \right)^2 .$$

Para minimizar este error cuadrático total (o equivalentemente el error cuadrático medio, $\phi/N$), derivamos parcialmente la función $\phi$ con respecto a cada una de las variables $k_r$ con $r = 0, 1, \dots, m$, e igualamos a cero cada una de esas derivadas parciales. El sistema resultante es equivalente al siguiente

$$\begin{cases} A_{1;y} = k_0 + \sum_{j=1}^{m} k_j A_{1;\,x_j} \\ A_{1,1;y,x_r} = k_0 A_{1;\,x_r} + k_r A_{2;\,x_r} + \sum_{\substack{j=1 \\ j \neq r}}^{m} k_j A_{1,1;\,x_j,x_r} \\ r = 1, 2, \dots, m. \end{cases}$$

Este sistema de ecuaciones lineales tiene $m + 1$ ecuaciones con $m + 1$ incógnitas. También puede expresarse del modo siguiente más simplificado

$$\begin{cases} A_{1;y} = k_0 + \sum_{j=1}^{m} k_j A_{1;\,x_j} \\ A_{1,1;y,x_r} = k_0 A_{1;\,x_r} + \sum_{j=1}^{m} k_j A_{1,1;\,x_j,x_r} \\ r = 1, 2, \dots, m. \end{cases}$$

Matricialmente se expresa de este modo

$$\boldsymbol{a} = \boldsymbol{kA}.$$

Donde

$$\boldsymbol{a} = \boldsymbol{a}_{1\times(m+1)} = (A_{1;y} \quad A_{1,1;y,x_1} \quad \cdots \quad A_{1,1;y,x_m}),$$

$$\boldsymbol{k} = \boldsymbol{k}_{1\times(m+1)} = (k_0 \quad k_1 \quad \cdots \quad k_m).$$

Y finalmente,

$$\boldsymbol{A} = \boldsymbol{A}_{(m+1)\times(m+1)} = \begin{pmatrix} 1 & A_{1;\,x_1} & \cdots & A_{1;\,x_m} \\ A_{1;\,x_1} & A_{1,1;\,x_1,x_1} & \cdots & A_{1,1;\,x_1,x_m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1;\,x_m} & A_{1,1;\,x_m,x_1} & \cdots & A_{1,1;\,x_m,x_m} \end{pmatrix}.$$

Esta matriz $\boldsymbol{A}$ depende exclusivamente de la información auxiliar de las variables explicativas del modelo de regresión lineal multivariante. La solución del sistema se obtiene del modo

$$\boldsymbol{k} = \boldsymbol{a}\boldsymbol{A}^{-1}.$$

Y las soluciones estimadas insesgadamente, $\widehat{\boldsymbol{k}}$, requieren estimar insesgadamente cada una de las componentes del vector $\boldsymbol{a}$, en muestreo irrestricto aleatorio por las medias muestrales correspondientes, es decir, mediante el vector estimado insesgadamente componente a componente $\widehat{\boldsymbol{a}}$ obtenemos las estimaciones insesgadas de los valores óptimos del ajuste lineal multivariante. En concreto, lo formalizamos del modo

$$\widehat{\boldsymbol{k}} = \widehat{\boldsymbol{a}}\boldsymbol{A}^{-1}.$$

Es preciso aclarar que cada modelo estimado depende directamente de la muestra seleccionada, y que habrá tantos modelos estimados como muestras distintas (para los mismos estimadores de $\widehat{\boldsymbol{a}}$), pero en promedio las estimaciones en $\widehat{\boldsymbol{k}}$ son insesgadas para las componentes respectivas del vector óptimo $\boldsymbol{k}$, que es único salvo casos triviales como el de que algunas de las

variables auxiliares coincidan entre sí o alguna fuera constante, etc. de modo que la matriz $\boldsymbol{A}$ no tuviera inversa.

Un ejemplo de aplicación de este tipo de regresión lineal multivariante objetiva es el que nos provee de un estimador insesgado de la media poblacional $\bar{y} = (1/N)\sum_{i=1}^{N} y_i$ aprovechando la característica del modelo consistente en que minimiza el error cuadrático total poblacional. En concreto, el estimador de la media poblacional $\bar{y}$ es

$$\hat{\bar{y}} = \hat{\boldsymbol{k}}\bar{\boldsymbol{x}}^t = \hat{\boldsymbol{a}}\boldsymbol{A}^{-1}\bar{\boldsymbol{x}}^t.$$

Donde $\bar{\boldsymbol{x}}^t$ es la matriz del vector columna de dimensiones $(m+1)\times 1$, que es la matriz traspuesta de la matriz $\bar{\boldsymbol{x}} = (1 \quad \bar{x}_1 \quad \bar{x}_2 \quad \cdots \quad \bar{x}_m)_{1\times(m+1)}$. Con $\bar{x}_r = (1/N)\sum_{i=1}^{N} x_{r,i}$, que es la media poblacional de la variable auxiliar $r$ -ésima $(r = 1, 2, \ldots, m)$. Dicho estimador $\hat{\bar{y}}$ es insesgado y óptimo, para distribución libre, para ajustar el modelo óptimo (de mínimo error cuadrático total poblacional), pero en general podría ser supuestamente sesgado para estimar la media poblacional $\bar{y}$.

Que el componente $A_{1,1;y,x_r}$ puede estimarse insesgada y óptimamente por $a_{1,1;y,x_r} = (1/n)\sum_{i=1}^{n} y_i x_{r,i}$ puede demostrarse de este modo.

$$E(y_i x_{r,i}) = \bar{y}\cdot\bar{x}_r + Cov(y_i, x_{r,i})$$

$$= \bar{y}\cdot\bar{x}_r + \frac{1}{N}\sum_{i\in U} y_i(x_{r,i} - \bar{x}_r).$$

El estimador insesgado e invariante por permutaciones es

$$\widehat{(A_{1,1;y,x_r})} = \widehat{E(y_\iota x_{r,\iota})} = \bar{y}_{\boldsymbol{s}}\cdot\bar{x}_r + a_{1,1;y,x_r} - \bar{y}_{\boldsymbol{s}}\cdot\bar{x}_r$$

$$= a_{1,1;y,x_r} = (1/n)\sum_{i=1}^{n} y_i x_{r,i} = (1/n)\sum_{i\in \boldsymbol{s}} y_i x_{r,i}.$$

En realidad, lo que ocurre es que hemos tratado de minimizar

$$\sum_{i=1}^{N} e_i^2,$$

en lugar de minimizar

$$\sum_{i=1}^{N} e_i^2$$

sujeto a que $\bar{e} = 0$, donde $\bar{e} = (1/N)\sum_{i=1}^{N} e_i$ es el error medio poblacional. Si hacemos esto último, el lagrangiano es

$$L = \sum_{i=1}^{N} e_i^2 + \lambda \sum_{i=1}^{N} e_i$$

y depende también de todos los coeficientes del ajuste lineal, es decir de $k_0, k_1, \dots, k_m$. Su resolución nos da las ecuaciones

$$\frac{\partial L}{\partial k_0} = 2N\left(-A_{1;\,y} + k_0 + \sum_{r=1}^{m} k_r A_{1;\,x_r}\right) - \lambda N = 0$$

$$\frac{\partial L}{\partial k_j} =$$

$$2N\left(-A_{1,1;y,x_j} + k_0 A_{1;\,x_j} + \sum_{r=1}^{m} k_r A_{1,1;\,x_r,x_j}\right) - \lambda N A_{1;\,x_j} = 0$$

$$j = 1, 2, \dots, m$$

Despejando el multiplicador de Lagrange $\lambda$ resulta ser de la primera ecuación $\lambda = 0$, pues la restricción

$$A_{1;y} = k_0 + \sum_{r=1}^{m} k_r A_{1;\,x_r}$$

obliga a este resultado. Resolviendo el sistema de ecuaciones resultante de esta simplificación, que no es más que el sistema inicial considerado sin restricción, determinamos los coeficientes $k_0, k_1, \dots, k_m$ óptimos sujetos a la restricción, que son los mismos ya obtenidos anteriormente. Así se garantiza el ajuste óptimo de error medio poblacional cero, y óptimo en el sentido de mínimo error cuadrático total poblacional. Por tanto se trata de un ajuste de mínima varianza $V(e) = A_{2;e} = A_{1,1;e,e}$ puesto que $A_{1;e} = \bar{e} = 0$. Como consecuencia, se puede asegurar que el estimador $\hat{\bar{y}}$ propuesto es insesgado para estimar la media poblacional $\bar{y}$. Pero además, al ser invariante ante permutaciones de los identificadores de la muestra aleatoria simple con reemplazamiento, el estimador $\hat{\bar{y}}$ es insesgado y uniformemente de mínima varianza para estimar la media poblacional $\bar{y}$ para distribución libre (Zacks, 1971, p. 150), en las condiciones dadas de unas mismas variables auxiliares disponibles. Una consecuencia inmediata es que hemos encontrado un estimador insesgado y de mínima varianza uniformemente mejor que el estimador de regresión lineal clásico cuando se dispone de una variable auxiliar. Lo mismo se puede decir cuando se dispone de un número finito de variables auxiliares.

Veamos ahora la estimación y el contraste de hipótesis del "error cuadrático medio del ajuste lineal multivariante óptimo objetivo en poblaciones finitas". Tal error cuadrático medio se puede expresar del modo

$$ECM = V(e) = \frac{1}{N}\sum_{i=1}^{N} e_i^2 - \bar{e}^2 = E(e^2) = \overline{e^2},$$

pues el error medio poblacional del ajuste óptimo vimos que es

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = E(e) = 0.$$

Aquí

$$e_i = y_i - k_0 - \sum_{r=1}^{m} k_r x_{r,i}.$$

Siendo $k_r = k_{r,\acute{o}pt}$ los valores óptimos del ajuste.

También obtenemos el "error cuadrático medio del ajuste" con los valores estimados sin sesgo $k_r = \hat{k}_{r,\acute{o}pt}$ (con $r = 0, 1, 2, \dots, m$ ), siendo $m$ el número de variables auxiliares o explicativas, y $x_{r,i}$ el valor de la variable auxiliar $x_r$ en la unidad $i$ (con $i = 1, 2, \dots, N$) de la población finita de tamaño $N$. El error cuadrático medio del ajuste, con los valores ajustados estimados insesgadamente $\hat{k}_{r,\acute{o}pt}$, da lugar a otros valores del error $\hat{e}$ pero su esperanza $E[E(\hat{e}|s)] = e$, y por tanto promediando en toda la población finita concluimos que $E(\hat{e}) = E(e) = 0$. Ahora es

$$\hat{e}_i = y_i - \hat{k}_{0,\acute{o}pt} - \sum_{r=1}^{m} \hat{k}_{r,\acute{o}pt} x_{r,i}.$$

El $ECM$ del ajuste óptimo teórico es $ECM = V(e)$. Tenemos entonces que

$$V(e) = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N} \{E[E(\hat{e}_i|s)]\}^2 = E\{[E(\hat{e}_i)]^2\} =$$

$$\frac{1}{N}\sum_{i=1}^{N} [E(\hat{e}_i)]^2 = \frac{1}{N}\sum_{i=1}^{N} E\left(\hat{e}_i^2\right) - \frac{1}{N}\sum_{i=1}^{N} V(\hat{e}_i).$$

De las expresiones anteriores es posible estimar sin sesgo el error cuadrático medio del ajuste óptimo teórico y el error cuadrático medio del ajuste concreto realizado con una muestra aleatoria simple sin reemplazamiento $s$ de tamaño $n$.

Para ello seleccionamos tres muestras independientes por muestreo aleatorio simple sin reemplazamiento, de tamaño común $n$: $s$, $s'$ y $s''$. Con las dos primeras muestras realizamos dos ajustes lineales multivariantes objetivos, y con la tercera muestra observamos los "errores" en cada ajuste anteriormente realizados con $s$ y $s'$ mediante las respectivas estimaciones insesgadas $\hat{k}_{r,\acute{o}pt}$ y $\hat{k}'_{r,\acute{o}pt}$, "errores" que denotamos por $\hat{e}_i$ y $\hat{e}'_i$ respectivamente, para toda unidad $i \in s''$. En esta tercera muestra $s''$ estimamos sin sesgo el "promedio del error al cuadrado" $E(\hat{e}_i^2)$ por $(\hat{e}_i^2 + \hat{e}'^2_i)/2$, y estimamos sin sesgo la "varianza del error" $V(\hat{e}_i)$ por $(\hat{e}_i - \hat{e}'_i)^2$. Así podemos estimar el error cuadrático medio óptimo teórico del ajuste lineal multivariante objetivo, mediante el estimador insesgado

$$\hat{V}(e) = \frac{1}{2n}\sum_{i\in s''}(\hat{e}_i^2 + \hat{e}'^2_i) + \frac{1}{n}\sum_{i\in s''}(\hat{e}_i - \hat{e}'_i)^2.$$

El "error cuadrático medio del ajuste concreto obtenido por una muestra aleatoria simple sin reemplazamiento $s$ de tamaño $n$" es el que denotamos

$$ECM(s) = \frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2 = E(\hat{e}^2).$$

Así, $ECM(s)$ se estima sin sesgo por

$$\widehat{ECM}(s) = \frac{1}{n}\sum_{i\in s''}\hat{e}_i^2.$$

A partir del estimador insesgado propuesto $\widehat{ECM}(s)$, es posible calcular su varianza del modo siguiente

$$V\left[\widehat{ECM}(s)\right] = \frac{N-n}{(N-1)n}\frac{1}{N}\sum_{i=1}^{N}\left[\hat{e}_i^2 - \widehat{ECM}(s)\right]^2 =$$

$$\frac{N-n}{(N-1)n}\frac{1}{N}\sum_{i=1}^{N}\left(\hat{e}_i^2 - \frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2\right)^2 =$$

$$\frac{N-n}{(N-1)n}\left[\frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^4 - \left(\frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2\right)^2\right] = \frac{N-n}{(N-1)n}V(\hat{e}^2).$$

Ya que la muestra $s''$ con que se estima $\widehat{ECM}(s)$ es seleccionada por muestreo aleatorio simple sin reemplazamiento de tamaño muestral efectivo prefijado $n$. Son propiedades conocidas de este diseño muestral.

De la expresión obtenida anteriormente, tenemos su estimador insesgado por las propiedades del muestreo aleatorio simple sin reemplazamiento de tamaño $n \geq 2$, concretamente

$$\hat{V}\left[\widehat{ECM}(s)\right] = \frac{N-n}{Nn(n-1)}\sum_{i\in s''}\left(\hat{e}_i^2 - \frac{1}{n}\sum_{i\in s''}\hat{e}_i^2\right)^2.$$

Pues la cuasivarianza muestral de la variable $\hat{e}^2$ es un estimador insesgado de la cuasivarianza poblacional de la misma variable en el muestreo aleatorio simple sin reemplazamiento de tamaño $n$.

Hemos necesitado de la muestra independiente $s''$ para estimar insesgadamente dicha varianza $V\left[\widehat{ECM}(s)\right]$ porque el ajuste depende de $s$ y, por ello, si hubiéramos basado el estimador de la varianza en la cuasivarianza muestral de la muestra $s$ se

hubieran podido producir sesgos apreciables ya que los valores del error $\hat{e}$ en el estimador dependerían de las unidades de la muestra $s$ con las que hemos estimado las constantes óptimas $k_{r,\acute{o}pt}$ del ajuste con la muestra $s$.

Ya que el error cuadrático medio del ajuste con una muestra aleatoria simple sin reemplazamiento genérica $s$ de tamaño $n$, $ECM(s)$, coincide con la varianza del error extendido a todos los posibles ajustes con muestras aleatorias simples sin reemplazamiento $s$ independientes de tamaño muestral $n$, $V(\hat{e})$, de la desigualdad de Chebychev tenemos que

$$p\left\{\left|\widehat{ECM}(s) - ECM(s)\right| < \varepsilon\right\} \geq$$

$$1 - \frac{V\left[\widehat{ECM}(s)\right]}{\varepsilon^2} \cong 1 - \frac{\hat{V}\left[\widehat{ECM}(s)\right]}{\varepsilon^2}.$$

Por tanto, es posible obtener intervalos al nivel de confianza (con probabilidad) mayor o igual aproximadamente a $1-\alpha$ para la función paramétrica $ECM(s)$, pues sería

$$\varepsilon = \sqrt{\frac{\hat{V}\left[\widehat{ECM}(s)\right]}{\alpha}} = \sqrt{\frac{N-n}{\alpha N n(n-1)} \sum_{i \in s''} \left(\hat{e}_i^2 - \frac{1}{n}\sum_{i \in s''} \hat{e}_i^2\right)^2}.$$

En concreto, el intervalo de confianza es precisamente el intervalo abierto siguiente

$$I = (a, b) = \left(\widehat{ECM}(s) - \varepsilon, \widehat{ECM}(s) + \varepsilon\right).$$

Donde

$$a = \frac{1}{n}\sum_{i \in s''} \hat{e}_i^2 - \sqrt{\frac{N-n}{\alpha N n(n-1)} \sum_{i \in s''} \left(\hat{e}_i^2 - \frac{1}{n}\sum_{i \in s''} \hat{e}_i^2\right)^2}.$$

Y

$$b = \frac{1}{n}\sum_{i \in s''} \hat{e}_i^2 + \sqrt{\frac{N-n}{\alpha N n(n-1)} \sum_{i \in s''} \left( \hat{e}_i^2 - \frac{1}{n}\sum_{i \in s''} \hat{e}_i^2 \right)^2}.$$

Como consecuencia, es posible contrastar en base a dichos intervalos de confianza aproximados obtenidos, cualquier hipótesis nula simple del valor concreto que pudiera tomar el $ECM(s)$ del ajuste lineal multivariante objetivo en poblaciones finitas con la muestra aleatoria simple sin reemplazamiento $s$ de tamaño $n$, en base a una muestra aleatoria simple sin reemplazamiento $s''$, de tamaño $n$, independiente de la anterior ($s$).

La región de aceptación del contraste es el intervalo de confianza $I$ al mismo nivel de confianza $1-\alpha$, pues si el valor dado para el $ECM(s)$ en la hipótesis nula simple pertenece al intervalo de confianza $I$, se debe aceptar dicha hipótesis al nivel de confianza mayor o igual aproximadamente a $1-\alpha$.

Con todo lo expuesto, hemos visto que es posible "estimar insesgadamente" el error cuadrático medio óptimo teórico del ajuste de regresión lineal multivariante objetivo basándonos en dos muestras aleatorias simples sin reemplazamiento independientes de tamaño fijo común $n$, así como "estimar insesgadamente" el error cuadrático medio del ajuste estimado insesgadamente al ajuste óptimo con una muestra aleatoria simple sin reemplazamiento $s$ de tamaño $n$, obtener su varianza (del estimador del $ECM$) y estimar insesgadamente esta varianza.

Todo ello permite estimar puntualmente y por intervalo, así como contrastar hipótesis nulas simples sobre el valor numérico del error cuadrático medio del ajuste estimado con una muestra aleatoria simple sin reemplazamiento de tamaño $n$, en base a otra

muestra con el mismo diseño pero independiente de la anterior y del mismo tamaño, al nivel de confianza mayor o igual aproximadamente a $1-\alpha$.

Generalizaciones de estos resultados serían:

(1) Considerar que la muestra independiente $s''$ tenga un tamaño muestral fijo $c \geq 2$, pero no necesariamente igual al tamaño de la muestra del ajuste $n$. Para ello bastaría sustituir en las fórmulas de este ejercicio el valor de $n$ por el valor de $c$, con $2 \leq c \leq N$.

(2) Considerar en la estimación insesgada del error cuadrático medio para el ajuste óptimo téorico dos o más muestras aleatorias simples sin reemplazamiento con las que ajustar el modelo lineal multivariante insesgado. Esto tiene consecuencias en el estimador pues ahora depende de los errores en cada unidad por cada ajuste, que son dos como hemos considerado, pero en general pueden ser más de dos hasta tantos como posibles muestras aleatorias simples sin reemplazamiento de tamaño $n$, es decir, como las

$$\binom{N}{n}$$

muestras con dicho diseño muestral. Como además estas muestras se obtienen independientes, en realidad es un número infinito de posibles de ellas basadas en las $\binom{N}{n}$ distintas posibles y en todas sus posibles repeticiones a partir de ellas.

**Ejercicio 6.21.** Proponer un estimador óptimo en muestreo doble que usa diseño muestral aleatorio simple con reemplazamiento en cada fase, observando una variable auxiliar en la primera fase y la variable de interés en la segunda fase.

**Solución.** Entendemos por muestreo doble aquel procedimiento de muestreo que se desarrolla en las siguientes dos fases.

En una *primera fase* se selecciona una muestra aleatoria simple con reemplazamiento $\boldsymbol{s}$ de tamaño fijo $n$, a partir de una población finita $U$ de tamaño $N$, y se observa la variable auxiliar $x$ en las unidades seleccionadas. Sea el vector reordenado $(x_1, x_2, \dots, x_n)$ obtenido a partir de la muestra ordenada de datos $\big((k, x_k): k \in \boldsymbol{s}\big)$ de la variable auxiliar $x$. En dicho vector pueden aparecer observaciones repetidas ya que el muestreo es con reemplazamiento.

En una *segunda fase* se selecciona una muestra aleatoria simple con reemplazamiento de tamaño fijo $n'$, a partir del vector reordenado $(x_1, x_2, \dots, x_n)$ que contiene las $n$ observaciones de la variable auxiliar, y por tanto tiene un tamaño efectivo fijo que es un número natural "mayor o igual que 1" y "menor o igual que $n$" de unidades de la población finita. En esta submuestra de tamaño fijo $n'$ obtenida en la segunda fase observamos la variable de interés $y$ que recogemos en el vector reordenado que denotamos $(y_1, y_2, \dots, y_{n'})$ que puede contener observaciones de unidades repetidas también.

Queremos estimar la media poblacional de la variable de interés,

$$\bar{y} = \frac{1}{N}\sum_{i \in U} y_i$$

Para ello proponemos en el muestreo doble anteriormente descrito el estimador

$$\bar{y}_d = \hat{\boldsymbol{a}}\boldsymbol{A}^{-1}\bar{\boldsymbol{x}}^t$$

Este estimador ha sido propuesto anteriormente, y ahora el vector $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_{1\times(m+1)}$ no está tomado a partir de la población finita porque la variable auxiliar solo se conoce dentro de la muestra de

tamaño fijo $n$, sino que está tomado del vector reordenado $(x_1, x_2, \dots, x_n)$ como población de referencia obtenida en la primera fase. El número $m+1$ es el número de parámetros a ajustar en el modelo lineal

$$y_i = k_0 + \sum_{r=1}^{m} k_r f_r(x_i) + e_i$$

Aquí $k_0, k_1, \dots, k_m$ son las constantes o parámetros a ajustar, $f_r(x)$ es la función real de variable real $r$-ésima usada en el ajuste, y $e_i$ es el error del ajuste en la unidad poblacional $i \in U$.

En concreto,

$$\boldsymbol{a} = \boldsymbol{a}_{1\times(m+1)} = (A_{1;y} \quad A_{1,1;y,f_1(x)} \quad \cdots \quad A_{1,1;y,f_m(x)})$$

Siendo

$$A_{1,y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

Y para $r = 1, 2, \dots, m$,

$$A_{1,1;y,f_r(x)} = \frac{1}{n}\sum_{j=1}^{n} y_j f_r(x_j)$$

Que son estimables insesgadamente y de mínima varianza para distribución libre respectivamente por

$$a_{1,y} = \frac{1}{n'}\sum_{i=1}^{n'} y_i$$

Y para $r = 1, 2, \dots, m$,

$$a_{1,1;y,f_r(x)} = \frac{1}{n'}\sum_{i=1}^{n'} y_i f_r(x_i)$$

La matriz cuadrada $\boldsymbol{A} = \boldsymbol{A}_{(m+1)\times(m+1)}$ resulta ser

$$\boldsymbol{A} = \begin{pmatrix} 1 & A_{1;\, f_1(x)} & \cdots & A_{1;\, f_m(x)} \\ A_{1;\, f_1(x)} & A_{1,1;\, f_1(x),f_1(x)} & \cdots & A_{1,1;\, f_1(x),f_m(x)} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1;\, f_m(x)} & A_{1,1;\, f_m(x),f_1(x)} & \cdots & A_{1,1;\, f_m(x),f_m(x)} \end{pmatrix}$$

Esta matriz $\boldsymbol{A}$ depende exclusivamente de la información auxiliar de las variables explicativas del modelo de regresión lineal multivariante. Por ejemplo,

$$A_{1,1:\, f_r(x),f_s(x)} = \frac{1}{n}\sum_{j=1}^{n} f_r(x_j) f_s(x_j)$$

Para todo $r, s = 1, 2, \dots, m$.

Finalmente, el vector $\overline{\boldsymbol{x}} = \overline{\boldsymbol{x}}_{1\times(m+1)}$ resulta ser el siguiente

$$\overline{\boldsymbol{x}} = (1 \quad A_{1;\, f_1(x)} \quad A_{1;\, f_2(x)} \quad \cdots \quad A_{1;\, f_m(x)})$$

Donde para $r = 1, 2, \dots, m$,

$$A_{1;\, f_r(x)} = \frac{1}{n}\sum_{j=1}^{n} f_r(x_j)$$

Probamos en el Ejercicio anterior que el ajuste proporciona un estimador insesgado para $\bar{y}_n$, que es la media muestral de la variable de interés en la primera fase. Además "la varianza de tal estimador" en la segunda fase, $V_2(\bar{y}_d)$, minimiza para distribución libre la varianza vectorial de cualquier estimador del vector $(m+1)$-dimensional $\boldsymbol{a}$, que permite estimar el ajuste de modo insesgado y de mínima varianza para distribución libre del modelo general lineal propuesto.

Es sencillo comprobar entonces que el estimador $\bar{y}_d$ es insesgado para la media poblacional $\bar{y}$ de la variable de interés:

$$E(\bar{y}_d) = E_1[E_2(\bar{y}_d)] = E_1(\bar{y}_n) = \bar{y}$$

Además, teniendo en cuenta el teorema de Madow, su varianza verifica que

$$V(\bar{y}_d) = E_1[V_2(\bar{y}_d)] + V_1[E_2(\bar{y}_d)] =$$

$$E_1[V_2(\bar{y}_d)] + V_1(\bar{y}_n) = E_1[V_2(\bar{y}_d)] + \frac{\sigma_y^2}{n}$$

Aquí la media muestral en la primera fase $\bar{y}_n$ es estimador insesgado para $\bar{y}$, y de mínima varianza para distribución libre (Zacks, 1971, p. 150), $\sigma_y^2$ es la varianza poblacional para la variable de interés, $E_1[V_2(\bar{y}_d)]$ minimiza el valor esperado de las posibles varianzas con dicho ajuste lineal para distribución libre, y por tanto $V(\bar{y}_d)$ alcanza el mínimo de cualquier estimador con el ajuste lineal dado para distribución libre. Es decir, el estimador $\bar{y}_d$ es óptimo en este sentido para distribución libre en el muestreo doble usando muestreo aleatorio simple con reemplazamiento en ambas fases y con submuestreo en la segunda fase. También la media muestral es insesgada y de varianza mínima para estimar la media poblacional en el muestreo aleatorio simple con reemplazamiento de tamaño muestral fijo, entre los estimadores lineales, cuando la población tiene varianza finita, como es el caso de cualquier variable de interés uniforme discreta asociada a una población finita. Este es un ejercicio sencillo usando la técnica de los multiplicadores de Lagrange para minimizar la varianza del estimador lineal sujeto a que sea insesgado.

Una consecuencia directa de este Ejercicio es que el estimador propuesto para la media poblacional a partir del modelo general lineal propuesto aquí, o a partir de $m$ variables auxiliares conocidas de antemano como hicimos en el Ejercicio anterior, es

estimador insesgado de mínima varianza para distribución libre con el estimador insesgado de mínima varianza ajustado al modelo lineal. Es decir, es óptimo en dicho sentido descrito.

Una crítica que se hacía al muestreo aleatorio simple con reemplazamiento de tamaño muestral fijo $n$ era que el muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n$ proporciona al estimador media muestral insesgación y más precisión que el anterior. Esta crítica carece de interés práctico al observar que con muestreo aleatorio simple con reemplazamiento se obtienen muestras con menor coste esperado que con muestreo aleatorio simple sin reemplazamiento del mismo tamaño muestral $n$, ya que las unidades pueden aparecer repetidas en el diseño muestral con reemplazamiento y como consecuencia el tamaño efectivo no es fijo sino menor igual a $n$, y por tanto el coste esperado es menor o igual al coste del tamaño efectivo fijo $n$.

El modelo general lineal que hemos propuesto con una variable auxiliar no tiene que tener solo dos parámetros de ajuste a minimizar, como sería el caso de ajustar una recta de $y$ sobre $x$, ni siquiera tiene que ser un polinomio necesariamente. Un ejemplo teórico supuesto distinto a estos podría ser el siguiente

$$y_i = k_0 + k_1 e^{x_i} + k_2 \frac{1}{x_i + 2} + e_i$$

Donde en este modelo lineal con $m = 2$ tenemos las funciones

$$f_1(x) = e^x$$

Y

$$f_2(x) = \frac{1}{x + 2}$$

También es posible ajustar un modelo lineal multivariante si en la primera fase se hubieran observado $m$ variables auxiliares en la muestra aleatoria simple con reemplazamiento de tamaño fijo $n$. Incluso pueden ajustarse otras posibilidades funcionales a partir de las variables auxiliares observadas en la primera fase del muestreo doble. El razonamiento es similar al propuesto en este Ejercicio o bien en líneas semejantes a las expuestas en el Ejercicio anterior.

Obviamente el modelo concreto que se proponga en cada caso tiene que tener una gran fiabilidad basada en la experiencia, es decir, que ha de ser propuesto por expertos en el tipo de datos manejados y con experiencia en el área de trabajo al que se aplica el modelo. En este sentido, existe la posibilidad de que dos o más expertos distintos propongan distintos modelos lineales concretos, y es entonces cuando se tiene que llegar a un consenso o acuerdo del modelo más conveniente para el fin que nos proponemos. Una propuesta de solución de consenso es que cada experto, de los $m$ disponibles, aporte su función de ajuste según su conocimiento, y que el modelo con los datos se encargue de seleccionar el mejor ajuste lineal de dichas funciones.

El hecho de habernos referido al "muestreo de poblaciones finitas" se debe a que es en poblaciones finitas donde tiene sentido hablar de muestreo aleatorio simple con reemplazamiento con unidades reales identificadas y accesibles. Hablar de muestreo aleatorio simple con reemplazamiento de una población infinita limita a muestras artificiales obtenidas con un ordenador y sin salir del mismo, es decir las unidades no pueden estar identificadas todas con el medio o los medios de acceso físico para observar los datos auxiliares o de interés en todas sus unidades.

La conclusión final es que queda resuelto un problema de optimización en la estimación de la media poblacional en muestreo doble con muestreo aleatorio simple con reemplazamiento en las

dos fases y submuestreo en la segunda fase, haciendo uso de tres procedimientos de optimización, dos de ellos en estimación para distribución libre (opcionalmente de estimación insesgada de mínima varianza para la media poblacional con población de varianza finita, entre todos los estimadores lineales) y el otro en el ajuste del modelo general lineal en poblaciones finitas.

Una posible crítica a esta resolución del problema de optimización es que si el tamaño poblacional $N$ es conocido, el conjunto de poblaciones finitas con distribución uniforme discreta con $N$ valores posibles de la variable de interés es mucho más concreta que el conjunto de todas las posibles poblaciones teóricas como presupone el método de optimización para distribución libre, por lo que un estimador óptimo para distribución libre puede no serlo para el conjunto reducido de poblaciones finitas de distribución uniforme discreta con $N$ valores de la variable de interés. De hecho, como justifica Ruiz Espejo (1987c), no existe tal estimador "insesgado y uniformemente de mínima varianza", ni siquiera "uniformemente de mínimo error cuadrático medio", en el modelo de población finita fijada, que es un modelo diferente pero más próximo al modelo de muestreo doble con submuestreo, con observación de una variable auxiliar en la primera fase de muestreo.

**Ejercicio 6.22.** Proponer un estimador insesgado óptimo en muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$, haciendo uso de una una variable auxiliar $x$ de media poblacional $\bar{X} = (1/N)\sum_{i=1}^{N} x_i$.

**Solución.** Entendemos por estimador de regresión lineal clásico para la media poblacional $\bar{Y} = (1/N)\sum_{i=1}^{N} y_i$ uno que tiene la forma del tipo

$$t = \bar{y}_s + b_s(\bar{X} - \bar{x}_s).$$

A continuación vamos a minimizar su error cuadrático medio, obteniendo el valor óptimo teórico de $b = b_{\text{mín}}$, así como el estimador óptimo teórico e insesgado $t_{\text{mín}}$ y su varianza mínima teórica $V(t_{\text{mín}})$. Finalmente obtenemos un estimador insesgado óptimo $t'_{\text{mín}}$ práctico que aproxima al teórico de varianza mínima.

Para obtener el valor teórico mínimo de $b = b_{\text{mín}}$, minimizamos el error cuadrático medio del siguiente modo. Sea

$$\phi = \sum_{s \in S} [\bar{Y} - \bar{y}_s - b(\bar{X} - \bar{x}_s)]^2$$

$$= \sum_{s \in S} (\bar{Y} - \bar{y}_s)^2 + \sum_{s \in S} b^2(\bar{X} - \bar{x}_s)^2 - 2\sum_{s \in S} b(\bar{Y} - \bar{y}_s)(\bar{X} - \bar{x}_s).$$

Minimizamos la función $\phi$ derivando con respecto a la variable $b$ e igualamos a cero para obtener el punto crítico de mínimo global.

$$\frac{d\phi}{db} = \sum_{s \in S} 2b(\bar{X} - \bar{x}_s)^2 - 2\sum_{s \in S} (\bar{Y} - \bar{y}_s)(\bar{X} - \bar{x}_s) = 0.$$

De donde

$$b = \frac{\sum_{s \in S} (\bar{Y} - \bar{y}_s)(\bar{X} - \bar{x}_s)}{\sum_{s \in S} (\bar{X} - \bar{x}_s)^2}$$

$$= \frac{\sum_{s \in S} \bar{Y}(\bar{X} - \bar{x}_s)}{\sum_{s \in S} (\bar{X} - \bar{x}_s)^2} - \frac{\sum_{s \in S} \bar{y}_s(\bar{X} - \bar{x}_s)}{\sum_{s \in S} (\bar{X} - \bar{x}_s)^2}$$

$$= \frac{Cov(\bar{y}_s, \bar{x}_s)}{V(\bar{x}_s)}.$$

Al suponer que $b$ es una constante, el estimador $t_s$ sería insesgado para estimar $\bar{Y}$. Pero si esta variable $b$ fuera aleatoria, digamos $b_s$, el valor mínimo de $b = b_{\text{mín}}$ que haría la mínima varianza del estimador $t_s$ sería

$$b_{\text{mín}} = \frac{Cov(\bar{y}_s, \bar{x}_s)}{V(\bar{x}_s)}.$$

El estimador óptimo teórico y su varianza se obtienen sustituyendo el valor óptimo de $b = b_{\text{mín}}$ en el estimador, por lo que tenemos que el estimador teórico insesgado de mínima varianza es

$$t_{\text{mín}} = \bar{y}_s + \frac{Cov(\bar{y}_s, \bar{x}_s)}{V(\bar{x}_s)}(\bar{X} - \bar{x}_s).$$

Pero no puede usarse porque el numerador es a su vez una función paramétrica que no es conocida en el muestreo. Su varianza mínima se obtiene de este modo

$$V(t_{\text{mín}}) = V(\bar{y}_s) + \frac{[Cov(\bar{y}_s, \bar{x}_s)]^2}{[V(\bar{x}_s)]^2} V(\bar{x}_s)$$

$$-2\frac{Cov(\bar{y}_s, \bar{x}_s)}{V(\bar{x}_s)} Cov(\bar{y}_s, \bar{x}_s)$$

$$= V(\bar{y}_s) - \frac{[Cov(\bar{y}_s, \bar{x}_s)]^2}{V(\bar{x}_s)}.$$

Un estimador práctico insesgado y de mínima varianza sería el siguiente

$$t'_{\text{mín}} = \bar{y}_s + \frac{\widehat{Cov}(\bar{y}_s, \bar{x}_s)}{V(\bar{x}_s)}(\bar{X} - \bar{x}_s) + \frac{\widehat{Cov}\left[\widehat{Cov}(\bar{y}_s, \bar{x}_s), \bar{x}_s\right]}{V(\bar{x}_s)}$$

$$= \bar{y}_s + \frac{\frac{1}{n^2}\sum_{i \in s} y_i (x_i - \bar{X})}{V(\bar{x}_s)}(\bar{X} - \bar{x}_s) + \frac{\frac{1}{n^3}\sum_{i \in s} y_i (x_i - \bar{X})^2}{V(\bar{x}_s)}.$$

Este estimador es válido para muestreo aleatorio simple con reemplazamiento de tamaño fijo $n$, que es el estimador "insesgado óptimo" o "insesgado de mínima varianza" para la media poblacional $\bar{Y}$ usando la información auxiliar $x$ ya que resulta invariante ante permutaciones en el orden de la muestra ordenada $\boldsymbol{s}$.

Para muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n$ tendremos el estimador de regresión lineal corregido insesgado propuesto por Ruiz Espejo (2013b), ya que el

razonamiento hasta la penúltima igualdad es totalmente similar para muestras no ordenadas equiprobables como es el caso de muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo $n$.

Los tres últimos estimadores insesgados de covarianzas pueden obtenerse razonadamente de la muestra aleatoria simple con reemplazamiento de tamaño fijo $n$, y también el estimador insesgado óptimo proporcionado por $b_{\boldsymbol{s}} = \widehat{Cov}(\bar{y}_{\boldsymbol{s}}, \bar{x}_{\boldsymbol{s}})/V(\bar{x}_{\boldsymbol{s}})$ de la función paramétrica $b_{\text{mín}}$.

# Capítulo 7

# Diseño de probabilidades desiguales

En este capítulo estudiamos los métodos más importantes de muestreo con probabilidades desiguales: el muestreo con probabilidades proporcionales al tamaño con reemplazamiento debido a Hansen y Hurwitz (1943), el esquema de muestreo con probabilidades proporcionales al tamaño sin reemplazamiento debido a Sánchez-Crespo (1980), y el muestreo con probabilidades de inclusión proporcionales al tamaño que utiliza el estimador de Horvitz y Thompson (1952). En todos ellos se proporciona un estimador insesgado de la media poblacional, y un estimador insesgado de su varianza.

## 7.1 Diseño de Hansen y Hurwitz

Es un diseño ordenado $TF(n)$ que asigna una probabilidad $p_k$ de seleccionar la unidad $k$ en cada una de las $n$ selecciones independientes, donde el resultado de la $i$-ésima selección es el $i$-ésimo componente de la muestra ordenada o vector de unidades $\boldsymbol{s}$. Los valores $p_k$ son números positivos conocidos tales que

$$\sum_{k=1}^{N} p_k = 1.$$

No es un diseño de tamaño efectivo fijo, pues pueden aparecer repetida una o varias unidades en la muestra ordenada. En un esquema de urna con bolas numeradas, sería el caso de $M$ bolas de las cuales $M_k$ de ellas tienen la anotación $k$, y por tanto

$$M_k = Mp_k \ (k = 1, 2, \dots, N).$$

Se realizan con el diseño de Hansen y Hurwitz selecciones con reemplazamiento, es decir, una vez observada la bola extraída se reincorpora a la urna antes de la siguiente selección. Observar que en el caso particular en que $p_k = 1/N \ (k = 1, 2, \dots, N)$ el diseño es el ya estudiado *mas*. A veces la situación inicial es de disponer una variable auxiliar $x_k > 0 \quad (k = 1, 2, \dots, N)$ y se asignan probabilidades de selección

$$p_k = \frac{x_k}{\sum_{i \in U} x_i} = \frac{x_k}{N\bar{x}} \ \ (k = 1, 2, \dots, N).$$

Si $x_k$ es un número entero positivo para todo $k \in U$, $n_k = x_k$ puede ser la composición de la urna para seleccionar la muestra ordenada con este diseño de Hansen y Hurwitz. Las probabilidades de inclusión son ahora

$$\pi_k = 1 - (1 - p_k)^n,$$

y si $k \neq m \in U$,

$$\pi_{km} = 1 - (1 - p_k)^n - (1 - p_m)^n + (1 - p_k - p_m)^n,$$

que generalizan el diseño *mas*.

