



Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

Librería de muestreo con R: samplingR

Autor:

D. Javier Estévez Asensio

Tutores:

**D. Alfonso Gordaliza Ramos,
D. Francisco Rodríguez Redondo**

Año 2023

Abstract

El muestreo estadístico es una técnica usada en multitud de ámbitos, desde el empresarial hasta el sanitario o el político, permitiendo obtener de forma eficiente y económica información sobre una población completa para realizar una toma de decisiones.

En este trabajo, se ha creado una librería de funciones relacionadas con el muestreo estadístico en el programa de software libre R con el objetivo de subirlo al repositorio oficial *CRAN* [1] y ampliar las funcionalidades que ofrecen otras librerías ya publicadas con temática similar.

Statistical sampling is a technique used in a multitude of fields, from business to healthcare and politics, allowing for efficient and cost-effective gathering of information on a complete population to aid in decision-making.

A library of statistical sampling related functions has been developed on R free software program with the goal of uploading it to the *Comprehensive R Archive Network (CRAN)* [1] to extend the functionalities offered by other similarly themed libraries already published on it.

Agradecimientos

Agradezco a mi familia, por brindarme una educación completa y apoyarme desde el primer día hasta el último.

A Francisco, por dejarlo todo para ayudarme siempre que lo he necesitado.

Y por último, me gustaría agradecer a la promoción 2018/2023 de InDat por unos años maravillosos junto a ellos.

Tabla de contenidos

1. Introducción	1
1.1. Tipos de muestreo	1
1.2. La estadística oficial	2
1.3. Muestreo en R	3
1.4. Objetivos	3
1.5. Estructura de la memoria	3
2. Creación de una librería de funciones en R	5
2.1. Creación de un paquete local	5
2.2. Documentación de funciones	6
2.3. Librería devtools	7
2.4. Creación de viñetas	8
2.5. Petición de subida a CRAN	9
3. Muestreo aleatorio simple	12
3.1. Selección de muestras	12
3.2. Estimadores lineales insesgados	14
3.2.1. Muestreo aleatorio simple sin reemplazamiento	14
3.2.2. Muestreo aleatorio simple con reemplazamiento	15
3.3. Estimadores de la varianza	15
3.3.1. Muestreo aleatorio simple sin reemplazamiento	15
3.3.2. Muestreo aleatorio simple con reemplazamiento	16
3.4. Tamaño de la muestra	16
3.4.1. Error de muestreo dado	16
3.4.2. Error de muestreo y coeficiente de confianza dados	17
3.4.3. Error relativo de muestreo dado	17
3.4.4. Error relativo de muestreo y coeficiente de confianza dados	17
3.4.5. Muestreo aleatorio simple sin reemplazamiento	18
3.4.6. Muestreo aleatorio simple con reemplazamiento	18
3.5. Dominios de estudio	18
3.6. Aplicaciones en la librería samplingR	19
4. Muestreo estratificado	21
4.1. Selección de la muestra	21
4.1.1. Afijación uniforme	22
4.1.2. Afijación proporcional	22
4.1.3. Afijación de mínima varianza	22

4.1.4.	Afijación óptima	22
4.2.	Estimadores lineales insesgados	23
4.3.	Estimación de varianzas	24
4.3.1.	Muestreo sin reemplazamiento	24
4.3.2.	Muestreo con reemplazamiento	24
4.4.	Tamaño de la muestra	25
4.4.1.	Afijación proporcional	25
4.4.2.	Afijación de mínima varianza	25
4.4.3.	Afijación óptima	25
4.4.4.	Muestreo sin reemplazamiento	26
4.4.5.	Muestreo con reemplazamiento	27
4.5.	Tamaño de muestra dado un presupuesto	28
4.6.	Aplicaciones para el muestreo estratificado de la librería <code>samplingR</code> . .	29
5.	Muestreo sistemático	31
5.1.	Selección de la muestra	32
5.2.	Análisis de la varianza y correlación	33
5.3.	Estimadores lineales insesgados	34
5.4.	Estimadores de la varianza	35
5.5.	Aplicaciones para el muestreo sistemático de la librería <code>samplingR</code> . .	36
6.	Muestreo por conglomerados	38
6.1.	Selección de la muestra	39
6.2.	Estimadores lineales insesgados	39
6.3.	Estimación de varianzas	40
6.4.	Tamaño de la muestra	40
6.5.	Aplicaciones para el muestreo por conglomerados de la librería <code>samplingR</code>	41
7.	Conclusiones y líneas futuras	43
7.1.	Conclusiones	43
7.2.	Líneas futuras	44
9.	Bibliografía	45
10.	Anexos	46

