Facultad de Ciencias Sociales y Jurídicas

Grado en Administración y Dirección de Empresas

Universidad de Jaén

Facultad de Ciencias Sociales y Jurídicas

Trabajo Fin de Grado

DISEÑO MUESTRALES EN EL TIEMPO

Alumno: Gema María Ramos Reyes

Junio, 2022

RESUMEN

En el presente Trabajo de Fin de Grado se desarrolla la teoría del muestro en ocasiones sucesivas, complementándolo con los estimadores indirectos de doble muestreo (razón, regresión, diferencia y producto), con el fin de poder disminuir el error de muestreo al estimar, en esta ocasión, la media poblacional. Se incluye ejemplos prácticos para ayudar al lector a comprender dicha teoría, además de las funciones de los estimadores indirectos de doble muestreo en ocasiones sucesivas en el software estadístico R Studio.

ABSTRACT

The aim of this Final Degree Project is tos tate the theory of sampling on successive occasions, complementing it with the double sampling indirect estimators (ratio, regression, difference and product), in order to be able to reduce the sampling error when estimating, on this occasion, the population mean. Practical examples are included to help the reader understand this theory, as well as the functions of the double sampling indirect estimators on successive occasions in the statistical software R Studio.

ÍNDICE

1. INTRODUCCIÓN

Desde los inicios del muestreo, principios del siglo XX, el objetivo principal del mismo ha sido poder encontrar un diseño muestral que mejore la precisión de las estimaciones de los parámetros como la media, el total o la varianza.

Inicialmente, surgieron diseños como el muestro aleatorio simple (muestreo sistemático, estratificado, por conglomerados, etc.) y el muestro no probabilístico, denominados "Métodos directos". Estos diseños solo dependen de la variable a estudiar, pero en la mayoría de ocasiones se dispone de información acerca de una población de la que se tiene conocimiento y que está relacionada con la variable a estudiar, la cual, esta información, se introduce en el diseño denominada variable auxiliar. Tras realizar la fase de estimación mediante estimadores indirectos, se puede llegar a obtener estimaciones más precisas que con los métodos directos si la característica a estudiar está fuertemente ligada a la variable auxilia, surgiendo métodos indirectos como estimación por regresión, la estimación de razón, la de producto y por diferencias.

1.1. Muestreo doble

Los métodos indirectos anteriormente citados consideran que el total o la media poblacional de la variable auxiliar es conocido. Existen ocasiones en la que esta información la desconocemos, por lo que el conocimiento de su distribución no es la ideal para obtener buenas precisiones, utilizándose, entonces, el muestreo doble.

En el muestreo doble se realizan las estimaciones necesarias a la característica auxiliar a partir de muestras previas para obtener estimaciones eficientes de la variable auxiliar, aunque esto signifique la reducción de la muestra.

Pero, ¿compensaría la posible ganancia de precisión con el coste necesario para obtener aquella información a partir de una muestra grande?

Fernández y Mayor (1994) obtiene el decrecimiento de muestra $n_0 - n$, en función de c y c':

$$n_0 - n = \frac{c'}{c} n'$$

siendo c’ = coste por unidad de la muestra n’; c = coste por unidad de la muestra n.

Si la ganancia obtenida con la información complementaria permite compensar la pérdida dada y mejorar la precisión del estimador, el muestreo doble será ventajoso, es decir, la muestra obtenida con la variable auxiliar debe de ser mucho menor que el coste de la variable principal.

Técnica

Neyman (1938) (*Fernández y Mayor* (1994)) propuso, como muestreo doble, el siguiente proceso secuencial:

1. En una primera fase se obtiene una muestra de la población $U = \{u_1, u_2, \ldots, u_N\}$ obteniéndose una muestra de m’, de tamaño n’ según el diseño muestral d_1. Esta información, en el caso de la estimación indirecta, la utilizaremos como información auxiliar para todo el proceso.
2. En una segunda fase, tomando la primera fase como población, se obtiene una muestra m, de tamaño n, a partir de la muestra anterior, que ahora se actúa como población y como diseño muestral d_2. Con la información proporcionada por esta segunda muestra se podrá realizar las estimaciones del parámetro a estudiar utilizando la información auxiliar obtenida previamente.

Tabla 1. Fases muestreo doble

	Tamaño poblacional	**Tamaño muestral**
Primera fase	N	n’
Segunda fase	n’	n

Elaboración propia

Aplicaciones de los estimadores indirectos

Diseño muestral

Suponiendo que el diseño muestral de la primera fase, d', es un muestreo aleatorio simple MAS (N, n') y que el diseño muestral de la segunda fase, d, es MAS (n', n), notamos con

n, tamaño muestral en la segunda ocasión

$\hat{R} = \frac{\bar{y}}{\bar{x}'}$, estimador de razón

$S_x^2 = \frac{\sum(x_i - \bar{x}')^2 \cdot n_i}{n-1}$, cuasivarianza de la primera ocasión

$S_y^2 = \frac{\sum(y_i - \bar{y}')^2 \cdot n_i}{n-1}$, cuasivarianza de la segunda ocasión

$S_{yx} = \frac{\Sigma xy - n\bar{x}\bar{y}}{n-1}$, cuasidesviación típica de *y* sobre *x*

$\rho = \frac{s_{yx}}{s_y s_x}$, coeficiente de correlación de Pearson

Estimador de razón

Suponiendo que el diseño muestral de la primera fase, d', es un muestreo aleatorio simple MAS (N, n') y que el diseño muestral de la segunda fase, d, es MAS (n', n), el estimador de razón para media poblacional de la variable de estudio es

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \cdot \bar{x}' = \hat{R} \cdot \bar{x}'$$

con varianza

$$V(\bar{y}_R) = \frac{N - n'}{Nn'} S_y^2 + \frac{n' - n}{n'n} (S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

Estimador de regresión

Suponiendo que el diseño muestral de la primera fase, d', es un muestreo aleatorio simple MAS (N, n') y que el diseño muestral de la segunda fase, d, es MAS (n', n), el estimador de regresión para media poblacional de la variable de estudio es

$$\bar{y}_{Reg} = \bar{y} + k(\bar{x}' - \bar{x})$$

donde *k* es una constante prefija de antemano.