## 7.2 Estimador insesgado de Hansen y Hurwitz

El estimador más importante asociado a este diseño es el estimador de Hansen y Hurwitz (1943) también llamado estimador Hansen-Hurwitz, que se define así

$$t_{HH} = \sum_{k \in \boldsymbol{s}} \frac{y_k}{Nnp_k} = \sum_{k \in U} \frac{y_k e_k}{Nnp_k},$$

donde en una muestra ordenada $\boldsymbol{s}$ la unidad $k$ aparece un número de veces $e_k$ $(= 0, 1, 2, \dots, n)$ al ser $p_k > 0$ la probabilidad de selección constante de la unidad $k$ en las $n$ extracciones de una bola aleatoria. Es decir, $e_k$ es el número de veces que la unidad $k$ aparece en la secuencia de la muestra ordenada $\boldsymbol{s}$ con diseño de Hansen y Hurwitz. Al ser independientes las selecciones de la urna, el modelo creado es una distribución $N$ -dimensional $(e_1, e_2, \dots, e_k, \dots, e_N)$ que se distribuye multinomial de parámetros $n$ y $p_k$. Es decir,

$$E(e_k) = np_k,$$

$$V(e_k) = np_k(1 - p_k),$$

$$E(e_k^2) = V(e_k) + [E(e_k)]^2 = np_k - np_k^2 + n^2 p_k^2,$$

y si $k \neq m \in U$,

$$Cov(e_k, e_m) = -np_k p_m,$$

$$E(e_k e_m) = Cov(e_k, e_m) + E(e_k)E(e_m) = (n^2 - n)p_k p_m.$$

El estimador $t_{HH}$ es insesgado para la media poblacional, pues

$$E(t_{HH}) = E\left(\sum_{k \in U} \frac{y_k e_k}{Nnp_k}\right) = \sum_{k \in U} \frac{y_k}{Nnp_k} E(e_k) = \bar{y}.$$

## 7.3 Varianza del estimador Hansen-Hurwitz

Directamente

$$V(t_{HH}) = E(t_{HH}^2) - \bar{y}^2.$$

Pero

$$E(t_{HH}^2) = E\left(\sum_{k \in U} \frac{y_k^2 e_k^2}{N^2 n^2 p_k^2} + \sum_{k \neq m \in U} \frac{y_k y_m e_k e_m}{N^2 n^2 p_k p_m}\right) =$$

$$\frac{1}{N^2 n^2}\left[\sum_{k \in U} \frac{y_k^2}{p_k^2} E(e_k^2) + \sum_{k \neq m \in U} \frac{y_k y_m}{p_k p_m} E(e_k e_m)\right] =$$

$$\frac{1}{N^2 n^2}\left[\sum_{k \in U} \frac{y_k^2}{p_k^2}(np_k - np_k^2 + n^2 p_k^2) + \right.$$

$$\left.\sum_{k \neq m \in U} \frac{y_k y_m}{p_k p_m}(n^2 p_k p_m - np_k p_m)\right] =$$

$$\frac{1}{N^2}\left[\frac{1}{n}\sum_{k \in U} \frac{y_k^2}{p_k^2} p_k - \frac{1}{n}\sum_{k \in U} y_k^2 + \right.$$

$$\left.\sum_{k \in U} y_k^2 + \sum_{k \neq m \in U} y_k y_m - \frac{1}{n}\sum_{k \neq m \in U} y_k y_m\right] =$$

$$\frac{1}{N^2}\left(\frac{1}{n}\sum_{k \in U} \frac{y_k^2}{p_k^2} p_k + N^2 \bar{y}^2 - \frac{N^2 \bar{y}^2}{n}\right),$$

de donde

$$V(t_{HH}) = E(t_{HH}^2) - \bar{y}^2 = \frac{1}{N^2}\left(\frac{1}{n}\sum_{k \in U} \frac{y_k^2}{p_k^2} p_k - \frac{N^2 \bar{y}^2}{n}\right) =$$

$$\frac{1}{N^2 n}\left(\sum_{k\in U}\frac{y_k^2}{p_k^2}p_k - N^2\bar{y}^2\right) = \frac{1}{N^2 n}\sum_{k\in U}\left(\frac{y_k}{p_k} - N\bar{y}\right)^2 p_k.$$

## 7.4 Estimador insesgado de la varianza

El estimador insesgado de la varianza es

$$\hat{V}(t_{HH}) = \frac{\sum_{k\in \boldsymbol{s}}\left(\frac{y_k}{p_k} - Nt_{HH}\right)^2}{N^2 n(n-1)} = \frac{\sum_{k\in \boldsymbol{s}}\frac{y_k^2}{p_k^2} - nN^2 t_{HH}^2}{N^2 n(n-1)},$$

desarrollando el cuadrado y teniendo en cuenta que por la definición del estimador $t_{HH}$,

$$\sum_{k\in \boldsymbol{s}}\frac{y_k}{p_k} = nNt_{HH}.$$

También admite la expresión siguiente ya que la varianza es invariante por cambio de origen,

$$\hat{V}(t_{HH}) = \frac{\sum_{k\in \boldsymbol{s}}\left(\frac{y_k}{p_k} - N\bar{y}\right)^2 - n(Nt_{HH} - N\bar{y})^2}{N^2 n(n-1)},$$

o bien, desarrollando los cuadrados del numerador y simplificando teniendo en cuenta la definición del estimador $t_{HH}$, obtenemos el numerador de la segunda expresión del estimador insesgado de la varianza de $t_{HH}$. De este modo queda una nueva expresión de $\hat{V}(t_{HH})$, donde el primer sumando del numerador puede expresarse así

$$\sum_{k\in \boldsymbol{s}}\left(\frac{y_k}{p_k} - N\bar{y}\right)^2 = \sum_{k\in U}\left(\frac{y_k}{p_k} - N\bar{y}\right)^2 e_k.$$

Efectivamente, el estimador es insesgado porque

$$E\left[\hat{V}(t_{HH})\right] = \frac{1}{N^2 n(n-1)}\left[\sum_{k\in U}\left(\frac{y_k}{p_k} - N\bar{y}\right)^2 np_k - nN^2V(t_{HH})\right],$$

o bien,

$$E\left[\hat{V}(t_{HH})\right] = \frac{1}{N^2(n-1)}\left[N^2nV(t_{HH}) - N^2V(t_{HH})\right] = V(t_{HH}).$$

## 7.5 Estimador insesgado de Sánchez-Crespo

El estimador de Sánchez-Crespo es análogo al de Hansen-Hurwitz y también ahora la unidad $k$ $(k = 1, 2, \dots, N)$ tiene asociadas $M_k = Mp_k$ bolas con su identificador dentro de la urna. El diseño de Sánchez-Crespo varía en que la selección de bolas para establecer la secuencia $\boldsymbol{s}$ de unidades en la muestra ordenada, se hace sin reemplazamiento. Con este esquema de muestreo, se define como $e_k$ el número de veces que la unidad $k$ es extraída de la urna, siendo el vector $N$-dimensional $(e_1, e_2, \dots, e_k, \dots, e_N)$ una variable aleatoria hipergeométrica generalizada cuya función de cuantía es

$$p(e_1, \dots, e_N) = \frac{\binom{M_1}{e_1}\dots\binom{M_N}{e_N}}{\binom{M}{n}},$$

y cuyos principales momentos son

$$E(e_k) = np_k,$$

$$V(e_k) = \frac{M-n}{M-1}np_k(1-p_k),$$

y para $k \neq m \in U$,

$$Cov(e_k, e_m) = -\frac{M-n}{M-1} np_k p_m.$$

De estos momentos podemos comprobar que el estimador de Sánchez-Crespo

$$t_{SC} = \sum_{k \in \boldsymbol{s}} \frac{y_k}{Nnp_k} = \sum_{k \in U} \frac{y_k e_k}{Nnp_k}$$

es diferente al de Hansen-Hurwitz $t_{HH}$, que aunque se escriban igual, el significado de la muestra $\boldsymbol{s}$ ha cambiado y la distribución de las multiplicidades $e_k$ ha cambiado consecuentemente. A pesar de haber cambiado la distribución, la esperanza de $e_k$ sigue siendo igual, y por tanto la demostración de la insesgación del estimador $t_{SC}$ no queda afectada en nada sustancial, y ambos estimadores son insesgados para estimar la media poblacional $\bar{y}$ con sus correspondientes diseños muestrales.

La varianza del estimador Sánchez-Crespo puede obtenerse ahora así

$$V(t_{SC}) = V\left(\sum_{k \in U} \frac{y_k e_k}{Nnp_k}\right) =$$

$$\frac{1}{N^2 n^2}\left[\sum_{k \in U} \frac{y_k^2}{p_k^2} V(e_k) + \sum_{k \neq m \in U} \frac{y_k y_m}{p_k p_m} Cov(e_k, e_m)\right] =$$

$$\frac{M-n}{M-1}\frac{1}{N^2 n^2}\left[\sum_{k \in U} \frac{y_k^2}{p_k} n(1-p_k) - \sum_{k \neq m \in U} y_k y_m n\right] =$$

$$\frac{M-n}{M-1}\frac{1}{N^2 n}\left(\sum_{k \in U} \frac{y_k^2}{p_k^2} p_k - N^2 \bar{y}^2\right) = \frac{M-n}{M-1} V(t_{HH}) \leq V(t_{HH}).$$

Un estimador insesgado de la varianza de $t_{SC}$ con el mismo diseño muestral de Sánchez-Crespo es

$$\hat{V}(t_{SC}) = \frac{M-n}{M}\frac{1}{N^2 n(n-1)}\sum_{k\in \boldsymbol{s}}\left(\frac{y_k}{p_k} - Nt_{SC}\right)^2.$$

En efecto, de modo similar a lo que hicimos con el estimador insesgado de la varianza de $t_{HH}$, ahora tenemos que

$$\hat{V}(t_{SC}) = \frac{M-n}{M}\frac{\sum_{k\in U}\left(\frac{y_k}{p_k} - N\bar{y}\right)^2 e_k - n(Nt_{SC} - N\bar{y})^2}{N^2 n(n-1)},$$

de donde

$$E\left[\hat{V}(t_{SC})\right] = \frac{M-n}{M}\frac{1}{N^2(n-1)}\left[N^2 nV(t_{HH}) - N^2 V(t_{SC})\right] =$$

$$\frac{M-n}{M(n-1)}\left[\frac{(M-1)n}{M-n} - 1\right]V(t_{SC}) = V(t_{SC}).$$

## 7.6 Muestreo con probabilidades de inclusión

Este diseño puede ser introducido para muestras ordenadas o muestras no ordenadas. Básicamente consiste en proporcionar un estimador insesgado de la media poblacional, que llamamos estimador de Horvitz y Thompson (1952), y otros estimadores insesgados de la varianza del estimador anterior. El estimador Horvitz-Thompson se define indistintamente para diseño no ordenado

$$t_{HT} = \sum_{k\in s}\frac{y_k}{N\pi_k} = \sum_{k\in U}\frac{y_k e_k}{N\pi_k},$$

y para diseño ordenado también

$$t_{HT} = \sum_{k \in r(s)} \frac{y_k}{N\pi_k} = \sum_{k \in U} \frac{y_k e_k}{N\pi_k},$$

donde en ambos casos $e_k$ es una variable aleatoria indicador que toma valor 1 si la unidad $k$ pertenece a la muestra, y toma valor 0 si dicha unidad no pertenece a la muestra. Por tanto $e_k$ no recoge el efecto de la multiplicidad de una unidad en la muestra ordenada, sino solo su pertenencia o no a la muestra. Es sencillo ver que

$$E(e_k) = 1 \cdot p(k \in s) + 0 \cdot p(k \notin s) = \pi_k,$$

y como $e_k^2 = e_k$,

$$V(e_k) = E(e_k^2) - [E(e_k)]^2 = \pi_k - \pi_k^2 = \pi_k(1 - \pi_k),$$

y si $k \neq m \in U$,

$$E(e_k e_m) = 1 \cdot p(k \text{ y } m \in s) + 0 \cdot p(k \text{ o } m \notin s) = \pi_{km},$$

y

$$Cov(e_k, e_m) = E(e_k e_m) - E(e_k)E(e_m) = \pi_{km} - \pi_k \pi_m.$$

Por tanto, el estimador $t_{HT}$ es insesgado para estimar la media poblacional pues

$$E(t_{HT}) = E\left(\sum_{k \in U} \frac{y_k e_k}{N\pi_k}\right) = \sum_{k \in U} \frac{y_k E(e_k)}{N\pi_k} = \bar{y}.$$

La varianza del estimador $t_{HT}$ se obtiene así

$$V(t_{HT}) = \frac{1}{N^2} V\left(\sum_{k \in U} \frac{y_k e_k}{\pi_k}\right) =$$

$$\frac{1}{N^2}\left[\sum_{k \in U} V\left(\frac{y_k e_k}{\pi_k}\right) + \sum_{k \neq m \in U} Cov\left(\frac{y_k e_k}{\pi_k}, \frac{y_m e_m}{\pi_m}\right)\right] =$$

$$\frac{1}{N^2}\left[\sum_{k\in U}\frac{y_k^2}{\pi_k^2}V(e_k)+\sum_{k\neq m\in U}\frac{y_k y_m}{\pi_k\pi_m}Cov(e_k,e_m)\right]=$$

$$\frac{1}{N^2}\left[\sum_{k\in U}\frac{y_k^2}{\pi_k^2}\pi_k(1-\pi_k)+\sum_{k\neq m\in U}\frac{y_k y_m}{\pi_k\pi_m}(\pi_{km}-\pi_k\pi_m)\right].$$

Si $\pi_{km}>0$ para todos los pares $k\neq m\in U$, entonces un estimador insesgado de la varianza del estimador Horvitz-Thompson es

$$\hat{V}(t_{HT})=\frac{1}{N^2}\left[\sum_{k\in s}\frac{y_k^2}{\pi_k^2}(1-\pi_k)+\sum_{k\neq m\in s}\frac{y_k y_m(\pi_{km}-\pi_k\pi_m)}{\pi_k\pi_m\pi_{km}}\right],$$

que admite la expresión siguiente en función de la variable aleatoria indicador

$$\hat{V}(t_{HT})=\frac{1}{N^2}\left[\sum_{k\in U}\frac{y_k^2}{\pi_k^2}(1-\pi_k)\,e_k+\right.$$

$$\left.\sum_{k\neq m\in U}\frac{y_k y_m(\pi_{km}-\pi_k\pi_m)}{\pi_k\pi_m\pi_{km}}e_k e_m\right].$$

Obviamente, las propiedades de las esperanzas matemáticas de los indicadores y sus productos son conocidas, por lo que sustituyéndolas podemos concluir que

$$E\left[\hat{V}(t_{HT})\right]=V(t_{HT}).$$

Otro estimador insesgado de la varianza del estimador Horvitz-Thompson es el proporcionado por Yates y Grundy (1953) que podemos expresar así (variando $k$ y $m$)

$$\hat{V}_{YG}(t_{HT})=\frac{1}{N^2}\sum_{k<m\in s}\frac{\pi_k\pi_m-\pi_{km}}{\pi_{km}}\left(\frac{y_k}{\pi_k}-\frac{y_m}{\pi_m}\right)^2=$$

$$\frac{1}{N^2}\sum_{k<m\in U}\frac{\pi_k\pi_m-\pi_{km}}{\pi_{km}}\left(\frac{y_k}{\pi_k}-\frac{y_m}{\pi_m}\right)^2 e_k e_m.$$

## 7.7 Ejercicios resueltos

**Ejercicio 7.1.** Se selecciona una muestra ordenada de tamaño fijo $n=4$ con diseño de probabilidades proporcionales al tamaño con reemplazamiento. La muestra seleccionada es $(4,3,5,7)$ y los valores de la variable de interés observada en dichas unidades han resultado ser ordenadamente $(33,21,15,9)$. Estimar la media poblacional con el estimador Hansen-Hurwitz y para el estimador Sánchez-Crespo (si la selección hubiera sido sin reemplazamiento), así como sus varianzas con la misma muestra ordenada de tamaño fijo. Como datos del problema, el tamaño poblacional es 20, y las probabilidades de primera selección son $p_4=p_3=1/20$ y $p_5=p_7=1/40$. El número de bolas en la urna antes de la primera selección es de 80.

**Solución.** El estimador insesgado de la media poblacional con el estimador Hansen-Hurwitz y Sánchez-Crespo coinciden para la misma muestra aunque el diseño muestral habría cambiado de uno a otro estimador. Con los datos del problema,

$$t_{HH}=t_{SC}=26.5$$

En cuanto a los estimadores insesgados de la varianza, cambian de valor numérico además del diseño muestral, quedando

$$\hat{V}(t_{HH})=105.8;\ \hat{V}(t_{SC})=100.5$$

**Ejercicio 7.2.** Se obtiene una muestra con diseño de probabilidades de inclusión. Esta muestra seleccionada resulta ser reducida la muestra no ordenada $\{2, 1\}$ y los valores observados son $y_2 = 4$ e $y_1 = 8$. Si las probabilidades de inclusión son

$$\pi_1 = \frac{1}{3}, \qquad \pi_2 = \frac{2}{3}, \qquad \pi_{1,2} = \frac{2}{9},$$

estimar insesgadamente la media poblacional y estimar sin sesgo la varianza de tal estimador, si el tamaño poblacional es $N = 4$.

**Solución.** El estimador insesgado de la media poblacional es el estimador Horvitz-Thompson, que para los datos recibidos la estima en

$$t_{HT} = 7.5$$

Y un estimador insesgado de su varianza es por el primer método

$$\hat{V}(t_{HT}) = 23.5,$$

mientras que por el estimador de Yates y Grundy sería

$$\hat{V}_{YG}(t_{HT}) = 0.$$

**Ejercicio 7.3.** De una población finita de tamaño $N$, se selecciona una muestra ordenada de modo que la primera selección se realiza con probabilidades proporcionales al tamaño indicado por la variable auxiliar positiva $x$, y las $n - 1$ restantes unidades se obtienen con probabilidades iguales sin reemplazamiento. Demostrar que la probabilidad de selección de una muestra $s$ de tamaño efectivo fijo $n$ es proporcional a la media muestral $\bar{x}_s$.

**Solución.** La probabilidad de seleccionar la unidad $i = 1, 2, \dots, N$ en la primera selección de la muestra ordenada, es

$$p_1(i_1) = \frac{x_{i_1}}{N\bar{x}}.$$

La probabilidad de seleccionar la unidad $i_2 \neq i_1$ en la segunda selección de la muestra ordenada es

$$p_2(i_2|i_1) = \frac{1}{N-1}.$$

La probabilidad de seleccionar la unidad $i_3 \neq i_1, i_2$ en la tercera selección de la muestra ordenada es

$$p_3(i_3|i_1, i_2) = \frac{1}{N-2}.$$

Así llegaremos a que la probabilidad de que la unidad $i_n \neq i_1, i_2, \dots, i_{n-1}$ sea seleccionada en la $n$-ésima selección de la muestra ordenada es

$$p_n(i_n|i_1, i_2, \dots, i_{n-1}) = \frac{1}{N-(n-1)}.$$

Por todo ello, la probabilidad de seleccionar la muestra ordenada $\boldsymbol{s} = (i_1, i_2, i_3, \dots, i_n)$ es

$$p(\boldsymbol{s}) = p_1(i_1)p_2(i_2|i_1)p_3(i_3|i_1, i_2)\cdots p_n(i_n|i_1, i_2, \dots, i_{n-1}) =$$

$$\frac{x_{i_1}}{N\bar{x}}\frac{1}{N-1}\frac{1}{N-2}\cdots\frac{1}{N-(n-1)} = \frac{x_{i_1}}{\bar{x}}\frac{(N-n)!}{N!}.$$

Considerando ahora muestras conjunto o no ordenadas, la probabilidad de seleccionar la muestra $s = \{i_1, i_2, i_3, \dots, i_n\}$ es

$$p(s) = \sum_{j=1}^{n}\frac{x_{i_j}}{\bar{x}}\frac{(N-n)!}{N!}(n-1)! =$$

$$\frac{\bar{x}_s}{\bar{x}}\frac{(N-n)!\,n!}{N!} = \frac{\bar{x}_s}{\bar{x}\binom{N}{n}}.$$

Este diseño muestral ordenado, debido a Midzuno (1951), y su diseño muestral no ordenado deducido proporcionan estimadores de razón insesgados para estimar la media poblacional $\bar{y}$.

**Ejercicio 7.4.** Demostrar que el estimador

$$t = \frac{1}{N^2}\sum_{i\,\in\, s}\frac{y_i}{\pi_i} + \frac{1}{N^2}\sum_{i\,\in\, s}\sum_{j\neq i\,\in\, s}\frac{y_j}{\pi_{ij}}$$

es insesgado para estimar la media poblacional $\bar{y}$, cuando $s$ es una muestra conjunto o no ordenada, y las probabilidades de inclusión de primer y de segundo órdenes son positivas.

**Solución.** Denotamos por $e_i$ a la variable aleatoria que toma valor 1 si $i \in s$, y toma valor 0 si $i \notin s$. Entonces podemos escribir el estimador $t$ de la forma

$$t = \frac{1}{N^2}\sum_{i\,\in\, U}\frac{y_i}{\pi_i}e_i + \frac{1}{N^2}\sum_{i\,\in\, U}\sum_{j\neq i\,\in\, U}\frac{y_j}{\pi_{ij}}e_i e_j,$$

de donde

$$E(t) = E\left(\frac{1}{N^2}\sum_{i\,\in\, U}\frac{y_i}{\pi_i}e_i + \frac{1}{N^2}\sum_{i\,\in\, U}\sum_{j\neq i\,\in\, U}\frac{y_j}{\pi_{ij}}e_i e_j\right) =$$

$$\frac{1}{N^2}\sum_{i\,\in\, U}\frac{y_i}{\pi_i}E(e_i) + \frac{1}{N^2}\sum_{i\,\in\, U}\sum_{j\neq i\,\in\, U}\frac{y_j}{\pi_{ij}}E\left(e_i e_j\right) =$$

$$\frac{1}{N^2}\sum_{i\in U} y_i + \frac{1}{N^2}\sum_{i\in U}\sum_{j\neq i\in U} y_j = \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U} y_j =$$

$$\frac{1}{N^2} N^2 \bar{y} = \bar{y}.$$

**Ejercicio 7.5.** Justificar que el estimador

$$\widehat{\sigma^2} = \frac{1}{2N^2} \sum_{i \in s} \sum_{j \neq i \in s} \frac{\left(y_i - y_j\right)^2}{\pi_{ij}}$$

es insesgado para la varianza poblacional, donde el diseño muestral verifica que las probabilidades de inclusión de segundo orden son positivas.

**Solución.** Como vimos en el Ejercicio 1.7,

$$\sigma^2 = \frac{1}{2N^2} \sum_{i \in U} \sum_{j \neq i \in U} \left(y_i - y_j\right)^2,$$

por lo que un estimador insesgado de $\sigma^2$ es

$$\widehat{\sigma^2} = \frac{1}{2N^2} \sum_{i \in U} \sum_{j \neq i \in U} \frac{\left(y_i - y_j\right)^2}{\pi_{ij}} e_i e_j =$$

$$\frac{1}{2N^2} \sum_{i \in s} \sum_{j \neq i \in s} \frac{\left(y_i - y_j\right)^2}{\pi_{ij}},$$

siendo $e_i$ la variable aleatoria indicador de la unidad $i$ en la muestra, es decir $e_i$ toma valor 1 si $i \in s$, y toma valor 0 si $i \notin s$. También hemos denotado por $\pi_{ij} = E\left(e_i e_j\right)$ a la probabilidad de inclusión de las unidades $i$ y $j$ en la muestra. Este estimador se debe a Yates y Grundy (1953).

**Ejercicio 7.6.** Haciendo uso del estimador de Hansen y Hurwitz, proponer un estimador insesgado del tamaño de la población finita $N$. Obtener la varianza del estimador, y deducir la desigualdad entre la media armónica y la media poblacional cuando la variable considerada es positiva.

**Solución.** Tomando como probabilidad de selección de la unidad $i$ al tamaño relativo positivo, al ser $x$ una variable positiva,

$$p_i = \frac{x_i}{N\bar{x}},$$

tenemos que

$$\sum_{i=1}^{N} p_i = 1.$$

El estimador insesgado del tamaño poblacional, es el estimador usual de Hansen y Hurwitz del total de la variable unidad, es decir

$$\widehat{N} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{p_{k_j}},$$

donde $k_j$ es la unidad seleccionada en la $j$-ésima selección de la muestra ordenada o secuencia. La varianza de este estimador es

$$V(\widehat{N}) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{1}{p_i} - N\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{1}{p_i} - N^2\right).$$

Como la varianza de $\widehat{N}$ es siempre positiva o cero, deducimos que

$$\sum_{i=1}^{N} \frac{1}{p_i} \geq N^2,$$

de donde,

$$\bar{x} \sum_{i=1}^{N} \frac{1}{x_i} \geq N,$$

o bien,

$$\frac{1}{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i}} \leq \bar{x},$$

es decir, la media armónica de valores positivos es menor o igual a la media aritmética de esos mismos valores.

**Ejercicio 7.7.** En el procedimiento de muestreo aleatorio con probabilidades distintas de selección sin reemplazamiento debido a Sánchez-Crespo, obtener la composición de la urna con bolas repetidas o no repetidas de la misma unidad de la población que guardando la proporción de bolas inicial, sea más eficiente para que el estimador de Sánchez-Crespo sea el más preciso entre sus posibles para estimar la media poblacional.

**Solución.** Si la composición de la urna es de $M_k$ bolas con el mismo identificador de la unidad $k$ $(k = 1, 2, \dots, N)$, en total hay

$$\sum_{k=1}^{N} M_k = M$$

bolas en la urna entre las unidades de la población finita de tamaño $N$. Obtendremos el máximo común divisor $m$ del conjunto $\{M_k : k = 1, 2, \dots, N\}$. La composición de la urna que resulta más

precisa es la que asigna a la unidad $k$ de la población finita un número $M_k/m$ de bolas con su identificador $k$. En total habrá

$$\sum_{k=1}^{N} \frac{M_k}{m} = \frac{M}{m}$$

bolas en la urna más precisa. De este modo el número total de bolas en la urna será el mínimo posible que puede proveer un procedimiento y un estimador de Sánchez-Crespo para estimar con su esquema la media poblacional y que respete la proporcionalidad de los valores enteros $M_k$. Como la varianza del estimador de Sánchez-Crespo es

$$V(t_{SC}) = \frac{M-n}{M-1} V(t_{HH})$$

proporcional a la varianza del estimador de Hansen-Hurvitz, por lo que la varianza se minimiza cuando la fracción

$$\frac{M-n}{M-1} = f(M)$$

es la menor posible con $M \geq n \geq 2$. En efecto, derivando la función $f$ respecto a $M$, tenemos

$$f'(M) = \frac{M-1-(M-n)}{(M-1)^2} = \frac{n-1}{(M-1)^2} > 0.$$

Por lo tanto, la función $f$ es creciente y por esto se hace mínima para el menor valor posible de $M$, es decir cuando el máximo común divisor de los $M_k$ sea 1. En el caso de que el máximo común divisor fuera un entero $m > 1$, la composición de la urna sería $M_k/m$ bolas como la multiplicidad de la unidad $k$ en la urna en extracciones sin reemplazamiento de mínima varianza para el estimador de Sánchez-Crespo, que respeta la proporción $M_k/M$

para la unidad $k$ entre las posibles multiplicidades en la urna para las unidades de la población finita.

**Ejercicio 7.8.** Una muestra ordenada seleccionada de una población finita es $\boldsymbol{s} = (1, 2, 1)$, con los valores observados $y_1 = 1$, e $y_2 = 2$. Si el diseño es muestreo aleatorio simple con reemplazamiento de tamaño fijo 3, de una población finita de tamaño 4. Se pide: Estimar la media poblacional con el estimador Horvitz-Thompson. Estimar sin sesgo la varianza del estimador anteriormente usado.

**Solución.** El estimador Horvitz-Thompson es

$$t_{HT} = \sum_{k \in r(\boldsymbol{s})} \frac{y_k}{N\pi_k} = \frac{3 \cdot 64}{229} = \frac{192}{229}$$

donde

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^n = 1 - \left(\frac{3}{4}\right)^3 = \frac{229}{256}.$$

Un estimador insesgado de la varianza del estimador anterior es

$$\hat{V}(t_{HT}) =$$

$$\frac{1}{N^2}\left[\sum_{k \in s} \frac{y_k^2}{\pi_k^2}(1 - \pi_k) + \sum_{k \neq m \in s} \frac{y_k y_m (\pi_{km} - \pi_k \pi_m)}{\pi_k \pi_m \pi_{km}}\right] =$$

$$\frac{135 \cdot 256 \cdot 9 + 128(72 \cdot 256 - 229^2)}{16 \cdot 9 \cdot 229^2}$$

donde

$$\pi_{km} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n = \frac{9}{32}.$$

**Ejercicio 7.9.** Proponer un estimador insesgado de la media poblacional basado en el estadístico media-de-razones con esquema muestral de probabilidades desiguales. Calcular su varianza y estimar insesgadamente esta varianza.

**Solución.** El estimador insesgado media-de-razones es

$$t_{MR} = \frac{\bar{x}}{n}\sum_{i\in \boldsymbol{s}} \frac{y_i}{x_i}$$

tanto para el esquema muestral de Hansen y Hurwitz, como para el esquema muestral de Sánchez-Crespo. Bastaría con que en el primer esquema muestral tomáramos $p_i = x_i/N\bar{x}$ $(i = 1, 2, \dots, N)$ como probabilidad de seleccionar la unidad $i$ en cada extracción, y en el segundo esquema muestral $M_i$ es el mínimo número natural proporcional a $x_i > 0$, para todo $i = 1, 2, \dots, N$; por esta razón, 1 es el máximo factor común de $\{M_i : i = 1, 2, \dots, N\}$ con la proporcionalidad $M_i \propto x_i$ $(i = 1, 2, \dots, N)$.

De la teoría vista, deducimos que la varianza del estimador $t_{MR}$ es

$$V_{HH}(t_{MR}) = \frac{1}{n}(A_{2-1}A_{01} - A_{10}^2),$$

donde

$$A_{kj} = \frac{1}{N}\sum_{i\in U} y_i^k x_i^j$$

es el momento poblacional no central de órdenes $k$ y $j$. De modo similar, tenemos

$$V_{SC}(t_{MR}) = \frac{M-n}{M-1} V_{HH}(t_{MR})$$

donde $M = \sum_{i=1}^{N} M_i$.

El estimador insesgado de esta varianza para el primer esquema muestral es

$$\hat{V}_{HH}(t_{MR}) = \frac{1}{n-1}\left(\frac{\bar{x}^2}{n}\sum_{i\in U}\frac{y_i^2}{x_i^2}e_i - t_{MR}^2\right) =$$

$$\frac{\sum_{i\in U}\left(\bar{x}\frac{y_i}{x_i} - t_{MR}\right)^2 e_i}{n(n-1)},$$

donde $e_i$ sigue la distribución multinomial de parámetros $n$ y $p_i$ $(i = 1, 2, \dots, N)$. Para el segundo esquema muestral es

$$\hat{V}_{SC}(t_{MR}) = \frac{M-n}{M}\frac{1}{n-1}\left(\frac{\bar{x}^2}{n}\sum_{i\in U}\frac{y_i^2}{x_i^2}e_i - t_{MR}^2\right) =$$

$$\frac{M-n}{M}\frac{1}{n(n-1)}\sum_{i\in U}\left(\bar{x}\frac{y_i}{x_i} - t_{MR}\right)^2 e_i$$

donde ahora $e_i$ sigue el esquema muestral de una distribución hipergeométrica de parámetros $N$, $n$, y $M_i$ $(i = 1, 2, \dots, N)$.

**Ejercicio 7.10.** Proponer una selección de una muestra de unidades de una población finita que en cada selección se obtiene una unidad $i$ con probabilidad proporcional (e independiente) a cierta cantidad positiva $x_i$. Explicar la selección de la muestra en el caso de que la muestra sea sin reemplazamiento de unidades.

**Solución.** Consiste en dividir o clasificar el intervalo $[0, 1]$ en tantos subintervalos como es el tamaño poblacional $N$. Así, llamando a

$$p_i = \frac{x_i}{\sum_{j=1}^{N} x_j}$$

El primer subintervalo sería

$$[0, p_1)$$

El segundo subintervalo sería

$$[p_1, p_1 + p_2)$$

El tercer subintervalo sería

$$[p_1 + p_2, p_1 + p_2 + p_3)$$

Y así sucesivamente hasta el último o $N$-ésimo, que sería

$$\left[\sum_{i=1}^{N-1} p_i, 1\right]$$

Seguidamente se seleccionan un grupo de dígitos, que siguiendo al número 0., estuviese en uno de los subintervalos descritos. La primera unidad seleccionada sería la que su identificador indica la posición del subintervalo seleccionado. Las siguientes selecciones de unidades de la muestra se obtienen de modo similar a como hemos obtenido la primera, con los sucesivos dígitos generados aleatoriamente.

En el caso sin reemplazamiento de unidades se procede similarmente, pero desechando unidades ya extraidas antes.

# Capítulo 8

# Muestreo por conglomerados

Un conglomerado es una clase o parte de una clasificación de la población finita en que se divide dicha población. La diferencia entre un conglomerado y un estrato es que el primero se selecciona aleatoriamente, mientras que el segundo se selecciona con seguridad, es decir se incluye con seguridad en la muestra aunque no sea en la totalidad de sus unidades. Cuando hablamos de grupos en el muestreo posagrupado, los grupos seleccionados en la primera fase, se muestrean en la segunda fase con seguridad, por lo que reciben el nombre de estratos también.

Si tenemos $L$ conglomerados, cada uno de ellos contiene varias unidades elementales de la población finita. Llamando $i$ al conglomerado $i$-ésimo $(i = 1, 2, \dots, L)$ que contiene $N_i$ unidades (secundarias) de la población finita, el número total de unidades de la población finita, número total de unidades secundarias o elementos de la población, o tamaño de la población finita es

$$N = \sum_{i=1}^{L} N_i.$$

En el muestreo unietápico por conglomerados (o muestreo por conglomerados sin submuestreo), se seleccionan $n$ conglomerados de entre los $L$ que constituyen el colectivo, y dentro de cada uno de estos $n$ conglomerados se observan todas las unidades secundarias que contienen. De este modo, los

conglomerados son las unidades de muestreo y las unidades secundarias son las unidades de observación ya que es de las unidades secundarias de donde se obtiene u observa la información de la variable de interés.

Denotamos por $y_{ij}$ a la variable de interés observada en la unidad $j$-ésima del conglomerado $i$-ésimo $(j = 1, 2, \dots, N_i; i = 1, 2, \dots, L)$. La media del conglomerado $i$ $(i = 1, 2, \dots, L)$ es

$$\bar{y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij},$$

y el total del conglomerado $i$ es $N_i\bar{y}_i$. La media poblacional es

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{L} N_i\bar{y}_i.$$

La cuasivarianza del conglomerado $i$ es

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(y_{ij} - \bar{y}_i\right)^2 = \frac{N_i}{N_i - 1}\sigma_i^2,$$

y la cuasivarianza poblacional es

$$S^2 = \frac{1}{N - 1}\sum_{i=1}^{L}\sum_{j=1}^{N_i}\left(y_{ij} - \bar{y}\right)^2 = \frac{N}{N - 1}\sigma^2.$$

En estas condiciones, el análisis de la varianza o variación total se puede descomponer en la variación dentro de conglomerados y la variación entre conglomerados, de modo igual al muestreo estratificado,

$$\sigma^2 = \sum_{i=1}^{L} \frac{N_i}{N} \sigma_i^2 + \sum_{i=1}^{L} \frac{N_i}{N} (\bar{y}_i - \bar{y})^2.$$

El “coeficiente de correlación intraconglomerados” es

$$\delta = \frac{\sum_{i=1}^{L} \frac{\sum_{k \neq m}^{N_i} (y_{ik} - \bar{y})(y_{im} - \bar{y})}{L N_i (N_i - 1)}}{\sigma^2},$$

que es un indicador del grado de homogeneidad de los conglomerados, donde en el denominador aparece la varianza poblacional.

## 8.1 Muestreo por conglomerados de igual tamaño

En el caso en que los conglomerados sean del mismo tamaño,

$$N_i = \bar{N} \ \ (i = 1, 2, \dots, L),$$

la media muestral en muestreo por conglomerados unietápico es

$$\bar{y}_c = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\bar{N}} \sum_{j=1}^{\bar{N}} y_{ij} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i,$$

donde $n$ es el tamaño muestral, $1 \leq n \leq L$, o número de unidades primarias o conglomerados seleccionados en la muestra. Este estimador es insesgado para estimar la media poblacional $\bar{y}$, pues en este caso la media de las medias de los conglomerados coincide con la media poblacional. En efecto, ya que $N = L\bar{N}$,

$$\frac{1}{L} \sum_{i=1}^{L} \bar{y}_i = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{\bar{N}} \sum_{j=1}^{\bar{N}} y_{ij} = \frac{1}{N} \sum_{i=1}^{L} \sum_{j=1}^{\bar{N}} y_{ij} = \bar{y}.$$

La varianza de $\bar{y}_c$ se obtiene directamente tanto para diseño *mas* como para diseño *mia*. Para diseño *mas*,

$$V(\bar{y}_c) = \frac{\sigma^2_{\bar{y}_i}}{n}.$$

Para diseño *mia*,

$$V(\bar{y}_c) = \frac{L-n}{L}\frac{S^2_{\bar{y}_i}}{n}.$$

Pero desarrollando $\sigma^2_{\bar{y}_i}$ tenemos

$$\sigma^2_{\bar{y}_i} = \frac{1}{L}\sum_{i=1}^{L}(\bar{y}_i - \bar{y})^2 = \frac{1}{L\bar{N}^2}\sum_{i=1}^{L}\left[\sum_{j=1}^{\bar{N}}(y_{ij} - \bar{y})\right]^2 =$$

$$\frac{1}{L\bar{N}^2}\sum_{i=1}^{L}\left[\sum_{j=1}^{\bar{N}}(y_{ij} - \bar{y})^2 + \sum_{k\neq m}^{\bar{N}}(y_{ik} - \bar{y})(y_{im} - \bar{y})\right] =$$

$$\frac{1}{\bar{N}}\sigma^2 + \frac{\bar{N}-1}{\bar{N}}\sigma^2\delta = \frac{\sigma^2}{\bar{N}}[1 + (\bar{N}-1)\delta],$$

lo que nos permite expresar de otros modos la varianza del estimador media muestral de las medias de los conglomerados unietápicos. En concreto, haciendo uso de la aproximación

$$(L-1)\bar{N} \approx N - 1,$$

es común ver la fórmula aproximada con diseño *mia*,

$$V(\bar{y}_c) \approx \frac{N-n\bar{N}}{N-1}\frac{\sigma^2}{n\bar{N}}[1 + (\bar{N}-1)\delta],$$

que permite comparar fácilmente su varianza con la de la media muestral de igual tamaño muestral si no hubiera conglomerados. En general se aprecia que la ganancia en precisión es mayor cuando

$\delta \approx -1/(N-1)$. Si $\delta \approx 0$, el estimador por conglomerados de igual tamaño unietápico tiene similar precisión que el estimador media muestral de igual tamaño muestral. Si $\delta > 0$, el muestreo por conglomerados de igual tamaño unietápico tiene menor precisión que el estimador media muestral de igual tamaño muestral. Si $\delta < 0$, tiene mayor precisión. Por lo tanto, lo ideal es agrupar unidades secundarias que sean diferentes o heterogéneas entre sí según la variable de interés dentro de cada conglomerado.

La estimación del total poblacional, de la proporción poblacional, y del porcentaje poblacional, siguen en líneas semejantes. La determinación del tamaño muestral para obtener un error absoluto máximo $e$ para un nivel de confianza $1-\alpha$ se resuelve de modo similar usando la desigualdad de Chebychev, pero ahora aparecen dos funciones paramétricas desconocidas: $\sigma^2$ y $\delta$.

En concreto, con diseño *mas*,

$$n = \frac{\sigma^2[1+(\overline{N}-1)\delta]}{\alpha e^2 \overline{N}},$$

mientras que con diseño *mia*,

$$n \approx \frac{L}{\frac{\alpha e^2 (L-1)\overline{N}}{\sigma^2[1+(\overline{N}-1)\delta]}+1}.$$

## 8.2 Muestreo sistemático

En el caso en que el tamaño poblacional $N$ sea divisible por el tamaño muestral $n$, sea $L = N/n$. En el muestreo sistemático, existirán $L$ muestras conjunto o no ordenadas distintas de tamaño efectivo $n$ que se seleccionan del siguiente modo:

a) Se selecciona una unidad entre las $L$ primeras de la población finita, y cada una de las primeras $L$ unidades con probabilidad $1/L$.
b) Las restantes $n-1$ unidades de la muestra son las que ocupan los lugares relativos idénticos en los $n-1$ restantes grupos de $L$ unidades de la población finita.

Habrá entonces $L$ muestras posibles,

$$s_i = \{i, L+i, 2L+i, \dots, N-L+i\}, i = 1, 2, \dots, L;$$

con una probabilidad de selección igual a $1/L = n/N$.

Algunas ventajas de este método de muestreo:

a) La muestra se extiende a toda la población.
b) Puede recoger el efecto de estratificación debido al orden en que se numeran las unidades de la población finita.
c) Es de aplicación y comprobación sencillas.

Algunos inconvenientes del muestreo sistemático:

a) En caso de periodicidad de la variable de interés, podría aumentar la varianza del estimador media muestral.
b) El problema teórico que se presenta en la estimación de las varianzas, pues no existen estimadores insesgados de la varianza de la media muestral con muestreo sistemático de arranque simple, salvo con el apoyo de otra muestra.

Si se selecciona la muestra $s_i$ con probabilidad $1/L$, tendremos como estimador la media muestral

$$\bar{y}_{s_i} = \frac{1}{n}\sum_{j=1}^{n} y_{i+L(j-1)} \quad (i = 1, 2, \dots, L),$$

que es insesgado para estimar la media poblacional $\bar{y}$, pues

$$E(\bar{y}_{s_i}) = \frac{1}{L}\sum_{i=1}^{L}\bar{y}_{s_i} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{n}\sum_{j=1}^{n} y_{i+L(j-1)} = \frac{1}{N}N\bar{y} = \bar{y}.$$

Como $L$ es el número de muestras posibles, su varianza será

$$V(\bar{y}_{s_i}) = \frac{1}{L}\sum_{i=1}^{L}(\bar{y}_{s_i} - \bar{y})^2 = \sigma_{bs}^2,$$

que es la variabilidad entre (del inglés "between") conglomerados o muestras, o bien,

$$NV(\bar{y}_{s_i}) = \sum_{i=1}^{L}\sum_{j=1}^{n}(\bar{y}_{s_i} - \bar{y})^2 = N\sigma^2 - \sum_{i=1}^{L}\sum_{j=1}^{n}[y_{i+L(j-1)} - \bar{y}_{s_i}]^2,$$

haciendo uso del análisis de la varianza. Por tanto,

$$V(\bar{y}_{s_i}) = \sigma^2 - \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{n}[y_{i+L(j-1)} - \bar{y}_{s_i}]^2 = \sigma^2 - \sigma_{ws}^2,$$

donde

$$\sigma_{ws}^2 = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{n}[y_{i+L(j-1)} - \bar{y}_{s_i}]^2 = \frac{1}{L}\sum_{i=1}^{L}\sigma_i^2$$

es la variación dentro (del inglés "within") de conglomerados o muestras. De la fórmula de la varianza, podemos comparar la precisión del muestreo sistemático con la de otros métodos de muestreo de igual tamaño muestral.

También podemos considerar que el muestreo sistemático es un muestreo por conglomerados de igual tamaño unietápico o sin submuestreo. En él se selecciona un solo conglomerado, y aplicando entonces los resultados del muestreo por conglomerados unietápico, tenemos que $L$ es el número de conglomerados, el

tamaño muestral coincide con el tamaño de un conglomerado $n = \bar{N}$, y el tamaño muestral de los conglomerados seleccionados es 1.

