Estadística Española • Volumen 60, número 195 / 2018, pp. 61-84

Recientes frutos en bioestadística

Mariano Ruiz Espejo
Universidad Católica San Antonio de Murcia

Resumen

Presentamos un conjunto de procedimientos bioestadísticos objetivos que consideramos de gran utilidad para la metodología actual. Entre ellos están la desmitificación de la distribución normal y otras similares como hipótesis de trabajo, el tratamiento objetivo de la no respuesta, el análisis de la varianza objetivo, y el ajuste lineal multivariante objetivo.

Palabras clave: bioestadística objetiva, distribución normal, no respuesta, análisis de la varianza, ajuste lineal multivariante.

Clasificación AMS: 62D05, 62-02, 62J10, 62P10, 97M60.

Recent fruits in biostatistics

Abstract

We present a set of objective biostatistical procedures that we consider of great utility for actual methodology. Between them they are the demythification of the normal distribution and other similar ones as work hypothesis, the objective treatment of the nonresponse, the objective analysis of variance, and the objective multivariate linear fit.

Keywords: objective biostatistics, normal distribution, nonresponse, analysis of variance, multivariate linear fit.

AMS classification: 62D05, 62-02, 62J10, 62P10, 97M60.

1. Introducción

En este artículo resumimos unas aportaciones del autor de los últimos años que considero que abren puertas a una investigación más rigurosa en bioestadística y en sus vertientes médica, psicológica, etc. por lo que pueden considerarse de un alto interés práctico por su objetividad en aspectos que incumben a todos como es el cuidado de la salud con medios técnicos correctos de autentificación estadística como apoyo a otras ciencias y con el fin común de actuar con el fin de preservar la salud de los seres humanos.

Otras aplicaciones posibles de estas aportaciones están en la agricultura, alimentación, estadística oficial, sociología, etc.

2. Hacia una bioestadística más objetiva

El objeto de esta sección es indicar un abuso muy frecuente en el uso de la Bioestadística en Medicina y Ciencias de la Vida, el cual está presente en muchas publicaciones en nuestros días.

Lo refiero a la suposición manipulada y reductiva de que los datos obtenidos en un estudio estadístico experimental están originados por observaciones independientes e idénticamente distribuidas de una distribución o población Normal pero sin comprobar estas suposiciones en la práctica. De este modo, el estudio es hipotético y podría no ser objetivo para su propósito.

En tales suposiciones hay una falsa convicción muy posiblemente de que la naturaleza está trabajando para nosotros o que todas nuestras suposiciones teóricas a nivel conceptual de abstracción (como la suposición de normalidad en los datos obtenidos, y como que la muestra está obtenida por observaciones independientes e idénticamente distribuidas de una población normal) serían correctas porque la naturaleza reproducirá estas suposiciones naturalmente en los estudios médicos experimentales.

Estas últimas suposiciones son hipotéticas solamente y sus hipótesis no están demostradas en cada estudio rigurosamente. Por esta razón, las conclusiones derivadas de tales suposiciones podrían estar confinadas a un estudio teórico pero sin implicaciones reales y objetivas. Esta es la situación actual generalizada en medicina experimental en nuestros días. Una razón es que los paquetes estadísticos comercializados para analizar los datos están basados en tales suposiciones teóricas.

La experimentación clínica ha sido estudiada desde un punto de vista bioético por Ciccone (2006) y Sgreccia (2012). Para que la teoría estadística funcione en experimentación clínica, la selección aleatoria de unidades observacionales debe ser hecha con selección probabilística. No vale tomar el primer paciente a mano, el segundo, etc. como si ellos fueran el orden de selección en la muestra. Esto sería una burda manera de seleccionar la muestra, porque estos pacientes no serían seleccionados probabilísticamente. Además la posible selección de la muestra de una *población normal* es realizable solamente con procedimientos de simulación artificial con ordenador, la muestra no sería del mundo real en el cual estamos interesados en inferir sobre un parámetro de ella.

Por todas estas razones concluimos que la base de poblaciones normales en medicina clínica es una burda aproximación y no es objetiva para el mundo práctico real. Otra razón es que para obtener una muestra aleatoria probabilística de una población, es necesario disponer de un marco de la población (por ejemplo, pacientes con una enfermedad, todos los voluntarios disponibles, etc.). Un marco de unidades de una población es siempre finito porque no podemos tener a mano un marco de infinitas unidades con el modo de acceso a ellas para observarlas y tomar los datos.