El valor mínimo de la varianza se obtendrá para un valor concreto de la constante *k*. Dicha constante, su valor mínimo, es:

$$k = \frac{S_{yx}}{S_x^2}$$

La varianza vale

$$V(\bar{y}_{Reg}) = \frac{N - n'}{Nn'} S_y^2 + \frac{n' - n}{n'n} S_y^2 (1 - \rho^2)$$

2. MUESTREO SUCESIVO EN DOS OCASIONES

2.1. Concepto teórico

En el análisis de una muestra es importante destacar el instante de tiempo al que hacen referencia los resultados muestrales, debido a que los elementos de la población pueden verse modificada, es decir, puede existir nuevos individuos que entren a formar parte de la población (nacimientos) o, por el contrario, dejar de hacerlo (muertes).

Las encuestas continuas son utilizadas en situaciones en donde la composición de la población es alterada, por lo que es necesario utilizar dos o más instantes de tiempo para obtener una validez en los resultados, teniendo como principal ventaja poder reflejar con la máxima rapidez los cambios de una característica poblacional, a través de la diferencia de nivel de una característica entre la ocasión actual y la siguiente.

Así mismo, las circunstancias de la encuesta y las características que se pretende estimar son determinantes para elegir un tipo de diseño muestral. Entre ellas, existen varias posibilidades:

Tabla 2. Tipos de muestreo

	Definición
Muestreo repetido	Extraer una nueva muestra en cada ocasión
Muestreo panel	Utilizar la misma muestra en todas las ocasiones
Muestreo en ocasiones sucesivas	Realizar un **reemplazamiento parcial** de unidades de una ocasión a otra

Elaboración propia

Según las circunstancias del estudio, la utilización de un muestreo u otro diferirá. En este sentido, para poder utilizar la información muestral de la ocasión precedente, es necesario tener una muestra que cuente, en los dos periodos de tiempo sucesivos, con algunos elementos comunes, a lo que se denomina solapamiento parcial.

Este solapamiento puede lograrse a través de la rotación de la muestra. *Wilks* (*Jiménez,* 1967) define este concepto como el proceso de eliminar ciertos elementos antiguos de la muestra y añadir nuevos, pudiendo, estos, haber estado antes en la muestra.

Utilizar el reemplazamiento parcial de la muestra conlleva algunas ventajas:

- Reduce costes, utilizar una muestra completamente nueva en cada ocasión es excesivamente costoso.
- Aumenta la precisión de los estimadores.
- La permanencia indefinida de las mismas unidades en la muestra puede crear problemas y reducir la eficiencia de los estimadores.

En la práctica, este concepto se indica a través del índice de solapamiento, *p*, la proporción de muestra que tienen en común dos instantes de tiempo consecutivos. En cambio, a la proporción en la que las muestras no son comunes se denomina fracción de reemplazamiento, *q*.

El Instituto Nacional de Estadística (INE), por ejemplo, realiza la Encuesta de Población Activa trimestralmente en el cual se utiliza un sistema en el que cada unidad elemental permanece durante seis trimestres en la muestra, contando con un índice de solapamiento entre trimestres (p=0.83) consecutivos del 83%, es decir, la muestra entre dos trimestres consecutivos es común en un 83% y difiere en un 17%.

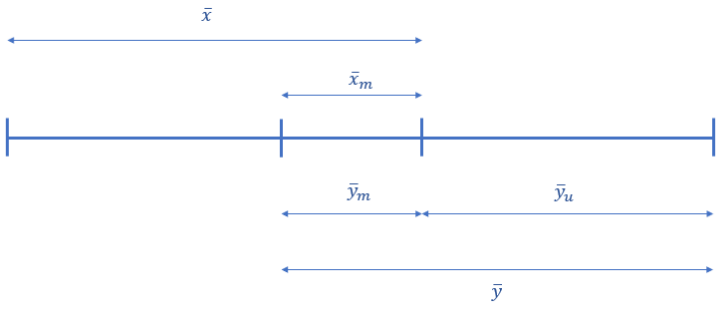
2.2. Diseño muestral

Suponemos que se realiza un muestreo aleatorio simple con reemplazamiento parcial, en donde se dispone información de la primera ocasión sobre una variable auxiliar, *x*, variable que debe de estar correlada con la variable de estudio, *y*.

Sea *N* el tamaño de la población, se realiza un muestreo doble donde se extrae una muestra aleatoria simple de tamaño *n'* en la primera ocasión. En una segunda ocasión se extrae una muestra aleatoria simple de tamaño *n*, en la cual se realiza otro muestreo en el que se extrae una muestra de tamaño *m* (muestra apareada) común en ambas fases, y las restantes $u = n - m$ (muestra no apareada) tomadas del universo $n' - m$ que queda tras omitir las *m* unidades.