## 8.3 Muestreo por conglomerados de tamaño desigual

Cuando los conglomerados son de tamaño desigual, es decir los $N_i$ son distintos o no todos iguales, entonces podemos denotar

$$y_i = \sum_{j=1}^{N_i} y_{ij} = N_i \bar{y}_i$$

al total del conglomerado $i$-ésimo $(i = 1, 2, \dots, L)$ de tamaño $N_i$ y de media $\bar{y}_i$ . Dada una muestra de $n$ unidades primarias o conglomerados de los $L$ que componen la población, una estimación insesgada del total poblacional, $N\bar{y}$, de la variable de interés $y$ es

$$t = \frac{L}{n} \sum_{i=1}^{n} y_i = L\bar{y}_t,$$

siendo

$$N\bar{y} = \sum_{i=1}^{L} y_i = L\bar{y}_T.$$

Su varianza para diseño *mas* es

$$V(t) = L^2 V(\bar{y}_t) = L^2 \frac{\sigma_{y_i}^2}{n} = \frac{L}{n} \sum_{i=1}^{L} (y_i - \bar{y}_T)^2,$$

que es estimable sin sesgo por

$$\hat{V}(t) = L^2 \frac{s_{y_i}^2}{n} = \frac{L^2}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{y}_t)^2.$$

Su varianza para diseño *mia* es

$$V(t) = L^2 \frac{L-n}{L-1} \frac{\sigma_{y_i}^2}{n} = L^2 \frac{L-n}{L} \frac{S_{y_i}^2}{n} = L(L-n) \frac{S_{y_i}^2}{n},$$

que es estimable sin sesgo por

$$\hat{V}(t) = \frac{L(L-n)}{n} s_{y_i}^2 = \frac{L(L-n)}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{y}_t)^2.$$

Otros métodos de estimación en el caso de muestreo por conglomerados unietápico de tamaño desigual, usando diseños de probabilidades desiguales, son semejantes a los ya vistos para unidades de la población finita en una sola etapa.

Así, por ejemplo, si $p_i = N_i/N$ es la probabilidad de seleccionar el conglomerado $i$, el estimador Hansen-Hurwitz del total poblacional es

$$t = \sum_{i \in \boldsymbol{s}} \frac{y_i}{np_i},$$

siendo $\boldsymbol{s}$ la muestra ordenada obtenida por diseño de probabilidades proporcionales al tamaño del conglomerado con reposición. Como vimos, este estimador es insesgado para el total poblacional $N\bar{y}$, su varianza es

$$V(t) = \frac{1}{n} \sum_{i \in U} p_i \left( \frac{y_i}{p_i} - L\bar{y}_T \right)^2,$$

y un estimador insesgado de la varianza del estimador $t$ es

$$\hat{V}(t) = \frac{1}{n(n-1)} \sum_{i \in \boldsymbol{s}} \left( \frac{y_i}{p_i} - t \right)^2.$$

También podemos usar el estimador Sánchez-Crespo en similares condiciones. Entonces, el esquema muestral es sin reposición, el estimador es el análogo

$$t = \sum_{i \in \boldsymbol{s}} \frac{y_i}{n p_i},$$

su varianza es

$$V(t) = \frac{M-n}{M-1} \frac{1}{n} \sum_{i \in U} p_i \left( \frac{y_i}{p_i} - L\bar{y}_T \right)^2,$$

y un estimador insesgado de esta varianza es

$$\hat{V}(t) = \frac{M-n}{M} \frac{1}{n(n-1)} \sum_{k \in \boldsymbol{s}} \left( \frac{y_k}{p_k} - t \right)^2.$$

Otro estimador posible sin reposición es el proporcionado por el estimador Horvitz-Thompson del total poblacional $N\bar{y}$ siguiente

$$t = \sum_{i \in r(\boldsymbol{s})} \frac{y_i}{\pi_i},$$

que es insesgado para estimar el total poblacional, su varianza es

$$V(t) = \sum_{i \in U} \frac{y_i^2}{\pi_i} (1 - \pi_i) + \sum_{k \neq m \in U} \frac{y_k y_m}{\pi_k \pi_m} (\pi_{km} - \pi_k \pi_m),$$

y un estimador insesgado de esta varianza es

$$\hat{V}(t) = \sum_{i \in r(\boldsymbol{s})} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + \sum_{k \neq m \in r(\boldsymbol{s})} \frac{y_k y_m}{\pi_k \pi_m} \frac{\pi_{km} - \pi_k \pi_m}{\pi_{km}}.$$

## 8.4 Submuestreo con conglomerados de igual tamaño

Se produce la situación de "submuestreo" cuando en una primera etapa se seleccionan $n$ unidades primarias o conglomerados con diseño *mas*, y después en una segunda etapa se selecciona un número determinado de subunidades o unidades secundarias o finales con diseño *mas* de cada uno de los conglomerados seleccionados en la primera etapa. Vamos a ver en esta sección el caso en el que las unidades de primera etapa son de igual tamaño, y en segundo lugar, veremos en la sección siguiente, el caso que se presenta cuando las unidades de primera etapa son de tamaño desigual.

Denotando por $y_{ij}$ al valor de la variable de interés de la $j$-ésima subunidad en la $i$-ésima unidad primaria, la media muestral por subunidad en la $i$-ésima unidad primaria es

$$\bar{y}_{\boldsymbol{s}(i)} = \frac{1}{m} \sum_{j=1}^{m} y_{ij},$$

donde $m$ es el tamaño muestral de la submuestra en el conglomerado $i$. La media global de muestra por subunidades es

$$\bar{\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{\boldsymbol{s}(i)},$$

donde $n$ es el tamaño muestral en la primera etapa o número de conglomerados que se seleccionan en la muestra. La varianza entre medias de unidades primarias es

$$\sigma_1^2 = \frac{1}{L}\sum_{i=1}^{L}(\bar{y}_i - \bar{y})^2.$$

La media de varianzas de unidades secundarias dentro de unidades primarias es

$$\sigma_2^2 = \frac{1}{L\bar{N}}\sum_{i=1}^{L}\sum_{j=1}^{\bar{N}}\left(y_{ij} - \bar{y}_i\right)^2 = \frac{1}{L}\sum_{i=1}^{L}\sigma_{2i}^2.$$

Entonces, si $n$ es el número de unidades primarias seleccionadas, y $m$ es el número de unidades secundarias o subunidades por conglomerado seleccionado en la segunda etapa, y las muestras extraídas por diseño *mas*, el estimador $\bar{\bar{y}}$ es insesgado para estimar la media poblacional $\bar{y}$, y su varianza es

$$V(\bar{\bar{y}}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{mn}.$$

En efecto,

$$E(\bar{\bar{y}}) = E_1[E_2(\bar{\bar{y}})] = E_1\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right) = \frac{1}{L}\sum_{i=1}^{L}\bar{y}_i = \bar{y}.$$

Y la varianza es

$$V(\bar{\bar{y}}) = V_1[E_2(\bar{\bar{y}})] + E_1[V_2(\bar{\bar{y}})].$$

$$E_2(\bar{\bar{y}}) = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i,$$

$$V_1[E_2(\bar{\bar{y}})] = \frac{\sigma_1^2}{n}.$$

$$V_2(\bar{\bar{y}}) = \frac{1}{n^2}\sum_{i=1}^{n} V_2[\bar{y}_{s(i)}] = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\sigma_{2i}^2}{m} = \frac{1}{nm}\frac{1}{n}\sum_{i=1}^{n}\sigma_{2i}^2,$$

$$E_1[V_2(\bar{\bar{y}})] = \frac{1}{nm}E_1\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_{2i}^2\right) = \frac{1}{nm}\sigma_2^2.$$

Luego,

$$V(\bar{\bar{y}}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{nm}.$$

Un estimador insesgado de esta varianza es

$$\hat{V}(\bar{\bar{y}}) = \frac{s_1^2}{n} + \frac{\bar{N}-1}{nm\bar{N}}s_2^2,$$

donde

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\bar{y}_{s(i)} - \bar{\bar{y}}\right]^2$$

y

$$s_2^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[y_{ij} - \bar{y}_{s(i)}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}s_{2i}^2.$$

En efecto,

$$(n-1)s_1^2 = \sum_{i=1}^{n}\left[\bar{y}_{s(i)} - \bar{\bar{y}}\right]^2 = \sum_{i=1}^{n}\bar{y}_{s(i)}^2 - n\bar{\bar{y}}^2,$$

por tanto

$$(n-1)E_2(s_1^2) = \sum_{i=1}^{n} \bar{y}_i^2 + \sum_{i=1}^{n} \frac{1}{\bar{N}m}\sigma_{2i}^2 -$$

$$n\left(\frac{1}{n}\sum_{i=1}^{n} \bar{y}_i\right)^2 - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\bar{N}m}\sigma_{2i}^2,$$

pues

$$E_2\left[\bar{y}_{s(i)}^2\right] = \left\{E_2\left[\bar{y}_{s(i)}\right]\right\}^2 + V_2\left[\bar{y}_{s(i)}\right]$$

y

$$E_2(\bar{\bar{y}}^2) = [E_2(\bar{\bar{y}})]^2 + V_2(\bar{\bar{y}}).$$

Luego,

$$(n-1)E_2(s_1^2) = \sum_{i=1}^{n}\left(\bar{y}_i - \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i\right)^2 + \frac{n-1}{\bar{N}mn}\sum_{i=1}^{n} \sigma_{2i}^2$$

y promediando sobre la primera etapa de diseño *mia*,

$$E[(n-1)s_1^2] = E_1[(n-1)E_2(s_1^2)] = (n-1)\sigma_1^2 + \frac{n-1}{\bar{N}m}\sigma_2^2.$$

Por tanto, $s_1^2$ es un estimador de $\sigma_1^2$ con un sesgo $\sigma_2^2/(\bar{N}m)$. Como $s_2^2$ es un estimador insesgado de $\sigma_2^2$, podemos proponer como estimador insesgado de la varianza de $\bar{\bar{y}}$ a

$$\hat{V}(\bar{\bar{y}}) = \frac{s_1^2}{n} + cs_2^2,$$

donde $c$ es una constante que se obtiene al asegurar que tal estimador sea insesgado. En efecto, de que

$$E\left[\hat{V}(\bar{\bar{y}})\right] = V(\bar{\bar{y}}),$$

obtenemos que

$$c = \frac{\bar{N} - 1}{nm\bar{N}},$$

de donde concluimos que el estimador insesgado buscado es

$$\hat{V}(\bar{\bar{y}}) = \frac{s_1^2}{n} + \frac{\bar{N} - 1}{nm\bar{N}} s_2^2.$$

La distribución de la muestra en las dos etapas se puede obtener al admitir la función de coste del tipo

$$C = c_1 n + c_2 nm,$$

es decir, el coste total es la suma de los costes proporcionales al número $n$ de unidades primarias seleccionadas por el coste $c_1$ de seleccionar una unidad primaria, y al número $nm$ de unidades secundarias seleccionadas por el coste $c_2$ de observación por unidad secundaria. Como tenemos la varianza

$$V(\bar{\bar{y}}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{nm},$$

para minimizar esta varianza para el coste total fijo $C$, tenemos el lagrangiano

$$L^* = V(\bar{\bar{y}}) + \lambda(C - c_1 n - c_2 nm)$$

que se resuelve así,

$$\frac{\partial L^*}{\partial n} = -\frac{1}{n^2}\sigma_1^2 - \frac{1}{n^2}\frac{\sigma_2^2}{m} - \lambda(c_1 + c_2 m) = 0$$

$$\frac{\partial L^*}{\partial m} = -\frac{1}{nm^2}\sigma_2^2 - \lambda c_2 n = 0.$$

Luego,

$$\lambda = -\frac{1}{n^2(c_1 + c_2 m)}\sigma_1^2 - \frac{1}{n^2 m(c_1 + c_2 m)}\sigma_2^2 =$$

$$-\frac{1}{c_2 n^2 m^2}\sigma_2^2,$$

o bien, multiplicando por $n^2 m^2 c_2(c_1 + c_2 m)$, tenemos

$$m^2 c_2 \sigma_1^2 + m c_2 \sigma_2^2 = (c_1 + c_2 m)\sigma_2^2,$$

que reordenando y simplificando resulta la ecuación de segundo grado en $m$

$$m^2 c_2 \sigma_1^2 - c_1 \sigma_2^2 = 0,$$

que resolviendo queda el valor óptimo teórico

$$m = \sqrt{\frac{c_1 \sigma_2^2}{c_2 \sigma_1^2}},$$

pues la raíz negativa no la consideramos ya que $m > 0$. Hemos dicho que es un óptimo teórico porque depende de funciones paramétricas que son desconocidas sin un censo. Finalmente

$$n = \frac{C}{c_1 + c_2 m}.$$

También es posible estudiar el caso de submuestreo con unidades de primera etapa iguales en tamaño, cuando el diseño básico usado es el diseño *mia* en la primera y segunda etapas. En este caso el estimador media de las medias muestrales $\bar{\bar{y}}$ es también insesgado y un estimador sin sesgo de su varianza es

$$\hat{V}(\bar{\bar{y}}) = \frac{L - n}{Ln} s_1^2 + \frac{n}{L}\frac{\bar{N} - m}{\bar{N} m n} s_2^2.$$

El tamaño muestral óptimo teórico de $m$, sujeto a un coste prefijado $C = c_1 n + c_2 nm$, es ahora

$$m = \sqrt{\frac{c_1 S_2^2}{c_2 \left( S_1^2 - \frac{S_2^2}{\bar{N}} \right)}},$$

donde

$$S_1^2 = \frac{1}{L-1} \sum_{i=1}^{L} (\bar{y}_i - \bar{y})^2$$

y

$$S_2^2 = \frac{1}{L(\bar{N}-1)} \sum_{i=1}^{L} \sum_{j=1}^{\bar{N}} \left( y_{ij} - \bar{y}_i \right)^2 = \frac{1}{L} \sum_{i=1}^{L} S_{2i}^2.$$

## 8.5 Submuestreo y conglomerados de tamaño desigual

Veamos un tratamiento a esta situación. La unidad primaria $i$ se selecciona con probabilidades proporcionales a $p_i > 0$, con

$$\sum_{i=1}^{L} p_i = 1.$$

Además podemos suponer que las $n$ selecciones se hacen con reemplazamiento. La submuestra es de tamaño $m_i$ subunidades de la unidad primaria $i$, con diseño *mas*. Si la unidad primaria $i$ se selecciona más de una vez, se restituye la totalidad de la submuestra seleccionada independientemente de tamaño $m_i$ unidades secundarias con diseño *mas* con reemplazamiento. Un estimador insesgado del total poblacional $N\bar{y}$ es

$$t = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\bar{y}_{s(i)}}{p_i} = \frac{1}{n}\sum_{i=1}^{L}\frac{N_i\bar{y}_{s(i)}}{p_i}e_i.$$

En efecto,

$$E(t) = E_1\left\{\frac{1}{n}\sum_{i=1}^{L}\frac{N_iE_2\left[\bar{y}_{s(i)}\right]}{p_i}e_i\right\} = \frac{1}{n}\sum_{i=1}^{L}\frac{N_i\bar{y}_i}{p_i}E_1(e_i) = N\bar{y},$$

siendo $e_i$ el número de veces que la unidad primaria $i$ es seleccionada en la muestra, con $E_1(e_i) = np_i$ $(i = 1, 2, \dots, L)$. La varianza de $t$ se obtiene haciendo uso del Teorema de Madow,

$$V(t) = E_1V_2(t) + V_1E_2(t).$$

$$V_2(t) = \frac{1}{n^2}\sum_{i=1}^{L}\frac{N_i^2}{p_i^2}e_i^2V_2\left[\bar{y}_{s(i)}\right] = \frac{1}{n^2}\sum_{i=1}^{L}e_i^2\frac{N_i^2}{p_i^2}\frac{\sigma_i^2}{m_i},$$

siendo $\sigma_i^2$ la varianza dentro del conglomerado $i$-ésimo, donde ahora $i = 1, 2, \dots, L$.

$$E_1V_2(t) = \frac{1}{n^2}\sum_{i=1}^{L}E_1\left(e_i^2\right)\frac{N_i^2}{p_i^2}\frac{\sigma_i^2}{m_i} =$$

$$\sum_{i=1}^{L}\frac{1 - p_i + np_i}{np_i}N_i^2\frac{\sigma_i^2}{m_i}.$$

También,

$$E_2(t) = \frac{1}{n}\sum_{i=1}^{L}\frac{N_iE_2\left[\bar{y}_{s(i)}\right]}{p_i}e_i = \frac{1}{n}\sum_{i=1}^{L}\frac{N_i\bar{y}_i}{p_i}e_i,$$

y

$$V_1E_2(t) = \frac{1}{n^2}\sum_{i=1}^{L}\frac{N_i^2\bar{y}_i^2}{p_i^2}V_1(e_i) = \sum_{i=1}^{L}\frac{N_i^2\bar{y}_i^2}{np_i}(1-p_i).$$

Luego,

$$V(t) = \sum_{i=1}^{L}\frac{1-p_i+np_i}{np_i}N_i^2\frac{\sigma_i^2}{m_i} + \sum_{i=1}^{L}\frac{N_i^2\bar{y}_i^2}{np_i}(1-p_i).$$

Y un estimador insesgado de esta varianza es

$$\hat{V}(t) = \sum_{i=1}^{L}\frac{1-p_i+np_i}{n^2p_i^2}N_i^2\frac{s_i^2}{m_i}e_i +$$

$$\sum_{i=1}^{L}\frac{N_i^2}{n^2p_i^2}(1-p_i)\left[\bar{y}_{s(i)}^2 - \frac{s_i^2}{m_i}\right]e_i,$$

o bien, simplificando,

$$\hat{V}(t) = \sum_{i=1}^{L}\left[\frac{N_i^2s_i^2}{np_im_i} + \frac{N_i^2(1-p_i)\bar{y}_{s(i)}^2}{n^2p_i^2}\right]e_i,$$

donde $s_i^2$ es la cuasivarianza muestral dentro del conglomerado $i$.

## 8.6 Ejercicios resueltos

**Ejercicio 8.1.** Un establecimiento comercial dispone de 1500 facturas que recogen los ingresos durante un mes de trabajo; se desea estimar la media por factura mediante muestreo sistemático, tomando un arranque aleatorio entre las primeras 15 facturas. Por la experiencia pasada se sabe que el coeficiente de correlación intraconglomerados es aproximadamente 0.05. Se desea saber si la

varianza del estimador usual en muestreo sistemático es más del doble que la varianza de la estrategia "diseño *mas*, media muestral" con idéntico tamaño muestral.

**Solución.** El tamaño de la muestra sistemática es

$$n = \frac{N}{L} = \frac{1500}{15} = 100,$$

que coincide con el tamaño del conglomerado $\bar{N}$. La varianza de la media muestral con muestreo sistemático es

$$V(\bar{y}_c) = \frac{\sigma^2}{n}[1 + (n-1)\delta] = V(\bar{y}_s)[1 + (n-1)\delta] = 5.95V(\bar{y}_s)$$

que es mayor que el doble de $V(\bar{y}_s)$.

**Ejercicio 8.2.** Con el fin de estimar la calidad de cierta marca de cerillas, se examina la producción que está empaquetada en cajas de 50 fósforos. El número de cajas producido es de 300. Para estimar la proporción de cerillas defectuosas, se prueban 5 cajas de modo destructivo y la proporción estimada de unidades defectuosas por caja en las 5 muestreadas es de 0.04. ¿Cuál será la varianza de este estimador si la proporción muestral estima bien la proporción poblacional, y el coeficiente de correlación intraconglomerados es 0?

**Solución.**

$$V(\hat{P}) \approx \frac{N - n\bar{N}}{N-1}\frac{PQ}{n\bar{N}}[1 + (\bar{N}-1)\delta] \approx 0.00015$$

**Ejercicio 8.3.** En el problema anterior, ¿cuál será el tamaño muestral de cajas o unidades primarias para que el error máximo de muestreo sea igual a $0.001$ para un nivel de confianza del $90\%$?

**Solución.**

$$n \approx \frac{NPQ}{(N-1)\bar{N}\left[\frac{\alpha e^2}{1+(\bar{N}-1)\delta}+\frac{PQ}{N-1}\right]} \approx 612,$$

que como es superior a las $300$ disponibles, sería necesario un muestreo exhaustivo de la totalidad de la producción, pero al ser un proceso destructivo se hace desaconsejable el estudio.

**Ejercicio 8.4.** Averiguar el tamaño muestral $n$ necesario para asegurar que el estimador $\bar{y}_c$, de la media poblacional $\bar{y}$, con conglomerados del mismo tamaño $30$, fijado un error absoluto máximo de muestreo de $0.05$ para un nivel de confianza del $95\%$. El tamaño poblacional es de $30000$ unidades. Además la experiencia en estudios anteriores nos da una estimación de la varianza poblacional de $0.15$ y una estimación del coeficiente de correlación intraconglomerados de $0.1$.

**Solución.**

$$n \approx \frac{N\sigma^2}{(N-1)\bar{N}\left[\frac{\alpha e^2}{1+(\bar{N}-1)\delta}+\frac{\sigma^2}{N-1}\right]} \approx 136.$$

**Ejercicio 8.5.** En una empresa industrial se empaquetan los productos en lotes de $10$ unidades, produciéndose diariamente $2000$ lotes. Con el fin de estimar la calidad del producto se procede a la estimación de la media poblacional de cierta característica de

interés, en muestreo bietápico, seleccionando 20 lotes con diseño *mas*, y dentro de cada lote extraído se examinan 3 subunidades con diseño *mas*. Estimar sin sesgo la varianza del estimador usual

$$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{\boldsymbol{s}(i)},$$

sabiendo que de la muestra obtenemos que

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\bar{y}_{\boldsymbol{s}(i)} - \bar{\bar{y}}\right]^2 = 0.3$$

y

$$s_2^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[y_{ij} - \bar{y}_{\boldsymbol{s}(i)}\right]^2 = 3.$$

**Solución.**

$$\hat{V}(\bar{\bar{y}}) = \frac{s_1^2}{n} + \frac{\bar{N}-1}{nm\bar{N}}s_2^2 = \frac{0.3}{20} + \frac{9}{600}3 = 0.015 + 0.045 = 0.06$$

**Ejercicio 8.6.** En una primera fase se selecciona con diseño *mas* de tamaño muestral 5 una muestra ordenada. En una segunda fase, de la muestra anterior se selecciona una submuestra con diseño *mas* de tamaño muestral 2. Obtener la varianza de la media muestral de los datos observados en la segunda fase, en función de la varianza poblacional.

**Solución.** En la primera fase seleccionamos una muestra de tamaño $n = 5$, y denotamos por $\bar{y}'$, $\bar{S}^2$ y $\bar{\sigma}^2$ a la media muestral, cuasivarianza muestral y varianza muestral respectivamente en la primera fase sin submuestreo. En la segunda fase se submuestrea

la muestra de la primera fase con el nuevo tamaño muestral 2, cuya media muestral $\bar{y}$, es de la que se nos pide su varianza. Usando el Teorema de Madow en dos fases,

$$V_2(\bar{y}) = \frac{\bar{\sigma}^2}{m} = \frac{n-1}{mn}\bar{S}^2,$$

de donde

$$E_1V_2(\bar{y}) = \frac{n-1}{mn}E_1(\bar{S}^2) = \frac{n-1}{mn}\sigma^2 = \frac{2}{5}\sigma^2.$$

También,

$$E_2(\bar{y}) = \bar{y}',$$

de donde

$$V_1E_2(\bar{y}) = V_1(\bar{y}') = \frac{\sigma^2}{n} = \frac{\sigma^2}{5}.$$

Luego,

$$V(\bar{y}) = E_1V_2(\bar{y}) + V_1E_2(\bar{y}) = \frac{3}{5}\sigma^2,$$

siendo $\sigma^2$ la varianza poblacional.

**Ejercicio 8.7.** ¿Cuándo se puede llamar una partición en conglomerados con el nombre de estratificación?

**Solución.** Cuando se seleccionan todos los conglomerados para ser observados total o parcialmente por muestreo.

**Ejercicio 8.8.** Proponer un estimador insesgado de la varianza poblacional en muestreo por conglomerados sin submuestreo, es

decir cuando todas las unidades secundarias pertenecientes a los conglomerados seleccionados son observadas.

**Solución.** Si el procedimiento de selección de unidades primarias en la muestra de conglomerados $s_1$ tiene probabilidades de inclusión $\pi_h$ y $\pi_{hg}$ positivas para cualesquiera unidades primarias $h \neq g$, el estimador insesgado de la varianza poblacional $\sigma^2$ es

$$\widehat{\sigma^2} = \sum_{h \in s_1} W_h \frac{\sigma_h^2}{\pi_h} + \sum_{h \in s_1} \sum_{g \neq h \in s_1} W_h W_g \frac{(\mu_h - \mu_g)^2}{\pi_{gh}},$$

siendo $W_h$ el tamaño relativo del conglomerado $h$, $\mu_h$ la media del conglomerado $h$, y $\sigma_h^2$ la varianza del conglomerado $h$.

Si $L$ es el número de conglomerados, y $e_h$ es el indicador del conglomerado $h$ en la muestra, tenemos que

$$E\left(\widehat{\sigma^2}\right) =$$

$$\sum_{h=1}^{L} W_h \frac{\sigma_h^2}{\pi_h} E(e_h) + \sum_{h=1}^{L-1} \sum_{g=h+1}^{L} W_h W_g \frac{(\mu_h - \mu_g)^2}{\pi_{gh}} E(e_h e_g) =$$

$$\sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L-1} \sum_{g=h+1}^{L} W_h W_g (\mu_h - \mu_g)^2 = \sigma^2,$$

por lo que el estimador propuesto es insesgado para estimar la varianza poblacional. Hemos usado la notación usual en muestreo estratificado que es válida porque los conglomerados son una clasificación de la población finita, como lo son los estratos.

**Ejercicio 8.9.** Seleccionamos una muestra sistemática $s$ de tamaño $m$, de una población finita $U$ de tamaño $N$, y posteriormente

seleccionamos una muestra irrestricta aleatoria $r$ de tamaño $n < N - m$ de entre las unidades de $U - s$. En tal caso, el estimador para la media poblacional usado es para una constante $\beta$,

$$t = \beta\bar{y}_s + (1-\beta)\bar{y}_r.$$

Comprobar que es insesgado, y obtener un estimador insesgado de su varianza cuando es posible.

**Solución.** Para ver que es insesgado $t$, procedemos de esta manera:

$$E(t) = EE[\beta\bar{y}_s + (1-\beta)\bar{y}_r|s] =$$

$$\beta E(\bar{y}_s) + (1-\beta)E[E(\bar{y}_r|s)] =$$

$$\beta\bar{y} + (1-\beta)E(\bar{y}_{U-s}) =$$

$$\beta\bar{y} + (1-\beta)E\left(\frac{N\bar{y} - m\bar{y}_s}{N-m}\right) =$$

$$\beta\bar{y} + (1-\beta)\bar{y} = \bar{y}.$$

La varianza de $\bar{y}_r$ la obtenemos por la fórmula de Madow,

$$V(\bar{y}_r) = E[V(\bar{y}_r|s)] + V[E(\bar{y}_r|s)],$$

donde

$$V[E(\bar{y}_r|s)] = V(\bar{y}_{U-s}) = \frac{m^2}{(N-m)^2}V(\bar{y}_s).$$

Ahora tenemos que

$$V(t) = \beta^2 V(\bar{y}_s) + (1-\beta)^2 V(\bar{y}_r) + 2\beta(1-\beta)Cov(\bar{y}_s, \bar{y}_r).$$

Como,

$$Cov(\bar{y}_s, \bar{y}_r) = E\{E[(\bar{y}_s - \bar{y})(\bar{y}_r - \bar{y})|s]\} =$$

$$E[(\bar{y}_s - \bar{y})E(\bar{y}_r - \bar{y}|s)] =$$

$$E[(\bar{y}_s - \bar{y})(\bar{y}_{U-s} - \bar{y})] =$$

$$E\left\{(\bar{y}_s - \bar{y})\left[-\frac{m}{N-m}(\bar{y}_s - \bar{y})\right]\right\} =$$

$$-\frac{m}{N-m}V(\bar{y}_s).$$

Por tanto,

$$V(t) = \beta^2 V(\bar{y}_s) +$$

$$(1-\beta)^2\left\{E\left[V(\bar{y}_r|s) + \frac{m^2}{(N-m)^2}V(\bar{y}_s)\right]\right\} -$$

$$\frac{2\beta(1-\beta)m}{N-m}V(\bar{y}_s) =$$

$$\left[\beta - (1-\beta)\frac{m}{N-m}\right]^2 V(\bar{y}_s) + (1-\beta)^2 E[V(\bar{y}_r|s)].$$

Así, como la media muestral en el muestreo sistemático no tiene un estimador insesgado de su varianza $V(\bar{y}_s)$, para $\beta = m/N$, podemos escribir

$$V(t) = \left(1 - \frac{m}{N}\right)^2 E[V(\bar{y}_r|s)],$$

que solo en este caso admite el estimador insesgado

$$\hat{V}(t) = \left(1 - \frac{m}{N}\right)^2 \hat{V}(\bar{y}_r|s \text{ fijada}) =$$

$$\left(1 - \frac{m}{N}\right)^2 \frac{N-m-n}{(N-m)n}s_r^2 = \frac{(N-m)(N-m-n)}{nN^2}s_r^2,$$

donde $s_r^2$ es la cuasivarianza muestral de la variable $y$ en la muestra irrestricta aleatoria $r$ dentro de $U - s$. Este estimador recibe el nombre de Rana y Singh por ser ellos sus descubridores.

**Ejercicio 8.10.** En el muestreo sistemático de doble arranque, consideramos como estimador de la media poblacional $\bar{y}$ a la media aritmética de las medias muestrales de tamaño $n = N/L$ de las dos muestras sistemáticas $s_1$ y $s_2$ obtenidas por muestreo irrestricto aleatorio de tamaño 2 de entre las $L$ unidades primarias. Obtener un estimador insesgado de la varianza poblacional.

**Solución.** Obtenemos una muestra irrestricta aleatoria $s$ de tamaño 2 de entre las $L$ unidades primarias. Llamando $\bar{y}_{s_1}$ e $\bar{y}_{s_2}$ a las medias muestrales de las dos muestras sistemáticas de tamaño $n$ cada una, el estimador insesgado de la media poblacional propuesto es

$$\bar{y}_s = \frac{\bar{y}_{s_1} + \bar{y}_{s_2}}{2} = \frac{\mu_{i_1} + \mu_{i_2}}{2},$$

donde $\mu_{i_j}$ es la media del conglomerado $j$-ésimo obtenido por muestreo aleatorio simple sin reemplazamiento de tamaño 2 de entre los $L$ conglomerados o unidades primarias correspondientes a las muestras sistemáticas seleccionadas. Obviamente,

$$E(\bar{y}_s) = \frac{1}{L}\sum_{h=1}^{L} \bar{y}_h = \bar{y}.$$

Un estimador insesgado de la varianza poblacional $\sigma^2$, aprovechando el resultado del Ejercicio 8.8, es

$$\widehat{\sigma^2} = \frac{1}{L}\sum_{h \in s} \frac{\sigma_h^2}{\pi_h} + \frac{1}{L^2}\sum_{h \in s}\sum_{g>h \in s} \frac{\left(\bar{y}_h - \bar{y}_g\right)^2}{\pi_{hg}},$$

en donde, por ser obtenidas las unidades primarias por muestreo irrestricto aleatorio de 2 unidades de entre las $L$ posibles,

$$\pi_h = \frac{\binom{L-1}{1}}{\binom{L}{2}} = \frac{2}{L},$$

y si $h \neq g = 1, 2, \ldots, L$, tenemos

$$\pi_{hg} = \frac{\binom{L-2}{0}}{\binom{L}{2}} = \frac{2}{L(L-1)}.$$

**Ejercicio 8.11.** En las condiciones del ejercicio anterior, obtener un estimador insesgado de la varianza del estimador propuesto de la media poblacional, que dependa de las varianzas $\sigma_h^2$ de las muestras sistemáticas $h \in s$.

**Solución.** Obtenemos en primer lugar la varianza del estimador propuesto,

$$V(\bar{y}_s) = \frac{1}{4} V\left(\bar{y}_{s_1} + \bar{y}_{s_2}\right) =$$

$$\frac{1}{4}\left[V\left(\bar{y}_{s_1}\right) + V\left(\bar{y}_{s_2}\right) + 2Cov\left(\bar{y}_{s_1}, \bar{y}_{s_2}\right)\right],$$

donde

$$V\left(\bar{y}_{s_1}\right) = \sigma^2 - \frac{1}{N}\sum_{h=1}^{L}\sum_{i \in s_h}\left(y_i - \bar{y}_{s_h}\right)^2 =$$

$$\sigma^2 - \frac{1}{N}\left(N\alpha_2 - n\sum_{h=1}^{L}\bar{y}_{s_h}^2\right) = \sigma^2 - \alpha_2 + \frac{1}{L}\sum_{h=1}^{L}\bar{y}_h^2 =$$

$$\frac{1}{L}\sum_{h=1}^{L}\bar{y}_h^2-\left(\frac{1}{L}\sum_{h=1}^{L}\bar{y}_h\right)^2=\frac{1}{L}\sum_{h=1}^{L}(\bar{y}_h-\bar{y})^2=$$

$$\sigma^2-\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2.$$

$$V(\bar{y}_{s_2})=EV(\bar{y}_{s_2}|s_1)+VE(\bar{y}_{s_2}|s_1).$$

$$V(\bar{y}_{s_2}|s_1)=\frac{\sigma_{U-s_1}^2}{n}=\frac{\sum_{i\in U-s_1}(y_i-\bar{y}_{U-s_1})^2}{(N-n)n}=$$

$$\frac{\sum_{i\in U-s_1}y_i^2-\frac{1}{N-n}\left(\sum_{i\in U-s_1}y_i\right)^2}{(N-n)n},$$

de donde

$$EV(\bar{y}_{s_2}|s_1)=\frac{E\left[N\alpha_2-na_{2s_1}-\frac{(N\bar{y}-n\bar{y}_{s_1})^2}{N-n}\right]}{(N-n)n}=$$

$$\frac{(N-n)\alpha_2-\frac{1}{N-n}\left[N^2\bar{y}^2-2Nn\bar{y}^2+n^2E(\bar{y}_{s_1}^2)\right]}{(N-n)n}=$$

$$\frac{\alpha_2}{n}-\frac{N^2\bar{y}^2-2Nn\bar{y}^2+n^2\left(\sigma^2-\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2+\bar{y}^2\right)}{(N-n)^2n}=$$

$$\frac{\alpha_2}{n}-\frac{(N-n)^2\bar{y}^2+n^2\left(\sigma^2-\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2\right)}{(N-n)^2n}=$$

$$\frac{\sigma^2}{n}-\frac{n}{(N-n)^2}\left(\sigma^2-\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2\right).$$

También,

$$E(\bar{y}_{s_2}|s_1) = \bar{y}_{U-s_1} = \frac{N\bar{y} - n\bar{y}_{s_1}}{N-n}.$$

$$VE(\bar{y}_{s_2}|s_1) = \frac{n^2}{(N-n)^2} V(\bar{y}_{s_1}) =$$

$$\frac{n^2}{(N-n)^2}\left(\sigma^2 - \frac{1}{L}\sum_{h=1}^{L}\sigma_h^2\right).$$

Finalmente,

$$Cov(\bar{y}_{s_1}, \bar{y}_{s_2}) = ECov(\bar{y}_{s_1}, \bar{y}_{s_2}|s_1) + Cov\left[E(\bar{y}_{s_1}|s_1), E(\bar{y}_{s_2}|s_1)\right]$$

$$= 0 + Cov\left(\bar{y}_{s_1}, \frac{N\bar{y} - n\bar{y}_{s_1}}{N-n}\right) =$$

$$-\frac{n}{N-n} V(\bar{y}_{s_1}) =$$

$$-\frac{n}{N-n}\left(\sigma^2 - \frac{1}{L}\sum_{h=1}^{L}\sigma_h^2\right).$$

Por lo que hemos completado la varianza del estimador media aritmética de las medias muestrales de las dos secuencias sistemáticas correspondientes a los dos arranques aleatorios sin reemplazamiento. En concreto, simplificando,

$$V(\bar{y}_s) = \frac{1}{4}\left[\frac{N^2(n+1) + N(-4n^2 - 2n) + 4n^3}{(N-n)^2 n}\sigma^2 + \right.$$

$$\left.\frac{-N^2 + 4Nn - 4n^2 + n}{(N-n)^2}\left(\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2\right)\right].$$

Para elaborar un estimador insesgado de la varianza de $\bar{y}_s$ bastará sustituir, en la fórmula de su varianza obtenida, los parámetros por sus estimaciones insesgadas. Del ejercicio anterior

tenemos que la varianza poblacional $\sigma^2$ es estimable sin sesgo y obtuvimos la expresión exacta de este estimador $\widehat{\sigma^2}$. También, un estimador insesgado del parámetro

$$\frac{1}{L}\sum_{h=1}^{L}\sigma_h^2$$

es la media aritmética de las varianzas de cada muestra sistemática $s_1$ y $s_2$, es decir

$$\frac{\sigma_{s_1}^2+\sigma_{s_2}^2}{2}.$$

Finalmente, sustituyendo estas dos estimaciones insesgadas en lugar de los parámetros correspondientes, tenemos el estimador insesgado de la varianza $V(\bar{y}_s)$, y que denotamos $\hat{V}(\bar{y}_s)$.

Observar que en el muestreo sistemático de arranque simple, la varianza poblacional no admite estimador insesgado porque existen al menos un par de unidades distintas de la población $i$ y $j$ con probabilidad de inclusión $\pi_{ij}=0$, y como consecuencia no es posible construir un estimador insesgado de la varianza de la media muestral con dicho diseño muestral sistemático de arranque simple.

**Ejercicio 8.12.** En el muestreo por conglomerados de igual tamaño sin submuestreo, con selección de conglomerados por diseño de muestreo aleatorio simple con reemplazamiento, proponer un estimador insesgado de la varianza del estimador usual de la media poblacional.

**Solución.** El estimador usual de la media poblacional en muestreo por conglomerados de igual tamaño sin submuestreo es

$$\bar{y}_c = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

donde $n$ es el tamaño muestral de conglomerados de igual tamaño seleccionados, e $\bar{y}_i$ es la media del conglomerado $i$ -ésimo seleccionado en la muestra. La varianza del estimador $\bar{y}_c$ es

$$V(\bar{y}_c) = \frac{\sigma^2_{\bar{y}_i}}{n} = \frac{1}{nL}\sum_{i=1}^{L} (\bar{y}_i - \bar{y})^2$$

siendo $L$ el número de conglomerados en la población. Así, el estimador insesgado de la varianza $V(\bar{y}_c)$ es

$$\hat{V}(\bar{y}_c) = \frac{1}{n(n-1)}\sum_{i=1}^{n} (\bar{y}_i - \bar{y}_c)^2$$

debido a que la cuasivarianza muestral en el muestreo aleatorio simple con reemplazamiento es un estimador insesgado de la varianza poblacional.

# Capítulo 9

# Ética y filosofía del muestreo

En este capítulo presentamos elementos básicos de los métodos de inferencia, sus inicios, su utilidad y se aportan argumentos que orientan a la búsqueda de un método de inferencia estadística objetivo y a su fundamentación.

## 9.1 Introducción

Consideramos de interés aquellas variables observables objetivamente definidas y en las que no haya ambigüedades a la hora de ser interpretadas por el consultado y por quien anota el dato de las posibles unidades medidas, observadas o encuestadas al efectuar sus respuestas. Para el observador, estadístico o encuestador lo importante, además de amar, honrar y respetar al observado o encuestado por su dignidad como ser humano y persona, son las observaciones reales en las unidades de la población y de la variable de interés que han sido definidas con claridad para cada estudio concreto. De ellas, mediante el tratamiento estadístico descriptivo o inferencial, se podrán extraer conclusiones que justifiquen otras decisiones o acuerdos asociados y consecuentes que promuevan un mayor bienestar social, personal, y de trascendencia humana.

El método racional en el que basamos las hipótesis y las tesis o conclusiones es de tipo lógico basado en verdades reveladas

comunes y coincidentes en la cultura judeocristiana y aplicado a las metodologías estadísticas. Cualquier variable de tipo de salud, social, económica, etc. puede ser estudiada definiendo adecuadamente las unidades de la población que interesa estudiar y definiendo la variable de interés con nitidez y el instante o periodo temporal de referencia.

La hipótesis de trabajo es sencilla. Básicamente puede resumirse en que el hecho demostrado matemáticamente de que 'A implica B', no necesariamente implica lógicamente que '(No A) implica B'. En general, nos referimos a 'A' como el conjunto de premisas de las que matemáticamente deducimos que implican 'B' que es un conjunto de conclusiones o tesis. Sin embargo, aunque la primera implicación matemática se demuestra muy concienzudamente en la ciencia estadística, alguna o algunas de las premisas que constituyen 'A' no se comprueban de modo alguno al aplicarse a estudios prácticos concretos. Luego, para estos estudios prácticos, algunas de las condiciones o premisas podrían no estar en 'A', o lo que es lo mismo podrían estar en '(No A)'.

Entonces, haber demostrado que 'A implica B', si en el caso práctico las premisas 'A' no han sido comprobadas o aseguradas, la condición de hecho podría ser en realidad '(No A)', con lo que la demostración matemática sería inútil a efectos de justificar 'B' o '(No B)', especialmente si no se ha demostrado que '(No A) implica B' o bien '(No A) implica (No B)'.

En muchos libros se supone o acepta implícitamente que una vez demostrado el resultado matemático 'A implica B', todos los ejemplos de aplicación siguientes se ajustan a las hipótesis o premisas 'A'. Suponer que 'A' es cierto en el ejemplo no es suficiente en la práctica de la inferencia, sino que es necesario

saberlo y poder comprobarlo, no solo suponerlo, lo cual es más difícil de lo que parece a primera vista.

En las ciencias naturales, de la salud, sociales, económicas, etc. la realidad es que no suele ser posible afirmar si se da ‘A’ o ‘(No A)’, por lo que no podemos concluir que se dará ‘B’ al poder ocurrir ‘(No A)’, a pesar de demostrar que ‘A implica B’. Una de las formas de concluir ‘B’ como cierta siempre sería demostrar que tanto partiendo de ‘A’ como de ‘(No A)’ llegamos a la misma conclusión ‘B’, pero esto no suele justificarse en la mayor parte de inferencias estadísticas más tradicionales o clásicas.

Las deducciones matemáticas tienen gran fuerza y justifican el modo de abordar los problemas de inferencia estadística en la práctica. Pero de nada sirve demostrar matemáticamente si a la hora de aplicarlo no sabemos si las premisas del razonamiento son o no son verificadas en los ejemplos prácticos de los que se desea información veraz estadística. Para asegurar su objetividad, el método estadístico debe estar demostrado matemáticamente, y además ser objetivo y seguro en su utilización práctica al verificar la realidad material de trabajo las hipótesis del método estadístico.

Los métodos estadísticos que no permiten conocer si sus hipótesis son realistas en la práctica podrían interesar a ciertas ciencias matemáticas abstractas, o como ilustración orientativa de sus utilidades potenciales, pero siendo honradamente realistas, si existen otros métodos estadísticos que se ajusten a las condiciones naturales y reales de presentación de los datos a observar, no cabe duda que estos métodos son prioritarios ante aquellos que solo tienen rigor matemático pero no rigor objetivo en su aplicación en realidades constatables.

Así pues no nos interesan métodos potencialmente utilizables, sino aquellos que con seguridad son correctamente aplicados.