Índice de figuras

2.1. Estructura de un proyecto en R.	5
2.2. Uso de funciones de devtools desde la interfaz gráfica.	8
2.3. Ejemplo de resultados usando la herramienta Winbuilder.	11
3.1. Ejemplo de tabla de números aleatorios.	13
5.1. Selección de una muestra sistemática 1 en k.	32

Índice de tablas

3.1. Tamaños de muestra en el muestreo aleatorio simple sin reemplazamiento.	18
3.2. Tamaños de muestra en el muestreo aleatorio simple con reemplazamiento.	18
4.1. Tamaños de muestra con afijación proporcional en el muestreo sin reemplazamiento	26
4.2. Tamaños de muestra con afijación de mínima varianza en el muestreo sin reemplazamiento	26
4.3. Tamaños de muestra con afijación óptima en el muestreo sin reemplazamiento	27
4.4. Tamaños de muestra con afijación proporcional en el muestreo con reemplazamiento	27
4.5. Tamaños de muestra con afijación de mínima varianza en el muestreo con reemplazamiento	28
4.6. Tamaños de muestra con afijación óptima en el muestreo con reemplazamiento	28
5.1. Tabla ANOVA	33

Capítulo 1

Introducción

Para multitud de análisis económicos, demoscópicos, sociosanitarios, etc., resulta de interés conocer el valor de determinados parámetros poblacionales (medias, proporciones, totales, ...), que comúnmente son desconocidos. Ya sea el salario promedio de las familias españolas o el total de accidentes en el sector de la construcción en un determinado territorio, es necesario invertir recursos en realizar un estudio para poder hacer el análisis de los datos de interés en busca de información útil.

El muestreo permite realizar un análisis sobre una subpoblación o muestra de la población objetivo a estudiar, obteniendo estimadores que se aproximen al valor real que se conseguiría al realizar el estudio sobre la población completa, con el aliciente de hacerlo de forma eficiente y con un uso de recursos reducido, lo que supone un método muy conveniente para organizaciones de todo tipo.

El muestreo estadístico parte de la premisa, apoyada en los resultados de la Estadística Matemática, de que al tomar una muestra aleatoria de la población las propiedades de la primera deben ser extrapolables a la segunda, ya que la muestra constituirá una representación a pequeña escala de la población completa. Por otro lado, para que el estudio sea adecuado debe asegurarse que la muestra contenga suficientes individuos de la población para que la información obtenida sea significativa, en el sentido de que los márgenes de error de las estimaciones sean aceptables. Cuando se combinan ambas cosas, es decir, una muestra seleccionada por un procedimiento aleatorio convenientemente diseñado y de un tamaño suficiente para que los márgenes de error sean aceptables, decimos que la muestra es representativa.

1.1. Tipos de muestreo

Existen diferentes tipos de muestreo según la forma que se considere para seleccionar la muestra de estudio. Entre los más destacados se encuentran:

- Muestreo aleatorio simple: Es el tipo de muestreo más simple. En él la muestra se obtiene seleccionando individuos de la población al azar hasta llegar al

tamaño de muestra requerido. La selección puede realizarse con o sin reemplazamiento.