El objetivo del muestreo sucesivo es obtener un estimador óptimo, en nuestro caso, el estimador de la media, combinando dos estimadores de las medias: un estimador indirecto de doble muestreo de la parte apareada de la muestra, y un estimador simple de la media de la parte no apareada.

Población N



Suponiendo que el diseño muestral de la primera ocasión, d', es un MAS (N, n') y que el diseño muestral de la segunda ocasión, d, es un MAS (n', m), notamos con

n el tamaño muestral total en cada ocasión

m, el tamaño muestral de aquellas unidades encuestadas en ambas ocasiones (muestra pareada)

u, el tamaño muestral de aquellas unidades no comunes en las dos ocasiones

$$u = n - m$$

$\bar{X}(\bar{Y})$ media poblacional en la primera (segunda) ocasión

$\bar{x}_m(\bar{y}_m)$ media muestral apareada en la primera (segunda) ocasión

$\bar{x}_u(\bar{y}_u)$, media muestral no apareada en la primera (segunda) ocasión

$p = \frac{m}{n}$, índice de solapamiento

$q = \frac{u}{n}$, fracción de reemplazamiento, tal que $p + q = 1$

ρ, coeficiente de correlación poblacional

$\Delta = \frac{S_x}{X} / \frac{S_y}{Y}$, razón del coeficiente de variación de la primera ocasión respecto de la segunda ocasión

$Z = \Delta(2\rho - \Delta)$

S_x^2 (S_y^2) cuasivarianza poblacional en la primera (segunda) ocasión.

Según (*Sen, Sellers y Smith* (1975)) las esperanzas de las medias muestrales apareadas y no apareadas son iguales:

$$E(\bar{x}_m) = E(\bar{x}_u) \qquad E(\bar{y}_m) = E(\bar{y}_u)$$

Puesto que las muestras son aleatorias, las medias muestrales serán estimadores insesgados de la media poblacional para cada ocasión:

$$E(\bar{x}_m) = x'_m \qquad E(\bar{y}_m) = x'_m$$
$$E(\bar{x}_m) = x'_m \qquad E(\bar{y}_m) = x'_m$$

2.3. Estimador de cambio

Como se ha mencionado anteriormente, el muestreo sucesivo combina dos estimadores, uno indirecto de la parte común y otro directo de la parte no común de la muestra.

Rodríguez, E. M. A., & Luengo, A. V. G. (2002) llegan a desarrollar este estimador, llegando a la simplificación del estimador cuando $S_x = S_y = S$

$$\Delta'_2 = \frac{p}{1-q\rho}(\bar{y}_m - \bar{x}_m) + \frac{q(1-\rho)}{1-q\rho}(\bar{y}_u - \bar{x}_u)$$

con varianza

$$V(\Delta'_2) = 2(1-\rho)\left(\frac{S^2}{n(1-q\rho)} - \frac{S^2}{N}\right)$$

2.4. Estimador de la suma de las medias

Hansen y otros (1953) obtienen el estimador óptimo de $\bar{Y} + \bar{X}$, el cual, dado el caso de que las varianzas sean iguales $\sigma_x = \sigma_y = \sigma$, el estimador de la suma de las medias es el siguiente:

$$z_2 = \frac{p}{1+q\rho}(\bar{x}_m + \bar{y}_m) + \frac{q(1+\rho)}{1+q\rho}(\bar{x}_u + \bar{y}_u)$$

con varianza

$$V(z_2) = \frac{2(1+\rho)}{1+q\rho}\frac{\sigma^2}{n}$$

2.5. ¿Cómo extraer las muestras en ocasiones sucesivas? Ejemplo práctico

Se expone a continuación, un ejemplo para entender mejor la extracción de la muestra a través de muestreo sucesivo:

Supongamos que un administrador de recursos forestales está interesado en estimar el número total de árboles muertos en un área de 200 parcelas y usa los datos del año anterior en 18 parcelas muestreadas con m.a.s. Luego se toma muestras al azar de 8 parcelas de estas primeras 18 parcelas y se realiza un recuento en estas últimas 8 parcelas.

Sea X el número de árboles muertos en la parcela en enero del año 2021 e Y el número de árboles muertos en enero de la parcela en el año 2022.

Tabla 3. Fases de muestreo

	Tamaño poblacional	**Tamaño muestral**
Primera fase (enero 2021)	N = 200	n' = 18
Segunda fase (enero 2022)	n' = 18	n = 8

Elaboración propia

Una vez muestreada la muestra *n*, se escoge al azar 5 árboles comunes en las dos fases, siendo esta $m = 5$, muestra apareada, y $u = 3$, muestra no apareada, muestreada del universo $n' - m$.

$n = 8$ → seleccionadas a través de m.a.s. en la población n', representadas en color amarillo

$m = 5$ → seleccionadas a través de m.a.s en la población n., representadas en negrita y letra roja.

$u = 3$ → seleccionadas a través de m.a.s. en la población $n' - m$, representadas en negrita y letra azul

Tabla 4. Muestra ejemplo 1

Nº parcelas muestreadas	Árboles muertos en enero 2021 (X)	Árboles muertos en enero 2022 (Y)
1	5	9
2	7	5
3	10	11
4	6	6
5	7	9
6	9	9
7	3	5
8	6	7
9	8	7
10	11	15
11	5	2
12	9	8
13	12	10
14	13	9
15	3	1
16	20	17
17	15	11
18	4	9

Elaboración propia

Por lo que las muestras en la segunda ocasión (Y) quedarían de la siguiente forma:

y = {5, 6, 9, 5, 2, 10, 1, 17}

y_m = {5, 6, 5, 10, 17}

y_u = {11, 7, 1}

3. ESTIMADORES INDIRECTOS EN MUESTREO SUCESIVO EN DOS OCASIONES

3.1. Método de razón

La estimación por razón mejora la precisión del estimador simple cuando existe una correlación positiva entre la variable auxiliar y la variable de estudio.