Además, de una población normal o gaussiana el número de unidades sería infinito, lo que es una hipótesis no real. De una población infinita es posible seleccionar una muestra artificial, pero es incorrecto decir que "la muestra de pacientes que tengo en mi estudio médico es una muestra aleatoria probabilística de la población de pacientes" porque el azar simple no garantiza una selección probabilística. Y, si no tenemos selección probabilística, los teoremas de inferencia estadística no funcionan en la práctica.

La única solución a estos dilemas es basar las inferencias en poblaciones finitas de pacientes o voluntarios (Ruiz Espejo, 2013, 2014b). Cualquier otra solución incluye errores burdos que no son controlados por la estadística objetiva. Este es el caso de la mayor parte de experimentos clínicos en la medicina actual.

3. Causas de la inconsistencia de la hipótesis normal

En esta sección, el autor indica un conjunto de causas de la inconsistencia de la hipótesis normal en bioestadística médica:

1. Una población infinita, como toda distribución normal con desviación estándar positiva, no puede ser listada en un marco de todas sus unidades, usualmente pacientes o voluntarios. La única posibilidad para hacer un marco poblacional es desde una población finita disponible, por ejemplo, el conjunto de pacientes de una enfermedad en un país o en un hospital.
2. Sin un marco poblacional, es imposible la selección de una muestra probabilística natural. La posibilidad de seleccionar datos artificiales en una muestra de una población normal no resuelve el problema para una población finita de pacientes.
3. Sin un marco poblacional, es imposible asegurar observaciones independientes e idénticamente distribuidas de una población normal porque no se puede controlar en la práctica estas propiedades estadísticas. Sin marco, el investigador no puede acceder a una determinada unidad o paciente porque no se tiene la posibilidad real de localizar y observar tal paciente si este no está identificado y accesible.
4. La única posibilidad que conozco para hacer útil un marco poblacional finito de pacientes para inferencias estadísticas objetivas es con unidades (pacientes, voluntarios, etc.) identificados y accesibles de una determinada población de ellos.
5. No existe muestra probabilística imparcial solamente vía selección al azar de unidades, lo cual es el caso de observaciones tratadas con supuestas poblaciones normales para las que la selección es a propósito, intencional, o por fácil disponibilidad usualmente.
6. La selección artificial de la muestra con propiedades probabilísticas y estadísticas determinadas puede ser realizable en el caso de poblaciones finitas. La muestra no sería artificial porque la selección artificial es para los identificadores de las unidades en la muestra, pero los datos estadísticos serían obtenidos por observación de unidades reales previamente identificadas.

7. Con marcos de poblaciones finitas sería posible identificar y acceder a las unidades seleccionadas de la muestra. Con poblaciones infinitas, no existe un marco para identificar y acceder a la muestra seleccionada de unidades a ser observadas. Por esto, una población infinita es una suposición que no puede ser comprobada en poblaciones naturales.
8. Con muestreo de poblaciones finitas es posible disponer del marco poblacional, identificar y acceder a las unidades de la población y a las unidades observadas en la muestra, por ejemplo con confidencialidad de pacientes en consulta, y con tratamiento anónimo de los datos estadísticos naturales observados en ella.
9. La distribución de una variable de una población finita es "uniforme discreta", y generalmente no es simétrica y no es mesocúrtica como lo sería una distribución normal teórica.
10. El *teorema central del límite* clásico que asegura una distribución asintótica normal para el estadístico *media aritmética muestral*, no es seguro en la práctica porque necesita las hipótesis de distribución independiente e idéntica, pero éstas no son comprobables si la población supuesta es infinita y si la población real es natural, debido a la no identificabilidad de todas las unidades de una población infinita a ser seleccionadas en la muestra. Otro argumento fue proporcionado por Plane y Gordon (1982).

Desde estos argumentos hemos visto la clara debilidad e inconsistencia de basar la bioestadística médica en hipótesis normales, cuando existe otro método más objetivo para estos propósitos, como es el muestreo de poblaciones finitas (Ruiz Espejo 2013, 2014a, 2014b; Cassel *et al.*, 1977).