- Muestreo estratificado: La población se divide en grupos lo más homogéneos posibles con respecto a una determinada característica, pero heterogéneos entre ellos. Los estratos no deben solaparse entre ellos. Una vez realizada la estratificación, se toma una muestra representativa de cada estrato para conformar la muestra poblacional.
- Muestreo sistemático: Se divide la población en un número de zonas del mismo tamaño (tantas como unidades tiene la muestra) y se selecciona un elemento de una de ellas. Una vez seleccionada se toman del resto de zonas las unidades en la misma posición que la seleccionada originalmente, formando una denominada *muestra de 1 en k*. Esta técnica tiene la ventaja de aplicarse de forma sencilla y bajo ciertas premisas, presentar errores de muestreo menores que en situaciones anteriores.
- Muestreo por conglomerados: La población se divide en grupos heterogéneos pero lo más homogéneos posibles entre ellos. La muestra se selecciona tomando un grupo aleatorio de conglomerados. Al realizar la estimación se analizan todos los individuos dentro de cada conglomerado (si utilizamos un muestreo monoetápico). En otras ocasiones también se puede realizar un nuevo submuestreo dentro de los conglomerados elegidos en la primera etapa.

1.2. La estadística oficial

Una de las formas más importantes en la toma de datos dentro de la producción estadística oficial consiste en recoger la información mediante encuestas muestrales, es decir una investigación parcial de la población finita a través de una encuesta. Por otro lado hay que tener en cuenta que una encuesta muestral cuesta menos que un censo y casi se puede afirmar que es mucho más acurada.

Ante la necesidad de información, tanto por parte de la sociedad como de los responsables políticos, en muchos países se han constituido legalmente los denominados *institutos nacionales de estadística* cuya finalidad es la de proporcionar información estadística lo suficientemente fiable sobre la situación de cada país. En estos centros administrativos las encuestas son una parte importante de su actividad. En el caso concreto de España, el Instituto Nacional de Estadística (INE) se rige por la ley 12/1989 del 9 de Mayo de la Función Estadística Pública, modificada por la disposición 11311, ley 13/2022.

Por lo tanto, los INEs producen regularmente estadísticas sobre características y actividades nacionales importantes, incluyendo la demografía, agricultura, población activa (EPA), salud y condiciones de vida, industria y comercio. Para conseguir estos objetivos utilizan de una forma muy técnica y ambiciosa diferentes técnicas de muestreo con utilidades ya existentes o ampliando la teoría de muestreo disponible

en cada momento.

En el caso concreto del INE en España, y salvo los trabajos censales o la recogida de información basada en registros administrativos, su basta producción estadística se basa en trabajos sobre muestras, elaborados con diferentes diseños muestrales, y entre los que destacan encuestas muy útiles en nuestra sociedad, como pueden ser la Encuesta de Población Activa (EPA), Encuesta de Presupuestos Familiares (EPF), etc. Es por lo tanto este Organismo Administrativo uno de los centros de referencia nacionales a la hora de aplicar las diversas técnicas de muestreo hoy conocidas.

1.3. Muestreo en R

En el lenguaje de programación R ya existen ciertas librerías públicas de funciones relacionadas con el muestreo. Las más destacadas por su completitud serían *sampling* [2] y *TeachingSampling* [3]. Viendo su documentación y funcionamiento uno puede darse cuenta que, en la primera, la estimación de la varianza se calcula usando el método de Deville o el estimador general de Horwitz-Thompson en lugar de realizar una estimación específica según el tipo de muestreo y parámetro a estimar. En la segunda librería sí se realiza la estimación de varianzas de manera específica.

En ambos casos al tomar las muestras o realizar estimaciones se requiere de un tamaño de muestra que no es posible estimar con ninguna función implementada en estos paquetes. Esta última cuestión resulta de gran importancia si a la hora de realizar el muestreo es necesario cumplir requisitos específicos como un error inferior a una tolerancia dada o no exceder un coste de estudio determinado.