El estimador de la media por razón es:

$$\bar{y}_r = w\bar{y}'_m + (1-w)\bar{y}_u$$

El estimador dado, es la combinación de los estimadores de una parte apareada (m unidades) y una no apareada (u unidades), $\bar{y}_m$ y $\bar{y}_u$, respectivamente:

Tabla 5. Estimadores de la parte apareada y no apareada del método de razón

	Estimador de Y	Varianza
Parte apareada	$\bar{y}'_m = \frac{\bar{y}_m}{\bar{x}_m}\bar{x}' = \hat{R}\bar{x}'$	$S_y^2 \frac{n' - u'Z}{mn'}$
Parte no apareada	$\bar{y}_u$	$\frac{S_y^2}{u}$

Elaboración propia

El valor óptimo de *w* que minimiza la varianza de $\bar{y}_r$ vale

$$w_{opt} = \frac{V(\bar{\boldsymbol{y}}_{\boldsymbol{u}})}{V(\bar{\boldsymbol{y}}_{\boldsymbol{u}}) + \boldsymbol{V}(\bar{\boldsymbol{y}}'_{\boldsymbol{m}})} = \frac{\theta p}{\theta - Z(\theta - p)(1-p)}$$

y la varianza de $\bar{y}_r$ es:

$$V(\bar{y}_r) = w^2 V(\bar{y}'_m) + (1-w)^2 V(\bar{y}_u)$$

que, si sustituimos *w* por w_{opt}, obtenemos la varianza mínima:

$$V(\bar{y}_r)_{min} = \frac{S_y^2(n' - u'Z)}{n'n - Zu'u}$$

Ganancia de precisión de $\bar{\boldsymbol{y}}_{\boldsymbol{r}}$ sobre $\bar{\boldsymbol{y}}$:

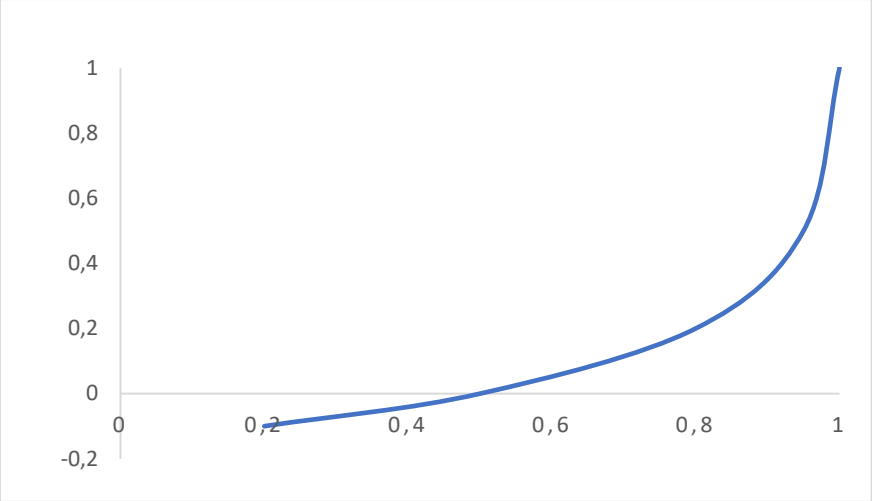
La ganancia de precisión relativa de un estimador respecto a otro la hallaremos dividiendo la varianza del segundo menos la varianza del primero entre la varianza del primero, pudiendo representar gráficamente la curva dada. Para que la ganancia del primero sobre el segundo sea positiva, nos debe dar un valor entre 0 y 1, el cual, si se multiplican por 100, obtendremos las ganancias en porcentajes.

Entonces, la ganancia de precisión de $\bar{y}_r$ sobre $\bar{y}$, la cual denominaremos como G1, se obtienen dividiendo**:**

$$G_1 = \frac{V(\bar{y}) - V(\bar{\boldsymbol{y}}_{\boldsymbol{r}})}{V(\bar{\boldsymbol{y}}_{\boldsymbol{r}})} = \frac{Zp(\theta - p)}{\theta - Z(\theta - p)}$$

La figura 3.1. muestra la ganancia de precisión G1, en función de ρ

Figura 3.1. Ganancia de precisión G1, en función de ρ



Elaboración propia

Por lo que, como demuestran *Sen, Sellers y Smith* (1975), el estimador de razón será más preciso que $\bar{y}$ siempre que $\rho > \frac{\Delta}{2}$., es decir siempre que $Z > 0$. La ganancia de precisión aumentará conforme aumenten los valores de θ y ρ, y Δ se aproxime a ρ.

3.2. Método de regresión

El objetivo del método de regresión es proporcionar el estimador óptimo para la media en la segunda ocasión, combinando un estimador de regresión de la media de la parte apareada con la media de la parte no apareada, para poder así, minimizar la varianza.