4. Estimación insesgada con no respuesta

Esta sección está dedicada a los nuevos frutos en inferencia estadística objetiva los cuales pueden ser usados en bioestadística médica. En secciones anteriores he explicado la necesidad de una bioestadística más objetiva y las causas de la inconsistencia de la suposición normal para tratar datos estadísticos biomédicos. Mi consejo es tomar en cuenta la objetividad de la inferencia estadística basada en muestras probabilísticas de una población finita de unidades identificadas y accesibles (Thompson, 2012; Ruiz Espejo, 2013).

Sin embargo, mientras las inferencias clásica o bayesiana en bioestadística no resuelven el problema de la no respuesta en las unidades muestreadas, por ejemplo cuando un paciente abandona el servicio de consulta de un médico, la inferencia en poblaciones finitas ha dado varias soluciones estadísticas a este problema. Las principales referencias de soluciones matemáticas están proporcionadas en dos artículos de Ruiz Espejo (2011, 2015b).

En concreto, cuando en una "muestra aleatoria simple con reemplazamiento" seleccionada aparecen algunas que no responden o no pueden ser observadas, es posible tener una estimación insesgada de la "media de la población finita de la variable de interés" que es el objetivo del estudio médico bioestadístico, e incluso es posible tener una estimación insesgada de la varianza del primer estimador insesgado.

La solución pasa por la disponibilidad de una submuestra probabilística (una "muestra aleatoria simple con reemplazamiento") de los no respondientes en la primera "muestra aleatoria simple con reemplazamiento" sobre todo el marco de la población finita. Este procedimiento podría ser implementado siguiendo un número reducido de estos no respondientes en las nuevas consultas médicas.

De tal modo, es realizable estimar con intervalos de confianza objetivos aproximados la función paramétrica de interés "media de la población finita" (Ruiz Espejo, 2013). Esta solución fue iniciada en un artículo de Hansen y Hurwitz (1946) con un estimador insesgado de la media poblacional, concretamente (siendo $\bar{y}_{(1)}$ la media muestral de respuestas, $\bar{y}_{(2)}$ la media submuestral de respuestas obtenidas de la submuestra de la muestra de no respuestas, $w_1 = n_1/n$ es la proporción muestral de respuestas de entre la muestra inicial, y $w_2 = 1 - w_1 = n_2/n$ es la proporción muestral de unidades que no responden en la muestra inicial)

$$\bar{y}_{nr} = w_1\bar{y}_{(1)} + w_2\bar{y}_{(2)},$$

y es en los últimos años y ahora cuando soluciones metodológicas objetivas han sido dadas, por ejemplo con el *estimador insesgado de la varianza* del estimador de Hansen y Hurwitz (1946) proporcionado en Ruiz Espejo (2011) ante el problema de la no respuesta,

$$\hat{V}(\bar{y}_{nr}) = \frac{1}{m-1}\left\{\sum_{h=1}^{2} w_h\widehat{\sigma_h^2} + \sum_{h=1}^{2} w_h\left[\bar{y}_{(h)}^2 - \hat{V}\left(\bar{y}_{(h)}\right)\right] - \bar{y}_{nr}^2\right\}$$

$$+\frac{\widehat{\sigma_2^2}}{(m-1)n_{(2)}}(mw_2^2 - w_2),$$

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, $m_1 = n_1 \geq 2$ y $m_2 = n_2 \geq 2$; $m = m_1 + m_2$. De hecho, el número total de respuestas es $n_1 + n_{(2)}$. Además, el estimador $\hat{V}\left(\bar{y}_{(1)}\right) = s_1^2/n_1$, $n_{(2)}$ es el tamaño submuestral en el segundo estrato o número de respuestas de la submuestra de entre las unidades muestrales que no respondieron, y

$$\hat{V}\left(\bar{y}_{(2)}\right) = \frac{\widehat{\sigma_2^2}}{n_2} + \frac{s_{(2)}^2}{n_{(2)}} = s_{(2)}^2\left[\frac{1}{n_2 - 1} + \frac{1}{n_{(2)}}\right].$$

Pero el uso de estos métodos en medicina y ciencias de la salud, para mejorar la calidad y la objetividad de la metodología estadística aplicable, todavía no ha sido concretado. Unas referencias son el libro y el artículo de Ruiz Espejo (2017a, 2018).