1.4. Objetivos

El objetivo de este trabajo consiste en desarrollar una librería de funciones en R para aportar una nueva visión a la hora de trabajar el muestreo estadístico y ampliar la operatividad de librerías ya existentes con funciones dedicadas a obtener el mínimo tamaño de muestra necesario para cumplir con requisitos especificados por el usuario. El desarrollo de las funciones y código necesario para la implementación de este trabajo tendrá su apoyo en el fundamento teórico del libro *Técnicas de muestreo estadístico: teoría, práctica y aplicaciones informáticas* [4] del estadístico César Pérez.

1.5. Estructura de la memoria

Los contenidos de este trabajo se desarrollan en el siguiente orden:

El primer capítulo se trata del actual, y describe el marco teórico del problema propuesto y sus objetivos.

El segundo capítulo realiza una revisión del proceso de creación de una librería de funciones genérica en R hasta su subida a *CRAN* [1].

En los capítulos 3 a 6 se recordarán las principales técnicas de muestreo y cómo se han codificado las mismas en el lenguaje de programación utilizado en este trabajo.

Finalmente, en el capítulo 7, se llegará a las conclusiones del proyecto y se trazarán posibles líneas de trabajo futuras para la continuación de la materia.

En los anexos se incluye el manual oficial de uso de la librería creada y se proporciona, mediante una *vignette*, un ejemplo de un ejercicio práctico que puede ser resuelto de una forma ágil y cómoda utilizando la librería *samplingR* [5], con la finalidad de ver las ventajas que aporta su uso.

Todos los productos generados durante este TFG se encuentran disponibles en un repositorio GitHub disponible en la bibliografía [6].

Capítulo 2

Creación de una librería de funciones en R

La creación de una librería en R es uno de los pilares en los que se apoya este sistema, ya que al tratarse de software de código abierto son los usuarios finales los que deben mantener de manera gratuita estas aplicaciones.

Para conseguir esto, los administradores de R han creado una serie de reglas y pautas lo suficientemente precisas y concisas con la finalidad de que los trabajos que se almacenen y se pongan a disposición de los clientes mantengan unos estándares de calidad lo suficientemente elevados como para que los usuarios finales confíen y así puedan utilizar este sistema. A continuación se pasan a detallar los pasos a seguir para superar estos estándares de calidad marcados por los administradores de R.

2.1. Creación de un paquete local

El primer paso para la creación de una librería en RStudio es elegir la opción *Nuevo proyecto* en la pestaña *Archivo* y seleccionar *Nuevo directorio* y *Paquete de R*. Se abrirá una ventana donde se debe colocar el nombre del paquete e indicar la dirección del disco donde situarla. Opcionalmente, se pueden adjuntar archivos de funciones ya creados para ser incluidos en el paquete. Como resultado nos encontraremos una carpeta con el nombre de nuestra librería con la siguiente estructura :








 man	06/02/2023 18:06	Carpeta de archivos	
 R	06/02/2023 18:06	Carpeta de archivos	
 .Rbuildignore	06/02/2023 18:07	Archivo RBUILDIG...	1 KB
 .Rhistory	01/05/2023 9:37	Archivo de origen ...	10 KB
 DESCRIPTION	15/04/2023 20:53	Archivo	1 KB
 NAMESPACE	29/04/2023 16:50	Archivo	1 KB
 samplingR.Rproj	01/05/2023 9:36	R Project	1 KB

Figura 2.1: Estructura de un proyecto en R.

Los contenidos que figuran dentro de esta carpeta de forma resumida son los

siguientes:

- `man`: carpeta donde se encuentran los ficheros de documentación de las funciones.
- `R`: carpeta donde se encuentran los archivos con extensión `.R` de funciones.
- `DESCRIPTION`: archivo donde se almacena información importante sobre el paquete.
- `NAMESPACE`: archivo donde se encuentran los nombres de funciones públicas al usuario y menciones sobre el uso de otras librerías de funciones en caso de requerir de sus funciones en algún momento del desarrollo.
- Archivo `.Rproj`: archivo que se usa para abrir el proyecto en una nueva sesión R.

Dependiendo de los archivos a crear pueden generarse carpetas adicionales, como explicamos en la sección dedicada a la creación de viñetas 3.4 Una vez obtenida toda esta estructura, ya se podría crear un nuevo archivo de funciones dentro de la carpeta `R` con el código deseado.