Tabla 6. Estimadores de la parte apareada y no apareada del método de regresión

	Estimador de $\bar{Y}$	Varianza
Parte apareada	$\bar{y}'_m = \bar{y}_m + b(\bar{x} - \bar{x}_m)$	$\frac{\sigma^2(1-\rho^2)}{m} + \frac{\sigma^2}{n} = \frac{1}{W_m}$
Parte no apareada	$\bar{y}_u$	$\frac{\sigma^2}{u} = \frac{1}{W_u}$

Elaboración propia

Siendo $b = \frac{S_{ymxm}}{S^2_{xm}}$, el coeficiente de regresión de las unidades muestrales de la parte común de la ocasión segunda sobre la primera y $\bar{x} = p\bar{x}_m + q\bar{x}_u$.

Al combinar ambos estimadores, el estimador de la media de regresión quedaría:

$$\bar{y}_{Reg} = \frac{\left[\bar{y}_m + \rho\frac{S_y}{S_x}(\bar{x}' - \bar{x}_m)\right] + u\bar{y}_u}{t + u}$$

Siendo

$$t = \frac{1}{\frac{\rho^2}{n'} + \frac{1-\rho^2}{m}}$$

con varianza

$$V(\bar{y}_{Reg}) = \frac{S^2_{\bar{y}}(n' - \rho^2 u')}{n'n - \rho^2 u' u}$$

Ganancia de precisión, G_2, de $\bar{y}_{Reg}$ sobre $\bar{y}_r$

$$G_2 = \frac{V(\bar{y}_r) - V(\bar{y}_{Reg})}{V(\bar{y}_{Reg})} = \frac{p\theta(\theta - p)(\rho^2 - Z)}{\{\theta - Z(\theta - p)(1-p)\}\{\theta - \rho^2(\theta - p)\}}$$

Con esta fórmula se demuestra que el estimador de regresión es consistentemente mejor que el estimador de razón, excepto cuando $\rho = \Delta$, en cuyo caso $G_2 = 0$ y es indiferente utilizar uno u otro.

3.3. Método de producto

Si la correlación existente entre la variable auxiliar y la variable objeto es negativa, el método de razón no es tan eficiente, por lo que *Artés y otros* (1998) [5] construyen un estimador óptimo para estimar la media en la segunda ocasión combinando un estimador producto de doble muestreo para la parte apareada de la muestra, con una media simple basada en la parte no apareada de la muestra en la segunda ocasión.

Tabla 7. Estimadores de la parte apareada y no apareada del método de producto

	Estimador de Y	Varianza
Parte apareada	$\bar{y}'_m = \frac{\bar{x}_m}{\bar{x}}\bar{y}_m$	$S_y^2 \frac{n + uZ}{mn}$
Parte no apareada	$\bar{y}_u$	$\frac{S_y^2}{u}$

Elaboración propia

El estimador de la media de producto es

$$\boldsymbol{\bar{y}_{prod} = w\bar{y}'_m + (1 - w)\bar{y}_u}$$

donde

$$w_{opt} = \frac{V(\bar{y}_u)}{V(\bar{y}_u) + V(\bar{y}'_m)} = \frac{p}{1 + (1 - p)^2 Z}$$

Su varianza es, por tanto

$$V(\bar{y}_{prod}) = \frac{V(\bar{y}'_m)V(\bar{y}_u)}{V(\bar{y}'_m) + V(\bar{y}_u)} = \frac{S_y^2(n + uZ)}{n^2 + Zu^2}$$

Ganancia de precisión, G2, de $\bar{y}_{Reg}$ sobre $\bar{y}_{prod}$

$$G_2 = \frac{V(\bar{y}_{Prod}) - V(\bar{y}_{Reg})}{V(\bar{y}_{Reg})} = \frac{p\theta(\theta - p)(\rho^2 - Z)}{\{\theta - Z(\theta - p)(1 - p)\}\{\theta - \rho^2(\theta - p)\}}$$

3.4. Método de diferencia

Se considera un estimador de diferencia para la parte apareada de doble muestreo y un estimador simple para la parte no apareada, siendo estos:

Tabla 8. Estimadores de la parte apareada y no apareada del método de diferencia

	Estimador de $\bar{Y}$	Varianza
Parte apareada	$\bar{y}_{2D} = \bar{y}_m - (\bar{x}_m - \bar{x})$	$\frac{S^2}{m}\left(1 + \left(1 - \frac{m}{n}\right)(1 - 2\rho)\right)$
Parte no apareada	$\bar{y}_{2u} = \bar{y}_u$	$\frac{S_y^2}{u}$

Elaboración propia

Así, combinando ambas partes, se queda el siguiente estimador de diferencia:

$$\bar{y}_{Dif} = w\bar{y}_{2D} + (1 - w)\,\bar{y}_{2u}$$

siendo su varianza

$$V(\bar{y}_{Dif}) = \frac{S^2}{n}\frac{1 + Zq}{1 + Zq^2}$$

Ganancia de precisión, G, de $\bar{y}_{Dif}$ sobre $\bar{y}$:

$$G = \frac{2}{1 + \sqrt{2}\sqrt{1 - \rho}} - 1$$

3.5. Ejemplo práctico

Para evaluar los métodos propuestos se han utilizado los datos recogidos en una investigación sobre el alcoholismo en los jóvenes durante y post pandemia de la COVID-19. Dicho estudio se ha llevado a cabo sobre los 1.021 alumnos de la facultad de Ciencias Sociales y Jurídicas de la Universidad de Jaén durante el mes de abril de 2020 y el mes de abril de 2022.

El estudio se ha basado en un muestreo sucesivo para que nos proporcione estimadores más precisos de las variables estudiadas, y para ello, se muestreó dos conjuntos de muestras aleatorias independientes:
1.- una muestra de 10 alumnos seleccionados en la primera ocasión (abril de 2020) de la población de 39 alumnos muestreados en la primera ocasión.
2.- una muestra de 12 alumnos seleccionados en la segunda ocasión (abril de 2022) de entre los 29 alumnos que no formaron parte de la muestra apareada.