Como el "muestreo de población finita fijada" es un método objetivo de inferencia estadística, y estas investigaciones matemáticas están incluidas en esta metodología objetiva, es de sentido común implementar estos avances metodológicos en la práctica médica y clínica sobre otras inferencias estadísticas menos objetivas las cuales son

ampliamente más usadas en la actualidad pero con suposiciones subjetivas e incomprobables en sus modelos estadísticos.

Otra solución para la estimación insesgada de la varianza con no respuesta fue propuesta por Thompson (2012), concretamente para el "muestreo aleatorio simple sin reemplazamiento" básico como diseño muestral en la primera y en la segunda fases de muestreo.

Estos serían buenos medios para inferir sobre la media poblacional como función paramétrica objetivo de una investigación biomédica o clínica con no respuesta, es decir, buenos medios para y con buenos fines.

5. Análisis de la varianza objetivo

En esta sección presentamos un método objetivo de análisis de la varianza cuando se conocen el número de unidades disponibles sobre las que experimentar para cada tratamiento o celda con la técnica de estratificación de las unidades por tratamientos y por celdas en su caso. Los trabajos originales para ello se pueden consultar en Ruiz Espejo y Delgado Pineda (2008).

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 ε_{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

$$\bar{X}_{t\cdot} = \frac{1}{N_t} \sum_{i=1}^{N_t} X_{ti}$$

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 tomar n_t constante independientemente del tratamiento t, aunque no necesario. En estas condiciones, la descomposición del modelo estudiado de diseños experimentales de un factor será

$$X_{ti} = \bar{X} + (\bar{X}_{t\cdot} - \bar{X}) + (X_{ti} - \bar{X}_{t\cdot})$$

Donde $A = \bar{X}$, $B_t = \bar{X}_{t\cdot} - \bar{X}$ y $\varepsilon_{ti} = X_{ti} - \bar{X}_{t\cdot}$, 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 $t = 1, 2, \dots, T$,

$$\sum_{i=1}^{N_t} \varepsilon_{ti} = 0$$

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 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 tomar n_{ti} constante independientemente de la celda en que se experimente, aunque tampoco sería necesario.

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 ε_{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

$$s_{ti\cdot}^2 = \frac{1}{n_{ti}-1}\sum_{j=1}^{n_{ti}}\left(x_{tij} - \bar{x}_{ti\cdot}\right)^2$$

Un estimador insesgado de F_t es

$$\hat{F}_t = \bar{x}_{t\cdot\cdot} - \bar{x}_{st}$$

Donde

$$\bar{x}_{t\cdot\cdot} = \frac{1}{n_{t\cdot}}\sum_{i=1}^{I}\sum_{j=1}^{n_{ti}} x_{tij} = \frac{1}{n_{t\cdot}}\sum_{i=1}^{I} n_{ti}\bar{x}_{ti\cdot}$$

Y

$$n_{t\cdot} = \sum_{i=1}^{I} n_{ti}$$

Además se puede comprobar que

$$\sum_{t=1}^{T} N_{t\cdot}F_t = \sum_{t=1}^{T} N_{t\cdot}(\bar{X}_{t\cdot\cdot} - \bar{X}) = N(\bar{X} - \bar{X}) = 0.$$

La varianza de $\hat{F}_t$ se obtiene como sigue

$$V\left(\hat{F}_t\right) = \frac{1}{N^2}\left[(N - N_{t\cdot})^2 V(\bar{x}_{t\cdot\cdot}) + \sum_{k\neq t}^{T} N_{k\cdot}^2 V(\bar{x}_{k\cdot\cdot})\right]$$

Donde para $t = 1, 2, \ldots, T$,

$$V(\bar{x}_{t\cdot\cdot}) = \sum_{i=1}^{I}\left(\frac{n_{ti}}{n_{t\cdot}}\right)^2 V(\bar{x}_{ti\cdot})$$

También un estimador insesgado de la varianza es

$$\hat{V}\left(\hat{F}_t\right) = \frac{1}{N^2}\left[(N - N_{t\cdot})^2 \hat{V}(\bar{x}_{t\cdot\cdot}) + \sum_{k\neq t}^{T} N_{k\cdot}^2 \hat{V}(\bar{x}_{k\cdot\cdot})\right]$$