Algunas de las hipótesis potencialmente útiles pero sin valor social práctico es suponer que la población estadística es infinita o que su distribución de probabilidad es una determinada, sin posibilidad de comprobación posible en la práctica o con clara contradicción con los hechos conocidos como que la población es finita de hecho y de que la distribución objetiva de partida es uniforme discreta.

Como métodos estadísticos más objetivos están los "métodos de estadística descriptiva" (que cuando las poblaciones son grandes resulta costosa, lenta y con facilidad de introducir errores en los datos por la gran cantidad de ellos a manejar) y el "muestreo de poblaciones finitas fijadas" que permite controlar mejor algunas dificultades prácticas o costes inasumibles presentados en las estadísticas descriptivas y en los censos.

En la Unión Europea cada estado miembro tiene sus propias leyes y propias medidas desarrolladas independientes y acordadas a nivel político. En los últimos años se van dando pasos en el sentido de compartir las informaciones estadísticas de carácter social a nivel oficial, público y privado. En Estados Unidos se han desarrollado menos las políticas de protección social y carecen de la experiencia de las europeas que existen desde finales del siglo XIX.

Los procedimientos de recogida y tratamiento de la información estadística haciendo uso de tecnologías de la información y de telecomunicaciones hacen posible hoy el conocimiento de la realidad, coyuntura, y en ciertos casos incluso de supuesta previsión del bienestar social. Pero todo esto contrasta con la situación de desconocimiento, carencia de registros de nacidos, o del dudoso comportamiento de los funcionarios en algunos lugares del planeta. Es necesario un registro de los

ciudadanos identificados y localizables para realizar con unas mínimas garantías los estudios estadísticos inferenciales sin caer en tener que suponer que la realidad es como alguien supone sin base segura y cierta.

La ética aplicada crece progresivamente como consecuencia de los avances tecnológicos y científicos, y de la toma de decisiones y consensos políticos, sociales y económicos. La moralidad de estos avances está en la mesa de análisis y de discusión. En cualquier caso se trata de alcanzar un "bien común" entendido como "conjunto de aquellas condiciones de la vida social que permiten a los grupos y a cada uno de sus miembros conseguir más plena y fácilmente su propia perfección" (*Gaudium et Spes*, 26, 1). Todo ello comporta el respeto a la persona, el bienestar social y el desarrollo del grupo, así como la paz. Para ello la educación de la familia y la responsabilidad en el trabajo constituyen el medio por el que el hombre participa en el bien de los demás y de la sociedad (*Centesimus Annus*, 43).

Las ciencias sociales buscan a menudo el apoyo de los datos tomados de la realidad. Sus fuentes principales son las estadísticas y las encuestas, que al ser interpretadas y utilizar un criterio subjetivo propio, muchas veces de un mismo dato se llega a concluir una cosa o la contraria.

Las estadísticas en Estados Unidos según el Census Bureau (Schmidtz y Goodin, 2000) como en España (Ruiz Espejo, 1998b) apuntan a mejoras en la evolución de las rentas familiares en los años de final del siglo XX. En España se obtuvieron los datos por métodos de inferencia objetiva de poblaciones finitas.

Como científicos debemos aportar los mejores medios para describir la realidad, en aras de proponer políticas tanto públicas como privadas que ayuden al desarrollo del bienestar social en base a la responsabilidad personal. Uno de los procedimientos para

describir la realidad, aunque no sea el único, es el de los métodos estadísticos.

Cuando queremos conocer la situación de hecho de una población humana en un determinado instante de tiempo, existen muchas posibles teorías estadísticas que podrían ser aplicadas al caso desde un punto de vista teórico. Sin embargo, un somero análisis de la realidad a investigar hace que muchas de las teorías que se explican como posibles candidatas a explicar la realidad no asumen realidades evidentes que aparecen y que deberían de tenerse en cuenta para poder aplicarse a un caso concreto. Algunas teorías asumen hechos contradictorios con las realidades que pretenden investigar.

Muchas teorías estadísticas parten del hecho de que la población sobre la que se trata de inferir puede ser representada por una función de densidad que es conocida salvo algún o algunos parámetros a estimar. La función de densidad tiene propiedades matemáticas que permiten desarrollar características inferenciales propias válidas para el modelo teórico postulado o supuesto, pero no necesariamente válidos para otros casos o hipótesis. Es el caso de la inferencia llamada paramétrica clásica.

Los modelos de inferencia paramétrica clásica, inferencia bayesiana, inferencia no paramétrica, inferencia de distribución libre, modelos de población finita fijada, modelos superpoblacionales, etc. y posibles combinaciones de ellos son algunas de las hipótesis de trabajo estudiadas en contextos matemáticos, pero sin comprobar la teoría formal en la realidad a la que se aplique, es decir que las hipótesis sean consistentes con los hechos.

Un ejemplo muy presentado es el adoptado al tratar de inferir sobre una población estadística que de hecho es finita, tratándola

como si fuera infinita en el análisis inferencial. Lo evidente, en este caso, es decir que para concluir algo sobre la población real es un mal comienzo basarse en una falsedad aunque sea aproximativa. Esto es más grave si la falsedad se refiere a personas o grupos sociales.

Otro ejemplo de muchos modelos inferenciales es suponer que la distribución poblacional es conocida de antemano, hipótesis que podría ser válida sobre el papel o para ser simulada por el ordenador para obtener muestras de tal modelo teórico. Pero lo que importa es saber si tal modelo de distribución de la supuesta población se corresponde y existe en la realidad que se trata de analizar e inferir. De hecho, no he podido recabar ningún testimonio seguro del mundo natural (no simulado artificialmente) que pueda afirmar que sin lugar a dudas tal modelo ha sido comprobado en la realidad con plena seguridad. Para que esto último tenga sentido ha de aceptarse el hecho de la existencia de la probabilidad en la física, y en particular al caso al que trata de aplicarse la teoría estadística. La respuesta a esta cuestión es muy polémica y a mi modo de ver no resuelta con claridad y objetividad hasta ahora (Sprott, 2000).

A nivel científico, por un lado teórico me permite afirmar que las teorías estadísticas inferenciales son todas matemáticamente aceptables por los razonamientos lógicos que los sustentan, pero por el lado práctico dudo de la utilidad de muchas teorías al ser aplicadas en la práctica porque se aplican sin la seguridad de haber comprobado con certeza la adecuación de sus hipótesis a la realidad que tratan de inferir o estudiar con la mayor veracidad posible.

Cuando el objeto final del estudio es el propio hombre como su salud al estudiar la eficacia de tratamientos médicos, farmacéuticos, salud pública, etc. es importante decir la verdad y no basarse en falsos conceptos. Actualmente la población mundial

se aproxima a los 7000 millones de personas, pero nunca podríamos decir que hay infinitas personas, ni nunca las habrá en el mundo que conocemos. Este hecho evidente, tenido en cuenta en el análisis estadístico formal, hace que muchas de las teorías desarrolladas por los teóricos (en concreto, los métodos inferenciales para poblaciones infinitas) dejen de tener interés para los propósitos que nos planteamos, y nos hacen dirigirnos a modelos de poblaciones finitas aunque el número de sus unidades sean muy numerosas.

Los modelos de distribución de poblaciones infinitas han servido de estudio teórico facilitando el estudio de otros modelos de mayor aplicabilidad, pero sus conclusiones no dejan de ser un cúmulo de trabajos sobre el papel en libros o revistas alejados del interés de aclarar el estado en que se encuentra una población humana o social. Que la lógica utilizada en resolver cuestiones teóricas sea de alto nivel, sirve de poco si no da luz sobre el problema concreto a resolver. La ciencia tiene sentido cuando lo descubierto sirve para algo. Así los razonamientos matemáticos que se basan en el análisis infinitesimal estudiado por físicos y matemáticos clásicos no aportan siempre mayor claridad para inferir sobre poblaciones de la realidad natural. Un ejemplo de este tipo de libros es el de Cramér (1953).

Sorprende que este tipo de teorías sean explicadas en otros estudios universitarios, como medicina, ciencias empresariales, etc. con contenidos más específicos y objetivos reales concretos diferentes a los que pueden guiar un científico abstracto a quien le vale que tenga alguna lógica aunque sin utilidad práctica clara.

No se trata de aplicar una teoría sin más a unos datos cualesquiera. Sino de decidir con qué teoría podemos aclarar la situación real que se estudia para que sea de la mayor aplicabilidad,

desde un punto de vista honesto y clarificador de las realidades de las que deseamos recabar información para inferir estadísticamente su situación en determinado tiempo y lugar.

Existe una vía infalible para la ciencia consistente en la posibilidad de demostrar la falsedad de teorías mediante la contraevidencia, que permite al investigador honesto científicamente abandonar la teoría falsa. Las matemáticas pueden demostrar teorías inalterables con certeza. Los filósofos de la ciencia al comprender las dificultades reales para sostener científicamente verdades inalterables, introdujeron el concepto de teoría probable, que puede ser contrastada al conocer nuevos datos de la observación empírica. Sus teorías se basaron entonces en el concepto de la probabilidad aún sin saber probar su validez científica en la práctica como veremos más adelante.

Los métodos inferenciales tienen la ventaja de que para conocer un parámetro poblacional con determinada precisión no es necesario conocer la variable que aporta cada individuo o unidad poblacional. Basta observar la variable en una muestra de esas unidades seleccionadas aleatoriamente de la población finita, en muchos casos de una proporción de tamaño inferior al uno por mil. Para hacer esta selección de unidades de la muestra se requiere tener un censo actualizado en el momento de referencia de la encuesta o de las observaciones. Además se precisa que toda persona seleccionada en la muestra pueda ser identificada, localizada y observada. Otra condición es que la información recabada de cada unidad sea verdadera pues de otro modo debería ser posible inspeccionar los datos ofrecidos por cada encuestado o informante aunque sea solo a una submuestra aleatoria de los encuestados, y comprobar en ellos los datos sin error alguno (Ruiz Espejo, 1988a).

Los métodos de muestreo de poblaciones finitas fijadas aunque sean de gran tamaño permiten ahorrar tiempo material y presupuesto económico o trabajo humano para conocer con cierta precisión el parámetro requerido, que usualmente es la media poblacional de la variable de interés, aunque pueden estudiarse varias variables de interés con la misma muestra de unidades. Una pequeña parte de la población, llamada muestra, puede informar con precisión sobre la totalidad de la población.

En realidad los intentos por realizar censos son ciertos en los imperios egipcio y chino miles de años antes de Jesucristo, así como los imperios griego y romano. También hay referencias de censos del pueblo judío en tiempos de Moisés, David y Salomón.

Los conocimientos seguros anhelados por los científicos del siglo XV y del XVI pasaron a ser en muchos casos conocimientos probables o inciertos con el uso de métodos estadísticos o bien de un cierto nivel de confianza con el uso de la inferencia. Pero en nuestros días ya tenemos elementos para diferenciar métodos inferenciales que incorporan herramientas no objetivas que podrían hacer invalidar o tomar con cautela las conclusiones del estudio concreto científico que se basó en ellos.

El concepto de probabilidad surgió en el siglo XVII con las teorías del análisis combinatorio y sus aplicaciones a los juegos de azar, que con el tiempo trajo el desarrollo de teorías estadísticas inferenciales.

Los Padres Fundadores Americanos en 1790 fijaron la realización de los censos de población en cada Estado cada diez años. Las poblaciones censadas sirvieron para fijar la contribución financiera de cada estado a la Unión, así como para asignar el número de delegados que cada estado podía enviar a la Cámara de

Representantes en Washington, en inglés llamada “Hause of Representatives” (Anderson, 1988).

El origen metodológico de la sociología puede ser señalado en las estadísticas sociales de Adolfo Quételet en 1835, precursor de las descripciones y mediciones de fenómenos sociales con pretensiones de rigor científico y con técnicas de inferencia estadística.

En las décadas de 1870 a 1900 se crearon los “Labour Departments” (Ministerios de Trabajo) tras las recesiones económicas en los principales países industrializados, haciendo de la protección de los trabajadores su principal propósito, incluyendo la prioridad del apoyo estadístico. Así las encuestas de presupuestos familiares se desarrollaron entre los años 1850 y 1940, que concentraban sus intereses casi exclusivamente en familias de trabajadores.

Historiadores de la estadística han situado el comienzo del uso oficial de los estudios y encuestas por muestreo a fines del siglo XIX, en concreto el Gobierno noruego consideró en 1894 llevar a la práctica algunas políticas sociales nuevas como normativas de pensiones y seguros de enfermedad. Al requerir información más completa que la que se recogía en los censos, la Oficina de Estadística Noruega empezó a realizar encuestas por muestreo representativo, no de tipo probabilístico, a gran escala para informar a las políticas gubernamentales (Seng, 1951). A. N. Kiaer, director de la Oficina Estadística Noruega, desarrolló en la práctica estas encuestas en 1895.

Hacia 1920 Sir Ronald Fisher empezó a esbozar la teoría estadística de la contrastación de hipótesis, y unos diez años después Jerzy Neyman y Karl Pearson la dotaron de los instrumentos técnicos formales necesarios para su utilización generalizada que actualmente han sido cuestionadas. Tales teorías

reposan sobre el concepto de probabilidad cuya existencia real en el mundo no se ha podido demostrar. Sin embargo es un instrumento conceptual que permite avanzar en el llamado "conocimiento probable" basado en datos y experiencias, pero no es un "conocimiento seguro" ni absoluto, y por tanto no puede considerarse conocimiento realmente.

En la literatura sociológica, la contrastación de hipótesis aparece a mediados del siglo XX, un ejemplo es el libro de Goode y Hatt (1952). Según estos autores "las hipótesis han de ser empíricamente demostradas como probables o no probables", y en ello consiste la prueba, en contrastarlas con los hechos para concluir su aceptación o rechazo. La lógica de la prueba la atribuyen a los métodos de John Stuart Mill. Pero esta demostración o prueba no son seguras pues conllevan dos tipos de errores posibles: aceptar una hipótesis falsa, y rechazar una hipótesis cierta. Para ellos, la ciencia no consigue absolutos, sino que reduce la cantidad de incertidumbre. Esta manera de ver las cosas revela la desconfianza a la verdad cuando ésta existe.

En los años de 1970 la Organización para la Cooperación y Desarrollo Económico (OCDE) realizó algunos progresos de concreción en la elaboración de indicadores sociales en los países miembros de la organización, que tuvo su influencia en España. El Instituto Nacional de Estadística español ha publicado indicadores sociales en los años 1991, 1997, 1999, 2001, etc. Los sistemas integrados de estadísticas sociales como concepto europeo de la década de 1990, han sido presentados por estadísticos holandeses y escandinavos desde sus respectivos países (Van Tuinen, Altena e Imbens, 1994).

A nivel de Naciones Unidas se han publicado títulos en 1974, 1986, 1991 y 1996 sobre estadísticas sociales y otras disponibles pueden consultarse en www.un.org.

Los actuales países desarrollados tienen censos de población, vivienda, agricultura, industria, etc. como algo cotidiano para su buen funcionamiento. El Instituto Nacional de Estadística (INE) español va adaptando sus normas de trabajo y metodología estadística a las directrices de la Unión Europea, a través de su Oficina Estadística de las Comunidades Europeas (Eurostat). La Oficina del Censo de los Estados Unidos (U. S. Census Bureau) aborda a nivel de contribuciones personales por expertos estadísticos y de reuniones de profesionales los problemas surgidos por la creciente globalización y sus implicaciones estadísticas.

Además de un recuento de la población que recoge la dirección e identificación de los ciudadanos, el censo tiene un interés añadido como marco básico para poder extraer de la población muestras de personas de modo aleatorio y probabilístico. En España los censos de población se realizan con una periodicidad de diez años. Además existen bases de datos públicas e informatizadas en muchas áreas de interés social referidos a periodos más breves o incluso continuos.

Los métodos de muestreo han sido utilizados científicamente y socialmente a lo largo del siglo XX, desarrollándose los fundamentos que relacionan las muestras con la población. Así, el Instituto Internacional de Estadística reconoció estos métodos como instrumentos válidos de investigación y desde entonces son de amplio desarrollo y uso en el mundo estadístico científico y oficial. Un libro que recoge las principales aportaciones científicas del muestreo de poblaciones finitas es el de Ruiz Espejo (2013c).

Tiene sentido desarrollar métodos estadísticos y tecnologías informáticas objetivos que faciliten el seguimiento de variables de

interés social, de salud, etc. que informen para poder desarrollar políticas efectivas para paliar necesidades humanas con un fundamento sólido y no a ciegas. Los métodos estadísticos de investigación social científicamente correctos y de garantía, útiles y eficaces, coherentes en sus hipótesis con los objetivos reales de estudio, han sido objeto de estudio desde finales del siglo XIX. Tras varias décadas de investigación reflejada en revistas científicas de estadística, a fines de los años cuarenta del siglo XX aparecieron los primeros libros recopilatorios de métodos y teorías de muestreo de poblaciones finitas, tanto en Reino Unido como Estados Unidos, y posteriormente en otros países como Francia, España, India, Holanda, Italia, etc.

La globalización de los mercados y el desarrollo de la sociedad de la información son dos factores que afectan de modo creciente a los registros de empresas estadísticas (Nielsen y Plovsing, 1997). De hecho se plantea la necesidad de crear un registro satélite internacional para propósitos transnacionales. La integración eficiente de los estudios aportados por diferentes empresas o fuentes estadísticas independientes para contribuir a resultados conjuntos de mayor interés y calidad de los usuarios, ha sido tratado estadísticamente por Ruiz Espejo, Singh y Singh (2001).

Existen investigaciones que suponen que cada investigador puede aportar al análisis inferencial su propia idea subjetiva o “a priori” acerca de lo que hasta ahora nadie ha visto u observado completamente sobre cómo se comporta en realidad. Estas ideas subjetivas se plasman en la formulación de un modelo teórico de las posibles poblaciones estadísticas, de distribuciones poblacionales, o de alguno de sus parámetros. Casos particulares de esta situación se da en la práctica al emplear métodos de inferencia paramétrica clásica o de inferencia bayesiana. En ellas

las conclusiones siempre están influidas por la idea subjetiva o por las elecciones personales del investigador al aportar su modelo y opinión de cómo se comporta la realidad exterior a él y no conocida perfectamente por él.

Pensamos que aceptar subjetividades para describir o inferir sobre hechos objetivos, no es una vía aceptable pues lo subjetivo influye en el resultado, cuando el hecho objetivo no se altera sensiblemente por lo que piense de él un investigador. De aceptar distintas aportaciones subjetivas, obtendríamos con los mismos datos observados distintas tesis a veces incompatibles. La objetividad de los hechos no se altera por la idea personal que la redefina subjetivamente. No es posible aceptar que una sola realidad sea muchas cosas posibles incompatibles entre ellas dependiendo de la opinión del observador de la misma realidad única.

La utilización de cualquier método estadístico basado en hipótesis subjetivas solo podría considerarse como provisional, no como método objetivo explicativo de una única realidad.

Para terminar este capítulo anotamos argumentos de fe que sostienen el enfoque que hemos hecho. Las referencias pueden consultarse del texto de la *Biblia de Jerusalén* (1999).

*Éxodo* 20,16; *Deuteronomio* 5,20: “No darás testimonio falso contra tu prójimo.” (Revelaciones de Dios en el monte Sinaí y en el monte Horeb).

*Levítico* 19,11: “No hurtaréis; no mentiréis; no os engañaréis unos a otros.” (Prescripciones morales de Dios a Moisés, para la comunidad de los israelitas).

*Levítico* 19,35-36: “No cometáis injusticia ni en los juicios, ni en las medidas de longitud, de peso o de capacidad: tened balanza exacta, peso exacto, medida exacta y fanega exacta. Yo soy Yahvé

vuestro Dios, que os saqué del país de Egipto." (Revelación de Dios mientras los israelitas atravesaban el desierto).

*Judit* 9,11: "No está en el número tu fuerza, ni tu poder en los valientes, sino que eres el Dios de los humildes…" (Plegaria de Judit a Yahvé).

*Sabiduría* 11,20: "Pero tú regulaste todo con medida, número y peso." (Oración dirigida al Señor).

*Mateo* 5,17: "No penséis que he venido a abolir la Ley y los Profetas. No he venido a abolir, sino a dar cumplimiento." (Palabras de Jesús a sus discípulos, que confirman la Revelación de Dios en los montes Sinaí y Horeb, y en la travesía del desierto).

*Mateo* 15,19-20: "Porque del corazón salen… falsos testimonios… Eso es lo que contamina al hombre; …" (Doctrina de Jesús sobre lo puro y lo impuro).

*Mateo* 19,18; *Marcos* 10,19; *Lucas* 18,20: "… no levantarás falso testimonio, …" (Palabras de Jesús al joven rico).

*Juan* 17,19: "Y por ellos me santifico a mí mismo, para que ellos también sean santificados en la verdad." (Palabras de oración de Jesús dirigida a Dios Padre de petición por sus discípulos fieles).

No son las únicas revelaciones, pero las considero claves para hacer una inferencia estadística objetiva.

Es obvio que decir de una población humana que es como no es en realidad, es levantar un falso testimonio contra los individuos de la población finita.

También considerar que una persona puede responder a la misma pregunta con dos cantidades diferentes es mentir por parte del que responde o del que toma la observación y anota dos

respuestas numéricas del mismo dato, pues falta a la exactitud de la medida.

Por otro lado, decir que se selecciona una muestra que has encontrado de hecho como si fuera tomada con unas condiciones de aleatorización concretas sin posible comprobación de quien lo dice, es afirmar como cierto algo que es incierto, o lo que es lo mismo mentir o engañar o exponerse a ambas cosas sobre el procedimiento de selección.

## 9.2 Bases bibliográficas

En esta sección estudiamos métodos estadísticos y de muestreo de poblaciones finitas, así como las políticas editoriales en las publicaciones de docencia y de investigación estadística. Vemos los métodos objetivos de observación y de estimación estadística como es el muestreo de poblaciones finitas y recogemos las referencias bibliográficas de carácter internacional más destacadas.

Para poder abordar estudios estadísticos con garantías científicas y realistas, es necesario disponer de un marco constante o periódicamente actualizado y renovado de las personas, familias, empresas, industrias, etc. objeto de estudio sobre las que se pretende tener información rápida y a bajo coste, basándonos en la identificación, accesibilidad y recogida de la información cierta de aquellas unidades que hayan sido seleccionadas en la muestra en el caso de ser un estudio inferencial apropiado y ético.

En un estudio inferencial caben dos tipos de errores: observacionales y científicos. Los errores observacionales son aquellos que resultan de al menos una observación de un dato erróneo que se incluye como cierto o verdadero en el estudio estadístico. Este tipo de errores pueden ser reconducidos científicamente, por ejemplo inspeccionando una submuestra de la

muestra que contenía posibles datos erróneos y conociendo su verdadera magnitud (Ruiz Espejo, 1988a).

Otro tipo de error en la estadística aplicada consiste en aceptar planteamientos o hipótesis de estudio que no concuerdan con las realidades a las que se aplican, ya sea por suponer condiciones o premisas lógicas que no existen en la realidad, no se dan, o bien por la imposibilidad de saber o comprobar en la práctica que la suposición hecha es cierta o no entre la infinidad de este tipo de suposiciones que es posible hacer. Pues basarse en un error casi seguro no es buen fundamento científico para basar una investigación objetiva. A partir de una mentira casi segura, poca verdad puede deducirse salvo que hagamos mucho más efectivas las observaciones verdaderas que el modelo supuesto.

Los errores probables de un estudio inferencial es algo posible y real en los métodos estadísticos objetivos, pero también estos errores son controlables en gran medida y estimables sin sesgo basados en un concepto de probabilidad como instrumento de selección de la muestra y aprovechando eficientemente esta información muestral.

Otra cosa sería asumir errores evitables o basar el estudio en hipótesis inciertas o asumidas sin certeza posible en su objetividad por ser asumidas sin comprobación posible en la práctica y por tanto de dudosa adecuación.

La inmensa mayoría de métodos estadísticos que ocupan los contenidos de las revistas de investigación en el ámbito anglosajón pueden tener aspectos novedosos matemáticos pero no reúnen los requisitos evidentes presentados en la práctica al abordar estudios sociales, administrativos o de las estadísticas oficiales para reflejar los hechos reales que acaecen en las sociedades y que pretenden conocer inferencialmente con la mayor calidad y objetividad. Este

es uno de los criterios necesarios a tener en cuenta para la mejora de la calidad de las estadísticas sociales (Desrosières, 2000).

Actualmente la materia de muestreo de poblaciones finitas ha superado etapas de consolidación de técnicas especializadas desde los años 30 y 40 del siglo XX. Los avances teóricos y prácticos que permiten fundamentar sus bases científicas y desarrollar su matemática formal puede verse en distintos libros como los de Cassel, Särndal y Wretman (1977), Cochran (1977), Fuller (2009), Hedayat y Sinha (1991), Mirás Amor (1985), Ruiz Espejo (2013c), Tucker (1998), etc. Los planteamientos básicos son comunes para todos estos libros, partiendo de una población finita numerada con la condición de que sean identificables y localizables (accesibles y medibles u observables sin errores) físicamente si su indicador numérico fuese seleccionado en una muestra obtenida al azar por un procedimiento probabilístico, de entre todas las unidades que constituyen la población finita.

Es cierto que muchas poblaciones finitas evolucionan con el tiempo (hay nuevos nacidos y defunciones), pero en el planteamiento básico no consideramos estos cambios ya que si interesara estudiar la población finita en otro instante de tiempo el muestreo de poblaciones finitas como método objetivo sigue siendo válido en el nuevo instante o periodo temporal.

Por otro lado no aceptamos que de una misma unidad puedan aportarse más que un solo dato u observación numérica verdadera del mismo fenómeno, lo que no ocurre en algunas teorías que admiten que haya más de una posible respuesta pudiendo darse la invención de todos estos datos salvo uno a lo sumo, pues de este modo admitiríamos engaño, mentira, fraude o estafa si fueran datos económicos por ejemplo. Así preservamos el espíritu de la verdad en nuestro estudio sobre la variable de interés, que observamos de modo exacto.

Si una estimación insesgada tiene poca variabilidad es que es bastante exacta o casi sin error. La objetividad viene de considerar una *población finita fijada* que puede ser censada o numerada, y sus unidades son identificadas sin error antes de proceder a la selección de la muestra. La objetividad viene también de que no necesitamos suponer algo incierto como hipótesis de trabajo, como ocurre en otros tipos de inferencia. La objetividad surge también de que en el muestreo de poblaciones finitas fijadas, la aleatorización es un instrumento objetivo (y no supuesto) para seleccionar la muestra con determinadas condiciones; y no admitimos que la aleatorización sean unas propiedades matemáticas que se supone que la naturaleza de los datos obtenidos cumplen sin comprobación alguna, como ocurre en la mayor parte de las teorías clásicas. Ver, como ejemplos de inferencia clásica, Zacks (1971), Rohatgi (1984), Murgui Izquierdo y Escuder Vallés (1994), Casas Sánchez (1996), Stuart, Ord y Arnold (1999), Garthwaite, Jolliffe y Jones (2002), Lejeune (2010), Young y Smith (2010) y Olive (2014). No es correcto dar por cierto lo que es incierto, y menos cuando hablamos de personas o grupos sociales pues podría constituir un falso testimonio sobre personas o sociedades.

La mayor parte de las investigaciones de estadística matemática se desarrollan hasta la fecha a niveles de abstracción muy elevados, tanto que pierden el sentido de la realidad aun conservando la lógica en algún sentido. Tal vez se deba a que las decisiones sobre la publicación o no de cada aportación está en manos de profesores universitarios que priman los contenidos conceptuales teóricos de cierto nivel matemático como herramienta, y aceptable para su presentación en revistas o libros cuyas editoriales buscan una rentabilidad que se concentra especialmente en contenidos académicos más que en contenidos de

verdadero aprovechamiento práctico y ético. Los contenidos útiles en la práctica directa también son publicados pero en una proporción realmente limitada entre los efectivamente publicados a nivel mundial. Además el interés científico sobre el papel de las publicaciones científicas se reduce a indicadores del número de citas anuales de los artículos de revistas, con sus muy limitadas contabilidades por sus deficiencias en la práctica al no ser instrumentos fieles a lo que realmente es, y teniendo en cuenta el número medio de páginas por artículo publicado, en concreto para calcular el factor de impacto de las revistas científicas. Se valoran especialmente aquellos que incluyen tratamientos informáticos elaborados que precisan de programas de cálculo, gráficos de alta calidad, y todas aquellas aportaciones que hagan visibles de alguna manera las contribuciones teóricas o prácticas, lo que requiere unas inversiones en software estadístico o asimilables con unos intereses comerciales claros y un mercado de subvenciones oficiales poco claro y diáfano.

Autores que optan por aportaciones especialmente útiles y de aplicación inmediata útil, llegan a ser tratados por los comités de algunas publicaciones con mucha severidad, pues a la exigencia personal del autor por aportar instrumentos prácticos, se suma la exigencia editorial de mantener un alto nivel de abstracción y de aportación matemática del mismo nivel que a cualquier otro trabajo aspirante aunque no tenga éste utilidad social o práctica alguna. A veces las revistas exhiben su intención de publicar contenidos aplicados, pero en realidad sus temas son casi exclusivamente teóricos con alguna referencia a conceptos realmente aplicados, o tratan temas de actualidad científica pero sin aportar ninguna solución real a lo que es materia de interés para el bien humano o común además de los propios interesados inmediatos como son los autores, la universidad, la sociedad o la empresa o institución que financia la publicación.

Existe a mi juicio un exceso de respeto por contenidos clásicos aunque sean poco realistas y poco objetivos. Un ejemplo es el exceso de literatura entorno a la distribución normal. En mi opinión se dio excesiva relevancia al "teorema central del límite" que consiste en que la media muestral o media aritmética de un número de observaciones de un mismo fenómeno converge (cuando dicho número de observaciones aumenta hacia infinito y según distintos criterios de aleatorización) a la distribución normal. De este modo, se hizo recaer excesiva importancia a dicha distribución de probabilidad, pues en la teoría de diseño de experimentos anglosajón la hipótesis de normalidad es de partida y de llegada, y se aplica a otros estudios sociales y a un sin fin de aplicaciones de la estadística a pesar de que tal hipótesis de partida como las condiciones probabilísticas de la aleatorización con que se supone se extraen las observaciones sean por lo general incomprobables en la realidad a la que se pretende aplicar. Sin embargo también es posible estudiar diseño de experimentos desde una perspectiva objetiva basada en muestreo de poblaciones finitas fijadas (Ruiz Espejo, 2018f), que aporta objetividad en esta materia.

La habilidad semántica y dialéctica de muchos estadísticos profesionales ha hecho que sus afirmaciones sean en un tono ambiguo, sugiriendo que los datos observados se ajustan bien frecuentemente a una distribución normal, lo cual no significa que sea tal distribución sino que estadísticamente no hay razones significativas para rechazar la hipótesis de normalidad de los datos. Pero no se le escapa a cualquier estadístico inteligente que no rechazar una prueba dista mucho de asegurar que sea cierta. Por tanto hay razones también para dudar de la suposición de una hipótesis que no ha sido rechazada ante un test de "bondad del ajuste" de los datos a la distribución normal.

Las empresas e instituciones que se encargan de realizar sondeos o estudios por muestreo se ven más preocupadas por dar una apariencia científica si hay ficha técnica de sus estudios estadísticos que a proporcionarlos de hecho al diseñar, proyectar y realizar los métodos que pretenden hacer valer en tales investigaciones prácticas.

Las estadísticas oficiales realizadas se mueven a niveles de gran conformismo con las estructuras administrativas tradicionales de los registros de datos, una inercia que rara vez incorpora aportaciones técnicas y de verdadera investigación aplicada de los últimos tiempos. Las aportaciones de soluciones a problemas de índole técnico o científico planteados en la práctica de encuestas o muestreos, no siempre tienen eco en la práctica oficial o privada. Lo que no quiere decir que no sea deseable.

La teoría de muestreo de encuestas ha sido muy influida por los avances en tecnologías computacionales y de análisis de datos, no siempre de modo objetivo, que han sido desarrollados desde el siglo XX (Bellhouse, 2000).

Podríamos citar un gran número de libros editados sobre muestreo de poblaciones finitas y de recopilaciones de la materia en las últimas décadas. Muchos de ellos están recogidos en la tesis del autor (2003a). Por su trascendencia destacamos los de Hansen, Hurwitz y Madow (1953) que aportó bases matemáticas a su estudio, y el de Wolter (1985, 2007) que recopiló material para el análisis del error de muestreo en base a la estimación de la varianza de los estimadores de las funciones paramétricas. Ejemplos de aplicación de este libro son las metodologías originales de los trabajos de Ruiz Espejo (2013c) y de Ruiz Espejo, Delgado Pineda y Singh (2006).

Son muchos los autores (que omitimos) que también presentan enfoques complementarios en algunos casos sobre los

métodos de muestreo de poblaciones finitas de la mayoría de continentes.

La estadística explicada en las universidades españolas y en general de todo el mundo siguen una dirección influida por los avances de la matemática de los últimos siglos, en concreto del análisis infinitesimal, cálculo diferencial, análisis matemático, análisis funcional, etc. De este modo se expandieron estos conocimientos limitados por su subjetividad en la práctica a otras áreas de la ciencia como la medicina, la economía, la empresa, etc. Las aportaciones de cada ciencia solo servían para perfilar el tipo de ejemplos y ejercicios a los que se aplicaba la metodología estadística estándar que se consideraba común para todas las ramas del conocimiento científico sin hacer en muchos casos un análisis de la objetividad de sus procedimientos en cada caso práctico de estudio.

Es de reconocer las aportaciones de muchos matemáticos que sin disponer de procedimientos objetivos como hoy disponemos, han dado soluciones a muchos problemas surgidos en el campo práctico basándose en hipótesis o planteamientos próximos a las condiciones que de hecho aparecen en el contexto de la ciencia concreta a la que lo aplicaban.

Pero los métodos proporcionados por los censos, la estadística descriptiva (Mengal, 1999) y la inferencia objetiva de muestreo en poblaciones finitas fijadas (Ruiz Espejo, 2013c), han resultado ser los más consistentes, realistas, y por tanto más objetivos.

Sorprende que tanto los censos como la estadística descriptiva hayan sido excluidos de los estudios universitarios en facultades de ciencias matemáticas pues son conocimientos básicos, prácticos y fundamentales para desarrollar la inferencia

objetiva. Este tipo de estudios se relegan a personal técnico administrativo como rutinas de trabajo, mientras que las abstracciones matemáticas de cierto nivel más estériles en cuanto a su objetividad se circunscriben a estudios superiores de grado o doctorado y reciben por lo general los mejores reconocimientos y apreciaciones académicas en dichas facultades.

Una diferencia de la inferencia realizada en poblaciones finitas fijadas por muestreo, de otros tipos de inferencia, es que las unidades seleccionadas lo son con un "procedimiento controlado de aleatorización", y no "de origen supuesto" como hace la inferencia paramétrica clásica, bayesiana, no paramétrica, de distribución libre, etc. por lo general.

La selección controlada de la muestra en la inferencia objetiva puede realizarse por medio de tablas de números aleatorios (como se intentó inicialmente a principios del siglo XX), o bien con ordenadores que generen esos dígitos ejecutando programas informáticos.

De este modo se hace posible la "descripción y explicación de la realidad social objetivamente, sin deformarla con nuestros deseos o intuiciones personales", y así hacer posible en las ciencias humanas disociar la "pura observación" de la "valoración subjetiva" de los fenómenos sociales contemplados.

Los métodos estadísticos objetivos se basan en hechos y en datos de dichos hechos, por lo que describen la realidad o infieren sobre ella en base a observaciones y métodos objetivos. Los métodos inferenciales predictivos se basan en hipótesis sobre cómo se comporta el fenómeno estudiado ya sea a través de un modelo presupuesto y por tanto subjetivo o no seguro. Por tanto los métodos estadísticos predictivos tampoco superan las objeciones más elementales en busca de objetividad en el procedimiento, aunque puedan parecer más imaginativos y descomprometidos con

la búsqueda de la verdad. El conocimiento objetivo se fundamenta en las cosas que están ahí y son, al alcance total o parcial del investigador. No es algo en lo que baso supuestamente el razonamiento, sin esfuerzo en conocer y por conocer apoyándome en realidades.

Para impulsar nuestro conocimiento hay que estar abiertos a las aportaciones de otras tecnologías que pueden redirigir las investigaciones formales o técnicas, así como dar oportunidades a la imaginación constructiva.

En los tipos de inferencia diferentes del muestreo de poblaciones finitas fijadas con aleatorización controlada, el modelo distribucional asumido para la variable estadística o aleatoria de la población puede ser diferente de la supuesta o incluso no existir tal distribución que se presupone en la realidad. Dos argumentos suelen ser esgrimidos en este caso.

El primero consiste en decir que aunque la distribución poblacional sea desconocida, podría aceptarse mediante un contraste de hipótesis. Nuestra objeción es que aceptar un modelo no significa que sea el único aceptable para el mismo test y los mismos datos. Incluso puede no ser ninguno de los propuestos.

Otro argumento utilizable es que las observaciones si no existen en el caso concreto al que aplicar los métodos estadísticos, estas pueden generarse o simularse mediante un programa informático adecuado de selección de datos. Nuestra respuesta es entonces que los datos no son ya de una población natural y real, sino producidos artificialmente por un ordenador según unas instrucciones programadas, lo que reduciría el problema a un estudio didáctico o de simulación teórica sin implicación práctica social.

Estas consideraciones no hacen menoscabo del interés matemático y formal de los razonamientos que sostienen las muchas técnicas estadísticas a las que se dedican la edición de cientos de revistas periódicas en el mundo, así como libros y otros materiales especializados. Su interés parece dirigirse a fomentar, exhibir y aumentar la destreza científica de los investigadores en matemáticas o en la aplicación de las técnicas estadísticas, lo cual dista mucho de que su uso sea correcto en cualquier aplicación por el mero hecho de que sean consistentes matemáticamente.

La coherencia de todas las hipótesis con las realidades a las que se desea aplicar, es otro requisito imprescindible para el buen uso de las aportaciones matemáticas en el contexto aplicado a fenómenos no simulados sino reales y naturales. La deficiencia por la que no sean coherentes es quizás que el profesor que lo explica no siempre “está en” o “se pone en situación de” casos reales o con quienes tratan de aplicar sus aportaciones. Aceptar muchas técnicas estadísticas comporta un trasfondo de explicar fenómenos que no tienen por qué seguir sus leyes y sus reglas de discernimiento.

Una estrategia de muestreo en poblaciones finitas fijadas consiste en el par compuesto por el diseño probabilístico de selección de unidades y el estimador del parámetro poblacional del que se trata de inferir. En Ruiz Espejo (1997c, 2011a, 2015b) se presentan soluciones a problemas de este tipo de estrategias muestrales en la práctica.

El seguimiento de las realidades sociales es algo muy importante en la planificación de soluciones a las necesidades de la población extendida cada vez a áreas de mayor amplitud. Saber cuáles son los problemas es algo que se puede conseguir con métodos estadísticos, pero resolverlos es la parte principal que no puede ser atendida sin conocimientos objetivos del estado social.

Muchos estadísticos pueden avalar técnicas como capaces de aportar un método científico para el conocimiento de realidades. Lo que no es muy común es defender aquellas que, tras una reflexión constructiva por su objetividad y con la experiencia sincera del científico conocedor de la ciencia inferencial estadística y atento a la moral, se mantienen válidas ante las posibles objeciones legítimas que pudieran hacerse. Esto último es lo que pretendemos aportar en este capítulo como fruto de nuestra reflexión.

El resultado es la proliferación de libros con grandes abstracciones y un empeño de los autores en convencer de que tales teorías son perfectamente aplicables a los datos que usualmente se manejan en la materia de fondo a la que se dirige. Sin embargo es fundamental la comprobación de las hipótesis de trabajo de los resultados matemáticos aplicables a las condiciones concretas de aplicación.

Aunque no toda metodología estadística aporta la misma claridad en el conocimiento social al ser aplicadas, pensamos que algunas de ellas son totalmente objetivas para este fin. Veremos cuáles son estas metodologías o métodos razonando los porqués de su utilidad real, es decir que sirven a su fin con objetividad. La lógica que empleamos no es solo de tipo matemático sino de comprobación de si las hipótesis empleadas en los teoremas y en los razonamientos matemáticos siguen siendo condiciones reales estudio en la práctica en la que se aplican.

Para un matemático, cualquier teorema bien demostrado es ciencia, pero para aplicarse a un caso práctico no todo teorema y sus premisas son adecuados y respetan la realidad del mundo natural al que se aplica. No carece de rigor matemático cada teorema demostrado, pero al inferir en un caso concreto puede

faltarse a la máxima coherencia y seriedad deseable si las hipótesis no pueden comprobarse que son verificadas en la práctica concreta.

El uso de métodos de investigación estudiados en el laboratorio matemático sin un fin útil práctico definido y que después se aplica sin fundamentar previamente en el método y en las hipótesis básicas las realidades a las que pretendemos dar luz, no contribuye sino a la confusión y a la creación de resultados sin base segura.

Con unos métodos de investigación adecuados a los aspectos de interés social y económico, es posible proyectar en base a informaciones fidedignas obtenidas cualquier tipo de política social y diseñar las disposiciones, convenios o pactos legales que den un carácter de derecho positivo a los compromisos, los acuerdos y las actuaciones consecuentes.

Los métodos estadísticos proporcionan instrumentos técnicos para detectar la evolución y el cambio de los hechos entre dos instantes o periodos de tiempo determinados y lugar concretos. Su uso ha sido importante en el U. S. Census Bureau y en otras instituciones oficiales de estadística.

Desde el punto de vista del desarrollo estadístico asistimos al esfuerzo de adaptación de los sistemas de información y de las metodologías estadísticas de las economías y estados en transición de los países del este de Europa tras su adhesión a la Unión Europea en las últimas décadas.

La sociedad global del bienestar no existe como realidad en la actualidad pero puede ser realizable en el futuro, y esta posibilidad de realidad es más deseable socialmente que la realidad global presente. Sin duda el papel de la estadística objetiva es clave en esa sociedad del bienestar y en la ya existente.

La estadística y en especial el muestreo de poblaciones finitas tendrán siempre un interés como ciencia aplicada, por resumir la información de los habitantes del planeta, además de hacerlo de modo económico al reducir los encuestados a una pequeña fracción de la población total. Además no producirá efecto cansancio de los respondientes y de los encuestadores como se daría si todas las personas fueran encuestadas repetidamente.

Sin embargo en España no se dispone de ficheros actualizados de todos los habitantes del país que permitan identificar y seleccionar muestras de ellos con fines sociales. En este terreno los especialistas en informática y sus soluciones técnicas tienen en su mano resolverlo alcanzando una administración informatizada. Aunque es posible ya una selección de muestras en algunas bases de datos a partir de ficheros parciales continuos o secciones censales periódicamente actualizadas.