Tabla 9. Fases del muestreo ejemplo 2

	Tamaño poblacional	**Tamaño muestral**
Primera fase	N = 1.021	n' = 39
Segunda fase	n' = 39	n =10+12

Elaboración propia

Para el presente estudio se ha tomado para ambas variables, *x* e *y* (primera, segunda ocasión), el número de veces que ha ingerido alcohol el alumno en los últimos 14 días, pero en ocasiones diferentes.
El procedimiento de estimación ha consistido en combinar los estimadores de las dos muestras independientes de alumnos: $\bar{y}_m$ y $\bar{y}_u$.

Datos

m = 10, muestra apareada

u = 12, muestra no apareada

n = 10 + 12 = 22, muestra total

x_m = {6, 2, 5, 9, 11, 1, 6, 0, 0, 3}

x_u = {1, 1, 9, 4, 4, 0, 8, 10, 1, 6, 5, 2}

y_m = {8, 4, 2, 9, 8, 0, 1, 7, 4, 6}

y_u = {1, 5, 5, 7, 0, 3, 10, 5, 9, 9, 8, 11}

$\bar{x}' = 6$

$\theta = \frac{n'}{n} = 1.772727$

$p = \frac{m}{n} = 0.4545455$

$$w_{opt} = \frac{\theta p}{\theta - Z(\theta - p)(1 - p)} = 0.3193916$$

Método de razón

1.- Calculamos estimadores de la media de la parte apareada y de la parte no apareada:

$\bar{y}'_m = \frac{\bar{y}_m}{\bar{x}_m}\bar{x}' = \hat{R}\bar{x}'$ = 6.837209

$\bar{y}_u$ = 6.083333

2.- Calculamos el estimador de la media por razón:

$\bar{y}_r = w\bar{y'}_m + (1 - w)\bar{y}_u = 2.819669$

3.- Calculamos varianza del estimador de la media por razón:

$V(\bar{y}_r)_{min} = \frac{S_y^2(n' - u'Z)}{n'n - Zu'u}$ = 0.635922

4.- Calculamos la ganancia de precisión de $\bar{y}_r$ sobre $\bar{y}$:

$$G_1 = \frac{Zp(\theta - p)}{\theta - Z(\theta - p)} = \frac{-1.0433 \cdot 0.4545(1.7727 - 0.4545)}{1.7727 - (-1.0433) \cdot (1.7727 - 0.4545)} = -0.198$$

Por lo que, como la ganancia de precisión presenta valores negativos, en este caso, con los datos dados, el estimador $\bar{y}_r$ no será más preciso que $\bar{y}$. Además, esta conclusión la podemos afirmar sabiendo que $Z = -1.0433 < 0$.

Método de regresión

Calculamos b:

$b = \frac{S_{y_m x_m}}{S_{x_m}^2}$ = 0.3770492

1.- Calculamos estimadores de la media de la parte apareada y de la parte no apareada:

$\bar{y}'_m = \bar{y}_m + b(\bar{x} - \bar{x}_m)$ = 4.889717

$\bar{y}_u$ = 6.083333

2.- Calculamos el estimador de la media por regresión:

$\bar{y}_{Reg} = \frac{\left[\bar{y}_m + \rho\frac{S_y}{S_x}(\bar{x}' - \bar{x}_m)\right] + u\bar{y}_u}{t+u}$ = 5.750135

3.- Calculamos varianza del estimador de la media por regresión:

$V(\bar{y}_{Reg}) = \frac{S_y^2(n' - \rho^2 u')}{n'n - \rho^2 u' u}$ = 0.4944343

Método de producto

1.- Calculamos estimadores de la media de la parte apareada y de la parte no apareada:

$\bar{y}'_m = \frac{\bar{x}_m}{\bar{x}} \bar{y}_m$ = 4.931277

$\bar{y}_u$ = 6.083333

2.- Calculamos el estimador de la media por producto:

$\bar{y}_{prod} = w\bar{y'}_m + (1-w)\bar{y}_u = 5.715376$

3.- Calculamos varianza del estimador de la media por producto:

$V(\bar{y}_r)_{min} = \frac{S_y^2(n+uZ)}{n^2+Zu^2}$ = 0.3184716

Método de diferencia

1.- Calculamos estimadores de la media de la parte apareada y de la parte no apareada:

$\bar{y}_{2D} = \bar{y}_m - (\bar{x}_m - \bar{x})$= 4.872727

$\bar{y}_u$ = 6.083333

2.- Calculamos el estimador de la media por diferencias:

$\bar{y}_{Dif} = w\bar{y}_{2D} + (1-w)\bar{y}_{2u} = 5.696676$

3.- Calculamos varianza del estimador de la media por diferencia:

$V(\bar{y}_{Dif}) = \frac{S^2}{n}\frac{1+Zq}{1+Zq^2} = 0.2868826$

4. FUNCIONES R STUDIO

Una vez estudiado el muestreo sucesivo en dos ocasiones, se ha procedido a implementar estos estimadores indirectos en ocasiones sucesivas en R Studio. Se ha realizado a través de funciones para poder, a través de ellas, encontrar la solución de cualquier problema o ejercicio dada esta solución.