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

$$\hat{V}(\bar{x}_{\cdot i \cdot}) = \sum_{t=1}^{T} \left(\frac{n_{ti}}{n_{\cdot i}} \right)^2 \hat{V}(\bar{x}_{ti \cdot})$$

Una estimación insesgada de la interacción $(FC)_{ti}$ es

$$\widehat{(FC)}_{ti} = \bar{x}_{ti \cdot} - \bar{x}_{t \cdot \cdot} - \bar{x}_{\cdot i \cdot} + \bar{x}_{st}$$

Y su varianza viene proporcionada por la expresión

$$V\left[\widehat{(FC)}_{ti}\right] =$$

$$V(\bar{x}_{ti \cdot}) - Cov(\bar{x}_{ti \cdot}, \bar{x}_{t \cdot \cdot}) - Cov(\bar{x}_{ti \cdot}, \bar{x}_{\cdot i \cdot}) + Cov(\bar{x}_{ti \cdot}, \bar{x}_{st}) +$$

$$V(\bar{x}_{t \cdot \cdot}) - Cov(\bar{x}_{t \cdot \cdot}, \bar{x}_{ti \cdot}) + Cov(\bar{x}_{t \cdot \cdot}, \bar{x}_{\cdot i \cdot}) - Cov(\bar{x}_{t \cdot \cdot}, \bar{x}_{st}) +$$

$$V(\bar{x}_{\cdot i \cdot}) - Cov(\bar{x}_{\cdot i \cdot}, \bar{x}_{ti \cdot}) + Cov(\bar{x}_{\cdot i \cdot}, \bar{x}_{t \cdot \cdot}) - Cov(\bar{x}_{\cdot i \cdot}, \bar{x}_{st}) +$$

$$V(\bar{x}_{st}) - Cov(\bar{x}_{st}, \bar{x}_{t \cdot \cdot}) - Cov(\bar{x}_{st}, \bar{x}_{\cdot i \cdot}) + Cov(\bar{x}_{st}, \bar{x}_{ti \cdot}).$$

Ahora se puede calcular todos los sumandos del segundo miembro de la expresión anterior. Conocemos los valores de las varianzas $V(\bar{x}_{ti\cdot})$, $V(\bar{x}_{t\cdot\cdot})$, $V(\bar{x}_{\cdot i\cdot})$ y

$$V(\bar{x}_{st}) = \frac{1}{N^2}\sum_{t=1}^{T}\sum_{i=1}^{I} N_{ti}^2 V(\bar{x}_{ti\cdot}).$$

Y de las covarianzas

$$Cov(\bar{x}_{ti\cdot}, \bar{x}_{t\cdot\cdot}) = Cov(\bar{x}_{t\cdot\cdot}, \bar{x}_{ti\cdot}) = \frac{N_{ti}}{N_{t\cdot}} V(\bar{x}_{ti\cdot}),$$

$$Cov(\bar{x}_{\cdot i\cdot}, \bar{x}_{ti\cdot}) = Cov(\bar{x}_{ti\cdot}, \bar{x}_{\cdot i\cdot}) = \frac{N_{ti}}{N_{\cdot i}} V(\bar{x}_{ti\cdot}),$$

$$Cov(\bar{x}_{t\cdot\cdot}, \bar{x}_{\cdot i\cdot}) = Cov(\bar{x}_{\cdot i\cdot}, \bar{x}_{t\cdot\cdot}) = \frac{N_{ti}^2}{N_{t\cdot} N_{\cdot i}} V(\bar{x}_{ti\cdot}),$$

$$Cov(\bar{x}_{ti\cdot}, \bar{x}_{st}) = Cov(\bar{x}_{st}, \bar{x}_{ti\cdot}) = \frac{N_{ti}}{N} V(\bar{x}_{ti\cdot}),$$

$$Cov(\bar{x}_{t\cdot\cdot}, \bar{x}_{st}) = Cov(\bar{x}_{st}, \bar{x}_{t\cdot\cdot}) = \frac{1}{N_{t\cdot} N}\sum_{i=1}^{I} N_{ti}^2 V(\bar{x}_{ti\cdot}),$$

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.