Es necesaria la interconexión entre teoría y aplicación práctica, conciliando las condiciones de aplicación práctica de un método y sus conceptos o hipótesis con la realidad que queremos conocer objeto de estudio. Así, los métodos objetivos para estudiar el comportamiento social no tienen que ser decisivos en las conclusiones del estudio, si aquellos se han fundamentado en datos ciertos.

Tras el próximo capítulo concluimos que algunas de las metodologías de inferencia estadística son mejores que otras para estimar parámetros poblacionales de variables cuantitativas fijas y observables de interés social. Esto no quiere decir que con tales métodos quede todo explicado a la luz de ciertas definiciones previas, o que no haya más que un método bueno para conocer la sociedad. Lo que sí quiere decir es que para ciertos parámetros poblacionales de interés social que pueden describirse mediante

variables cuantitativas y fijadas en cada unidad de la población, existen métodos claramente capaces de superar todas o más objeciones que otros muchos métodos estadísticos que ocupan un lugar importante en las investigaciones y publicaciones científicas actuales, pero no superan las mismas objeciones.

La lógica interna de una demostración matemática es más fácil de analizar que la verdad de las proposiciones prácticas.

## 9.3 Desarrollos estadísticos

En esta sección explicamos los argumentos fundamentales en que nos basamos para hacer una selección de métodos de estadística inferencial tal y como se desarrollan en los cursos universitarios de la materia.

Queremos hacer ver con claridad las razones que nos impulsan a dudar de ciertas metodologías en la práctica, argumentando lógicamente y respetando la verdad pues éste es el fin de un estudio estadístico inferencial, arrojar luz y claridad al fenómeno estudiado. También proponemos otras metodologías que superan tales condicionantes, por lo que nuestra intención es plenamente constructiva, veraz y racional.

Uno de los puntos de partida para valorar estas metodologías es la existencia real o no de la probabilidad en la naturaleza. No me refiero al azar. Sino al hecho o ilusión sin base real de encontrar indicios de que la probabilidad como concepto matemático acuñado en 1933 por Kolmogorov sea revalidado en el mundo en que vivimos.

El hecho de que algún libro de física como el de Pécseli (2000) usen del concepto para explicar realidades no prueba su existencia real. En las ciencias sociales ocurre algo parecido.

Cuando realizamos un experimento y aun poniendo el máximo cuidado en controlar todas las circunstancias importantes, el resultado de tales casos varía de una observación a otra en una forma irregular que elude todo tipo de predicción sobre el resultado, y en este caso Cramér (1953) considera que la sucesión de experimentos son aleatorios. Cualquier registro sistemático de los resultados de sucesiones de experimentos constituye un conjunto de datos estadísticos relativos al fenómeno considerado. El objetivo de la estadística, para este autor, es investigar la posibilidad de extraer de los datos estadísticos inferencias válidas, elaborando los métodos mediante los cuales pueden obtenerse tales inferencias. Pero si las condiciones no fueran similares en cada experimento sino que fueran exactamente las mismas, ¿habría aleatoriedad o resultados diferentes en dos o más experimentos así realizados?

Cramér añade que debe modificarse toda teoría que no se ajuste a los hechos, como principio general de toda investigación científica que se denomine como tal. Este principio racional puede ser aplicado por la mayoría de escuelas de estadística, pues damos a continuación argumentos de hecho para que reconsideren sus estudios.

La mayoría de las investigaciones matemáticas de tipo estadístico en la actualidad utilizan hipótesis de partida en los razonamientos que no pueden ser comprobadas o corroboradas directamente “antes de” ni “durante” su aplicación al fenómeno que se estudia.

En concreto, se utiliza el concepto de distribución poblacional y se le asigna una distribución determinada salvo uno o varios parámetros desconocidos de la misma que corresponden a una clase de ellas del mismo tipo y distribución, pero de las que se trata de

estimar dichos parámetros para decir algo de las características a investigar (Ríos García, 1977).

La situación anterior es la más elemental presentada en la inferencia paramétrica clásica. La situación no varía mucho para otros tipos de inferencia como la bayesiana o la no paramétrica, donde las hipótesis formuladas pueden hacerse aún más incomprobables y abstractas, despegándose más de lo tangible, comprobable y controlable para poder partir de unas condiciones lo más realistas y objetivas posibles, como sería deseable y se requiere en la práctica.

Utilizar la inferencia estadística presupone la aceptación, como en la mayor parte de los métodos estadísticos, de que la población experimental (si existiera de hecho) de la que tomamos muestras, se distribuye según alguna teoría o modelo de distribución. Esto exigiría la comprobación práctica de unos axiomas matemáticos incomprobables a su vez que expliquen y hagan válida la inferencia estadística en tales ciencias experimentales.

Aceptar en la realidad el concepto de "población experimental" implica aceptar el concepto de la probabilidad, sobre la cual no hay evidencia física ni consenso entre los científicos de su existencia real. A pesar de ello, la simulación con ordenadores de este concepto y estos modelos ha permitido resolver cuestiones científicas de carácter matemático, como por ejemplo los métodos de Monte Carlo para el cálculo aproximado del número $\pi$ (Ríos García, 1977).

Como consecuencia, no podemos afirmar nada experimentalmente a partir de la estadística inferencial sin comprobar las hipótesis o axiomas que fundamentan los métodos estadísticos usados, que deben adecuarse a la realidad que experimentamos. Pero como esta adecuación última es

incomprobable en la mayor parte de los casos aplicados con las técnicas estadísticas inferenciales dichas, podemos concluir que la mayor parte de las investigaciones estadísticas de tipo matemático aportan además de elaborados razonamientos lógicos, poca o dudosa luz sobre las realidades a las que se aplican.

La hipótesis poblacional de normalidad, como axioma que ha ocupado y ocupa el centro de los modelos de distribución de los datos experimentales, está basada según sus defensores en alegar argumentos de tipo experimental. Este modelo de distribución poblacional no se suele justificar, pues es imposible de demostrar, pero suele ser admitido como argumento explicativo de realidades de tipo agrícola o biológico en diseño de experimentos desde su inicio con tal axioma de normalidad. El uso de la "distribución normal" en estudios aplicados es algo muy común. Esta distribución fue descubierta por De Moivre en 1733, como distribución límite de la distribución binomial, aunque su descubrimiento pasó inadvertido. Posteriormente Gauss en 1809 y Laplace en 1812 la redescubrieron. Sus obras en las que publicaron sus resultados fueron muy influyentes de modo que de modo casi axiomático sus seguidores consideraron que prácticamente cualquier distribución estadística en la práctica se acercaría a la distribución normal con solo disponer de un número grande de observaciones suficientemente precisas.

Así se pensaba que la desviación de cualquier variable aleatoria respecto a su media se consideraba como un "error" sujeto a la "ley de errores" que a su vez se expresaba tácitamente como asumible directamente por la distribución normal.

El "teorema central del límite" que asegura que la media aritmética de un gran número de variables aleatorias independientes e igualmente distribuidas, tiene una distribución

normal en el límite, y ampliaciones particulares posteriores de este teorema que aseguran este comportamiento para funciones más generales que la media aritmética, así como en distintas condiciones de variables dependientes, hicieron que muchos científicos creyeran en la "ley de errores" como algo casi natural; los experimentadores lo creyeron porque piensan que se trata de un teorema matemático, y los matemáticos lo creyeron por pensar que era un hecho experimental.

Sin embargo estas creencias no deben ser absolutas, ya que es difícil o casi imposible encontrar en la práctica exactamente las condiciones que garantizan matemáticamente esa conclusión, y además la experiencia de muchos científicos posteriores en distintos campos de conocimiento nos hace ver que la "ley de errores" no es ni mucho menos un absoluto, como puede verse en la distribución de rentas en Economía que son de tipo "asimétricas a la derecha" y no "normales y simétricas". Lo que significaría la falsación de la teoría que, según el filósofo Popper, por ello debería abandonarse como generalizable a cualquier fenómeno, y debiéndose demostrar en cada caso su idoneidad al fenómeno estudiado.

Las cuatro fases del proceso estadístico según Ríos García (1977), son: descripción, análisis, contraste de hipótesis y aplicación a la previsión. La primera fase tiene por finalidad presentar los datos observados de diversas maneras describiendo en todo momento la realidad constatable y objetiva mediante operaciones simples de tipo matemático. La fase de análisis (de construcción de un modelo teórico que permite enunciar una ley), y de contraste de hipótesis (con nuevas experiencias que pueden hacer confirmarla o rechazarla), corresponden a la estadística subjetiva e inductiva, en la que pueden obtenerse avances no necesariamente "seguros" en el conocimiento de los hechos, sino

solo “posibles” o “probables” y especialmente débiles cuando se incorporan teorías y modelos subjetivos para su obtención.

La fase de aplicación a la previsión, o utilización de la ley enunciada para anticipar los resultados de nuevas experiencias, podría tener utilidad en algún caso, pero existen riesgos en su mal uso práctico debido a la formulación de leyes incomprobables, o por su inexistencia en la realidad o su evolución a lo largo del tiempo. En tales casos la previsión es parcial o totalmente a ciegas, con sus consecuentes derivaciones que podrían falsear las predicciones.

De estas críticas está libre el muestreo de poblaciones finitas fijadas que permite evaluar inferencialmente situaciones sociales de hecho. Lo cual no elimina las limitaciones prácticas del mismo. Así, por ejemplo, no siempre es posible reunir efectivamente los requisitos imprescindibles para ser aplicado. No siempre es posible disponer de un listado completo de las unidades que componen la población. Esta situación, que podría presentarse, puede ser subsanada con leyes censales que exijan a los ciudadanos, empresas, pacientes, etc. su inscripción en los “registros oficiales” de los que se puedan obtener los listados siempre para beneficio de los propios registrados y de la comunidad.

La estadística matemática inferencial se basa en el concepto de probabilidad, que puede no existir en la realidad como ha afirmado el estadístico matemático italiano Bruno de Finetti (1974, 1975) en sus libros de teoría de la probabilidad.

De ser así, surgen dos opciones en la estadística práctica: conformarnos con lo que sabemos por métodos descriptivos y censales, o bien aprovechar los resultados matemáticos de estadística inferencial simulando (por ejemplo, con ordenadores) su existencia y reproduciendo “número aleatorios” seleccionados

de acuerdo a ese concepto teórico de probabilidad que, aunque no exista en realidad tal probabilidad, los "números aleatorios" permitan reproducir sus propiedades y aprovechar los resultados demostrados por estadísticos matemáticos para resolver cuestiones de tipo práctico. De no ser así, la posible existencia de la probabilidad aunque no haya sido probada tampoco asegura conocer su valor axiomático exacto para cada suceso que nos interese en la práctica, debido a la imposibilidad real de conocerla en su valor numérico exacto y, esto, si fuera objetivo e independiente del tiempo. Si su valor exacto es desconocido, pocas leyes podríamos aplicar en la práctica con la mínimas garantías de seguridad en que las distribuciones probabilísticas utilizadas inferencialmente sean las verdaderas o adecuadas en cada caso concreto en que pretendamos usarlas con fines de utilidad práctica, como es en la investigación social y biomédica.

A pesar de no saber en realidad si existe o no en la práctica el concepto de probabilidad, los científicos han dado lugar a muchas maneras de interpretarla o de definirla. La axiomática de Kolmogorov parte ya de su existencia, y regula las condiciones mínimas que debe cumplir tal concepto, unas propiedades lógicas derivadas de las propiedades límite de otro concepto que sí es medible, la frecuencia relativa de un suceso.

El concepto de probabilidad de un suceso es el de frecuencia relativa del mismo suceso y su límite al realizar una sucesión de experimentos en idénticas condiciones. Al ser cada experimento independiente de los anteriores y posteriores, para cada número finito de experiencias existe una frecuencia relativa de ocurrencia del suceso, pero en realidad nunca se conocerá el límite de la sucesión de experimentos al no poder realizar el cómputo final de la frecuencia relativa de los infinitos experimentos necesarios para obtener el límite de tal sucesión de frecuencias producidas en las sucesivas experimentaciones acumuladas.

De este modo, aunque la probabilidad de un suceso existiese, tal probabilidad no podrá ser conocida experimentalmente, sino solo por aproximaciones que nos puedan proporcionar las frecuencias relativas de dicho suceso en un número finito de experiencias observables. De aquí que la probabilidad, aun existiendo supuestamente, no será posible conocerla con exactitud de la experimentación.

Otro concepto de probabilidad es el de "probabilidad intuitiva, lógica o necesaria" debida a George Boole y propuesta por él como una generalización de la lógica, trata de medir la relación entre dos proposiciones concretas, una de las cuales no es consecuencia lógica de la otra.

El concepto de probabilidad utilizado por la relación "apuesta/premio" como cociente entre dos cantidades económicas, de las que el denominador es una cantidad objetiva, y el numerador es subjetivo para cada jugador o apostante, es otra definición subjetiva de probabilidad muy conocida entre jugadores y economistas.

La confianza de un individuo en la realización de un suceso es utilizado en la teoría clásica de la probabilidad, como ocurre en la teoría bayesiana, debida al pastor protestante Thomas Bayes. Para la teoría probabilística e inferencial bayesiana, la asignación de probabilidades o de distribuciones "a priori" es algo que aporta el propio investigador estadístico quien realiza unas valoraciones generalmente subjetivas, que en muchos casos o en la mayoría son inasumibles por otros investigadores aun compartiendo la misma metodología bayesiana, y por todos aquellos que creen en la objetividad de las probabilidades y de sus distribuciones, si ambas existieran.

Las teorías subjetivas de la probabilidad tienen sentido para el sujeto o individuo que aporta su idea u opinión “a priori” sobre la probabilidad de los sucesos, y que puede modificar tal idea u opinión al incorporar nuevos resultados experimentales sobre el mismo suceso.

Para todos estos, o su gran mayoría, no tiene sentido plantear un concepto objetivo de la probabilidad de un suceso compartido por todos, sino que es más bien un instrumento personal más que aspira a ser utilizado como herramienta en el proceso de análisis teórico y formal con la posibilidad de incorporar nueva experimentación. Pero esta experimentación no se obtiene siempre por métodos objetivos tampoco, sino que pueden contener un sesgo intencional e incluso no probabilístico en la obtención de los datos ya que las unidades no estarían identificadas en algunos casos y no serían accesibles con igual o supuesta probabilidad de cada observación en la práctica.

De lo anterior podemos concluir que la estadística moderna que (con las excepciones de la estadística descriptiva y de la inferencia en poblaciones finitas fijadas, con probabilidades simuladas por ordenador para la selección de unidades) está basada en el concepto de probabilidad intrínseca en la naturaleza de los datos, puede considerarse una construcción lógica pero con pies de barro al apoyarse en concepciones de la probabilidad de los que no hay garantías de su existencia. Por tanto, las consecuencias de las teorías inferenciales que se fundamentan en ellas no pueden ser de una garantía como si tales conceptos hubieran sido demostrados y comprobados en la práctica.

En el siglo XIX, la calidad en la estadística se entendía como la consecuencia de la seguridad y la evidencia de naturaleza exhaustiva en la actividad de la recolección de los datos censales

de los cuerpos oficiales, concepción inspirada por la propia "teoría legal" que subyace en dicha filosofía de actuación.

En los años 20 del siglo XX el estado del bienestar introdujo métodos de dirección estadísticos en el sentido actuarial de la palabra (Desrosières, 1997), lo que dio legitimidad social a los datos elaborados estadísticamente vinculando cambios en el Estado y en la concepción de la estadística como ciencia.

En la década de 1930, el estadístico americano W. E. Deming fue quien introdujo el cálculo probabilístico en la estadística oficial, concretamente en las primeras investigaciones por encuestas de empleo y desempleo (Anderson, 1988). Hasta entonces las muestras se seleccionaban con criterios de representatividad o proporcionalidad (incluyendo el azar no probabilístico), pretendiendo en todo caso que la muestra fuera una miniatura de la población sobre la que se quería inferir, pero sin utilizar de hecho el propio concepto de probabilidad que ya utilizaban y manejaban matemáticos, filósofos, lógicos, etc.

En la década de los años 1940, Deming usó las mismas técnicas para el desarrollo del "control de la calidad" en la producción industrial, manejando el muestreo aleatorio y la verificación de defectos en los artículos producidos en serie en Estados Unidos, y posteriormente en Japón y Europa con su "quality movement" y "quality circles" en los años 1980, y la "total quality" y la "zero-defect" tan en boga las últimas décadas en las industrias automovilística y electrónica.

Una dificultad de los métodos inferenciales basados en la probabilidad es la de crear las condiciones experimentales para reproducir las mismas características en diferentes observaciones de un mismo fenómeno aleatorio. En un ejemplo físico, pensamos que reproducir las mismas características para evitar que el

movimiento de las estrellas y planetas no influyera en las leyes gravitacionales es prácticamente imposible o no está al alcance humano pues aunque pueda parecer imperceptible a nuestros ojos, son cambios reales y de dimensiones muy grandes. Este cambio en las causas (admitiendo la ley de la gravitación universal como un ejemplo), hace que sea suficiente en principio para producir efectos diversos en la experimentación sucesiva en la que el transcurso del tiempo tiene su importancia en el cambio de las condiciones externas.

Suponer que las distribuciones discretas, continuas y otras más generales, incluyendo mixturas de ellas, son el modelo poblacional objetivo de la investigación social no puede ser un hecho seguro, pues hemos visto que las poblaciones humanas son finitas, y si cada unidad tiene la misma probabilidad teórica (en el sentido de Kolmogorov) de ser seleccionada en cada selección, hace ver que la distribución en este caso es uniforme discreta y la muestra que origina los datos es una muestra aleatoria simple (ver el concepto de distribución uniforme discreta en el libro de Casas Sánchez y Santos Peñas, 1995). Básicamente consiste en una distribución discreta que concentra probabilidad igual positiva en un número finito de puntos de la recta real.

Así pues, la inferencia basada en modelos paramétricos o no paramétricos se hace imposible de llevar en condiciones objetivas aún en el caso más sencillo de muestras aleatorias simples. Con mayor razón, si las observaciones son dependientes o con mayor sofistificación, serán modelos más irreconocibles en la práctica desde la deseable objetividad.

En cualquier tipo de inferencia estadística tradicional se pretende conocer algo sobre la población completa de partida en base a una muestra de la misma población. En la estimación puntual, la muestra de observaciones se utiliza para aproximar uno

o varios parámetros desconocidos de la población aunque ésta pueda ser sospechada o conocida a excepción de uno o algunos de los parámetros que actúan como constantes desconocidas en la inferencia paramétrica o como distribuciones supuestas a su vez en la inferencia bayesiana.

En la estimación por intervalo, la muestra sirve para proporcionar un intervalo que contiene al supuesto valor del parámetro de interés con determinado nivel de confianza. También pueden estimarse por intervalo dos o más parámetros, dando lugar a dos o más intervalos de confianza.

En el contraste de hipótesis, se trata de decidir si se acepta o se rechaza una hipótesis relativa a uno o varios parámetros poblacionales con cierto nivel de confianza, basándose en una muestra aleatoria de datos procedentes de la observación de la misma población sobre la que se trata de inferir sus características.

En general, la estimación por intervalo y el contraste de hipótesis pueden realizarse a nivel teórico conociendo la distribución exacta o aproximada de algún estadístico o función de la muestra y que dependa del parámetro a inferir. Para conocer tal distribución del estadístico, es necesario conocer la distribución poblacional de partida. Aun cuando existiera esa distribución poblacional (cosa no garantizada por las razones de que podría no existir la probabilidad, ni es posible saber con certeza por lo general el tipo de la misma), la distribución del estadístico no es conocida ni suele ser segura ni comprobable en la práctica. Entre otras razones porque no hay modo posible conocido de garantizar que las observaciones hayan respetado rigurosamente las propiedades probabilísticas de la selección aleatoria de dichas observaciones.

En la inferencia paramétrica clásica se supone que la población se distribuye según cierto modelo de distribución o ley,

que determina su clase de distribuciones (por ejemplo: normal, uniforme, gamma, beta, etc.) antes de obtener las observaciones, salvo uno o varios parámetros a los que habría que estimar en base a los datos observados o experimentales procedentes de la misma población o de la clase de distribuciones poblacionales (clase que se suele considerar fijada en todo el proceso inferencial).

Una vez fijados los parámetros, determinan una única distribución de probabilidad de la clase de distribuciones. En la inferencia paramétrica clásica el problema se reduce a estimar el o los parámetros desconocidos y supuestamente fijos con la ayuda del modelo supuesto y de las observaciones que se toman. Así la distribución del estimador de cada parámetro depende del modelo supuesto, por tanto de su o sus parámetros, y de las observaciones, así como de la elección del estimador concreto o estimadores tomados.

Algunos criterios de selección de estimadores son el principio de máxima verosimilitud, el principio de suficiencia, el principio de completitud, etc. También existen otros diversos métodos estadísticos para la obtención de estimadores, como el método de los momentos, el método de los cuadrados mínimos, etc.

Una vez seleccionado el estimador por alguno de los criterios o métodos anteriores, puede estudiarse si verifican propiedades deseables como la insesgación, la varianza mínima uniformemente, la eficiencia asintótica, etc. que son de gran utilidad para apreciar el estimador según las propiedades que verifica. Libros como los de Cramér (1953), Ríos García (1977), Stuart y Ord (1994), Stuart, Ord y Arnold (1999), y Olive (2014) contienen elementos de todos estos extremos apuntados.

Otro tipo de inferencia es la no paramétrica, en la que la población objetivo no pertenece a un modelo dado exceptuando uno o varios parámetros que toman un único valor fijo y

desconocido cada uno de ellos, como suponíamos en la inferencia paramétrica clásica. Sino que ahora la población pertenece a una clase de variables aleatorias de un tipo más general, como podría ser el de las variables aleatorias con función de distribución continua, o con función de densidad continua, o con una función de densidad conocida salvo su media y su varianza (es decir, conocida salvo cambios de origen y escala) u otras muchas posibilidades en las que incluyan clases muy generales de distribuciones entre las que se supone se encuentra la población objetivo que es en concreto sobre la que queremos inferir.

Las restricciones realizadas para definir la clase de distribuciones posibles de la población objetivo pueden provenir de condiciones de buenas cualidades de facilidad en el manejo matemático por estar ya estudiadas sus propiedades inferenciales o de condiciones de origen o de propiedades de tipo matemático (como pueden ser la continuidad, derivabilidad sucesiva de las funciones, etc.). En este tipo de inferencia no paramétrica, al aumentar el número de posibles poblaciones es lógico que la población objetivo pueda estar mejor aproximada entre las posibles que en el caso paramétrico, aunque no siempre sería así.

Para la inferencia no paramétrica tampoco tenemos garantías de que las observaciones disponibles se hayan seleccionado según las condiciones de aleatorización supuestas sobre el papel, al igual que ocurría en la inferencia paramétrica clásica. La posible no existencia de la probabilidad como causa de los posibles datos, sigue pesando sobre este tipo de inferencia. También la imposible comprobación de que la selección probabilística supuesta se produce en la práctica, ya que no hay unidades identificadas y accesibles en general.

Un caso particular de inferencia no paramétrica es la inferencia de distribución libre. En este caso se supone que la población puede ser cualquiera (libre), desconocida y fija, sin limitaciones particulares como ocurre en la inferencia no paramétrica habitual.

Del mismo modo que la inferencia no paramétrica no puede mejorar desde un punto de vista práctico las inferencias por suponer hipótesis sin comprobación posible, como ocurre en la paramétrica clásica, la inferencia de distribución libre mejora siempre a la inferencia no paramétrica ya que los datos pueden ser mejor aproximados por cualquier posible población, mientras que en las otras inferencias el rango de poblaciones es menor entre las que dilucidar la mejor población concreta aproximada a los datos obtenidos. Otra deficiencia de la inferencia de distribución libre es que da la misma importancia a una distribución posible como es una distribución uniforme discreta, como a otra imposible que contradiga las condiciones prácticas.

La inferencia bayesiana parte de una distribución poblacional que como en los casos anteriores de inferencia paramétrica o no paramétrica se supone conocida de entrada salvo alguna o algunas constantes, cuya distribución es subjetiva aunque no podrán comprobarse estos extremos salvo que haya un control real de la distribución o distribuciones “a priori”.

Además se tiene la existencia supuesta de uno o más parámetros poblacionales desconocidos de antemano y que a su vez se supone que serían variables aleatorias con una determinada distribución de probabilidad subjetiva, cuya justificación no siempre es suficiente a juicio de muchos autores, como por ejemplo Ríos García (1977).

En realidad la lógica que soporta tal afirmación no difiere mucho del esgrimido en la adopción de una distribución de

probabilidad en la inferencia paramétrica, pues será en la mayoría de los casos una suposición incomprobable e injustificable en sus extremos, si bien puede tener algo de aproximación subjetiva que puede basarse en experiencias anteriores, pero insuficientes para poder afirmar con seguridad cuál es la distribución (O'Hagan, 1994).

Sin embargo la probabilidad condicional puede ayudar a conocer mejor procedimientos de la inferencia estadística objetiva (Ruiz Espejo y Singh, 2003).

En la inferencia en poblaciones finitas la variable de interés está fijada en cada una de las unidades de la población considerada y puede afirmarse que la distribución poblacional es uniforme discreta si el procedimiento de selección asigna la misma probabilidad a cada unidad de la población finita.

Así el argumento fundamental no es una suposición incomprobable sobre la naturaleza aleatoria en sí de la variable observada, sino más bien en los hechos de saber que la población es finita, que la variable se concreta en un valor fijo observable en cada unidad de la población finita y en que la aleatoriedad surge solo de la selección aleatoria y controlada (artificialmente) de la muestra de la población finita.

Tal aleatorización no es proporcionada implícitamente por los propios datos naturales a los que se accede, como ocurre en los otros métodos estadísticos inferenciales tratados como los paramétricos, no paramétricos, de distribución libre, o bayesianos.

La aleatorización en la inferencia en poblaciones finitas fijadas procede de la aplicación de métodos de muestreo con selección aleatoria y probabilística de unidades que puede simularse por ordenador que genere números aleatorios que

permitan obtener las unidades de la muestra (aleatoria antes de la selección, y fijada después) con la que, al ser observadas tales unidades de la muestra seleccionada, poder basar nuestra inferencia objetiva.

Así cualquier unidad de la población finita puede ser seleccionada en la muestra y observada, medida, encuestada e inspeccionada en su caso, para aportar su información cierta al estudio concreto para el que se requiere información. Esto no ocurre por lo general en otros tipos de inferencia, donde se accede a los datos disponibles, aunque no suponga un esfuerzo especial y requieran atenerse a unas propiedades determinadas de diseño probabilístico, por buscarlos de hecho donde los haya. En estos tipos de inferencia, sería una muestra de pacientes los que consultan en cierta semana a un médico. Pero en el muestreo de poblaciones finitas se requiere conocer la lista de pacientes, y de ella seleccionar la muestra por métodos probabilísticos y no de un mero azar que luego interpretemos que es una muestra aleatoria simple o con otro diseño de muestreo determinado sin hacer nada para asegurarlo en la práctica.

En la inferencia de los modelos superpoblacionales y en estudios analíticos, se supone que el dato fijo observado es una muestra aleatoria de tamaño uno de un modelo probabilístico que se supone actúa para generar los datos en cada unidad e inherente a la naturaleza de la misma. Este modelo puede ser común para todas las unidades o diferente dependiendo de la unidad o de la observación. Por ello este modelo no es seguro ni comprobable como hemos visto en otros tipos de inferencia, y puede ser un paso en el vacío que separa la teoría de la realidad concreta a la que se pretende inferir o aproximar inferencialmente.

Esto es cierto además porque de suponer un modelo así, cada unidad puede dar un dato distinto en cada ocasión en que se le

observe, pero contradice la realidad cuando en ella nos interesamos por hechos únicos y fijos (datos fijos o fijados), además de que son observables y medidos sin error. La posible suposición de que tales datos fijos son provenientes de un mismo modelo aleatorio es en sí misma equívoca y no se atiene a la exactitud de los hechos.

En los estudios sociológicos es muy común usar las fórmulas matemáticas para estimar y calcular errores pero estas fórmulas suponen que la selección de la muestra es de tipo probabilístico, algo que no suele ocurrir en el tipo de estudios o sondeos por cuotas (Martínez, 1999).

La estimación y el contraste de hipótesis propios de la inferencia con suposiciones que deforman la realidad, no hacen sino obstaculizar el conocimiento de métodos objetivos al ocuparles un tiempo precioso que se ha negado a éstos que no dan pasos en el vacío. El empeño en que los métodos inferenciales tienen que suponer que la población puede representarse por funciones de densidad en muchas aplicaciones no obedece más que a la conveniencia matemática para argumentar lógicamente el modelo a nivel teórico, más que en un verdadero conocimiento del caso concreto propiamente dicho del que se trata en la práctica.

El muestreo de poblaciones finitas con datos fijos y observados sin error es uno de los métodos estadísticos que gozan de la mayor objetividad y es uno de los procedimientos inferenciales de mayor uso en la estadística oficial en los países más desarrollados y democráticos. También es la base técnica para muchos indicadores oficiales del bienestar y en su recogida de datos.

Otras técnicas de inferencia se desarrollan en contextos de educación universitaria y de investigación teórica pues son aportaciones de menor importancia real por su trascendencia

aunque a veces de mayor impacto en investigación según criterios universitarios del ámbito inglés, lo que da índice de la disociación entre el mundo académico y la práctica de nivel objetivo.

En España, el "Instituto Nacional de Estadística", el "Centro de Investigaciones Sociológicas", y el "Instituto de Estudios Fiscales" son tres ejemplos importantes donde se diseñan y desde donde se efectúan estudios de carácter censal, estadístico, sociopolítico y socioeconómico. En Estados Unidos, los censos de población, de tipo electoral, y algunos otros más relacionados los realiza el "U. S. Census Bureau", y también tienen importancia los censos y recuentos de tipo laboral ("Labour Force").

En la mayor parte de los países existen fuentes estadísticas oficiales desagregadas en un conjunto de organismos oficiales a los que les afecta aunque, en el caso de Europa, con unas directrices nacionales y multinacionales dirigidas desde Eurostat, la "Oficina Estadística de las Comunidades Europeas", o desde la OCDE, la "Organización para la Cooperación y el Desarrollo Económico", para facilitar la comparabilidad interna y externa de los datos obtenidos entre diversos países o regiones, áreas geográficas o políticas.

Además existe un conjunto de empresas privadas que conjunta o individualmente colaboran en la realización de estudios estadísticos de predicción electoral, sondeos de opinión, investigación de mercados, etc. así como empresas que aportan tecnologías para la realización de estos estudios.

Como rama científica la inferencia estadística objetiva tiene los mismos fundamentos en todos los países donde se investigan estos métodos de alta calidad estadística. La recogida de datos se realiza en los estudios por observación física o registral, o por entrevista postal, presencial, por internet o telefónica. En algunos casos se procede a posteriori a efectuar inspecciones o

supervisiones de los datos recabados en la primera fase, para corregir posibles sesgos por posibles errores de medida o por efectos de la no respuesta (Ruiz Espejo, 1988a).

Los sondeos o muestreos por cuotas tan usados en centros oficiales de investigación de la opinión pública así como de empresas de estudios socioeconómicos y de sondeos de opinión, no tienen por lo general base probabilística y por tanto no son inferencia estadística objetiva aunque en muchos casos vistos se les de esa apariencia al presentar información de errores de muestreo o intervalos de confianza cuando éstos solo son posibles con selecciones controladas probabilísticas de las unidades de la población en base al marco actualizado de todas las unidades de la población. Si no hay probabilidad en la selección de la muestra, no puede hablarse de insesgación o de varianza de los estimadores o de nivel de confianza o de intervalos de confianza.

Otro tipo de estafa es realizar un muestreo sistemático o de otro tipo más complejo que incluye el azar en la selección de unidades, y presentar las conclusiones del estudio como si fuera hecho por muestreo aleatorio simple y utilizando sus fórmulas, cuando no fueron obtenidos los datos por este procedimiento de aleatorización.

La desconexión real entre las premisas exigidas en la teoría y las condiciones prácticas en que se realiza el estudio, le hace carecer de rigor y garantías para ser presentado como científico y como estudio objetivo. Esto es extensivo a otros tipos de materias estadísticas, como el modelo general lineal, los métodos de regresión, el análisis multivariante, la teoría estadística de la decisión, el diseño de experimentos, la biometría, etc. que por lo general necesitan mayor concordancia entre las hipótesis de trabajo

y la práctica, que las hipótesis aplicadas sean hechos objetivos y no meras suposiciones, etc.

Un ejemplo práctico en el que se puede ver por qué es tan importante comprobar las hipótesis formuladas matemáticamente, es el de Ruiz Espejo y Singh (2001), en el que se justifica cómo ante unos mismos datos observados, las diversas hipótesis que pueden formularse teóricamente de cómo surgen los datos generados, hacen seleccionar distintos estimadores insesgados a veces únicos, e incluso óptimos. En la práctica estas hipótesis de partida del modelo no suelen ser comprobadas ni comprobables, sino que queda en manos de la decisión tomada por el investigador encargado del estudio.

En otras palabras, ante un caso de duda, es el investigador o el estadístico quien desde su subjetividad o experiencia decide al final el estimador insesgado u óptimo siempre que tuviera razón al seleccionar el modelo subjetivo que propone como generador de los datos, de lo cual nunca sabremos la verdad con exactitud en la mayor parte de las inferencias estudiadas.

En este sentido, puede influir la “opinión” del experto en los resultados del estudio además de la propia “realidad” que genera los datos, se valora la opinión subjetiva de una persona como la verdadera realidad. Así en muchos tipos de inferencia se antepone la idea subjetiva de una persona a la verdad, lo cual sería una desorientación o perversión ante la información objetiva que se busca y es posible obtener.

Indicamos que en todos los tipos de inferencia explicados, salvo la inferencia objetiva en poblaciones finitas, no requieren de un listado identificador de todas las unidades de la población. El hecho de que en un estudio muestral se realice sin controlar efectivamente la selección probabilística de la muestra, hace debilitar en muchos aspectos la fuerza de la verdad y de la

comprobabilidad de las posibles respuestas que dan los encuestados y, por tanto, de las conclusiones del estudio.

## 9.4 Bioestadística

En esta sección, que inicialmente fue un trabajo académico, que ampliaba mis investigaciones, titulado *Investigación Ética y Bioestadística*, vamos a hacer un repaso a los diversos aspectos que consideramos mejorables o advertibles respecto a la ética del uso de la Bioestadística como medio para investigar o aproximar científicamente la eficacia de medios saludables en personas sanas así como de tratamientos o terapias curativas para los pacientes de diversas enfermedades.

Uno de los fines de la bioestadística es determinar si un tratamiento médico es más eficaz que otros ya disponibles.

La bioética se centra en el hombre-persona, mientras que la bioestadística se centra en la objetividad de lo que podemos conocer o inferir de unos datos experimentales. Por tanto tiene prioridad un trato digno con las personas antes que avanzar en el conocimiento científico (la caridad o el amor a las personas sobre la verdad o el conocimiento de las personas). Pero tampoco daríamos un trato digno a las personas si no tuviéramos un buen conocimiento científico para curar o paliar sus enfermedades o dolencias cuando aparecen.

El estudio de la estadística aplicada a la Biomedicina ha sido objeto de diversos libros en las últimas décadas. La estadística que emplean estos libros suelen ser de tipo inferencial basada en supuestas hipótesis de normalidad de los datos obtenidos, lo que permite aprovechar los métodos estadísticos llamados clásicos con dichas hipótesis. Algunas referencias de estos libros se encuentran

en las Referencias, y especialmente también en la bibliografía de la tesis doctoral en Sociología del autor (2003a).

Con métodos de muestreo y estimación en poblaciones finitas tenemos actualmente instrumentos para inferir objetivamente sobre parámetros de poblaciones finitas, como son todas las poblaciones humanas. Recurrir a lo que no es con la intención de afirmar cosas sobre lo que es, no es un camino correcto. Un ejemplo de ese tipo de abuso sería suponer que la población humana es infinita con la intención de concluir cosas sobre una población que sabemos que es finita. No sería ético. También hacemos ver que la ciencia debe basarse en hechos para concluir su tesis, pues construir una ciencia basada en suposiciones no comprobadas o no comprobables es una tarea sin fundamento práctico cuando lo que se desea con ella es concluir algo fiable de realidades, no de hipótesis supuestas. Para evitarlo la ciencia estadística ha desarrollado instrumentos cada vez más adecuados y objetivos para estimar parámetros poblacionales y contrastar hipótesis estadísticas como puede verse en el libro del autor (2013c) y en los primeros capítulos de este libro.

Básicamente decimos que no es posible hacer inferencias objetivas si no se reúnen estos requisitos: (a) Selección probabilística de las unidades u observaciones en la muestra; y para ello, es necesario que las unidades sean finitas, identificadas, y accesibles para obtener su dato verdadero. (b) El diseño muestral anterior debe completarse con el método de estimación insesgada del parámetro de referencia, y el método de estimación insesgada de la varianza del estimador anterior del parámetro de referencia, siempre que sea posible completar esto último.

Naturalmente lo deseable es mantener en salud a las personas desde su concepción hasta que esto sea posible. Para ello trasmitir la experiencia reflexiva de los padres adultos con bondad a los hijos menores desde pequeños es insustituible y una garantía de una

buena educación, así como la trasmisión de la fe, la sabiduría, unas virtudes y unos valores que otras instancias superiores no deberían imponerles sino facilitarles el libre ejercicio de su conciencia y su voluntad. Por tanto, no consideramos que el recurso al tratamiento farmacológico como primera instancia sea lo más adecuado sino que creemos que dar unas condiciones de educación por los padres asesorados por sacerdotes u otros profesionales para los hijos puede ser un medio pacífico y más efectivo para prevenir enfermedades e infecciones.

En este camino la bioestadística puede estudiar también modos de vida saludables y la conveniencia de ciertos hábitos buenos como puede ser hacer ejercicio físico. Una publicación que estudia estos aspectos y otros más profundos con base estadística es *Journal of Marriage and Family*, entre otras revistas sociológicas. La prevención siguiendo pautas de vida saludables es algo que debe conocerse y practicarse, también desde una perspectiva religiosa y/o sociológica, que podría ser corroborada por la estadística.

A veces se presenta la enfermedad y entonces es necesario recurrir al médico quien dispondrá de conocimientos y el apoyo de la estadística para confirmar los efectos beneficiosos de medicamentos como una solución no primaria pero sí al alcance ante una persistencia de la enfermedad.

La bioestadística, como ciencia experimental, no puede prescindir de la experimentación y la recogida de datos, pues éstas constituyen la frontera que diferencia a las ciencias empíricas de las que no lo son (Sgreccia, 2012). Un ejemplo es el caso de la penicilina que como antibiótico hace más de un siglo permitía curar diversas enfermedades como la pulmonía, pero que, con los años, la resistencia de los agentes causantes de algunas de esas

enfermedades, hacen de la penicilina ineficaz en ciertos casos infecciosos actuales. Por esto, la experimentación debe seguir buscando las nuevas causas de las enfermedades actuales y darles tratamientos contrastados para su curación según estudios recientes.

Llegado a este punto, vemos importante indicar que la estadística describe datos reales, o bien, con ellos trata de inferir objetivamente sobre los parámetros poblacionales. Es por tanto muy aventurado querer extrapolar el valor de los datos para predecir el futuro o para inferir sobre una población que se ha supuesto como posible generadora de los datos, pues en estos casos la estadística dejaría de ser un medio objetivo para convertirse en un medio subjetivo de análisis expuesto a más errores. En ese caso los métodos subjetivos añaden a las conclusiones errores debidos a las hipótesis añadidas con las que se razona para concluir unas estimaciones o un contraste sobre otra hipótesis estadística. Este es el caso de las inferencias clásica, bayesiana, y otras más, en cuyos fundamentos son necesarios suposiciones sobre la población y a veces sobre los parámetros.

Si queremos conocer el efecto benéfico del medicamento éste ha de ser probado en una muestra de la “población finita” compuesta por todos los enfermos que la padecen de esa población en un instante o en un periodo determinado. Daremos repaso a algunas de las normas que regulan estos estudios de experimentación con pacientes con la intención de aportar una visión científica objetiva que respetando los principios generales de autonomía de los pacientes, de modo que ayuden a curar a todos ellos por el uso de métodos estadísticos investigados en los últimos años, y estos nos informen correcta y adecuadamente de las realidades investigadas en los pacientes.

Además de posibles pacientes colaboradores con estas investigaciones, es bueno concienciar a otros (aunque sean dos o unos pocos) para que su negativa a colaborar inicial se transforme en colaboración efectiva, basada en la beneficencia y posibles recompensas personales, que permita concluir resultados objetivos estadísticamente de "todos los pacientes", que son a los que van dirigidos los esfuerzos curativos. De otro modo, limitándonos a los pacientes que voluntariamente quieran colaborar, los estimadores serían sesgados y no tendríamos medidas estimadas sin sesgo del error de muestreo que conllevarían, lo que limitaría el éxito conclusivo del estudio experimental. Así buscamos conseguir el bien de todas las personas y distinguir lo que es ciencia objetiva de lo que pueda ser pseudocientífico, o no complete todo el recorrido para garantizar su objetividad.

En concreto, no podría garantizarse que una muestra sea aleatoria simple solo por disponer un número de datos de la población. La población debe estar identificada por sus unidades (personas) y éstas deben ser accesibles para los observadores del estudio, en concreto a las personas seleccionadas en la muestra aleatoria según rigurosos métodos probabilísticos de obtención de la muestra. La muestra seleccionada es de identificadores, y por la accesibilidad de las unidades de la muestra observamos a los pacientes anónimos con dichos identificadores seleccionados.

Las inferencias clásica y bayesiana, entre otras basadas en supuestas poblaciones infinitas, no hacen uso de un procedimiento cuidadoso de selección de la muestra probabilística representativa en las poblaciones de personas, por lo que no podrán concluirse resultados objetivos con estos tipos de inferencia. Pueden suponer que la muestra ha sido seleccionada según un tipo de muestreo concreto, pero no garantizarlo en la práctica al no estar identificadas sus unidades. Suponer que una población es infinita

cuando en realidad es finita es otro error de entrada y de planteamiento que conllevaría posibles errores en las conclusiones consecuencia de racionalizar una o varias falsedades sin prestar atención al aspecto ético en la ciencia en todas sus fases.

Al tomar decisiones no solo elijo qué cosas quiero hacer, sino también qué clase de persona quiero ser. Ser mejor persona es superior éticamente a tener más o hacer más. Ser honesto en la ciencia es superior éticamente a publicar más o con más factor de impacto. Lo ideal sería que ambas cosas estuvieran relacionadas causa-efecto pero esto no es más que un deseo.