4.1. Método de razón

```
> E.Razon.S <- function(ym,yu,xm,xu,N,n1,mX1) { #ym = y Apareada; n1 = n'
+   m=length(ym) # tamaño de la muestra común en la segunda ocasión
+   ym.media = mean(ym) # media de y de la parte común de la muestra
+   xm.media = mean(xm) # media de x de la parte común de la muestra
+   x <- c(xm,xu) #x es el conjunto total de la parte apareada y de parte no apareada
+   y <- c(ym,yu)
+   mX=mean(x)
+   n = length(y) # tamaño de la muestra en la segunda muestra n = m+u
+   sx2 = var(x)
+   sy2 = var(y)
+   sx=sqrt(sx2)
+   sy=sqrt(sy2)
+   u = n-m # Muestra no común de la muestra en segunda ocasión
+   u1 = n1-m # Muestra no común de la muestra en la primera ocasión
+   delta = (sx/sum(x))/(sy/sum(y))
+   rho = cor(y,x)
+   Z = delta*(2*rho - delta)
+   e.R = ym.media/xm.media # Estimador de razón
+   teta=(n1/n)
+   p=(m/n) # Índice de solapamiento
+   w=(teta*p)/(teta-Z*(teta-p)*(1-p))
+
+   ##### ESTIMACIONES DE LA MEDIA:
```

```
+
+   ## Parte apareada:
+   e.media.apareada1 <- e.R*mX1
+
+   v.media.apareada1 <- sy2*((n1-u1*Z)/(m*n1))
+
+   ## Parte NO apareada:
+
+   e.media.Napareada <- mean(yu)
+
+   e.media.Napareada <- sy2/u
+
+   ## Estimador de razon:
+
+   e.media.R <- w*e.media.apareada1 + (1-w)*e.media.Napareada
+
+   v.media.R <- ((sy2*(n1-u1*Z))/(n1*n - Z*u1*u)) #varianza del
estimador de la media de razón
+
+
+   ##### ESTIMACIONES DEL TOTAL:
+
+   e.total.R <- N*e.media.R
+
+   v.total.R <- (N^2)*v.media.R
+
+   ###### Ganancia de precisión de est.media.Razon sobre
est.media.mas:
+
+   G1 <- (Z*p*(teta-p))/(teta - Z*(teta-p))
+
+   ##### Salida
+
```

```
+   list(Est.media.R=e.media.R, Varianza.media.R=v.media.R,
Est.total.R=e.total.R, Est.varianza.R=v.total.R,
Ganancia.pres.sobre.estmediaMas=G1)
+ }
```

4.2. Método de regresión

```
> E.Regresion.S <- function(ym,yu,xm,xu,N,n1,mX1) {
+   m=length(ym) # tamaño de la muestra común en la segunda ocasión
+   ym.media = mean(ym) # media de y de la parte común de la muestra
+   xm.media = mean(xm) # media de x de la parte común de la muestra
+   x <- c(xm,xu) #x es el conjunto total de la parte apareada y de parte
no apareada
+   y <- c(ym,yu)
+   mX = mean(x)
+   n = length(y) # tamaño de la muestra en la segunda ocasión n=m+u
+   sx2 = var(x)
+   sy2 = var(y)
+   sx=sqrt(sx2)
+   sy=sqrt(sy2)
+   u = n-m
+   u1 = n1-m
+   delta = (sx/sum(x))/(sy/sum(y))
+   rho = cor(y,x)
+   Z = delta*(2*rho - delta)
+   teta=(n1/n)
+   p=(m/n) #Índice de solapamiento
+   w=(teta*p)/(teta-Z*(teta-p)*(1-p))
+
+   b=cov(ym,xm)/var(xm)
+   ## Parte apareada:
+
+   e.media.apareada <- ym.media+b*(mX-xm.media)
+
+   ## Parte no apareada:
```

```
+
+   e.media.Napareada <- mean(yu)
+
+   ## Valor de t:
+
+   t = 1/((rho^2 / n1)+((1-rho^2)/m))
+
+   ## Estimación media por regresión:
+
+   e.media.Reg <- (t*(e.media.apareada + rho*(sy/sx)*(mX1-
xm.media)) + u*e.media.Napareada)/(t+u)
+
+   v.media.Reg <- (sy2*(n1-(rho^2)*u1))/((n1*n)-((rho^2)*u1*u))
#Estimación de la varianza del estimador de la media de regresión
+
+   ## Estimacion total por regresion:
+
+   e.total.Reg <- N*e.media.Reg
+
+   v.total.Reg <- (N^2)*v.media.Reg #Estimación de la varianza del
estimador del total de regresión
+
+   ## Ganancia precision sobre razon
+
+   G2 = ((p*teta)*(teta-p)*(rho^2 - Z))/((teta - Z*(teta-p)*(1-p))*(teta-
(rho^2)*(teta-p)))
+
+   ## salida
+
+   list(Est.media.Regresion=e.media.Reg, Est.varianza.media.Regresion
= v.media.Reg, Est.total.Regresion=e.total.Reg,
Est.varianza.total.Regresion=v.total.Reg,
Ganacia.Prec.Reg.sobre.Raz=G2)
+ }
```