6. Ajuste lineal multivariante objetivo

La teoría desarrollada en esta sección ha sido propuesta por Ruiz Espejo (2015a, 2015c). 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

$$\phi = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - k_0 - \sum_{r=1}^{m} k_r x_{r,i} \right)^2 .$$

Para minimizar este error cuadrático total (o equivalentemente el error cuadrático medio, ϕ/N), derivamos parcialmente la función ϕ 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

$$\begin{cases} A_{1;y} = k_0 + \sum_{j=1}^{m} k_j A_{1;\,x_j} \\ A_{1,1;y,x_r} = k_0 A_{1;\,x_r} + k_r A_{2;\,x_r} + \sum_{\substack{j=1 \\ j \neq r}}^{m} k_j A_{1,1;\,x_j,x_r} \\ r = 1, 2, \dots, m. \end{cases}$$

Aquí $A_{h,z}$ es el momento no central de orden h poblacional para la variable z, y $A_{1,1;y,x}$ es el momento no central conjunto de órdenes 1 y 1 poblacional de las variables y y x,

$$A_{h,z} = \frac{1}{N} \sum_{i=1}^{N} z_i^h,$$

$$A_{1,1;y,x} = \frac{1}{N} \sum_{i=1}^{N} y_i x_i$$

y

$$A_{1,1;\,x_j,x_r} = \frac{1}{N} \sum_{i=1}^{N} x_{j,i} x_{r,i}.$$

Este último 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. El vector $\widehat{\boldsymbol{a}}$ puede obtenerse directamente de modo insesgado y óptimo siguiendo los razonamientos expuestos por Ruiz Espejo *et al.* (2013; 2016) y Ruiz Espejo (2015d) en el caso de muestreo aleatorio simple con reemplazamiento.

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

$$\hat{\bar{y}} = \widehat{\boldsymbol{k}}\bar{\boldsymbol{x}}^t = \widehat{\boldsymbol{a}}\boldsymbol{A}^{-1}\bar{\boldsymbol{x}}^t.$$

Donde $\overline{\boldsymbol{x}}^t$ es la matriz del vector columna de dimensiones $(m+1)\times 1$, que es la matriz traspuesta de la matriz $\overline{\boldsymbol{x}} = (1 \quad \bar{x}_1 \quad \bar{x}_2 \quad \cdots \quad \bar{x}_m)_{1\times(m+1)}$. Con $\bar{x}_r = (1/N)\sum_{i=1}^{N} x_{r,i}$, que es la media poblacional de la variable auxiliar r-ésima $(r = 1, 2, \dots, m)$. Dicho estimador $\hat{\bar{y}}$ es insesgado para ajustar el modelo óptimo (de mínimo error cuadrático total poblacional), pero en general podría ser supuestamente sesgado para estimar la media poblacional $\bar{y}$. Además $\hat{\bar{y}}$ es insesgado óptimo en el sentido de distribución libre (Zacks, 1971, p. 150), es decir, es un estimador uniformemente de mínima varianza e insesgado (UMVU estimator) para distribución libre. El estimador clásico de mínimos cuadrados recogido en la expresión $(X^tX)^{-1}X^tY$ no garantizaba la insesgación del estimador.

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 λ 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 $\bar{y}$. Puede considerarse una válida alternativa al estimador insesgado, en las mismas condiciones de información auxiliar disponible, propuesto en Ruiz Espejo (2016).

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

$$V(e) = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N}\{E[E(\hat{e}_i|s)]\}^2 = E\{[E(\hat{e}_i)]^2\} =$$

$$\frac{1}{N}\sum_{i=1}^{N}[E(\hat{e}_i)]^2 = \frac{1}{N}\sum_{i=1}^{N} E(\hat{e}_i^2) - \frac{1}{N}\sum_{i=1}^{N} V(\hat{e}_i).$$

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

$$\hat{V}(e) = \frac{1}{2n}\sum_{i\in s''}\left(\hat{e}_i^2 + \hat{e}'^2_i\right) + \frac{1}{n}\sum_{i\in s''}(\hat{e}_i - \hat{e}'_i)^2.$$

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

$$\varepsilon=\sqrt{\frac{\hat{V}\left[\widehat{ECM}(s)\right]}{\alpha}}=\sqrt{\frac{N-n}{\alpha Nn(n-1)}\sum_{i\in s''}\left(\hat{e}_i^2-\frac{1}{n}\sum_{i\in s''}\hat{e}_i^2\right)^2}.$$

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 teórico 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.