La moral no puede ignorar o menospreciar las conclusiones científicas, y el científico debe tener en cuenta y practicar en su investigación las exigencias éticas, no siendo aceptables aquellos métodos e investigaciones que no tengan en cuenta la dignidad de las personas y la verdad. Moral y ciencia se complementan como la fe y la razón, y se condicionan mutuamente en el camino hacia el bien y la verdad (Trevijano Etcheverria, 2011).

En la lucha contra la enfermedad desde que la medicina es ciencia, el camino necesario para progresar y conseguir nuevas metas no es otro que la investigación y la experimentación, llevadas a cabo científicamente y no solo de forma observacional o empírica, de nuevos modos de intervención farmacológica o tecnológica como diagnóstico y terapia (Ciccone, 2006). La investigación biomédica en sujetos humanos constituye la fase final de un camino de investigación científica que comenzando en los laboratorios, sigue en los animales, para terminar en el hombre. Este es un momento importante y lleno de problemas éticos de la investigación y la experimentación consistentes en sucesivos intentos para comprobar si (y en qué medida) la nueva intervención médica que se está contrastando produce los efectos buscados en la investigación.

Por investigación se entiende cualquier actividad que se proponga adquirir verdad o nuevos conocimientos. Es científica cuando se lleva a cabo según la metodología de las ciencias modernas. Se llama biomédica a la investigación desarrollada en el ámbito de la salud y de la enfermedad, en el campo de las ciencias biológicas y tiene como fin el conocimiento de nuevas modalidades terapéuticas.

La experimentación clínica de los fármacos o terapias nuevos viene obligada también éticamente. Las hipótesis fundamentadas sobre los efectos beneficiosos que un nuevo fármaco o terapia prometen tener sobre un organismo humano tienen su base a veces en la experimentación en laboratorio y con animales. Pero esas hipótesis deben ser contrastadas, utilizando el fármaco o la terapia en seres humanos. Algunos riesgos son inevitables, ya que hay características biológicas individuales que tienen diversas reacciones en unos organismos a otros con el mismo tratamiento.

De acuerdo con Ciccone (2006) y Sgreccia (2012) los estudios clínicos generalmente se clasifican en cuatro fases. La segunda y la tercera fases se prueban en pacientes con la enfermedad antes de que el fármaco se pudiera comercializar. En todas las fases debe formularse el objetivo u objetivos, cuál es la pregunta que debe responderse, conocer los trabajos previos antes de comenzar para ver si es aconsejable el estudio y si su diseño es el adecuado, anular o reducir los sesgos, determinar el tamaño de la muestra de pacientes y cómo se han de seleccionar. Debe fijarse el parámetro por el que va a medirse la consecución del objetivo, así como otros parámetros a estimar pero que no definen la finalidad del ensayo clínico.

Como indican la Normas de Buena Práctica Clínica (2.3), los derechos, la seguridad y el bienestar de los sujetos del estudio son

consideraciones más importantes que deben prevalecer sobre intereses de la ciencia y de la sociedad (concretados en los principios de no maleficencia y de beneficencia en la actuación médica). De aquí una de las limitaciones más importantes del bioestadístico, quien por un lado debe proporcionar información o conocimiento inferencial objetivos y fiables, pero con el límite del consentimiento informado de los sujetos de investigación, quienes pueden salir de la experimentación en cualquier momento.

El respeto de la persona y la investigación científica son objeto de los puntos 2292 al 2296 del Catecismo de la Iglesia Católica (CIC).

Entre las normativas en materia de experimentación hemos seleccionado varias de ellas que orientarán nuestro trabajo en los temas que afectan conjuntamente a la ética y a la bioestadística, que desarrollaremos en los siguientes contenidos.

En los últimos años hemos hecho avances en las ciencias estadísticas, y vemos oportuno destacar los aspectos más relevantes y actuales de la estadística y en su utilización con el fin de mejorar la vida o aliviar los males, especialmente del ser humano. La Bioestadística es la ciencia estadística aplicada a la vida. Así los avances en la objetividad de la estadística tienen consecuencias en el conocimiento de los instrumentos bioestadísticos que dan luz sobre cuestiones como la prevención de la enfermedad, la enfermedad misma, los medicamentos que pueden curarla o tratarla, etc. sin perder de vista que la persona a la que se destinan estos estudios y conocimientos mejorados en definitiva es la persona humana que debemos considerar como un fin en sí mismo y, por su dignidad, darle el trato humano y respetuoso que le corresponde en todo momento.

El objetivo general de esta sección es presentar resumidamente los aspectos éticos relevantes en relación con la

salud y la experimentación en seres humanos, y el objetivo específico consiste en destacar las aportaciones recientes en el área de la Bioestadística para el bien primordial de la salud humana y el conocimiento científico que puede obtenerse en los estudios saludables descriptivos o inferenciales de hechos y datos y/o por observación experimental de pacientes y personas sanas.

El fin de la Bioestadística es aportar instrumentos científicos objetivos en la medida de lo posible como medios para resumir o inferir el conocimiento y la información relevante de experimentos observacionales, especialmente en seres humanos, y para concluir consecuencias en ellos.

Nuestro objetivo es describir los avances recientes en este área de la Bioestadística y tratar de compaginar un conocimiento más objetivo y veraz con el objetivo prioritario de respetar a las personas humanas y tratar de proporcionarles los medios mejores como consejos saludables para una vida sana, así como de la búsqueda, con ciencia objetiva, del mejor tratamiento posible de las enfermedades y dolencias cuando estas aparecen.

En este caso la Bioestadística es un medio, que debe ser bueno, es decir, ético, objetivo y eficiente basado en datos y en lo posible nunca en hipótesis supuestas y no comprobables sino sólo asumiendo condiciones de trabajo que sean hechos en la práctica real y concreta, para el tratamiento de la información estadística que proporciona un estudio o un ensayo ya sea de terapias saludables o curativas en seres humanos.

Como material de Bioestadística vamos a considerar la inferencia en poblaciones finitas, ya que cualquier población humana es finita en un instante dado, por ejemplo la población de pacientes afectados por determinada enfermedad. De este modo reconocemos la verdad de la realidad de pacientes en el modelo

estadístico con el que estudiarla. Por otro lado nos interesa estudiar hechos reales. Esto nos hace descartar como modelos todos aquellos que necesitan "suponer cómo es la realidad" en lugar de "reconocer su realidad" sobre el terreno. De este modo, podemos prescindir de modelos de inferencia clásica y bayesiana, y de la práctica totalidad de los modelos superpoblacionales, pues requieren sustituir realidades reconocibles por hipótesis teóricas no comprobadas ni comprobables.

Por todo ello, nos centramos en el modelo de muestreo y estimación en poblaciones finitas (pues son estas las que nos interesan en la práctica real y comprobada), con datos reales y objetivos (que pueden medirse sin error en cada unidad de la población finita y con ningún daño posible a las personas observadas), seleccionados por muestras de acuerdo a diseños o esquemas de muestreo que junto a un estimador asociado permiten obtener conclusiones inferenciales objetivas (Ruiz Espejo, 2013c).

Otros métodos estadísticos, como los explicados en los libros de Berger y Wong (2009), Good y Hardin (2006), Indrayan (2013), Kupper, Neelon y O'Brien (2011), Lejeune (2010), Olive (2014), Piantadosi (2005), van Belle y Kerr (2012), y de Winkel y Zhang (2007), son de inferencia clásica u otros métodos que suponen hipótesis subjetivas en sus modelos de análisis de los datos. Libros que han sido revisados por el autor, algunas de cuyas referencias están recogidas al final del libro. Mejoras en la objetividad de los métodos estadísticos de experimentación e inferencia son los trabajos de Ruiz Espejo y Delgado Pineda (2008) y de Ruiz Espejo (2013c).

## Declaración de Helsinki

La Declaración de Helsinki (DH) de la Asociación Médica Mundial (AMM) es la enumeración de los Principios éticos para las investigaciones médicas en seres humanos. Son 35 Principios de entre los que comentamos los que consideramos de mayor interés conjunto ético y bioestadístico.

En el Principio 12 se dice que la investigación médica en seres humanos debe conformarse con los principios científicos generalmente aceptados y debe apoyarse en un profundo conocimiento de la bibliografía científica, en otras fuentes de información pertinentes, así como en experimentos de laboratorio correctamente realizados y en animales, cuando sea oportuno.

En el Principio 16 se dice que la investigación médica en seres humanos debe ser llevada a cabo sólo por personas con la formación y calificaciones científicas apropiadas, que la investigación en pacientes o voluntarios sanos necesita la supervisión de un médico u otro profesional de la salud competente y calificado apropiadamente, y que la responsabilidad de la protección de las personas que toman parte en la investigación debe recaer siempre en un médico u otro profesional de la salud.

En el Principio 22 se indica que la participación de personas competentes en la investigación médica debe ser voluntaria, y que ninguna persona competente debe ser incluida en el estudio a menos que ella acepte libremente.

En el Principio 23 se dice que deben tomarse toda clase de precauciones para resguardar la intimidad de la persona que participa en una investigación y la confidencialidad de su información personal.

En el Principio 24 se dice que en la investigación médica en seres humanos competentes, cada individuo potencial debe recibir información adecuada acerca de los objetivos, métodos, fuentes de financiación, etc. de la investigación. La persona potencial debe ser informada del derecho de participar o no en la investigación y de retirar su consentimiento en cualquier momento, sin exponerse a represalias. Después que la información ha sido comprendida por el individuo, el médico u otra persona calificada apropiadamente debe pedir entonces, preferiblemente por escrito, el consentimiento informado y voluntario de la persona. Y si el consentimiento no se puede otorgar por escrito, el proceso para lograrlo debe ser documentado y atestiguado formalmente.

Por todo lo anterior, la Declaración de Helsinki (DH) se ocupa de los aspectos más importantes en relación con el consentimiento informado de los individuos que participen en una investigación, o del de su representante legal cuando corresponde recurrir a él. La posibilidad de que un individuo pueda retirarse de la investigación en cualquier momento, hace que los métodos inferenciales que puedan usarse para determinar la eficacia de los tratamientos experimentales puedan carecer de base objetiva sobre la que hacer las conclusiones al poder producirse la eventual no respuesta durante la investigación. Sin embargo, aparentemente en la inferencia clásica o bayesiana, en las que cuando se suponen poblaciones infinitas no se cuida la representatividad probabilística de la muestra a través del diseño muestral, puede parecer que sí pueden extraerse conclusiones pues todo se reduciría a obtener una muestra de un tamaño determinado sin comprobar su representatividad en la práctica. Este proceso requiere de otras hipótesis que hacen subjetiva y más alejada de la realidad las posibles conclusiones para la población investigada a partir de una muestra de ella, que ya no sería selección probabilística en la práctica, aunque sí lo pueda ser en su supuesto análisis estadístico

teórico. Lo que no garantiza objetividad en las conclusiones al perder todo rastro de control objetivo en la selección de la muestra en el estudio.

## Deontología médica

En este apartado vamos a comentar el Código de Deontología Médica (Guía de Ética Médica), del Consejo General de Colegios Oficiales de Médicos (2011), en los aspectos en los que la Bioestadística ha podido avanzar en sus conocimientos éticos.

El primer deber en la conciencia moral de cualquiera es formar una buena conciencia, es decir, estudiar, buscar la verdad, consultar con las personas prudentes para salir de dudas, perseverar, etc. Para actuar bien, en el sentido de deber moral, ha de ser en todos sus aspectos, sustancia y circunstancia. Si falla uno de ellos se pervierte su bondad. Las reglas del buen hacer en las acciones conforme a los imperativos de la razón, constituyen los deberes profesionales. Toda profesión honrada tiene la índole de servicio a Dios y a los demás. Ningún mandato moral preceptúa lo que hay que hacer para obtener tal o cual fin o bien, sino algo de debido cumplimiento. La ética cuenta, como referentes normativos, con la naturaleza (metafísica) y la razón.

El Código de Deontología Médica (CDM) a lo largo de un preámbulo, 21 capítulos, una disposición adicional y disposiciones finales, describe las normas cuyo incumplimiento supone incurrir en falta disciplinaria.

En el capítulo tercero del CDM se exponen las Relaciones del Médico con los Pacientes. En su Artículo 19.2 se dice que la historia clínica de un paciente para su análisis científico, estadístico, y con fines docentes y de investigación se respetará

rigurosamente la confidencialidad de los pacientes. Esto tiene la consecuencia de que cada paciente debería estar identificado por un número o clave que permita acceder a los datos estadísticos del paciente, pero respetando la confidencialidad de los mismos, es decir, guardando el anonimato del nombre y su persona que lo identificarían, para los efectos de la explotación estadística de sus datos. Esto puede hacerse mediante un carnet con el nombre y la clave o número del paciente que puede leer el médico en consulta, pero el médico investigador o el profesional estadístico solo accederían a la clave o número identificador y a los datos estadísticos del paciente ya de modo anónimo. Además esta identificación numérica es necesaria e imprescindible para poder seleccionar muestras representativas de acuerdo con un diseño muestral probabilístico para efectuar inferencias objetivas. Esto es una de las hipótesis de trabajo del libro del autor (2013b), que presenta los métodos inferenciales objetivos para poblaciones finitas.

En el capítulo cuarto del CDM se detalla la Calidad de la Atención Médica. En el Artículo 26.2 se dice explícitamente que no son éticas las prácticas carentes de base científica y que prometen a los enfermos la curación, los procedimientos ilusorios o insuficientemente probados que se proponen como eficaces, etc. A este respecto hay que indicar que base científica puede tener un método estadístico subjetivo, pero este será siempre éticamente inferior a un método estadístico objetivo con base científica y real.

El Artículo 30.1 indica que el secreto profesional debe ser la regla; no obstante se enumeran algunas excepciones. En el Artículo 30.1.g se indica que el médico deberá mantener el secreto aunque el paciente lo autorice. No obstante todo esto, entiendo que se guarda secreto profesional aun cuando se obtenga de las historias clínicas información anónima con fines estadísticos objetivos, cuya

finalidad sea mejorar los tratamientos a los enfermos o tendentes a su curación efectiva.

En el 59.4 se dice que el médico investigador tiene el deber de publicar los resultados de su investigación por los cauces normales de divulgación científica, tanto si son favorables como si no lo son. Advierte además que no es ética la manipulación o la ocultación de datos (para obtener beneficios personales o de grupo, o por motivos ideológicos). De este artículo, deducimos que la manipulación de los datos con métodos estadísticos subjetivos, como son la inferencia estadística clásica y la bayesiana entre otras, son menos éticas porque el modelo o los modelos supuestos aportan subjetividad al manipular los datos.

Una consecuencia de la voluntariedad para ser incluido entre los pacientes observables por experimentación (con la cual se proteje a los no voluntarios en su libre voluntad), deducida de los métodos objetivos de inferencia en poblaciones finitas, es que las conclusiones de un estudio estadístico basado en una muestra aleatoria de voluntarios sólo podrían inferirse a funciones paramétricas de la población de pacientes voluntarios. Así, si se respeta a todos los pacientes en su deseo de colaborar o no colaborar con la investigación, lo cual es ético y dignificante, también en honor a la verdad se puede deducir que las conclusiones del estudio estadístico no tendrían alcance sobre todos los pacientes a los que después se desease tratar con los resultados de la investigación. La razón es que puede haber diferencia entre el parámetro de la población de pacientes y el parámetro de la población de los pacientes voluntarios. Lo que podría concluirse para unos no tiene por qué concluirse para los otros. Esto se demuestra matemáticamente y puede resolverse también matemáticamente usando métodos objetivos de estimación

insesgada basada en hechos en el caso de que aparezca no respuesta (como ocurre en los no voluntarios).

Algunas referencias que presentan soluciones a este problema, que han sido estadísticamente investigadas recientemente, son las de Ruiz Espejo (2011a, 2015b) y de Thompson (2012). Con estos estudios adicionales se concluye que con la participación adicional de dos nuevos voluntarios elegidos con un diseño de muestreo aleatorio simple de entre los no voluntarios de la muestra de pacientes en un primer intento, pueden concluirse inferencias objetivas sobre toda la población de pacientes. Esto requiere convencer al menos a dos (o más) personas seleccionadas en una submuestra de entre las que no estaban dispuestas a participar en la investigación médica, por ejemplo proporcionándoles un seguro médico vitalicio, una pensión o una retribución económica por su participación acordes con las consecuencias causadas por su participación en la investigación. En cualquier caso, si se obtuviera la colaboración de un número mayor a dos entre los no voluntarios en un primer ofrecimiento seleccionados por el diseño muestral, los resultados de la investigación serían insesgados, más precisos y objetivos.

De este modo se conseguirían unas conclusiones válidas para todos los pacientes de la población, no solo válidas para los pacientes potencialmente voluntarios en una primera instancia de entre la población de pacientes (y así protegeríamos de hecho a todos los pacientes de la población, pues la investigación inferiría objetivamente sobre una función paramétrica basada en la información proporcionable por todos ellos). Si no se consiguiera la cooperación posterior de los dos o más primeros pacientes seleccionados en una submuestra de entre los que en la muestra inicial se manifestaran no dispuestos a ser objeto de investigación médica (pues de otro modo si no fueran los primeros, la eliminación de algunos seleccionados tiene el efecto de que el

diseño muestral así proporcionado haría sesgado el estimador, ver Ruiz Espejo, 1986b), no podrían extraerse consecuencias objetivas basadas en hechos más que de entre los pacientes potencialmente voluntarios, y por esto no podrían inferirse conclusiones objetivas para toda la población de pacientes. Todo lo que se afirmase sobre esta población de pacientes sería subjetivo, y por tanto menos válido científicamente que el método estadístico objetivo que hemos explicado y referenciado.

Lo que ha de hacerse estadísticamente, por tanto, es aprovechar la investigación estadística reciente sobre no respuesta de los autores citados, con los recursos y la persuasión para obtener la cooperación de pacientes de la muestra inicialmente de no voluntarios, y proceder a una recogida de datos (que puede ser en consulta médica, en encuesta, o bien telemática) y calcular unos estimadores de acuerdo con dichas investigaciones recientes que proporcionan métodos objetivos de inferencia estadística. Pues de otro modo, se seguirían procedimientos subjetivos de recogida de datos, análisis e inferencia que la ciencia verdadera no puede garantizar aunque hubiera en ellos algunos rasgos racionales incompletos o subjetivos.

En el capítulo veinte, Publicidad Médica, se insiste en su Artículo 65.3 que la publicidad médica deber ser objetiva, prudente y veraz, de modo que no levante falsas esperanzas o propague conceptos infundados.

### Buena práctica clínica

En este apartado voy a dar repaso a los aspectos relacionados con la Bioestadística entre las Normas de Buena Práctica Clínica (NBPC de Enero de 1997 y corregida en Julio de 2002).

En la Norma 6.4 se detalla el Diseño del Ensayo, y en la 6.5 la Selección y Retirada de Sujetos. En 6.7 la Valoración de la Eficacia, en donde deben especificarse los parámetros de eficacia (llamados “funciones paramétricas de eficacia” en la inferencia objetiva en poblaciones finitas) y otros aspectos relacionados. En 6.9 se trata de la Estadística. En 6.9.1 se refiere a la descripción de los métodos estadísticos que se usarán, incluyendo el calendario de todos los análisis intermedios, en 6.9.2 se refiere al tamaño muestral, pero en 6.9.3 se habla del nivel de significación que será utilizado, lo que hace implícito una referencia a la inferencia clásica que ha sido tan puesta en cuestión en los últimos años (ver por ejemplo Nuzzo, 2014) pues presupone sin demostración la normalidad de los datos, como si fueran de poblaciones infinitas, aspecto que es mejorado claramente por un enfoque que asegure un nivel de confianza mínimo estimado sin sesgo como se explica por ejemplo en Ruiz Espejo (2013c).

En 8.2.11 se refiere a documentar valores/rangos normales de procedimientos médicos/de laboratorio/técnicos/pruebas. En este punto se supone que el medio es la estadística clásica incidiendo en la distribución normal como medio de análisis supuesto, que no probado ni objetivo. En 8.2.18 habla de documentar el método de aleatorización de la población del ensayo. A este respecto ya hemos explicado que para que la muestra sea probabilísticamente representativa de la población de la que se extrae las unidades (de la muestra de la población) estas deben estar numeradas e identificadas para poder seleccionar una muestra aleatoria probabilística de acuerdo a las especificaciones necesarias del diseño muestral para concluir objetivamente resultados inferenciales.

En 8.3.6 se reincide en documentar los valores y rangos normales que se revisan durante el ensayo. Las mismas críticas son aplicables. En 8.3.14 se habla de documentar la confirmación de

los datos registrados. En 8.3.15 se trata de documentar las correcciones de los cuadernos de recogida de datos. En 8.3.21 habla de documentar que el investigador/institución guarda una lista confidencial de los nombres de todos los sujetos asignados con los números de inclusión al ensayo. Permite así al investigador/institución revelar la identidad de los sujetos e/o identificarlos anónimamente en el estudio. La Norma 8.3.22 exige documentar la inclusión cronológica de los sujetos por el número asignado en el ensayo. En 8.4.3 se exige listar completamente los códigos de identificación de los sujetos incluidos en el ensayo para el caso en que se requiera un seguimiento, guardando la lista de forma confidencial durante el periodo de tiempo acordado. En 8.4.4 se refiere a documentar el Certificado de Auditoría (CA). En 8.4.7 se exige documentar el informe final del investigador al Comité Ético de Investigación Clínica (CEIC) y a la autoridad reguladora.

## Investigación biomédica en seres humanos

Como indica el documento *Pautas Éticas Internacionales para la Investigación Biomédica en Seres Humanos* del Consejo de Organizaciones Internacionales de las Ciencias Médicas (CIOMS) en colaboración con la Organización Mundial de la Salud (OMS), el primer instrumento internacional sobre ética de la investigación médica –el Código de Nüremberg (CN)– fue promulgado en 1947 como consecuencia del juicio a los médicos que habían dirigido experimentos atroces en prisioneros y detenidos sin su consentimiento durante la segunda guerra mundial. Este Código protegía la integridad del sujeto de investigación y establecía condiciones para la conducta ética de la investigación en seres humanos, destacando su consentimiento voluntario para la investigación. La investigación en seres humanos debiera ser

realizada o supervisada sólo por investigadores debidamente cualificados y experimentados.

En la Pauta 1 se da la justificación ética de la investigación biomédica en seres humanos, que radica en la expectativa de descubrir nuevas formas de beneficiar la salud de las personas. Los investigadores y los patrocinadores deben asegurar que los estudios propuestos en seres humanos estén de acuerdo con "principios científicos generalmente aceptados" y se basen en conocimiento adecuado de la literatura científica pertinente. Pienso que el uso de una inferencia estadística objetiva como la que explico en este libro es la más ética entre las otras posibles inferencias subjetivas, que están más difundidas y aceptadas en la práctica pero sin base o fundamento ético que las avale.

La Pauta 2 dice que todas las propuestas para realizar investigaciones en seres humanos deben ser sometidas a uno o más comités de evaluación científica y de evaluación ética para examinar su mérito científico y aceptabilidad ética. En la evaluación se han de considerar las fuentes fiables de conocimiento en relación, no reduciéndolas a consideraciones teóricas, estadísticas o biológicas exclusivamente que podrían tener un efecto limitante y reductor para la curación de la enfermedad, sino que cualquier otra fuente de conocimiento científico o de experiencia sobre la materia a investigar debería tenerse en cuenta para un mejor análisis y perfilar los mejores tratamientos a comparar. Algunos ejemplos de este tipo de consideraciones son los recientes artículos Editorial (2014) en la revista científica *Nature*, y de Reardon (2014).

En la Pauta 3 se regula que los comités, tanto del país del patrocinador como en el país anfitrión, tienen la responsabilidad de realizar una evaluación científica y una ética, estando también facultados para rechazar propuestas de investigación que no

cumplan con sus estándares científicos o éticos. Lo razonable es que si los métodos estadísticos inferenciales son subjetivos, éstos pueden ser éticamente relegados o rechazados respecto a los métodos estadísticos inferenciales objetivos.

Para las investigaciones con placebo los métodos de muestreo de respuesta aleatorizada son los indicados en Chaudhuri (2011), Ruiz Espejo y Singh (2003), y Warner (1965). Estos tienen menor eficiencia que los métodos de muestreo tradicionales de respuesta directa, pero salvaguardan el anonimato de a qué pregunta se responde, es decir, la respuesta recogida como dato no sabe el encuestador o el investigador a qué tratamiento responde de una forma directa, con lo que se preserva la intimidad del encuestado o del voluntario observado. Aunque en el análisis posterior sí pueda extraer conclusiones inferenciales a partir de la respuesta aleatorizada obtenida de los sujetos de investigación.

La Pauta 12 trata de la distribución equitativa de cargas y beneficios en la selección de grupos de sujetos en la investigación, es decir que los grupos o comunidades invitados a participar en una investigación debieran ser seleccionados de tal forma que las cargas y beneficios del estudio se distribuyeran equitativamente, y que debe justificarse la exclusión de grupos o comunidades que pudieran beneficiarse al participar en el estudio. Desde un punto de vista bioestadístico, la exclusión de un grupo o comunidad hace que la población estadística sobre la que inferir se reduzca y que, como consecuencia, los parámetros sobre los que inferir han cambiado. Esto tiene como consecuencia, como ya hemos indicado, que los estimadores insesgados para la población investigada de hecho, sean sesgados para la población objetiva de todos los sujetos pacientes. Es decir, en la práctica se da el efecto de no respuesta y ésta deber ser estadísticamente tratada para concluir resultados objetivos en poblaciones finitas. A este

respecto insistimos en que dejar no representado en la población a investigar parte de los sujetos o pacientes, hace que los estimadores insesgados para las funciones paramétricas de la población observable, o tenida en cuenta para la selección de la muestra, sean sesgados para las mismas funciones paramétricas de la verdadera población finita objetivo de todos los pacientes sobre la que se pretende inferir. La solución a este problema ya ha sido comentada en los casos anteriores en los que nos hemos referido.

La Pauta 19 trata del derecho a tratamiento y compensación de sujetos perjudicados, en el primer caso se indica que tengan derecho a tratamiento médico gratuito de calidad por tal perjuicio, y a un apoyo económico o de otro tipo que pueda compensarlos equitativamente por cualquier menoscabo, discapacidad o minusvalía resultante, así como en caso de muerte.

## 9.5 Conclusiones

La investigación del conocimiento de las realidades sociales concretas es una necesidad imprescindible para la actuación coordinada en busca del efectivo bienestar social de los ciudadanos de una sociedad.

Por ello, los investigadores estadísticos pueden aportar soluciones eficaces en este proceso para la búsqueda de la verdad con métodos objetivos que conecten las realidades sociales con los medios lógicos adecuados para este fin.

La ciencia y la técnica, y en especial la inferencia estadística objetiva, son recursos valiosos cuando son puestos al servicio del hombre, promoviendo su desarrollo integral en beneficio de todos los ciudadanos que responsablemente aportan lo que está a su alcance con honradez.

La ciencia por sí sola no indica el sentido de la existencia y del progreso humano, especialmente cuando ésta no está ordenada al hombre y a sus valores morales como sentido de su finalidad, al mismo tiempo que somos conscientes de las limitaciones de toda aportación científica.

Difícilmente la ciencia y las tecnologías serán capaces de llegar a resolver los problemas de siempre o más acuciantes de la humanidad como son sus limitaciones básicas en la vida, aunque puedan hacerlos más tolerables, compatibles o con una mayor calidad de vida durante más tiempo y para muchos hombres, deseablemente su totalidad.

Los métodos estadísticos objetivos de investigación del bienestar social tienen que jugar un papel muy importante en esta labor. Nuestra aportación más importante en este sentido se concreta en discernir técnicas y metodologías científicas estadísticas más acertadas a estos fines.

La estadística descriptiva objetiva y los censos son los mejores métodos posibles para la realización de estudios sociales basados en hechos reales, desde el punto de vista de recabar información completa sobre las realidades a conocer de una determinada población finita para ser realistas, ya sea humana o de cualquier otro carácter de interés social.

Pero la información exhaustiva puede ser de un coste tan elevado que los estudios por muestreo de dichas poblaciones finitas proporcionen informaciones veraces y eficaces para los fines propuestos a un coste razonable, para el aumento de la calidad de los datos recogidos y de su tratamiento estadístico, y de la limitación del esfuerzo y trabajo necesario para disponer de información fiable suficiente, así como de la reducción del tiempo necesario para elaborar estas informaciones estadísticas con

respecto al que llevarían los censos de observación exhaustiva en toda la población finita.

Si a esto unimos la posibilidad de dar medidas del error en las estimaciones como consecuencia del uso de la aleatoriedad probabilística para la selección de la muestra, podemos concluir que los métodos de muestreo de selección probabilística de poblaciones finitas fijadas, aun no siendo infalibles, pueden ser de gran ayuda por sus buenas aproximaciones de sus estimadores a los parámetros poblacionales con el nivel de confianza mínimo aproximado deseado.

La estadística descriptiva, los censos informatizados y los muestreos con fines sociales y de conocimiento descriptivo de las sociedades tienen un valor relevante a escala local, regional y nacional en muchos países y a escalas mayores. También es un deseable objetivo a alcanzar en el futuro la consecución de informaciones estadísticas del estado del bienestar a nivel planetario.

Los métodos de muestreo de poblaciones finitas fijas son los métodos de inferencia más objetivos porque requieren el menor número de hipótesis sin posible comprobación y, las hipótesis que hace son realistas, de hecho o comprobables (Ruiz Espejo, 2013c). Además conserva una gran flexibilidad por sus presupuestos económicos reducidos y su rapidez en la puesta en práctica de la recogida de datos, a diferencia de los censos.

Otros métodos de inferencia requieren aportaciones subjetivas en los modelos y no identifican las unidades de la población lo que hace imposible garantizar un control de la selección probabilística de acuerdo al diseño de muestreo o incorporan elementos subjetivos en el análisis que puede hacer que pierda validez en sus conclusiones si queremos que éstas sean objetivas.

La inferencia estadística objetiva que hemos descrito se orienta desde un principio a recabar información sobre hechos determinados de las unidades de una población finita con la mayor coherencia. Otros métodos de inferencia pueden contener mayor complejidad lógica pero no están adaptados al fin que buscamos, que es información objetivamente obtenida y veraz.

En estos métodos objetivos cabe la posibilidad de salvaguardar la confidencialidad o privacidad de las respuestas de los encuestados con métodos de respuesta aleatorizada, debida a Warner (1965) y analizados por Ruiz Espejo y Singh (2003). También admiten el uso de información auxiliar objetiva de las unidades de la población finita (Ruiz Espejo, 1998c).

La inferencia estadística objetiva requiere menor número de hipótesis, y éstas son más asumibles con las condiciones de hecho desde el punto de vista de su aplicabilidad. Por ello, los resultados que se deducen son más objetivos.

La existencia del concepto de probabilidad en el mundo real es algo discutido. Pero aunque no exista, su concepto puede simularse informáticamente para poder seleccionar de modo controlado según las características del diseño muestral como si tal probabilidad existiera, deduciéndose las propiedades matemáticas de la inferencia estadística objetiva.

Algo diferente ocurre cuando se supone que la naturaleza de los datos presentados y recogidos para su análisis estadístico son de tipo probabilístico, lo que supone afirmar que la naturaleza se comporta como las leyes de probabilidad de las hipótesis de tales métodos de inferencia estadística. Esto supone que la naturaleza se comportaría como han pensado los matemáticos al avanzar en sus estudios teóricos, algo muy aventurado de reconocer y que escapa a las posibilidades limitadas de los hombres hoy por hoy y de sus

conocimientos seguros. No existe garantía de la existencia del modelo probabilístico generador de los datos en la misma naturaleza, ni tampoco de su conocimiento explícito de tal modelo.

De aquí que algunos autores hayan estudiado métodos estadísticos robustos tratando de suplir esas carencias de los métodos de inferencia basado en modelos subjetivos, cuando estos modelos no tienen por qué ser verdaderos. Pero siguen adoleciendo de que las propiedades de la aleatorización que generan los datos son supuestas y no objetivas o controladas (Ruiz Espejo, 1990).

Debemos reconocer que en la puesta en práctica de la inferencia estadística objetiva pueden presentarse situaciones no previstas en la teoría. Así, si un censo no contiene alguna o algunas unidades de la población, el marco censal como listado de las unidades de la población no es un perfecto punto de partida para seleccionar la muestra. Pero su influencia, si no se presenta un número apreciable de casos de omisión o multiplicidad, por lo general es mínima a efectos prácticos siguiendo los protocolos de registro tradicionales.

Conclusiones finales:

1. La globalización exige una nueva sociedad fundada sobre la ciudadanía global que desemboque en una sociedad a escala mundial, y una de cuyas bases sea la mejora continua del bienestar social.
2. Esta necesaria sociedad del bienestar debe marcar políticas y directrices basadas en hechos ciertos y reales, lo que exige la adopción de métodos estadísticos y de inferencia objetivos y fiables que reflejen e instituyan la realidad social en la que actuar con medidas concretas.
3. Aunque ciertamente definimos y comprendemos el significado del concepto de probabilidad, no hay pruebas de que exista en el mundo natural. En concreto, es prácticamente

imposible reproducir con exactitud las mismas condiciones de experimentación para observar datos de un mismo fenómeno. Esto es especialmente cierto en fenómenos de tipo social y del bienestar alcanzado por una población humana finita o de grupos humanos (en un instante dado) que están en constante evolución y desarrollo.

4. Si la probabilidad no existiera, solo los datos censales y la estadística descriptiva permitirían conocer los hechos observados de una realidad social.
5. Aunque la probabilidad no existiera, es posible reproducir con ordenador las mismas características del concepto matemático de la probabilidad, en cuanto a la generación de muestras aleatorias en condiciones predeterminadas probabilísticas para la selección de unidades de la población finita. Haciendo uso de las propiedades de un estimador, podemos inferir sobre parámetros o funciones paramétricas poblacionales de modo objetivo.
6. Todos los demás métodos estadísticos de inferencia desarrollados en la actualidad, suponen no solo la existencia de la probabilidad en la realidad de los hechos naturales (algo no demostrado), sino que los datos observados responden a algún modelo subjetivo de probabilidad que el investigador supone que es el que explica dichos hechos, pero sin capacidad de aportar una demostración que compruebe esto.
7. Los métodos estadísticos de inferencia restantes basados en el concepto de probabilidad como inherente a los hechos naturales no dejan de ser meras aportaciones teóricas e imaginarias en cuanto a su aplicación a datos naturales. Entre estas incluimos a la inferencia paramétrica clásica, la bayesiana, la no paramétrica, la de distribución libre, la basada en modelos de superpoblación, etc. de los que no

negamos su aportación lógica pero sí dudamos seriamente de su aportación objetiva en la práctica.

8. En muchos casos las hipótesis poblacionales contradicen los hechos conocidos; por ejemplo, suponer que el tamaño de la población es infinito cuando sabemos que es finito, o bien que las respuestas a una pregunta pueden ser varias cuando solo una es la respuesta verdadera. De hecho, de una mentira como base de un argumento poca verdad se puede deducir.
9. El conocimiento estadístico de las realidades sociales requiere el uso de métodos objetivos y de medios adecuados a sus fines respetando la moralidad y la ética en todo el proceso.

Por todo ello, sugiero la necesidad y conveniencia de incluir en la docencia e investigación universitaria métodos de inferencia en poblaciones finitas fijadas, y el esfuerzo por aplicar estos métodos en los organismos internacionales como la ONU, la OCDE, y en la Unión Europea y España.

Consideramos también necesario no abandonar el estudio de la estadística descriptiva y de la inferencia estadística objetiva en los estudios universitarios en todos los niveles de educación.

La alternativa sería equiparar y dar la misma dignidad a la "ciencia verdadera" que a la "pseudociencia falsa" o posiblemente falsa, ésta última basada en sofismas, premisas inciertas o falsas, argumentos inválidos aunque estén revestidos de apariencias engañosas como son las falacias, etc.

Este libro ofrece razonamientos válidos que refutan tesis y muchos razonamientos inválidos en la ciencia práctica, y deja al descubierto las falacias presentes en algunos argumentos, así como las muchas premisas inciertas o falsas en que se basan la mayor parte de las teorías inferenciales estadísticas. Entre ellas están los "sofismas *a priori*" porque el defecto está al comienzo, es decir,

antes de empezar a razonar; y también los "sofismas de *prejuicio*" que parten de la aseveración de algo que se da por cierto sin que esté comprobado ni demostrado. Son los casos de la inferencia bayesiana y de la inferencia clásica, paramétrica, no paramétrica, semiparamétrica, superpoblacional, etc.

Dado que las premisas son el fundamento de la conclusión, el sofisma de falsa premisa ha recibido asimismo el nombre de "error fundamental". Algunos autores lo han denominado modernamente como "sofisma de simple inspección", porque para impugnarlos no es necesario revisar los argumentos, los razonamientos o las inferencias, sino que bastaría observar las premisas y detectar la falsedad de una de ellas; pero dicha denominación es inadecuada porque muchas veces la falsedad de la premisa no puede descubrirse mediante la mera inspección.

A veces se pretende demostrar una conclusión en base a la opinión de una o varias personas cualificadas sobre el asunto que se discute. Se denomina "autoridad" a una persona o conjunto de personas cualificadas para el conocimiento acerca de algo. Un tipo de sofisma consiste en tomar una "proposición como verdadera en sí misma", prescindiendo de toda prueba, por el solo hecho de que fue afirmada por una "autoridad". El "argumento de autoridad" es legítimo para apoyar conclusiones probables, pero es una falacia cuando se pretende que sea suficiente para obtener una conclusión rigurosamente demostrada. Así, si sabemos que una proposición fue sostenida por Aristóteles, San Agustín, Santo Tomás de Aquino, Leibnitz, o cualquiera de los grandes pensadores, podemos considerarla como *probablemente verdadera*, pero para tener la certeza de su verdad necesitamos una demostración suficiente. La cualificación de una autoridad en unas competencias concretas, no garantizan que se sea competente en cualquier otro tema en el que pueda hacer afirmaciones al margen de su especialidad de

conocimiento. Suele suceder que la autoridad científica ganada por una persona en determinada disciplina, se traslade ilegítimamente a otros ámbitos del conocimiento en los que no ha acreditado conocimientos reconocidos.

Creo que de nada sirve decir que la población es de un modo determinado supuesto, si en la práctica no tenemos un listado de las unidades de la población o no pudiéramos acceder a observar el dato en la unidad seleccionada en la muestra probabilística. Porque suponer que la naturaleza es aleatoria y que trabaja incondicionalmente para nosotros proporcionándonos la muestra directamente sin hacer nosotros nada por ello sino solo suponer unas hipótesis de cómo es seleccionada, resulta ingenuo confiar que nuestras hipótesis son respetadas por la naturaleza sin que nosotros ni nadie pueda comprobar que es así en la práctica de un estudio inferencial con datos del mundo real. Este punto debe hacer reflexionar a los teóricos de la inferencia estadística y consecuentemente descarten "modelos de supuestas propiedades", que nadie puede comprobar que éstas se cumplen en la práctica y en muchos casos son impracticables por las exigencias de las hipótesis supuestas, pero que son necesarias que se cumplan teórica y prácticamente para poder afirmar algo con dichos modelos supuestos y dichos datos seleccionados de la realidad física u observacional de donde se obtienen de acuerdo o no a dichas reglas lógicas.

Todos estos razonamientos tienen su sentido dentro de un marco legal que dispone la colaboración de las personas o ciudadanos, de los hogares, de las empresas y sociedades, etc. con el estado de una nación y con los estudios por muestreo basados en sus censos o bases de datos. En este sentido se elaboran leyes que obligan a los ciudadanos o a otros grupos sociales a colaborar en la elaboración de los censos o en encuestas oficiales. A estas encuestas van dirigidas las reflexiones de este libro.

Existen otro tipo de encuestas en las que los participantes colaboran voluntariamente tras ser informados del interés de la misma, por lo que su selección en la muestra no es aleatoria sino por adhesión particular a los fines e interés del estudio. A estos estudios los llamamos “muestreos de participación voluntaria”. Tienen su interés por ejemplo en estudios sociológicos, psicológicos, etc. En ellos el encuestado es el que elige participar o no en el estudio y por tanto la inferencia estadística tiene poco que aportar en este caso, pero sí la estadística descriptiva.

En otros casos el estudio por muestreo es por selección opinática. En estos la muestra es seleccionada por el investigador que realiza el estudio, según criterios que tienen unos argumentos de representatividad o de tipo técnico que lo hacen viable.

Estos tipos de estudio por muestreo no siempre son aleatorios de modo probabilístico, por lo que en estos casos tanto en los estudios de participación voluntaria de los encuestados, en los de selección opinática del investigador, como en aquellos en que se produzcan ambos hechos de “voluntariedad de encuestados” y de “selección por opinión del investigador” no entran dentro de los estudios de inferencia estadística objetiva al que nos hemos referido en este libro salvo que la selección de la muestra pueda realizarse por métodos de muestreo probabilístico sobre un marco o censo de todas las unidades identificadas, aunque se respete su anonimato, de la población finita sobre la que se desea realizar la inferencia estadística.

Como primer resultado, hemos de advertir que el uso de la inferencia clásica y de la inferencia bayesiana, y otras basadas en modelos superpoblacionales entre otros métodos estadísticos, no son una vía segura ni objetiva para realizar inferencias basadas en

hechos y fundamentos ciertos, pues a la variabilidad propia de la inferencia hay que añadir otras dos fuentes de error en sus análisis:

1. Tener que suponer como cierto lo que no es o puede no ser la población (por ejemplo, asumir como hipótesis la normalidad de los datos observados en la experimentación, que es un reduccionismo). Una hipótesis supuesta no puede considerarse como un hecho si aquélla no se demuestra y no se prueba su veracidad.
2. Descontrol en la selección de la muestra aleatoria, ya que al no identificar numeradamente a las unidades de la población, la accesibilidad a las mismas queda trastocada y/o parcialmente obstaculizada, favoreciendo la mayor representatividad de unas unidades sobre otras o de unas observaciones sobre otras pero sin tenerlo en consideración en el análisis estadístico clásico, bayesiano, no paramétrico, o superpoblacional de esos datos. Las hipótesis del diseño muestral deben demostrarse en la práctica, no solo suponerlas y tratarlas como hechos que no son. No se puede confundir una "hipótesis supuesta" con un "hecho verdadero".

Por otro lado, sí es objetivo partir de la base cierta de que la población humana en estudio en un instante dado es finita, y su tamaño poblacional $N$ (el número de personas a las que se dirige el estudio) puede ser conocido de antemano (Ruiz Espejo, 2013c). También es un elemento de análisis objetivo que las personas o sujetos que forman parte de la población finita estén identificados y tengamos el medio de acceder a ellos, ya sea por medio de la consulta médica o cuando su colaboración sea requerida tras manifestar su consentimiento informado a participar en el estudio experimental. Ambos aspectos son considerados en el estudio de la inferencia en poblaciones finitas con datos fijos observados, pues en casos de discrepancias en las observaciones estas son

consecuencia de fallos en la medición del dato o en respuestas defectuosas, maliciosas, o engañosas en el sujeto.