2.2. Documentación de funciones

Un apartado muy importante y que constituye buena práctica a la hora de escribir una nueva función es la documentación. En R resulta particularmente útil definirla ya que gracias a la librería *roxygen2* [7] la documentación incluida en las funciones será traducida a archivos legibles por el usuario que aparecerán en la sección *Ayuda* al acceder a la documentación de dicha función, lo que facilita al usuario final comprender mejor cuál es el objetivo de esa función. Para empezar a documentar se debe introducir el carácter `#'` y utilizar un indicador adecuado, empezando con `@`, siendo los más usados los siguientes:

- `title`: Descripción general de la función. Indica el propósito de la función de forma breve.
- `description`: Ampliación del propósito de la función.
- `param`: Primero indica el nombre del parámetro dentro de la función y seguido de un espacio lo que representa para la función. Debe haber tantos como parámetros tenga la función.
- `return`: Indica si la función devuelve un objeto del tipo de objeto que se trata y sus contenidos si es conveniente.

- `details`: Sirve para ampliar información acerca de la función que no tenga cabida en el apartado de descripción como funcionamiento interno de la función o valores específicos de parámetros.
- `references`: En caso de querer aportar referencias teóricas sobre las que se fundamenta el código realizado.
- `examples`: Permite escribir un caso de uso de la función para referencia del usuario.
- `importFrom`: Indica que se ha hecho uso de una función externa durante el desarrollo de la función. Se utiliza escribiendo el nombre de la librería y el de la función utilizada separadas por espacios.
- `export`: Hace pública la función al usuario.

En ocasiones, habrá funciones de uso recurrente que se usan para comodidad a la hora de desarrollar las funciones principales, y que no es necesario ponerlas a disposición de usuario final. Es una buena práctica que estas funciones “auxiliares” sean documentadas como cualquier otra de nuestra librería.

Sin embargo, esta práctica sin más, implica su publicación al usuario general y por lo tanto aparecerán en el manual y serán accesibles por los usuarios finales. Si no queremos que esto ocurra, se deberá eliminar el indicador `@export` y sustituirlo por `@noRd`. De esta forma disponemos de la documentación de la función en el fichero de código, pero no aparecerá en el manual ni en la sección *Ayuda*. Tampoco será accesible por el usuario final, al menos no de la forma convencional, es decir con el nombre de la función o bien con el formato `librería::función`.

Es posible acceder a ellas utilizando tres puntos dobles, por ejemplo, `samplingR:::all01list` nos permite utilizar una función auxiliar que se utiliza en este trabajo en el control de funciones para comprobar que todos los valores de todas las entradas de una lista son ceros o unos. Es un caso no deseable pero es la forma más parecida de lograr el equivalente a funciones denominadas “privadas” de otros lenguajes de programación.

2.3. Librería devtools

Devtools [8] es una herramienta que facilita la creación y desarrollo de librerías en R. Proporciona funciones de gran utilidad y que son muy recomendables a la hora de fijar un flujo de trabajo en el desarrollo. Entre ellas podemos destacar las siguientes:

- `document()`: traduce los comentarios en formato Roxygen2 [7] de tus funciones a archivos `.md` de ayuda y crea el archivo `NAMESPACE` actualizado. La primera vez que se ejecuta es conveniente borrar el archivo `NAMESPACE` creado por defecto para que escriba los nombres de tus funciones.

- `check()`: hace un test automatizado similar al que se ejecuta cuando realizas tu petición de subida a CRAN [1]. Informa de errores, avisos y notas en el paquete. También existen funciones más específicas para comprobar concretamente la documentación o si cumple los requisitos para sistemas operativos Mac o Windows. Si se utiliza RStudio también es posible hacerlo en la parte superior derecha de la interfaz, donde suele encontrarse las variables de entorno. Pestaña Build \Rightarrow Check
- `build()`: genera un archivo comprimido con extensión `.tar.gz` preparado para ser instalado o subido a CRAN [1]. Al igual que con `check()` también existen métodos para generar el manual de uso de la librería o viñetas de las que se hablará más adelante. En la interfaz de RStudio Build \Rightarrow More \Rightarrow Build Source Package.