Finalmente indicamos que el ajuste lineal multivariante objetivo tiene un valor aproximativo, pero no predictivo. Un ejemplo es que el valor predictivo de una unidad de la muestra seleccionada es ya conocido con exactitud, mientras que su aproximación mediante el ajuste lineal multivariante puede contener errores que pueden ser salvables por la observación de la unidad en la muestra.

7. Conclusiones

Hemos resumido un conjunto de aportaciones recientes a la estadística objetiva que consideramos de alto interés en áreas clave para el desarrollo de la sociedad. El reto es poner en práctica estos desarrollos especialmente en el sistema de salud y en la estadística oficial. Confío en que la profesionalidad de los estadísticos perciba las ventajas de estos procedimientos que superan con holgura a los métodos clásicos que pueden quedar como objetos de museo en la historia de la ciencia. Un compendio de todas estas y otras técnicas objetivas es el de Ruiz Espejo (2017b).

Referencias

CASSEL, CLAES-MAGNUS; SÄRNDAL, CARL-ERIK; & WRETMAN, JAN HAKAN (1977). «*Foundations of Inference in Survey Sampling*». Wiley. New York, NY.

CICCONE, LINO (2006). «*Bioética. Historia. Principios. Cuestiones*», Segunda Edición. Palabra. Madrid.

HANSEN, M. H.; & HURWITZ, W. N. (1946). «The problem of nonresponse in sample surveys». *Journal of the American Statistical Association* 41, 517-529.

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.

RUIZ ESPEJO, MARIANO (2011). «An objective solution to the problem of unbiased estimation with nonresponse». *Statistical Reports* 13, 1-2.

RUIZ ESPEJO, MARIANO (2013). «*Exactitud de la Inferencia en Poblaciones Finitas*». Bubok. Madrid.

RUIZ ESPEJO, MARIANO (2014a). «*Fundamentos de la Inferencia Estadística Objetiva*», Tercera Edición. Lulu Press. Raleigh, NC.

RUIZ ESPEJO, MARIANO (2014b). «*Investigación Ética y Bioestadística*», Segunda Edición. Lulu Press. Raleigh, NC.

RUIZ ESPEJO, MARIANO (2015a). «Estimación insesgada del error cuadrático medio del ajuste lineal multivariante objetivo». *Statistical Reports* 22, 1-7.

RUIZ ESPEJO, MARIANO (2015b). «Estimación insesgada objetiva para no respuesta». *Estadística Española* 57 (186), 29-37.

RUIZ ESPEJO, MARIANO (2015c). «Regresión lineal multivariante objetiva en poblaciones finitas». *Statistical Reports* 21, 1-12.

RUIZ ESPEJO, MARIANO (2015d). «Sobre estimación insesgada óptima del cuarto momento central poblacional». *Estadística Española* 57 (188), 287-290.

RUIZ ESPEJO, MARIANO (2016). «Estimación de regresión multivariante insesgada». *Estadística Española* 58 (190), 123-131.

RUIZ ESPEJO, MARIANO (2017a). «*Bioestadística Ética*». Lulu Press. Raleigh, NC.

RUIZ ESPEJO, MARIANO (2017b). «*Ciencia del Muestreo*». Bubok. Madrid.

RUIZ ESPEJO, MARIANO (2018). «Tratamiento científico de la no respuesta en encuestas». *Statistical Reports* 29, 1-6.

RUIZ ESPEJO, MARIANO; DELGADO PINEDA, MIGUEL; & NADARAJAH, SARALEES (2013). «Optimal unbiased estimation of some population central moments». *Metron* 71, 39-62.

RUIZ ESPEJO, MARIANO; DELGADO PINEDA, MIGUEL; & NADARAJAH, SARALEES (2016). «Optimal unbiased estimation of some population central moments». *Metron* 74, 139.

RUIZ ESPEJO, MARIANO; & DELGADO PINEDA, MIGUEL (2008). «Analysis of variance experimental designs with checkable hypothesis: a reflection». *Statistical Reports* 4, 1-21.

SGRECCIA, ELIO (2012). «*Manual de Bioética I. Fundamentos y Ética Biomédica*». Biblioteca de Autores Cristianos. Madrid.

THOMPSON, STEVEN K. (2012). «*Sampling*», Tercera Edición. Wiley. Hoboken, NJ.

ZACKS, SHELEMYAHU (1971). «*The Theory of Statistical Inference*». Wiley. New York, NY.