4.3. Método de producto

```
> E.Producto.S<- function(ym,yu,xm,xu,N,n1,mX1) {
+
+   m = length(ym) # tamaño de la muestra común en la segunda ocasión
+   ym.media = mean(ym)
+   xm.media = mean(xm)
+   x <- c(xm,xu) #x es el conjunto total de la parte apareada y de parte
no apareada
+   y <- c(ym,yu)
+   n = length(y) # tamaño de la muestra en la segunda ocasión n=m+u
+   sx2 = var(x)
+   sy2 = var(y)
+   sx=sqrt(sx2)
+   sy=sqrt(sy2)
+   u = n-m
+   u1 = n1-m
+   delta = (sx/sum(x))/(sy/sum(y))
+   rho = cor(y,x)
+   Z = delta*(2*rho - delta)
+   teta=(n1/n)
+   p=(m/n) # Índice de solapamiento
+   w=(teta*p)/(teta-Z*(teta-p)*(1-p))
+
+   ## Parte apareada:
+
+   e.media.apareada1 <- (xm.media/mX)*ym.media
+
+   ## Parte no apareada:
+
+   e.media.Napareada <- mean(yu)
+
+   ## Estimacion de la media:
+
+   e.media.Prod <- w*e.media.apareada1 + (1-w)*e.media.Napareada
```

```
+
+   v.media.Prod <- (sy2*(n+u*Z))/(n^2 + Z*u^2)
+
+   ## Estimacion del total
+
+   e.total.Prod <- N*e.media.Prod
+
+   v.total.Prod <- (N^2)*v.media.Prod
+
+   ## Ganancia precision de regresion sobre producto:
+
+   G2 <- (p*teta*(teta-p)*(rho^2 + Z))/((teta-Z*(teta-p)*(1-p)))*((teta-
(rho^2)*(teta-p)))
+
+   ## salida
+
+   list(Est.media.Producto=e.media.Prod, Est.varianza.media.Producto =
v.media.Prod, Est.total.Producto=e.total.Prod,
Est.varianza.total.Producto=v.total.Prod,
Ganacia.Prec.Reg.sobre.Raz=G2)
+
+ }
```

4.4. Método de diferencias

```
> E.Diferencia.S <- function(ym,yu,xm,xu,N,n1,mX1){ #ym = y
Apareada; n1 = n'
+   m = length(ym)
+   ym.media = mean(ym)
+   xm.media = mean(xm)
+   x <- c(xm,xu) #x es el conjunto total de la parte apareada y de parte
no apareada
+   y <- c(ym,yu)
+   n = length(y)
```

```
+  sx2 = var(x)
+  sy2 = var(y)
+  sx=sqrt(sx2)
+  sy=sqrt(sy2)
+  u = n-m
+  u1 = n1-m
+  delta = (sx/sum(x))/(sy/sum(y))
+  rho = cor(y,x)
+  Z = delta*(2*rho - delta)
+  teta=(n1/n)
+  p=(m/n) #índice de solapamiento
+  q=(u/n) #Fracción de reemplazamiento
+  w=(teta*p)/(teta-Z*(teta-p)*(1-p))
+  s2=var(ym)
+
+  ## Parte apareada:
+
+  e.media.apareada <- ym.media-(xm.media - mX)
+
+  ## Parte no apareada:
+
+  e.media.Napareada <- mean(yu)
+
+  ### Estimacion de la media:
+
+  e.media.D <- w*e.media.apareada+(1-w)*e.media.Napareada
+
+  v.media.D <- (s2*(1+Z*q))/(n*(1+Z*q^2))
+
+  ### Estimacion del total:
+
+  e.total.D <- N*e.media.D
+
+  v.total.D <- (N^2)*v.media.D
```

```
+
+   ### salida:
+
+   list(Est.media.Diferencia=e.media.D, Est.varianza.media.Diferencia =
v.media.D, Est.total.Diferencia=e.total.D,
Est.varianza.total.Diferencia=v.total.D)
+
+ }
```

5. CONCLUSIONES

Durante la realización del trabajo se ha tenido como prioridad la fácil comprensión del mismo, siempre y cuando, el lector cuente con conocimientos básicos de estadística y matemáticas.

Tras su realización, se demuestra que el muestreo sucesivo es una técnica que reduce el error de muestreo del estimador de la media considerablemente, estimador el cual se ha enfocado el trabajo, siempre y cuando la variable auxiliar *x* esté correlada con la variable de estudio *y*.

El objetivo del muestreo sucesivo es obtener el estimador óptimo combinando dos estimadores de las medias: un estimador indirecto de doble muestreo para la parte apareada (parte común de la muestra en ambas ocasiones) y un estimador simple para la parte no apareada (parte no común de la muestra), obteniéndose estimadores de la media mediante métodos de regresión, razón, diferencia y producto. Aquí surge entonces, dos nuevos conceptos: el índice de solapamiento, *p*, y la fracción de reemplazamiento, *q*.

Estos índices indican, en proporción, la parte común y no común, p y q respectivamente, que la muestra presenta, siendo entonces

$p = \frac{m}{n}$,

$q = \frac{u}{n}$, tal que $p + q = 1$

Además, en el presente trabajo se ha implementado, mediante R Studio, las funciones de los estimadores indirectos en muestreo sucesivo en dos ocasiones, para que se pueda realizar las operaciones dadas en este tipo de muestra para obtener resultados fiables y concisos.

6. BIBLIOGRAFÍA

- [LIBRO] RODRÍGUEZ, Eva María Artés; LUENGO, Amelía Victoria García. *Diseños muestrales en el tiempo*. Universidad Almería, 2002.

- [ARTÍCULO] Rodríguez, E. M. A., García, M. D. M. R., & Cebrián, A. A. (1999). Aportaciones al muestreo sucesivo. *Metodología de Encuestas*, *1*(1), 19-28.

- *Estimadores indirectos III: Muestreo doble*. Apuntes de Técnicas Avanzas de Muestreo por María Virtudes Alba Fernández.