Al realizar inferencias objetivas en las que se presenta una muestra de sujetos representativa de una población de interés (por ejemplo, todos los pacientes de una enfermedad en el mundo en cierta fecha concreta), si queremos observar a todos los seleccionados en una muestra, por el consentimiento informado la muestra inicialmente se reducirá a voluntarios de entre los seleccionados. Pero para que el estudio tenga validez objetiva es necesario obtener una muestra de entre los no voluntarios en el primer intento de búsqueda de consentimiento informado. Así, obteniendo la colaboración de una submuestra de entre los no voluntarios en un primer intento, pero que sean voluntarios en un segundo intento, podemos concluir el estudio con estimaciones insesgadas de la función paramétrica de eficacia "media poblacional" de la variable de interés, y además tenemos la estimación insesgada de la varianza de la anterior estimación insesgada, lo que nos permitirá obtener intervalos de confianza estimados de la media poblacional, y como consecuencia contrastar hipótesis sobre dicha media poblacional. Los trabajos iniciales en los que se presentan estas posibles mejoras son los de Ruiz Espejo (2011a, 2013d, 2015b) y de Thompson (2012).

De este modo se obtendrían conclusiones válidas para toda la población de pacientes o sujetos, no limitándonos exclusivamente al estrato o subpoblación de voluntarios en el primer intento (lo que conllevaría sesgos de estimación, y la imposibilidad práctica de obtener conclusiones objetivas sobre toda la población de los posibles pacientes), que es como hasta ahora se había regulado su participación en los estudios experimentales. Del modo que hemos sugerido, se evitaría la hipótesis subyacente en la inferencia clásica, bayesiana o superpoblacional (y otras muchas consideradas

teóricamente, como la inferencia no paramétrica, semiparamétrica, etc.) de que la media obtenida en el estrato de voluntarios es la misma que en el estrato de no voluntarios, hipótesis que reduce las conclusiones del estudio a los voluntarios. Y, en cualquier caso, debería de demostrarse que tal hipótesis de igualdad de medias en ambos estratos fuera cierta para que los estudios sean algo más objetivos con dichos tipos de inferencia, cosa que no se hace con la normativa actual, y así las conclusiones de la investigación no se podrían afirmar de toda la población compuesta por los voluntarios y los no voluntarios. Así pues, serían estudios teóricos bajo supuestas hipótesis no comprobadas.

La confidencialidad de los sujetos que participan en la investigación puede conseguirse identificando al sujeto que participa con un número o clave que conste en su historia clínica, y ésta esté desprovista de cualquier otra identificación que revelara la persona concreta de la que se trata. Esta es una hipótesis de trabajo para la inferencia en poblaciones finitas, como se presenta en Ruiz Espejo (2013c). Solo el médico tiene acceso a aportar los datos de su paciente, a quien conoce personalmente en consulta o en seguimiento, pero su historia médica sería confidencial y anónima salvo en el identificador del paciente que sería una clave de acceso a su historia clínica y a sus datos observados e información auxiliar.

Como conclusión, el conocimiento científico necesario para descubrir nuevos medicamentos o terapias requeriría, desde un punto de vista bioestadístico, la colaboración de sujetos voluntarios para el estudio experimental y la colaboración de otros sujetos no voluntarios inicialmente de entre los que fueron seleccionados aleatoria y probabilísticamente entre todos los posibles sujetos de la muestra inicial, y que no se mostraron dispuestos a colaborar para el fin de conocer los efectos de un nuevo tratamiento terapéutico o médico para cierta enfermedad concreta o para la

prevención de la misma. Pero que después acceden a colaborar como voluntarios en condiciones más ventajosas en una submuestra de la muestra de pacientes no voluntarios en el primer intento de obtener su consentimiento informado.

Se presenta por tanto un dilema: desde un punto de vista respetuoso de la dignidad de los sujetos humanos, es necesario el consentimiento informado de éstos; pero en el caso de que hubiera no consentimiento por parte de los sujetos, no podrían garantizarse estadísticamente las propiedades del nuevo tratamiento para la totalidad de los pacientes salvo que un número de éstos (de tamaño muestral mayor o igual a dos) seleccionados aleatoria y probabilísticamente colaboren en el estudio en un segundo intento de obtener su consentimiento informado para participar en el estudio experimental.

Una alternativa es suponer que los sujetos voluntarios y los no voluntarios son equivalentes o que sus parámetros que miden la eficacia del tratamiento no varían de uno a otro estrato o dominio, algo que en teoría general sería falso. Pero para garantizar estadísticamente la efectividad del tratamiento esta suposición debe ser demostrada y no solo supuesta. Demostrarlo sería más complicado (en realidad requeriría realizar un censo de todos los sujetos y todos deberían ser voluntarios, lo que sería una contradicción con que hay algunos que no dan su consentimiento) que trabajar con la hipótesis liberadora de que ambos estratos podrían tener diferentes medias y varianzas, y entonces para concluir resultados para toda la población finita de sujetos bastaría seleccionar una muestra aleatoria y probabilística de cada uno de los estratos (el de los voluntarios y el de los no voluntarios) independientemente en cada uno de ellos. De este modo aunque las funciones paramétricas (medias poblacionales de cada dominio) indicadoras de la eficacia del tratamiento fueran iguales o no lo

fueran en todos los casos, los métodos estadísticos que hemos sugerido para la no respuesta permiten extraer conclusiones para toda la población de sujetos, no solo para los voluntarios como ocurriría en el caso general en que no considerásemos la no respuesta o no consentimiento iniciales como conformadores del estudio experimental.

En efecto, llamando a la media muestral de voluntarios $\bar{y}_{1,\nu}$, donde $\nu$ es el número de voluntarios a partir de la muestra irrestricta aleatoria de tamaño $n$ obtenida de toda la población finita de tamaño $N$, su esperanza matemática sería

$$E\left(\bar{y}_{1,\nu}\right) = E\left[E\left(\bar{y}_{1,\nu}|\nu\right)\right] = \sum_{\nu=0}^{n} E\left(\bar{y}_{1,\nu}|\nu\right)p(\nu).$$

Si el tamaño muestral $n$ $(1 \leq n \leq N)$ es menor o igual al tamaño del estrato de no respuesta o de no voluntarios, que llamamos $N_2$, es decir $n \leq N_2$, como la probabilidad $p(\nu = 0) > 0$ pues se puede dar el caso en que la muestra irrestricta aleatoria esté toda en el segundo estrato, entonces la esperanza matemática $E\left(\bar{y}_{1,\nu}|\nu = 0\right)$ no existe al no estar definida dicha media muestral pues la muestra es de cero unidades, y por tanto la esperanza matemática $E\left(\bar{y}_{1,\nu}\right)$ no existe.

Pero si el tamaño muestral $n > N_2$, entonces el número de voluntarios en la muestra recorre los valores siguientes $\nu = n - N_2, n - N_2 + 1, \ldots, \text{mín}\{n, N_1\}$, siendo $N_1$ el tamaño del estrato de voluntarios de la población finita, es decir $N_1 = N - N_2$. Y, en este caso, $p(\nu = 0) = 0$ pues siempre habrá voluntarios en la muestra irrestricta aleatoria, y como consecuencia para todos los valores de $\nu$ de dicho recorrido la esperanza matemática $E\left(\bar{y}_{1,\nu}|\nu\right) = \bar{y}_1$, siendo $\bar{y}_1$ la media del estrato de voluntarios en la población, que puede ser distinta de la media poblacional $\bar{y}$. Así, en este caso,

deducimos que la esperanza matemática incondicional es también $E(\bar{y}_{1,\nu}) = \bar{y}_1$. Pero, en general, $\bar{y}_1 \neq \bar{y}$.

Aunque estas metodologías propuestas para tratar la no respuesta surgieron a nivel teórico, de "Teoría de Muestras" para ser aplicadas a datos recogidos en "Encuestas por Muestreo", hemos visto y así hemos explicado su utilidad como ciencia objetiva para tratar éticamente los datos estadísticos recogidos de acuerdo a un diseño o un esquema muestral en estudios experimentales, ya sean éstos observacionales (recogidos visual, auditiva, táctilmente…) o terapéuticos (como consejos de un padre, de un hermano o de un amigo, tratamientos psicológicos, etc. e incluso tratamientos médicos tradicionales o de terapias nuevas en fase de estudio).

En conclusión, es necesaria la colaboración de sujetos que inicialmente no consienten la experimentación en sí mismos, y que sean voluntarios a su vez en un segundo intento entre los elegidos aleatoriamente y probabilísticamente, para inferir conclusiones objetivas válidas para toda la población de sujetos de la eficacia del tratamiento. Esta conclusión no colisiona con el principio de autonomía y del respeto a la dignidad de los sujetos que no pueden ser obligados a consentir ser objeto de experimentación si no es informada, libre y voluntariamente, ya que esta dignidad de la persona es un principio superior al de querer obtener conocimiento científico y por tanto una verdad a cualquier coste, como el de no respetar la libertad de las seres humanos que no estuviesen dispuestos a asumir los riesgos de la experimentación informada en sí mismos. Los fines de la ciencia y el conocimiento no pueden imponerse a la voluntad de las personas, aunque esta voluntad conlleve que los tratamientos no puedan ser estudiados objetivamente según la ciencia estadística, y como consecuencia no

puedan ser estudiados con validez para toda la población de sujetos o pacientes.

Como, por otro lado, se regula que no están permitidos incentivos a la participación voluntaria en los estudios experimentales aparte de los ya considerados en justicia social, cabe preguntarse si considerar que el efecto de poder estudiar de modo objetivo el efecto de un tratamiento en toda la población de sujetos, no es suficiente razón para no incentivar con justicia social a la participación en el estudio de algunos de los no voluntarios iniciales. Pues entendemos que la objetividad científica que se obtuviera sería válida para toda la población de pacientes o de sujetos desde las fases dos y tres, por lo que pensamos que debe ser valorado también en justicia social la colaboración posterior de no voluntarios iniciales en estas fases con otras condiciones de participación revisadas al alza. Pues si fueron no voluntarios con las compensaciones que se propusieron y que aceptaron los voluntarios iniciales pero otros rechazaron, con esas mismas compensaciones raramente accederían a participar en el estudio en otra fase si no se mejoran claramente las compensaciones en un segundo intento de consentimiento informado.

Una solución de compromiso a este dilema que sugerimos es proporcionar unas compensaciones responsables de acuerdo con los posibles daños como consecuencia de la participación en el estudio experimental (y no solo como contrato retribuido prefijado único “pase lo que pase”), que puede ir desde sufragar los gastos derivados de los desplazamientos y días laborales perdidos en el caso de no derivarse ningún efecto perjudicial en la salud del sujeto que participa, pasando por retribuciones gradualmente proporcionadas a los daños derivados o seguros médicos razonables por su contribución al estudio, hasta indemnizaciones y/o seguros médicos vitalicios y de otros tipos en casos de extremas consecuencias. Estas condiciones deben ser perfectamente

explicadas oralmente e informadas por escrito antes de obtener el consentimiento informado de los sujetos. También pueden ser mejoradas alguna o algunas de las cláusulas firmadas para los no voluntarios iniciales, para que puedan facilitar su participación en el estudio en un segundo intento si fueran seleccionados en la muestra de no voluntarios en el primer intento.

Todo ello sin descuidar minimizar los riesgos de la experimentación en seres humanos como es preceptivo en cualquier estudio de este tipo, hasta hacer de tales riesgos prácticamente nulos o despreciables a altos niveles de confianza.

Ciertamente, en la fase cuarta se puede obtener información de toda la población de pacientes sobre el medicamento en cuestión que trata de conocerse, pero sería un elemento de riesgo comercializar el medicamento sin experimentar entre los pacientes no voluntarios iniciales en las anteriores fases del estudio por las razones expuestas.

Por todo ello, no se trata de llegar a un consenso o acuerdo de subjetividades de o entre estadísticos, sino de apreciar sobre todo lo objetivo y lo cierto, por encima de lo que es opinable, subjetivo y posiblemente falso. La ética exige una ciencia objetiva y demostrada en sus premisas, planteamientos y argumentos, que garantice sin errores lo que afirma.

La ética también da prioridad a la dignidad de las personas, pacientes, voluntarios, etc. sobre cualquier avance de la ciencia en el conocimiento de esas personas. La caridad es una prioridad sobre la ciencia. La verdad es así por sí misma, y no es necesariamente producto de consenso ni de una mayoría. La dignidad humana exige sobreponer la caridad al avance de la ciencia, y anteponer la verdad objetiva a cualquier decisión colectiva o individual.

Cuando decimos que “no hay verdad sin caridad, ni caridad sin verdad” nos referimos a la verdad de la revelación católica. Llevar esta afirmación a una verdad del conocimiento de la ciencia en general y de las personas, no sería correcta sin caridad ante todo con las personas, pues no es lícito revelar las imperfecciones identificando al imperfecto, sin faltar a la caridad. Esta posible identificación queda entre el paciente y el médico.

De acuerdo con la identificabilidad de las unidades de la población finita, sería posible una inferencia estadística objetiva. De acuerdo con el respeto a la voluntariedad de ser accesible una unidad personal seleccionada en la muestra de individuos o personas, tendríamos la solución inferencial que hemos estudiado del muestreo con no respuesta. Pero si las unidades seleccionadas en la submuestra del “estrato de no respuesta en el primer intento” no acceden todas a ser investigadas, entonces no sería posible, hasta lo que hemos visto, una inferencia estadística objetiva basada en la “estimación insesgada de la media poblacional”, y de un estimador insesgado de la varianza de tal estimación insesgada. La solución sería de nuevo submuestrear la submuestra del estrato de no respuesta en los dos primeros intentos, lo que parece de bastante complejidad. Pero pensar que en un tercer intento de obtener todas las respuestas (prefijadas en número), cuando ya ha habido no respuesta en dos intentos anteriores y se han presentado no voluntarios en dichas dos oportunidades previas, no resulta inteligente pensar que en una tercera oportunidad todos los muestreados sean finalmente voluntarios. Por este motivo, reiterar submuestras de no voluntarios recurrentes carece de sentido práctico, además de las molestias que ocasionarían a los propios encuestables o experimentables, y a los propios encuestadores o experimentadores.

Otra alternativa posible es el uso de la estadística descriptiva objetiva de los voluntarios investigados, teniendo cuidado en que

los datos sean verdaderos como mejor medio que obtener muchos datos que puedan ser errados, evitando variables no definidas claramente o difusas y que puedan dar lugar a respuestas ambigüas o no únicas ante un mismo hecho observado.

El aspecto que no se puede olvidar es el debido respeto y el debido amor a los encuestados u observados. Su libre voluntad, para participar o no en el estudio, ha de ser tenida en cuenta. No basta con una supuesta buena intención del experimentador o del encuestador. Un buen fin, como sería el conocimiento más perfecto y verdadero de una enfermedad o de un nuevo tratamiento médico o el estudio sociológico de una actitud ante determinada cuestión de una población humana, no puede llevarse a cabo con medios malos que no respeten o no amen a las personas de las que se obtendría la información para tal fin. El fin bueno debe obtenerse con medios buenos, pues de otro modo se pervierte la pretendida bondad de la investigación por muestreo. Una cita bíblica que puede ayudar a entenderlo es *Romanos* 3,8.

De este modo, muchas investigaciones por muestreo pueden ser objetivas, pero en algunos casos requiere la libre colaboración de voluntarios, y esta libertad humana que no es predecible ni modelizable objetivamente puede impedir la objetividad estadística inferencial de las conclusiones de un estudio.

Por tanto, es posible una inferencia estadística objetiva cuando las unidades de la población finita están identificadas, aunque fueran personas anónimas, y todas las unidades a observar fueran accesibles en el caso de ser seleccionadas en la muestra.

En otros casos, como en el que se presenta no respuesta, hay métodos objetivos que podrían funcionar con mejores medios y recursos asistenciales, pero no queda garantizada esta objetividad inferencial debido a la libertad de las personas a ser observadas, al

poder faltar parte de la información que la muestra seleccionada probabilísticamente exigiría acceder teóricamente.

Hay métodos estadísticos que son objetivos tanto en el sustrato matemático como en la puesta en práctica de la metodología sobre el terreno. Nos referimos también a que substituir realidades objetivas por hipótesis ideales es cambiar ciencia objetiva por ideología. Pues una idea sin comprobación posible como causa del análisis estadístico da lugar a ideología y no precisamente a una ciencia objetiva.

Inferencia clásica. Presupone un modelo poblacional que en muchos ejemplos no puede comprobarse en la práctica, por lo que puede ser o es ideología y no ciencia objetiva. En este caso la idealización consiste en sustituir la verdadera población que existe en el mundo real por la idea que el investigador pueda hacerse subjetivamente usando funciones matemáticas que supuestamente aproximan la realidad pero sin la posibilidad de comprobar fehacientemente su ajuste correcto a la realidad.

Inferencia bayesiana. Introduce prejuicios como la llamada distribución *a priori*, para concluir estimaciones sesgadas, donde no las tendría la inferencia clásica. Ahora la ideología se introduce al sustituir un valor real de una población, que en principio puede ser desconocido, por una distribución subjetiva que representa la idea que tiene el investigador bayesiano del parámetro desconocido que pretende estimar. Es decir, el investigador sustituye una realidad concreta desconocida para él por una distribución ideal subjetiva y supuesta por él en el análisis inferencial.

Estadística descriptiva. No está a salvo de posibles manipulaciones tampoco este tipo de estadística. Un ejemplo es, en el caso de un histograma, sustituir la media de la variable estadística en un intervalo por el punto medio del intervalo; de este modo, al promediar los puntos medios por las frecuencias relativas

de los intervalos, el valor medio de la variable estadística queda afectada por la idea de que el punto medio del intervalo representa a la media de la distribución en dicho intervalo. Lo correcto sería ponderar las medias parciales de la variable en cada intervalo por sus respectivas frecuencias relativas y al sumar todos estos productos obtendríamos sin error la media de la variable estadística completa. El idealismo, en este caso, consiste en sustituir la media parcial de la variable en el intervalo por el punto medio del intervalo, haciendo perder información y sesgando el valor del parámetro media de la variable estadística si quisiéramos reconstruir el valor medio ponderando las medias de cada intervalo por sus frecuencias relativas y sumándolas todas ellas.

Otro ejemplo de idealismo es el que tiene lugar al usar la distribución normal en base al teorema central del límite. Si bien es cierto que la media aritmética de las observaciones obtenidas por muestreo aleatorio simple de una misma población con varianza finita, tiende a ser normal asintóticamente en distribución, no es menos cierto que la mayoría de aplicaciones de este teorema no comprueban en la práctica la hipótesis de partida que da validez al resultado, que la muestra sea en realidad una muestra aleatoria simple con reemplazamiento. Esto es observable en revistas de medicina basadas en datos de muestras al azar, pero no en muestras aleatorias simples, es decir, que en cada dato se recoge la variable de interés en un sujeto que es seleccionado independientemente con probabilidades iguales y con la misma distribución que la población de partida. Si no hay esta previa selección aleatoria simple, no puede hablarse después con garantía de que los datos elaborados sigan distribuciones aproximadamente normales, *ji*-cuadrado, *t* de Student, *F* de Snedecor, etc. en base al teorema central del límite ya que no se respeta en la práctica una hipótesis fundamental del teorema. En realidad lo que se hace es predecir

que una muestra seleccionada al azar va a proporcionarnos una muestra como si fuera aleatoria simple, lo cual puede ser intuitivo pero no se prueba racionalmente. Si una intuición tuviese un valor aproximativo, entonces estaríamos aproximándonos intuitivamente a la aproximación asintótica dada por el teorema central del límite. Y en este proceso de doble aproximación hemos perdido el hilo conductor racional en aras de una practicidad que no puede asegurar científicamente lo que afirma al final.

Muchas de estas idealizaciones se basan en ideas surgidas en el siglo XIX, en el que idealismo y el positivismo tuvieron gran aceptación (Izquierdo Urbina, 2015, p. 71), pero que dieron lugar a muchas ideologías científicas todavía en nuestros días, llenando pizarras y revistas científicas hasta la actualidad.

En un problema de inferencia el objetivo no es decir la verdad de un parámetro desconocido, sino de estimar tal parámetro con un error que tratamos de minimizar atendiendo a las condiciones específicas del problema. Se puede hablar de la verdad de usar un estimador que es óptimo o que es admisible dentro de un conjunto de dichos estimadores, pero no cabe esperar saber la verdad del parámetro con una mera estimación del mismo basada en una muestra de datos solamente. Sí sabremos que hemos estimado bien en las condiciones concretas optimizando el estimador, por ejemplo exigiendo que sea “insesgado”, es decir, que el promedio de sus probables estimaciones coincida con el parámetro (o función paramétrica) que deseamos conocer mediante el método inferencial. En realidad la insesgación es un requisito totalmente justificado y deseable, que las posibles estimaciones tengan por promedio exactamente el valor verdadero que pretendemos estimar. La minimización del error consiste en conseguir la mínima dispersión de las posibles estimaciones proporcionadas por una estrategia muestral compuesta por un diseño de muestreo y un estimador concreto que pertenece a una clase de ellos. El diseño

muestral asigna la probabilidad de cada posible muestra y el estimador es una función que depende de los datos observados en las unidades de la muestra, y depende también de los identificadores de las unidades seleccionadas en la muestra.

Estas son algunas de las cuestiones éticas además de las recientemente estudiadas por el autor y, más concretamente, en las planteadas en la bioestadística médica y en los estudios de salud pública. Queda de manifiesto que la estadística empleada en la mayoría de estudios médicos y de salud pública hasta fechas recientes adolecen de subjetivismo y se fundamentan en parte en el idealismo, por lo que distan de ser metodologías objetivas como sería deseable al tratar con seres humanos para no hacer *falsos testimonios* sobre el conjunto de pacientes o sobre los sujetos de los que se toman las observaciones o datos con fines estadísticos ya sean descriptivos o inferenciales.

Dos citas bíblicas que prohíben esta manera de proceder son Éxodo 20, 16 y Deuteronomio 5, 20 (Editorial DDB, 1999), en ambos casos recogiendo la palabra de Yahvé, Dios Padre de los cristianos. También Jesús confirmó el mandamiento de *no dirás falsos testimonios*, por ejemplo en Mateo 15, 19 (Editorial DDB, 1999).

Repasamos a continuación algunos procedimientos que considero han aportado objetividad a la estadística, no solo como razonamientos válidos, sino sobre todo como aprovechables en la práctica sin excesivos costes.

Lo que se pretende con un análisis estadístico objetivo es que lo que se afirma acerca de una metodología o de una estrategia de muestreo sea cierto, y no un cúmulo de aproximaciones en diversas fases o etapas a un método que desfigurarían las cualidades reales

de lo que realmente se hace con respecto a lo deseable teórica y objetivamente.

Las razones de la objetividad en estadística han sido expuestas con detalle en este libro. En él indico que no basta con tener una teoría razonable sino que todo el proceso de teoría y puesta en práctica debe ser correcto y sin saltos en el vacío. Debo decir que la inferencia clásica y la inferencia bayesiana tienen lagunas en el razonamiento o en la práctica como para que pudiéramos considerarlas objetivas en muchos casos que se ponen como ejemplos de su potencial científico. Un ejemplo de estas lagunas es que la muestra no se suele seleccionar de acuerdo a un diseño muestral previamente definido. Un libro que explica con algún detalle esta forma de seleccionar la muestra es el de Mirás Amor (1985).

También algunos métodos de tratamiento de los datos observados de una variable estadística en la estadística descriptiva adolecen de simplificaciones que no guardan o conservan todo el potencial informativo de los datos originales, en especial para los fines propuestos con el estudio estadístico.

Algunos de los primeros resultados probados sobre la existencia de estimadores insesgados uniformemente de mínima varianza, y de estimadores uniformemente de mínimo error cuadrático medio, han sido tratados –en el contexto de poblaciones finitas en el modelo objetivo de población finita fijada– por Ruiz Espejo (1987c).

Un ejemplo es el tratamiento objetivo de la no respuesta cuando en la muestra aparecen sujetos o unidades de las que no podemos obtener respuesta a pesar de haber sido seleccionadas en la muestra de acuerdo con un diseño muestral. Se han escrito libros y muchos artículos sobre el tratamiento de la no respuesta, pero desde los años 40 del siglo XX no se había resuelto el problema de

estimar sin sesgo la varianza del estimador insesgado para no respuesta propuesto por Hansen y Hurwitz (1946), y popularizado en el libro de Cochran (1977), de un modo objetivo y convincente. Este problema ha sido resuelto satisfactoriamente por Ruiz Espejo (2011, 2013d, 2013g, 2015b) y por Thompson (2012) bajo diversos esquemas o estrategias de muestreo.

Otro problema que ha sido resuelto satisfactoriamente desde la perspectiva de la estadística objetiva es el problema de inferencia en muestreo posagrupado (Ruiz Espejo *et al.*, 2006). También han sido caracterizados los diseños muestrales admisibles para el estimador Horvitz-Thompson por Ruiz Espejo (1987b). La optimalidad del muestreo aleatorio simple con reemplazamiento en la clase de todos los diseños ordenados posagrupados proporcionales al tamaño, y de tamaño fijo, para el estimador media muestral, ha sido probada por Ruiz Espejo (2008).

Un problema teórico resuelto que tiene implicaciones en la inferencia clásica y también en la inferencia estadística objetiva es el de estimación insesgada óptima de los momentos poblacionales más importantes. La primera solución a este problema en los momentos centrales poblacionales de orden cuatro se debe a Ruiz Espejo *et al.* (2013, 2016) y a Ruiz Espejo (2015h).

Otro problema sobre la protección de la intimidad en respuesta aleatorizada con distribución *a priori* objetiva, dada por el diseño muestral, ha sido estudiado por Ruiz Espejo y Singh (2003).

Sobre estimación lineal óptima a partir de medias muestrales independientes o incorrelacionadas se han resuelto algunos problemas en Ruiz Espejo *et al.* (1995), y generalizados en Ruiz Espejo *et al.* (2001).

También hemos probado la admisibilidad de un estimador de regresión lineal corregido insesgado sobre el estimador de regresión lineal clásico, y justificamos la existencia de estimadores concretos de regresión multivariante insesgados (Ruiz Espejo, 2016a, 2016c). Y en Ruiz Espejo (2015h) proporciono estimadores insesgados, así como estimadores insesgados de sus varianzas en algunos casos, a partir del estadístico media-de-razones, lo que son unas soluciones objetivas interesantes en el caso de disponer de una variable estadística altamente correlacionada con la variable de interés en estudio.

Otro estimador insesgado de la "varianza del estimador insesgado" en muestreo sistemático de doble arranque, también de modo objetivo, ha sido proporcionado por Ruiz Espejo (2014b). Y en Ruiz Espejo (1997f) se prueba la unicidad de la estrategia de Zinger con varianza estimable insesgadamente.

Otros estimadores insesgados con criterios objetivos han sido propuestos recientemente por Ruiz Espejo (2018b, 2018f, 2018j) ya sea para la media poblacional así como para la varianza.

Lo que he pretendido hacer ver en este resumen es que no basta un racionalismo cualquiera en el estudio y en la investigación estadística, sino que también es necesaria una visión rica y completa de los matices que hacen que una investigación o una enseñanza sean realizables en la práctica. Sin perder de vista que los resultados han de exponerse de un modo correcto en el fondo, en la forma, en lo lógico y en lo práctico. Si además se hace todo esto amablemente, creo que se ha llegado a un estado de madurez en la ciencia estadística.

Cualquier intuición o racionalismo reductivo de los problemas estadísticos que no superen todos estos elementos básicos de racionalidad, practicidad y de buen espíritu dejarían incompletas las aportaciones a la ciencia, aunque rellenen muchas

páginas con gran exposición de fórmulas o tablas complicadas, porque los argumentos no se sostendrían ante un examen mínimamente minucioso de su verdadera utilidad para el fin que se proponen.

Como primera cosecuencia, se puede llegar a afirmar que una gran parte de investigaciones estadísticas, publicadas incluso en revistas con factor de impacto estadístico ampliamente reconocido, no alcanzan algunos de los estándares que hemos sugerido en este artículo de exposición para tratar los datos estadísticamente en su descripción o para fines inferenciales.

Como segunda consecuencia, si se pretende alcanzar una educación y una investigación que pueda llamarse ciencia a todas luces, es necesario aunar esfuerzos para proporcionar materiales didácticos en estadística menos idealistas, así como promocionar a editores de publicaciones de ciencia que tengan un currículum investigador acorde con los argumentos que he expuesto. Pero mientras que los motivos editoriales se orienten más con una visión de negocio que de integridad en la ciencia dudo que llegue a verse una estadística de calidad que podamos llamar ciencia y no meras ideas sueltas, sin conexión real y racional entre lo que pretendidamente afirman y lo que realmente se ha hecho al hacer tal afirmación; es decir, un fraude científico en algunos o en muchos casos.

Otra fuente de información son las encuestas. Hay muchas teorías matemáticas y estadísticas que darían soporte científico a las investigaciones por encuestas realizadas en la sociedad, pero para un científico es claro que para que cualquiera de estas teorías puedan dar algún fruto de veracidad los datos recogidos han de ser verdaderos y ciertos, al menos en la fase confirmatoria final. Dicho de otro modo, sería inútil tomar datos falsos si se pretende que la

encuesta arroje algo de luz sobre una cuestión política o sociológica.

De aquí que la persona a la que va dirigida una encuesta sea una persona que responda la verdad o, al menos, que el dato sea tomado de cada persona o unidad (empresa, confesión religiosa, industria, universidad, parroquia, etc.) observada por un encuestador sea fiel a la verdad que observa y posteriormente anota o registra como respuesta.

Un ejemplo sería el de un médico que lleva cuenta de las enfermedades que se presentan en su consulta. No sería necesario que el paciente diga la enfermedad sino que el mismo médico consultado puede conocer la enfermedad y, en un caso extremo, informar de que desconoce la enfermedad o el mal que le han consultado.

En la presente sección veremos las posibles actuaciones ante el desarrollo de una encuesta desde un punto de vista profesional, ético y moral.

Los estudios que conducirían a un soporte científico correcto del análisis inferencial, es decir, de lo que puede afirmarse inductivamente de la población con los datos obtenidos de una muestra aleatoria probabilística seleccionada de la población, requieren unas propiedades científicas que han de ser respetadas en el procedimiento de selección de la muestra y de estimación.

Por ejemplo, si queremos usar al final unas fórmulas que dan la medida del error de muestreo aleatorio simple con reemplazamiento de una estimación puntual de un “parámetro poblacional” (como sería el caso de “la media poblacional de la variable consultada” en la encuesta), hemos de cuidar que la selección de personas o unidades reproduzca con evidencia la equiprobabilidad e independencia probabilística en las distintas

sucesivas selecciones de personas o unidades de la población para formar parte de la muestra aleatoria simple con reemplazamiento.

De otro modo, las fórmulas serían inútiles ya que el fundamento con el que se obtiene teóricamente la fórmula no se ha cumplido en el caso concreto al que se pretende aplicar dicha fórmula. Se debe respetar la independencia estadística en las sucesivas selecciones de unidades de la muestra, que es una hipótesis necesaria para la que fórmula final de la varianza de la estimación tenga sentido, por ejemplo.

Desde un punto de vista moral no sería correcto afirmar con el uso de tal fórmula que una estimación puntual tiene determinado error de muestreo si en realidad no se han cumplido las condiciones para las cuales la fórmula tiene sentido. Esto es lo que ocurre en la mayor parte de las investigaciones médicas, psicológicas, sociológicas, económicas, etc. que he conocido: que se presentan como científicas unas conclusiones que moral, científica y éticamente no serían de correcto recibo al no comprobar todas las hipótesis implícitas que supone la teoría que aporta tales estimaciones.

Del mismo modo, fórmulas que se han obtenido recientemente para sobrellevar el efecto de la no respuesta en una muestra inicial, pudiendo estimar sin sesgo el parámetro poblacional y la varianza del estimador para ello, presuponen que la encuesta al final obtiene todas las respuestas buscadas en un submuestreo de no respondientes. Esto es posible suponerlo a nivel teórico, pero la práctica es más ilustrativa de que no será siempre lo que ocurra pues hay que respetar la voluntad de los encuestados. Además no todo lo teóricamente pensado ni todo lo técnicamente posible son necesariamente moralmente aceptables. Y, en caso de dilema, el respeto a las leyes morales es una conducta superior a la

imposición de unas reglas científicas que no respetan a los encuestados.

Esto sería aplicable a los estudios teóricos que dan solución al problema de la estimación insesgada con no respuesta. Mi intención con estas investigaciones teóricas era proveer de un estimador insesgado de la varianza para el estimador insesgado de la media poblacional propuesto por autores americanos. Con ello demostraba que si tenían los datos para estimar insesgadamente la media poblacional, podemos disponer con la misma información de estimadores insesgados de la varianza de tal estimador. Pero con ello nunca pretendí reconocer que estas soluciones teóricas fueran necesariamente morales o éticas, ya que su uso condiciona la voluntad libre de los encuestados, al menos en una segunda fase de submuestreo. Sería contradictorio e inmoral que la ciencia teórica sirva para no respetar a las personas y su voluntad, su conciencia bien formada en definitiva, y su deseo y su propósito sincero de obrar bien y evitar el mal.

Es, por tanto, lamentable el abuso del uso de fórmulas que tienen su sentido en una correcta investigación teórica, pero cuando son aplicadas a casos prácticos que no pueden atenerse a las condiciones supuestas en la teoría o no se han preservado esas condiciones o incluso no se sabe si es así en la práctica real de la encuesta o estudio observacional, es mejor no afirmar lo que no se puede garantizar que sea cierto o al menos que cumpla unas propiedades estadísticas que la práctica realizada no puede confirmar al no cumplir las condiciones requeridas para que sea así.

La razón por la que las muestras aleatorias seleccionadas pueden ser no simples (es decir, seleccionadas con probabilidades iguales con reemplazamiento en extracciones independientes sucesivas) es porque en primer lugar la mayoría de encuestas no se seleccionan con un procedimiento correcto, por ejemplo, sin marco

poblacional, o no sorteando con equiprobabilidad entre las unidades para seleccionar los sucesivos encuestados, o no siendo independientes las sucesivas selecciones.

En general existen otros muchos procedimientos de selección de las unidades de la muestra con sus respectivos estimadores con los que realizar las inferencias estadísticas, pero en cada caso de estos se ha de cuidar que en la práctica se respete el diseño muestral concreto que se usa para realizar la inferencia así como el estimador concreto. Y todo esto no es inmediato ni cabe esperar que los datos recibidos de cualquier modo sean una muestra aleatoria probabilística tal y como el diseño muestral indica o presupone.

Lo que ocurre en general en la práctica es que las unidades que colaboran voluntariamente en aportar sus datos a la encuesta, lo hacen tras ser informados de la finalidad de la encuesta a realizar y aceptando su consentimiento informado a colaborar efectivamente en la toma de datos.

Algo parecido ocurre o debería ocurrir en los ámbitos médico y psicológico en los que puede experimentarse un tratamiento y obtener así una respuesta como consecuencia del tratamiento. En estos casos está regulado el consentimiento informado de los pacientes a propuesta de los médicos especialistas.

Además de las cuestiones éticas que deben superar todos los estudios experimentales con seres humanos, se añaden los planteados por el diseño muestral y la estimación concretas que permitirían hacer tal inferencia inductiva.

En todos estos casos, la muestra no es probabilística sino intencional y/o voluntaria en la que intervienen uno o muchos actores, por lo que en realidad depende más de voluntades humanas

que de un estricto azar controlado para poder hacer inferencias científicas según la teoría estadística inferencial.

Por todo ello, cuando la selección de la muestra no es por azar controlado probabilísticamente de acuerdo a un diseño muestral ordenado (o no ordenado) concreto, sino que intervienen otros factores individuales o personales voluntarios, dicha muestra puede tener un valor empírico pero no inferencial estadístico. La muestra sería una selección de la población, pero no permite hacer inferencias con rigor científico, aunque aparentemente se dispongan de todos los datos requeridos en algunas fórmulas para que pudiera realizar una posible inferencia. En cualquier caso esta inferencia no sería válida o no es garantizable en lo que pudiera afirmar sobre la población de la que se ha seleccionado la muestra intencional o de voluntarios.

La práctica moral exige la conformidad del encuestado en participar en el estudio de la encuesta, pues es razonable respetar la libertad de cada persona en dar unos datos personales sobre "quién es, lo que piensa, quiere, hace o tiene" sin forzar su voluntad en ningún momento. En este sentido, es razonable que el encuestado no responda, o deje de responder la pregunta o las preguntas que considere oportuno, o desista de seguir respondiendo en cualquier momento del cuestionario. Solo así se verán respetados sus derechos personales en todo momento y será libre de cooperar oportunamente con el estudio de la encuesta si así lo desea.

Cooperar en responder una encuesta es un acto de libertad. No puede imponerse la obligatoriedad de colaborar con un estudio ajeno sin contar con el beneplácito del posible encuestado. De otro modo los patrocinadores de la encuesta emplearían métodos coercitivos que no respetarían la voluntad de los encuestados y

forzarían a estos a colaborar aun estando disconformes, algo no aceptable en una sociedad libre.

Como vemos, tanto la verdad en las respuestas como la libertad en responder forman parte esencial de todo estudio por encuestas, sin las cuales no sería posible realizarlas y ni siquiera se podrían obtener unos resultados aprovechables sin esa verdad buscada, ni sería ético obtener las respuestas forzando voluntades en contra de su parecer para cooperar con los objetivos de la encuesta.

"No dirás falsos testimonios ni mentirás" enseña la moral católica, pero como hemos advertido, si no fueran verdaderas las respuestas la ciencia teórica relativa a la inferencia en poblaciones finitas no valdría para nada, así de simple y exigente. Pero aunque la moral católica no informe al investigador, sin ella carece de sentido toda investigación que no pretenda ser veraz. Por otras razones de no menos relevancia, otros tipos de inferencia estadística son más cuestionables.

Un ejemplo de ello es que cuando se supone teóricamente que la población es infinita (caracterizada por una función de densidad continua) sin serlo, como suele hacer la inferencia clásica, bayesiana, no paramétrica, etc. implícitamente lo que se está diciendo es que podemos prescindir de cualquier parte finita de la población en estudio porque la posible población infinita (caracterizada por una función de densidad continua) no se vería afectada por ello. Así, si la población es finita en realidad y prescindimos o eliminamos intencionalmente todas las unidades de la población finita, la inferencia con población teórica infinita no se vería afectada pero en realidad habríamos vaciado de sentido la inferencia estadística pretendida.

En realidad el mensaje que trasmite una inferencia en población infinita con función de densidad continua es que se puede eliminar cualquier parte finita de la población objetivo para el propósito inferencial, pero esta visión es errónea, como puede verse fácilmente que eliminando una parte de la población la media poblacional se ve afectada en poblaciones finitas, que son las que tienen algún interés práctico como para garantizar que las estipulaciones del diseño muestral pueden ser llevadas a la práctica de modo controlado. Esto sin entrar a valorar qué significado puede tener que el investigador pueda o pretenda eliminar parte de la población para hacer sus inferencias. Nada limpio como cabe suponer, y de hecho esto se hace en muchos estudios inferenciales de seres humanos.

Pero la consecuencia de este proceder es que se introducen sesgos en la estimación del parámetro media poblacional, mediante la media muestral de las “unidades no excluidas de la población”, que serían como el estrato de respuesta. Habría otro estrato de no respuesta compuesto por las unidades de la población que han sido excluidas y que, por tanto, no podrán responder. Y todo esto supuesto que todas las unidades no excluidas respondiesen. Los razonamientos para entender estas cosas pueden seguirse casi directamente de la teoría de la estimación puntual cuando se presenta la no respuesta.

Por las razones expuestas, cabe preguntarse si es posible compaginar la ciencia estadística teórica y la práctica real moral. La respuesta a simple vista parece que solo tiene una respuesta: o se respeta la voluntad de los encuestados para participar o no en la encuesta (en cuyo caso la ciencia estadística teórica sería inútil, aunque los patrocinadores de la encuesta hicieran gala de sus virtudes éticas o morales), o bien se obliga a los encuestados a responder y sin falta de respuesta ni de verdad en las mismas (lo que presupone un comportamiento legal y moral en todos los

encuestados, algo que parece ser imposible a juzgar por los expertos funcionarios encargados de este tipo de estudios en la práctica).

Obviamente la solución al dilema ciencia-moral, ya que parece que no pueden compatibilizarse simultáneamente, pasa por la realización de unos censos de colaboración legal obligatoria, y por unas encuestas postcensales de obligada participación por causas justificadas o, en su defecto, de muestras empíricas no probabilísticas que mostraran la situación de una o varias variables estadísticas sin pretensiones inferenciales ya que la inferencia solo se podría llevar a efecto con muestras probabilísticas de acuerdo a un diseño muestral como parte de una estrategia de muestreo unida a un estimador de la función paramétrica a estimar.

La mayor parte de estudios por muestreo que se realizan en la sociedad son de carácter no probabilístico ya que no disponen del marco poblacional necesario para seleccionar la muestra de acuerdo con un diseño muestral, y también cuando el marco poblacional se dispone es muy frecuente no hacer uso del mismo. Por ello la ciencia aplicable queda vacía de contenido en la práctica aunque se pretenda presentar unas conclusiones inferenciales pero sin cuidar todos los requisitos para que fuera una inferencia inductiva científica. Sería un fraude en la ciencia como resulta ser en muchos casos.

Y es que la moral no se limita a decir la verdad y a respetar la voluntad de los encuestados, se trata de un comportamiento de acuerdo a la ley divina y natural, y de la que forman parte aquellos dos requisitos como necesarios pero no suficientes para una actuación enteramente moral.

Como conclusión, recomiendo que en el caso de poder llevar a cabo una encuesta según prescribe la teoría estadística inferencial

(lo cual requiere el uso de poblaciones finitas y de un marco de la población) y las condiciones éticas y morales aplicables, lo razonable es poner en la práctica todas las condiciones para que tal teoría se respete en la práctica en todas sus hipótesis y en las conclusiones rigurosas que se deducen.

En otro caso, si la teoría estadística no puede llevarse cuidadosamente en la práctica, lo aconsejable es describir objetivamente la muestra empírica obtenida pero sin pretensión inferencial alguna que resultase engañosa. De este modo se preserva lo que consiste la ley moral natural: "haz el bien y evita el mal". Y también preservaría la moral cristiana en este aspecto: "no dirás falsos testimonios ni mentirás".

Para un mayor detalle de estos razonamientos pueden consultarse los libros indicados en las referencias.

Para concluir y para que sirva de reflexión a los lectores, os recomiendo los siguientes textos bíblicos, del Catecismo de la Iglesia Católica y del Código de Derecho Canónico:

*Colosenses* 2,8: "Mirad que nadie os esclavice mediante la vana falacia de una filosofía, fundada en tradiciones humanas, según los elementos del mundo y no según Cristo."

*1ª Timoteo* 6,20-21: "Timoteo, guarda el depósito. Evita las palabrerías profanas, y también las objeciones de la falsa ciencia; algunos que la profesaban se han apartado de la fe. La gracia con vosotros."

*Hebreos* 13,9: "No os dejéis seducir por doctrinas diversas y extrañas."

*2ª Pedro* 3,18: "Creced, pues, en la gracia y en el conocimiento de nuestro Señor y Salvador, Jesucristo. A él la gloria ahora y hasta el día de la eternidad. Amén."

*Catecismo de la Iglesia Católica* 2295: Las investigaciones o experimentos en el ser humano no pueden legitimar actos que en sí mismos son contrarios a la dignidad de las personas y a la ley moral. El eventual consentimiento de los sujetos no justifica tales actos. La experimentación en el ser humano no es moralmente legítima si hace correr riesgos desproporcionados o evitables a la vida o a la integridad física o psíquica del sujeto. La experimentación en seres humanos no es conforme a la dignidad de la persona si, por añadidura, se hace sin el consentimiento consciente del sujeto o de quienes tienen derecho sobre él.

*Código de Derecho Canónico*: los delitos contra la vida, la libertad y la dignidad de las personas tienen la misma categoría de los delitos de aborto o de homicidio, o de abuso de menores, en el Código de Derecho Canónico en el año 2021.

# Anexo I

# Distintos tipos de inferencia

1. Inferencia objetiva en poblaciones finitas fijadas

Se basa en disponer de una población finita conocida y de tamaño fijo, cuyas unidades o elementos son identificables y accesibles para poder recabar información de la variable de estudio en cada unidad seleccionada en la muestra. Por tanto la población debe estar listada por las unidades y su medio de localización o acceso a las mismas. Una muestra es un conjunto de unidades o una secuencia finita de ellas (aunque haya repeticiones), seleccionadas según un procedimiento probabilístico de obtención o diseño muestral. Si una unidad está en la muestra, debe ser observada su variable de interés y el dato recabado es aprovechado en la fase de estimación. Informaciones auxiliares pueden usarse en el diseño muestral o en el estimador o en ambos.

2. Inferencia paramétrica clásica

En la inferencia paramétrica clásica, la población está caracterizada por una función de densidad, de cuantía o de distribución, de la que se conoce su fórmula que depende de una o varias constantes desconocidas que aparecen en la fórmula, y se denominan parámetros. La inferencia consiste en aproximar dichos parámetros basándose en la propia fórmula subjetiva y en los datos recogidos en una muestra con determinadas especificaciones con las que se supone ha sido obtenida.

3. Inferencia bayesiana

En este caso la población sigue un modelo subjetivo similar al anteriormente descrito, pero los parámetros no se suponen fijos sino que a su vez se supone que son variables aleatorias con una distribución “a priori” supuestamente conocida por el investigador. En base a una muestra aleatoria de la población, la distribución “a priori” de los parámetros poblacionales se ve modificada o rectificada por otra distribución “a posteriori” de los parámetros una vez observada la supuesta muestra con aleatoriedad probabilística. El criterio de estimación puntual de los parámetros puede ser por varios procedimientos destacando el método de la máxima verosimilitud a posteriori.

4. Inferencia no paramétrica

Es similar a la inferencia paramétrica, si bien la descripción de la población no depende de constantes desconocidas sino de propiedades o cualidades que describen una variedad de poblaciones que las verifican. La inferencia no paramétrica trata de desvelar cuál de ellas es más acorde con los datos obtenidos en una supuesta muestra aleatoria de la población.

5. Inferencia de distribución libre

Puede considerarse un caso muy particular del anterior en el que la variedad de distribuciones poblacionales que supuestamente una de ellas es la cierta, se amplía a todas las posibles distribuciones.

6. Inferencia con modelos superpoblacionales

Los modelos superpoblacionales parten de una población finita de tamaño fijo de modo similar al modelo de inferencia objetiva, pero ahora cada unidad de la población puede ofrecer diversas respuestas a la misma pregunta u observación de la variable de interés, y que es ahora una variable aleatoria en cada unidad que puede ser modelizado por otros tipos de inferencia como la paramétrica clásica o la bayesiana.

Existen otros tipos de inferencia estadística derivadas de las anteriores, pero básicamente tienen los mismos o similares puntos débiles que los ya indicados y por lo que no son plenamente éticos como ciencia aplicada a seres humanos u otras unidades de la población que afecten a seres humanos o grupos de ellos.

# Anexo II

# Muestreo aleatorio simple

El muestreo aleatorio simple es el caso más sencillo del muestreo, aquél en que la observación es la que se toma de la unidad seleccionada en la correspondiente selección independiente con probabilidades iguales para cada unidad de la población finita, y para cierto tamaño de la muestra.

Básicamente la teoría que fundamenta la estimación insesgada de momentos poblacionales no centrales y centrales, así como sus varianzas puede verse en el artículo de Ruiz Espejo *et al.* (2013) y revisado posteriormente por Ruiz Espejo (2015h). Dodge y Rousson (1999) y Ruiz Espejo (1998b) fueron los que en sus trabajos iniciales aportaron las ideas para resolver los problemas de estimación insesgada con muestreo aleatorio simple con reemplazamiento.

Todos los estimadores insesgados referidos en este anexo son además de mínima varianza para distribución libre por ser invariantes ante permutaciones en el orden de las observaciones de la secuencia ordenada de la muestra aleatoria simple (Zacks, 1971, p. 150).

Una fuente de números aleatorios con reemplazamiento para seleccionar una muestra aleatoria simple con reemplazamiento de identificadores de unidades de una población finita puede obtenerse en la dirección web random.org.

# Referencias


AGENCIA ESPAÑOLA DE MEDICAMENTOS Y PRODUCTOS SANITARIOS (2002). *Normas de Buena Práctica Clínica.* Ministerio de Sanidad y Consumo. Madrid.

ANDERSON, M. (1988). *The American Census – A Social History.* Yale University Press. New Haven, CT.

ARDILLY, P. (2006). *Les Techniques de Sondage.* Technip. Paris.

ARDILLY, P.; TILLÉ, Y. (2006). *Sampling Methods: Exercises and Solutions.* Springer. New York, NY.

ARMITAGE, P. (1947). A comparison of stratified with unrestricted random sampling from a finite population. *Biometrika* **34**, 273-280.

ARNÁIZ VELLANDO, G. (1965). *Introducción a la Estadística Teórica.* Lex-Nova. Madrid.

ASOCIACIÓN MÉDICA MUNDIAL (2008). *Declaración de Helsinki: Principios Éticos para las Investigaciones Médicas en Seres Humanos*, adoptada y posteriormente enmendada. 59ª Asamblea General de la Asociación Médica Mundial. Seúl.

AZORÍN POCH, F.; SÁNCHEZ-CRESPO RODRÍGUEZ, J. L. (1994). *Métodos y Aplicaciones del Muestreo.* Alianza Universidad Textos. Madrid.

BARNETT, V. (2002). *Sample Survey Principles and Methods* (3rd edition). Arnold. London.

BASU, D. (1958). On sampling with and without replacement. *Sankhyā: The Indian Journal of Statistics, Series A* **20**, 287-294.

BASU, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā: The Indian Journal of Statistics, Series A* **31**, 441-454.

BAYLESS, D. L.; RAO, J. N. K. (1970). Estimators and variance estimators in unequal probability sampling. *Journal of the American Statistical Association* **65**, 1645-1667.

BELLHOUSE, D. R. (2000). Survey sampling theory over the twentieth century and its relation to computer technology. *Survey Methodology* **26**, 11-20.

BERGER, M. P. F.; WONG, W. K. (2009). *An Introduction to Optimal Designs for Social and Biomedical Research.* Wiley. Chichester.

BREWER, K. R. W.; HANIF, M. (1983). *Sampling with Unequal Probabilities.* Springer-Verlag. New York, NY.

BRUNT, L. (2001). The advent of the sample survey in the social sciences. *Journal of the Royal Statistical Society, Series D (The Statistician)* **50**, 179-189.

CASAS SÁNCHEZ, J. M. (1996). *Inferencia Estadística para Economía y Administración de Empresas.* Centro de Estudios Ramón Areces. Madrid.

CASAS SÁNCHEZ, J. M.; SANTOS PEÑAS, J. (1995). *Introducción a la Estadística para Economía y Administración de Empresas.* Centro de Estudios Ramón Areces. Madrid.

CASSEL, C. M.; SÄRNDAL, C. E.; WRETMAN, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley. New York, NY.

CATECISMO DE LA IGLESIA CATÓLICA (1993). *Catecismo de la Iglesia Católica* (3ª edición revisada). Asociación de Editores del Catecismo. Getafe.

CHAUDHURI, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys.* Chapman & Hall/CRC. Boca Raton, FL.

CHAUDHURI, A. (2014). *Modern Survey Sampling.* Chapman & Hall/CRC. Boca Raton, FL.

CHAUDHURI, A.; CHRISTOFIDES, T. C. (2013). *Indirect Questioning in Sample Surveys.* Springer. Heidelberg.

CHAUDHURI, A.; MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques.* Dekker. New York, NY.

CHAUDHURI, A.; STENGER, H. (1992). *Survey Sampling: Theory and Methods.* Dekker. New York, NY.

CHAUDHURI, A.; VOS, J. W. E. (1988). *Unified Theory and Strategies of Survey Sampling.* North-Holland. Amsterdam.

CICCHITELLI, G.; HERZEL, A.; MONTANARI, G. E. (1997). *Il Campionamento Statistico* (2a edizione). Il Mulino. Limena.

CICCONE, L. (2006). *Bioética. Historia. Principios. Cuestiones* (2ª edición). Ediciones Palabra. Madrid.

COCHRAN, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association* **37**, 199-212.

Cochran, W. G. (1977). *Sampling Techniques* (3rd edition). Wiley. New York, NY.

Cochran, W. G. (1980). *Técnicas de Muestreo.* CECSA. México.

Congregación para la Doctrina de la Fe (1987). *Instrucción DONUM VITAE. Sobre el Respeto de la Vida Humana Naciente y la Dignidad de la Procreación.* En: Vatican.va. Roma.

Consejo de Organizaciones Internacionales de las Ciencias Médicas (2002). *Pautas Éticas Internacionales para la Investigación Biomédica en Seres Humanos.* Organización Mundial de la Salud. Ginebra.

Consejo General de Colegios Oficiales de Médicos (2011). *Código de Deontología Médica. Guía de Ética Médica.* Organización Médica Colegial. Madrid.

Cox, D. R. (1952). Estimation by double sampling. *Biometrika* **39**, 217-227.

Cramér, H. (1953). *Métodos Matemáticos de Estadística.* Aguilar. Madrid.

Daly, J. F.; Eckler, R. A. (1962). Applications of electronic equipment to statistical data-processing in the U. S. Bureau of the Census. *Bulletin of the International Statistical Institute* **39**, 319-327.

De Finetti, B. (1974). *Theory of Probability.* Volume 1. Wiley. Chichester.

De Finetti, B. (1975). *Theory of Probability.* Volume 2. Wiley. Chichester.

DEMING, W. E. (1960). *Sample Design in Business Research.* Wiley. New York, NY.

DEMING, W. E. (1966). *Some Theory of Sampling.* Dover. New York, NY.

DEMING, W. E.; GLASSER, G. J. (1959). On the problem of matching lists by samples. *Journal of the American Statistical Association* **54**, 403-415.

DESABIE, J. (1962). *Théorie et Pratique des Sondages.* Institut Nationale de la Statistique et des Études Économiques. Paris.

DESROSIÈRES, A. (1997). The administrator and the scientist: how the statistical profession has changed. *Statistical Journal of the United Nations Economic Commission for Europe* **14**, 31-50.

DESROSIÈRES, A. (2000). Measurement and its uses: harmonization and quality in social statistics. *International Statistical Review* **68**, 173-187.

DODGE, Y.; ROUSSON, V. (1999). The complications of the fourth central moment. *The American Statistician* **53**, 267-269.

DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **15**, 262-269.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **16**, 152-164.

DUSSAIX, A. M.; GROSBRAS, J. M. (1992). *Exercices de Sondages.* Économica. Paris.

Dussaix, A. M.; Grosbras, J. M. (1993). *Les Sondages: Principes et Méthodes.* Presses Universitaires de France. Paris.

Editorial (2014). What lies beneath. *Nature* **507**, 273.

Editorial DDB (1999). *Biblia de Jerusalén en Letra Grande* (9ª edición revisada y aumentada). Desclée De Brouwer. Bilbao.

Fernández García, F. R.; Mayor Gallego, J. A. (1995a). *Ejercicios y Prácticas de Muestreo en Poblaciones Finitas.* Ediciones Universitarias de Barcelona. Barcelona.

Fernández García, F. R.; Mayor Gallego, J. A. (1995b). *Muestreo en Poblaciones Finitas: Curso Básico.* Ediciones Universitarias de Barcelona. Barcelona.

Fienberg, S. E.; Martin, M. E.; Straf, M. L. (1995). The Committee on National Statistics: fostering interactions between statisticians in Academia and Gobernment. *International Statistical Review* **63**, 257-269.

Foreman, E. K. (1991). *Survey Sampling Principles.* Dekker. New York, NY.

Foreman, E. K.; Brewer, K. R. W. (1971). The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **33**, 391-400.

Fuller, W. A. (2009). *Sampling Statistics*. Wiley. New York, NY.

Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika* **71**, 171-175.

Garthwaite, P.; Jolliffe, I.; Jones, B. (2002). *Statistical Inference* (2nd edition). Oxford University Press. New York, NY.

GLASSER, G. J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute* **30**, 28-32.

GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **17**, 269-278.

GOOD, P. I.; HARDIN, J. W. (2006). *Common Errors in Statistics (and How to Avoid Them)* (2nd edition). Wiley. Hoboken, NJ.

GOODE, W. J.; HATT, P. K. (1952). *Methods in Social Research.* McGraw-Hill. New York, NY.

GOURIEROUX, C. (1981). *Théorie des Sondages.* Économica. Paris.

GOVINDARAJULU, Z. (1999). *Elements of Sampling Theory and Methods.* Prentice Hall. Upper Saddle River, NJ.

GROSBRAS, J. M. (1987). *Méthodes Statistiques des Sondages.* Économica. Paris.

HALDANE, J. B. S. (1945). On a method of estimating frequencies. *Biometrika* **33**, 222-225.

HANSEN, M. H.; HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics* **14**, 333-362.

HANSEN, M. H.; HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* **41**, 517-529.

HANSEN, M. H.; HURWITZ, W. N.; MADOW, W. G. (1953). *Sample Survey Methods and Theory* (Volume II. *Theory*). Wiley. New York, NY.

HANURAV, T. V. (1966). Some aspects of unified sampling theory. *Sankhyā: The Indian Journal of Statistics, Series A* **28**, 175-204.

HARTLEY, H. O.; RAO, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics* **33**, 350-374.

HARTLEY, H. O.; ROSS, A. (1954). Unbiased ratio estimators. *Nature* **174**, 270-271.

HEDAYAT, A. S.; SINHA, B. K. (1991). *Design and Inference in Finite Population Sampling.* Wiley. New York, NY.

HENDRICKS, W. A. (1956). *The Mathematical Theory of Sampling.* Scarecrow Press. New Brunswick, NJ.

HIDIROGLOU, M. A. (2001). Double sampling. *Survey Methodology* **27**, 143-154.

HORVITZ, D. G.; THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

INDRAYAN, A. (2013). *Medical Biostatistics* (3rd edition). Chapman & Hall/CRC. Boca Raton, FL.

IOANNES PAULUS PP. II (1991). *Carta Encíclica CENTESIMUS ANNUS.* Libreria Editrice Vaticana. Roma.

IZQUIERDO URBINA, C. (2015). *Teología Fundamental* (4ª edición). EUNSA. Pamplona.

JESSEN, R. J. (1978). *Statistical Survey Techniques.* Wiley. New York, NY.

KISH, L. (1965). *Survey Sampling.* Wiley. New York, NY.

KISH, L. (1979). *Muestreo de Encuestas.* Trillas. México.

KUPPER, L. L.; NEELON, B. H.; O'BRIEN, S. M. (2011). *Exercises and Solutions in Biostatistical Theory.* Chapman & Hall/CRC. Boca Raton, FL.

LAHIRI, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute* **33** (2), 133-140.

LAVALLÉE, P. (2007). *Indirect Sampling.* Springer. New York, NY.

LEJEUNE, M. (2010). *Statistique. La Théorie et Ses Applications* (2ième edition). Springer. Paris.

LOHR, S. L. (2010). *Sampling: Design and Analysis* (2nd edition). Brooks Cole. Boston, MA.

MARTÍNEZ, V. (1999). Diseño de encuestas de opinión: barómetro CIS. *Qüestiió* **23**, 343-362.

MENGAL, P. (1999). *Statistique Descriptive Appliquée aux Sciences Humaines* (6ème edition). Verlag Peter Lang. Berne.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics* **2**, 99-108.

MIRÁS AMOR, J. (1985). *Elementos de Muestreo para Poblaciones Finitas*. Instituto Nacional de Estadística. Madrid.

MURGUI IZQUIERDO, J. S.; ESCUDER VALLÉS, R. (1994). *Estadística Aplicada. Inferencia Estadística.* Tirant lo Blanch. Valencia.

MURTHY, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā: The Indian Journal of Statistics, Series A* **18**, 379-390.

MURTHY, M. N. (1964). Product method of estimation. *Sankhyā: The Indian Journal of Statistics, Series A* **26**, 69-74.

MURTHY, M. N. (1967). *Sampling Theory and Methods.* Statistical Publishing Society. Calcutta.

NARAIN, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**, 169-174.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558-606.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33**, 101-116.

NIELSEN, P. B.; PLOVSING, J. (1997). Concepts used in statistical business registers in view of globalization and the information society. *International Statistical Review* **65**, 351-363.

NUZZO, R. (2014). Statistical errors. *Nature* **506**, 150-152.

OLIVE, D. J. (2014). *Statistical Theory and Inference.* Springer. Cham.

OLIVERA RAVASI, J. (2015). Aprendiendo a pensar: lógica de los sofismas (21 artículos). En: InfoCatolica.com. Pamplona.

O'HAGAN, A. (1994). *Kendall's Advanced Theory of Statistics.* Volume 2B. *Bayesian Inference.* Arnold. London.

Pablo, Obispo de la Iglesia Católica (1965). *Constitución Pastoral GAUDIUM ET SPES sobre la Iglesia en el Mundo Actual*. En: Vatican.va. Roma.

Pathak, P. K. (1962). On simple random sampling with replacement. *Sankhyā: The Indian Journal of Statistics, Series A* **24**, 287-302.

Pécseli, H. L. (2000). *Fluctuations in Physical Systems*. Cambridge University Press. Cambridge.

Piantadosi, S. (2005). *Clinical Trials. A Methodologic Perspective* (2nd edition). Wiley. Hoboken, NJ.

Plane, D. R.; Gordon, K. R. (1982). A simple proof of the nonapplicability of the Central Limit theorem to finite populations. *The American Statistician* **36**, 175-176.

Raj, D. (1968). *Sampling Theory*. McGraw-Hill. New York, NY.

Ramakrishnan, M. K. (1969). Some results on the comparison of sampling with and without replacement. *Sankhyā: The Indian Journal of Statistics, Series A* **31**, 333-342.

Rao, J. N. K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association* **58**, 202-215.

Rao, J. N. K. (1975). Unbiased variance estimation for multistage designs. *Sankhyā: The Indian Journal of Statistics, Series C* **37**, 133-137.

Rao, J. N. K.; Hartley, H. O.; Cochran, W. G. (1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **24**, 482-491.

RAO, J. N. K.; LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika* **71**, 387-395.

RAO, T. J. (1967). On the choice of a strategy for the ratio method of estimation. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **29**, 392-397.

REAL ACADEMIA ESPAÑOLA (1992). *Diccionario de la Lengua Española* (21ª edición). Espasa Calpe. Madrid.

REARDON, S. (2014). NIH rethinks psychiatry trials. *Nature* **507**, 288.

RÍOS GARCÍA, S. (1977). *Métodos Estadísticos* (2ª edición). Ediciones del Castillo. Madrid.

ROHATGI, V. K. (1984). *Statistical Inference.* Wiley. New York, NY.

RUEDA GARCÍA, M. M.; ARCOS CEBRIÁN, A. (1999). *Problemas de Muestreo en Poblaciones Finitas.* Grupo Editorial Universitario. Granada.

RUIZ ESPEJO, M. (1985). Equiprecisional allocation and optimum stratification. *Statistics* **16** (4), 559-562.

RUIZ ESPEJO, M. (1986a). Funciones paramétricas estimables en teoría de muestras. *Estadística Española* **28** (112-113), 69-73.

RUIZ ESPEJO, M. (1986b). Sesgo de no respuesta en el intento *n*. *Estadística Española* **28** (112-113), 75-78.

RUIZ ESPEJO, M. (1987a). A control in stratified sampling. *Statistics* **18** (2), 287-291.

RUIZ ESPEJO, M. (1987b). Diseños muestrales admisibles para el estimador Horvitz-Thompson. *Trabajos de Estadística* **2** (1), 45-50.

RUIZ ESPEJO, M. (1987c). Sobre estimadores UMV y UMECM en poblaciones finitas. *Estadística Española* **29** (115), 105-111.

RUIZ ESPEJO, M. (1988a). Estimación insesgada con observaciones erradas y no respuesta. *Trabajos de Estadística* **3** (1), 71-80.

RUIZ ESPEJO, M. (1988b). *Estudio de Modelos de Muestreo Aplicado*. Tesis Doctoral en Ciencias Matemáticas por la Universidad Complutense de Madrid. Madrid.

RUIZ ESPEJO, M. (1990). Una clase de estimadores de la media poblacional robustos e invariantes lineales. *Metron* **48** (1-4), 55-66.

RUIZ ESPEJO, M. (1991). El estimador producto generalizado. *Estadística Española* **33** (127), 285-290.

RUIZ ESPEJO, M. (1993). Nuevos estimadores de la varianza en poblaciones finitas. *Qüestiió* **17** (2), 203-219.

RUIZ ESPEJO, M. (1994). Distribución del número de unidades distintas en una muestra aleatoria simple con reemplazamiento de una población finita. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **49**, 117-118.

RUIZ ESPEJO, M. (1995). Una relación entre cuasicovarianzas muestral y poblacional. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **50**, 51-53.

RUIZ ESPEJO, M. (1996). Distribución del número de extracciones con reemplazamiento para obtener una muestra de costo fijo

de una población finita. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **51**, 77-79.

RUIZ ESPEJO, M. (1997a). Covariance of sample moments for finite populations. *Proceedings of the Royal Irish Academy, Section A – Mathematical and Physical Sciences* **97** (2), 163-167.

RUIZ ESPEJO, M. (1997b). El teorema de Rao-Blackwell en poblaciones finitas e implicaciones informativo-económicas. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **52**, 71-74.

RUIZ ESPEJO, M. (1997c). Nota sobre estrategias muestrales para estudios socioeconómicos en España. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **52**, 63-65.

RUIZ ESPEJO, M. (1997d). Sobre la cuasicovarianza muestral en el muestreo aleatorio simple. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **52**, 55-57.

RUIZ ESPEJO, M. (1997e). Una metodología para el control por muestreo de cuentas económicas. *Estudios de Economía Aplicada* **7** (2), 159-180.

RUIZ ESPEJO, M. (1997f). Uniqueness of the Zinger strategy with estimable variance: Rana-Singh estimator. *Sankhyā: The Indian Journal of Statistics, Series B* **59** (1), 76-83.

RUIZ ESPEJO, M. (1998a). Estrategias óptimas de muestreo: una revisión. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **53**, 207-210.

RUIZ ESPEJO, M. (1998b). *Renta Familiar Declarada Media por Comunidades Autónomas y España (1988-1992): Estimación y Aplicaciones*. Tesis Doctoral en Ciencias Económicas y Empresariales por la Universidad Nacional de Educación a Distancia. Madrid.

RUIZ ESPEJO, M. (1998c). Unbiased estimation using auxiliary information. *Statistica Applicata – Italian Journal of Applied Statistics* **10** (3), 433-436.

RUIZ ESPEJO, M. (2003a). *Observaciones a los Métodos Estadísticos de Investigación del Bienestar Social en el Marco Global*. Tesis Doctoral en Sociología por la Universidad Pontificia de Salamanca. Madrid.

RUIZ ESPEJO, M. (2003b). Review of *Sample Survey Theory: Some Pythagorean Perspectives* (by Paul Knottnerus; Springer-Verlag, New York, NY, 2003). *Biometrics* **59** (4), 1193.

RUIZ ESPEJO, M. (2007). Review of *Common Errors in Statistics (and How to Avoid Them)*, 2nd edn (by Phillip I. Good and James W. Hardin; Wiley, Hoboken NJ, 2006). *Journal of Applied Statistics* **34**, 366.

RUIZ ESPEJO, M. (2008). Optimality of srswr in the class of all ordered ppsm designs of fixed size for sample mean estimator. *Statistical Reports* **2**, 1-7.

RUIZ ESPEJO, M. (2011a). An objective solution to the problem of unbiased estimation with nonresponse. *Statistical Reports* **13**, 1-2.

RUIZ ESPEJO, M. (2011b). Review of *Statistique: La Théorie et Ses Applications*, 2ème édn (by Michel Lejeune; Springer, Paris, 2010). *International Statistical Review* **79**, 501.

RUIZ ESPEJO, M. (2012). Review of *Randomized Response and Indirect Questioning Techniques in Surveys* (by Arijit Chaudhuri; Chapman & Hall/CRC, Boca Raton, FL, 2011). *Biometrics* **68** (4), 1329-1330.

RUIZ ESPEJO, M. (2013a). *Ejercicios de Estadística*. Bubok. Madrid.

RUIZ ESPEJO, M. (2013b). Estimador de regresión lineal corregido insesgado. *Statistical Reports* **19**, 1-4.

RUIZ ESPEJO, M. (2013c). *Exactitud de la Inferencia en Poblaciones Finitas*. Bubok. Madrid.

RUIZ ESPEJO, M. (2013d). Objective unbiased variance estimation with nonresponse: a review. *Statistical Reports* **18**, 1-10.

RUIZ ESPEJO, M. (2013e). Review of *Design and Analysis of Experiments in the Health Sciences* (by Gerald van Belle and Kathleen F. Kerr; Wiley, Hoboken NJ, 2012). *Journal of Applied Statistics* **40** (12), 2778-2779.

RUIZ ESPEJO, M. (2013f). Review of *Sampling*, 3rd edn (by Steven K. Thompson; Wiley, Hoboken NJ, 2012). *Journal of Applied Statistics* **40** (4), 920-921.

RUIZ ESPEJO, M. (2013g). Una demostración del estimador insesgado de la varianza en presencia de no respuesta. *Statistical Reports* **17**, 1-5.

RUIZ ESPEJO, M. (2014a). *Investigación Ética y Bioestadística*. Trabajo Fin de Máster en Bioética por la Universidad Católica San Antonio. Murcia.

RUIZ ESPEJO, M. (2014b). Objective unbiased variance estimation with systematic sampling of double start. *Statistical Reports* **20**, 1-13.

RUIZ ESPEJO, M. (2014c). Review of *Handling Missing Data in Ranked Set Sampling* (by Carlos N. Bouza-Herrera; Springer, Heidelberg, 2013). *International Statistical Review* **82**, 157-158.

RUIZ ESPEJO, M. (2014d). Review of *Medical Biostatistics* (3rd edition by Abhaya Indrayan; Chapman & Hall/CRC, Boca Raton, FL, 2013). *Journal of Applied Statistics* **41** (4), 911.

RUIZ ESPEJO, M. (2015a). Estimación insesgada del error cuadrático medio del ajuste lineal multivariante objetivo. *Statistical Reports* **22**, 1-7.

RUIZ ESPEJO, M. (2015b). Estimación insesgada objetiva para no respuesta. *Estadística Española* **57** (186), 29-37.

RUIZ ESPEJO, M. (2015c). Regresión lineal multivariante objetiva en poblaciones finitas. *Statistical Reports* **21**, 1-12.

RUIZ ESPEJO, M. (2015d). Review of *Les Techniques de Sondage* (by Pascal Ardilly; Technip, Paris, 2006). *Technometrics* **57** (2), 292.

RUIZ ESPEJO, M. (2015e). Review of *Modern Survey Sampling* (by Arijit Chaudhuri; Chapman & Hall/CRC, Boca Raton FL, 2014). *Journal of Applied Statistics* **42** (4), 917-918.

RUIZ ESPEJO, M. (2015f). Review of *Practical Tools for Designing and Weighting Survey Samples* (by Richard Valliant, Jill A. Dever and Frauke Kreuter; Springer, New York NY, 2013). *Journal of Official Statistics* **31** (4), 813-815.

RUIZ ESPEJO, M. (2015g). Sampling schemes providing unbiased mean-of-the-ratios estimates: a review. *Estadística Española* **57** (187), 133-139.

RUIZ ESPEJO, M. (2015h). Sobre estimación insesgada óptima del cuarto momento central poblacional. *Estadística Española* **57** (188), 287-290.

RUIZ ESPEJO, M. (2016a). Estimación de regresión multivariante insesgada. *Estadística Española* **58** (190), 123-131.

RUIZ ESPEJO, M. (2016b). Estimadores producto generalizados corregidos insesgados. *Statistical Reports* **23**, 1-12.

RUIZ ESPEJO, M. (2016c). Unbiased corrected classic estimates. *Estadística Española* **58** (189), 43-56.

RUIZ ESPEJO, M. (2017a). *Bioestadística Ética.* Lulu Press. Raleigh, NC.

RUIZ ESPEJO, M. (2017b). *Ciencia del Muestreo.* Bubok. Madrid.

RUIZ ESPEJO, M. (2017c). Cuestiones éticas de la bioestadística médica objetiva. *Statistical Reports* **25**, 1-11.

RUIZ ESPEJO, M. (2017d). Estimación de la desviación estándar. *Estadística Española* **59** (192), 37-44.

RUIZ ESPEJO, M. (2017e). Estudios de salud pública como ciencia empírica objetiva de datos. *Statistical Reports* **26**, 1-9.

RUIZ ESPEJO, M. (2017f). *Fundamentos de la Inferencia Estadística Objetiva* (4ª edición). Lulu Press. Raleigh, NC.

RUIZ ESPEJO, M. (2017g). Review of *Handbook of Design and Analysis of Experiments* (Angela Dean, Max Morris, John Stufken and Derek Bingham, eds; Chapman & Hall/CRC, Boca Raton FL, 2015). *Statistical Reports* **24**, 1-2.

RUIZ ESPEJO, M. (2018a). Ciencia e idealismo en estadística. *Estadística Española* **60** (197), 325-332.

RUIZ ESPEJO, M. (2018b). Covarianza de la cuasivarianza y la media muestrales. *Estadística Española* **60** (196), 159-163.

RUIZ ESPEJO, M. (2018c). El consenso como medio de acuerdo de expertos. *Estadística Española* **60** (197), 333-341.

RUIZ ESPEJO, M. (2018d). Muestreo doble óptimo. *Estadística Española* **60** (196), 151-157.

RUIZ ESPEJO, M. (2018e). Recensión de *Ciencia del Muestreo* (por Mariano Ruiz Espejo; Bubok, Madrid, 2017). *Statistical Reports* **27**, 1-4.

RUIZ ESPEJO, M. (2018f). Recientes frutos en bioestadística. *Estadística Española* **60** (195), 61-84.

RUIZ ESPEJO, M. (2018g). Reseña de *Análisis del Impacto de los Programas de Mejora de la Calidad Educativa en Centros Escolares Públicos* (por Laura López-Torres, Diego Prior y Daniel Santín; Centro de Estudios Ramón Areces, Madrid, 2017). *Revista Complutense de Educación* **29** (2), 631-632.

RUIZ ESPEJO, M. (2018h). Tratamiento científico de la no respuesta en encuestas. *Statistical Reports* **29**, 1-6.

RUIZ ESPEJO, M. (2018i). Un método general de estimación insesgada de la varianza. *Estadística Española* **60** (195), 49-59.

RUIZ ESPEJO, M. (2018j). Una demostración sencilla de la varianza de la cuasivarianza muestral. *Estadística Española* **60** (196), 219-224.

RUIZ ESPEJO, M. (2019a). Creencias y práctica religiosa de los investigadores. *Statistical Reports* **30**, 1-17.

RUIZ ESPEJO, M. (2019b). Optimal unbiased estimation of some parametric functions. *Statistical Reports* **31**, 1-14.

Ruiz Espejo, M. (2020a). Convergencia en ley de una sucesión de variables aleatorias uniformes discretas a cualquier variable aleatoria. *Statistical Reports* **34**, 1-6.

Ruiz Espejo, M. (2020b). Encuestas: la ciencia teórica y la práctica moral. *Statistical Reports* **32**, 1-10.

Ruiz Espejo, M. (2021). Reflexiones éticas sobre la enseñanza universitaria de estadística. *Statistical Reports* **36**, 1-14.

Ruiz Espejo, M. (2022a). Algunas referencias sobre estimación insesgada óptima. *Statistical Reports* **37**, 1-3.

Ruiz Espejo, M. (2022b). Reseña de *La Calidad del Profesorado en la Adquisición de Competencias de los Alumnos. Un Análisis Basado en PIRLS-2011* (por Jorge Calero y J. Oriol Escardíbul; Centro de Estudios Ramón Areces, Madrid, 2017). *Statistical Reports* **39**, 1-3.

Ruiz Espejo, M. (2022c). Sobre la optimización del estimador de regresión lineal. *Statistical Reports* **48**, 1-4.

Ruiz Espejo, M. (2022d). Sobre la supuesta imposibilidad de investigar y practicar la fe. *Statistical Reports* **41**, 1-9.

Ruiz Espejo, M. (2024a). Contribuciones españolas al muestreo insesgado de poblaciones finitas. *Statistical Reports* **59**, 1-6.

Ruiz Espejo, M. (2024b). Contribuciones españolas al muestreo objetivo de poblaciones finitas. *Statistical Reports* **58**, 1-6.

Ruiz Espejo, M. (2025). Una demostración sencilla del estimador de regresión lineal insesgado óptimo. *Statistical Reports* **68**, 1-3.

Ruiz Espejo, M.; Delgado Pineda, M. (2006). Review of *Estimation in Surveys with Nonresponse* (by Carl-Erik Särndal & Sixten

Lundström; Wiley, Chichester, 2005). *Journal of the American Statistical Association* **101** (474), 854.

Ruiz Espejo, M.; Delgado Pineda, M. (2008a). Analysis of variance experimental designs with checkable hypothesis: a reflection. *Statistical Reports* **4**, 1-21.

Ruiz Espejo, M.; Delgado Pineda, M. (2008b). On the efficiency of stratified pps sampling. *Statistical Reports* **6**, 1-10.

Ruiz Espejo, M.; Delgado Pineda, M. (2008c). Stratified population variance estimation. *Statistical Reports* **5**, 1-12.

Ruiz Espejo, M.; Delgado Pineda, M.; Nadarajah, S. (2003). Estimation of finite population parameters with several realizations. *Statistical Papers* **44** (2), 267-278.

Ruiz Espejo, M.; Delgado Pineda, M.; Nadarajah, S. (2013). Optimal unbiased estimation of some population central moments. *Metron* **71** (1), 39-62.

Ruiz Espejo, M.; Delgado Pineda, M.; Nadarajah, S. (2016). Optimal unbiased estimation of some population central moments. *Metron* **74** (1), 139.

Ruiz Espejo, M.; Delgado Pineda, M.; Singh, H. P. (2006). Postgrouped sampling method of estimation. *Test* **15** (1), 209-226.

Ruiz Espejo, M.; Marqués Vilallonga, A. (2018). Review of *Principles of Scientific Methods* (by Mark Chang; Chapman & Hall/CRC, Boca Raton FL, 2015). *Journal of Applied Statistics* **45** (4), 775-776.

Ruiz Espejo, M.; Rueda García, M. M. (1993). Un esquema muestral sin reemplazamiento inmediato. *Revista de la*

*Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **48**, 145-151.

Ruiz Espejo, M.; Santos Peñas, J. (1989a). Estrategias intermedias de muestreo. *Estadística Española* **31** (121), 227-235.

Ruiz Espejo, M.; Santos Peñas, J. (1989b). Unbiased mean-of-the-ratios estimators. *Statistica* **49** (4), 617-622.

Ruiz Espejo, M.; Santos Peñas, J. (1990). Sampling design providing unbiased new product estimators. *Statistica* **50** (2), 285-288.

Ruiz Espejo, M.; Santos Peñas, J.; Rueda García, M. M. (1995). Optimum linear integration of independent sample means from finite populations. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **50**, 55-60.

Ruiz Espejo, M.; Singh, H. P. (2001). Unbiased and optimal linear estimators for some superpopulation models. *Revista de la Academia de Ciencias Exactas, Físicas, Químicas y Naturales de Zaragoza, Serie 2ª* **56**, 99-109.

Ruiz Espejo, M.; Singh, H. P. (2003). Protection of privacy with objective prior distribution in randomized response. *Statistica* **63** (4), 697-701.

Ruiz Espejo, M.; Singh, H. P.; Saxena, S. (2008). On inverse sampling without replacement. *Statistical Papers* **49** (1), 133-137.

Ruiz Espejo, M.; Singh, H. P.; Singh, R. (2001). Optimal unbiased linear integration of estimators. *Statistics and Risk Modeling* **19** (4), 373-394.

SAMPATH, S. (2001). *Sampling Theory and Methods.* Narosa. New Delhi.

SAMPFORD, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499-513.

SÁNCHEZ-CRESPO RODRÍGUEZ, J. L. (1976). *Muestreo de Poblaciones Finitas Aplicado al Diseño de Encuestas* (2ª edición). Instituto Nacional de Estadística. Madrid.

SÁNCHEZ-CRESPO RODRÍGUEZ, J. L. (1980). *Curso Intensivo de Muestreo en Poblaciones Finitas* (2ª edición). Instituto Nacional de Estadística. Madrid.

SÁNCHEZ-CRESPO RODRÍGUEZ, J. L.; DE PARADA HERRERO, J. (1990). *Ejercicios y Problemas Resueltos de Muestreo en Poblaciones Finitas.* Instituto Nacional de Estadística. Madrid.

SÄRNDAL, C. E.; LUNDSTRÖM, S. (2005). *Estimation in Surveys with Nonresponse.* Wiley. Hoboken, NJ.

SCHMIDTZ, D.; GOODIN, R. E. (2000). *El Bienestar Social y la Responsabilidad Individual.* Cambridge University Press. Madrid.

SCHUMAN, H.; SCOTT, J. (1987). Problems in the use of survey questions to measure public opinion. *Science* **280**, 957-959.

SEN, A. R. (1953). On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119-127.

SENG, Y. P. (1951). Historical survey of the development of sampling theories and practice. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **114**, 214-231.

SETH, G. R.; RAO, J. N. K. (1964). On the comparison between simple random sampling with and without replacement. *Sankhyā: The Indian Journal of Statistics, Series A* **26**, 85-86.

SGRECCIA, E. (2012). *Manual de Bioética I: Fundamentos y Ética Biomédica.* Biblioteca de Autores Cristianos. Madrid.

SGRECCIA, E. (2014). *Manual de Bioética II: Aspectos Médico-Sociales.* Biblioteca de Autores Cristianos. Madrid.

SINGH, R.; MANGAT, N. S. (1996). *Elements of Survey Sampling.* Kluwer. Dordrecht.

SOM, R. K. (1996). *Practical Sampling Techniques* (2nd edition). Dekker. New York, NY.

SPROTT, D. A. (2000). *Statistical Inference in Science.* Springer-Verlag. New York, NY.

STUART, A. (1984). *The Ideas of Sampling* (3rd edition). Oxford University Press. New York, NY.

STUART, A.; ORD, J. K. (1994). *Kendall's Advanced Theory of Statistics*. Volume 1. *Distribution Theory* (6th edition). Edward Arnold. London.

STUART, A.; ORD, J. K.; ARNOLD, S. F. (1999). *Kendall's Advanced Theory of Statistics.* Volume 2A. *Classical Inference and the Linear Model.* Arnold. London.

THIONET, P. (1953). *La Théorie des Sondages.* Imprimerie Nationale. Paris.

THIONET, P. (1958). *La Théorie des Sondages.* INSÉÉ. Paris.

THOMPSON, M. E. (1997). *Theory of Sample Surveys.* Chapman & Hall. London.

THOMPSON, S. K. (2012). *Sampling* (3rd edition). Wiley. Hoboken, NJ.

TILLÉ, Y. (2001). *Théorie des Sondages.* Dunod. Paris.

TILLÉ, Y. (2006). *Sampling Algorithms.* Springer. New York, NY.

TREVIJANO ETCHEVERRIA, P. (2011). Moral y ciencias. En: InfoCatolica.com. Pamplona.

TSCHUPROW, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2**, 461-493, 646-683.

TUCKER, H. G. (1998). *Mathematical Methods in Sample Surveys.* World Scientific. Singapore.

VAN BELLE, G.; KERR, K. F. (2012). *Design and Analysis of Experiments in the Health Sciences.* Wiley. Hoboken, NJ.

VAN TUINEN, H. K.; ALTENA, J. W.; IMBENS, H. (1994). Surveys, registers and integration in social statistics. *Statistical Journal of the United Nations Economic Commission for Europe* **11**, 321-356.

WARNER, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**, 63-69.

WILKS, S. S. (1960). A two-stage scheme for sampling without replacement. *Bulletin of the International Statistical Institute* **37** (2), 241-248.

WINKEL, P.; ZHANG, N. F. (2007). *Statistical Development of Quality in Medicine.* Wiley. Chichester.

WOLTER, K. M. (2007). *Introduction to Variance Estimation* (2nd edition). Springer. New York, NY.

WORLD HEALTH ORGANIZATION (2000). *Operational Guidelines for Ethics Committees that Review Biomedical Research.* World Health Organization. Geneve.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys* (4th edition). Griffin. London.

YATES, F.; GRUNDY, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **15**, 235-261.

YOUNG, G. A.; SMITH, R. L. (2010). *Essentials of Statistical Inference*. Cambridge University Press. Cambridge.

ZACKS, S. (1971). *The Theory of Statistical Inference.* Wiley. New York, NY.

ZARKOVICH, S. S. (1965). *Sampling Methods and Censuses.* Food and Agricultural Organization. Rome.