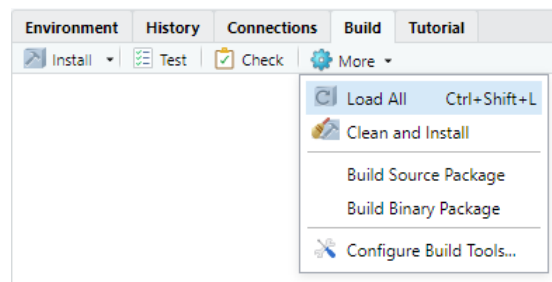


Figura 2.2: Uso de funciones de devtools desde la interfaz gráfica.

De forma excepcional y en pocas ocasiones, es posible que al realizar `check()` se reciba una nota indicando que ha sido imposible comprobar la hora. Es un error que se debe a que la fecha la verifica a través de *Word Clock API* [9] y el servicio está temporalmente inoperativo, por lo que puede ser ignorado o reintentarse más adelante.

2.4. Creación de viñetas

Generalmente, la documentación de las funciones ayuda a dar una idea general del uso de una función específica, pero no siempre va acompañada de ejemplos de uso o esta ayuda no queda lo suficientemente clara para comprender en detalle el resultado que se pueda obtener con esas funciones.

Una solución útil para hacer frente al problema anterior es la creación de viñetas, las cuales se diseñan con la finalidad de mostrar la utilidad de una librería o de parte de sus funciones en su conjunto mediante la resolución de un problema propuesto. Aparecen en la ayuda del paquete antes del listado de funciones y se pueden realizar múltiples viñetas para abarcar los casos de uso más importantes para aquellas librerías de mayor extensión.

Para generar una viñeta se puede usar el comando `usethis::use_vignette("nombre")`. Si se trata de la primera, se creará una carpeta en la raíz del directorio del paquete

con nombre *vignettes*, donde se almacenarán las viñetas creadas junto a un archivo *.gitignore* para evitar que los archivos creados pasen control de versiones, además de añadir las dependencias al archivo *DESCRIPTION*. Dentro de la carpeta aparecerá un archivo *.Rmd* (es decir, se trataría de un archivo de tipo Rmarkdown) con el nombre de nuestra viñeta. En él encontramos dos bloques de código: el primero destinado a incluir los metadatos de la viñeta, como su título y forma de presentación, que puede ser como un archivo HTML, PDF, un cuaderno de trabajo similar a los cuadernos Jupiter de Python, etc. El segundo bloque incluye los ajustes predeterminados de presentación del código y sus resultados. El último simplemente realiza la llamada de nuestra librería para poder empezar a trabajar.

2.5. Petición de subida a CRAN

En el supuesto de que se decida subir la librería al repositorio oficial para uso general de todos los usuarios, se deberá ir al sitio web dedicado a la realización de esta acción, accesible desde la página oficial. Previamente se deberá realizar algunas modificaciones a los contenidos del paquete para poder enviarlo.

El archivo *DESCRIPTION* contiene la información principal de la librería, tal como su nombre o las personas que lo han desarrollado. Debe ser rellenado con la información pertinente. Entre sus campos encontramos:

- *Package*, que especifica el nombre de la librería. Éste no puede coincidir con el de ningún otro subido a CRAN con anterioridad. Los nombres son case-insensitive, por lo que, por ejemplo, *sampligr* no será aceptado por coincidir con el nombre de la librería de este trabajo. La forma más cómoda de buscar un nombre es ir a *install* de la pestaña *Packages* de RStudio y escribir el nombre elegido, ya que con el autocompletado podrás ver los nombres con coincidencia parcial al texto introducido.
- *Title* muestra el título que aparecerá en la página general de ayuda del paquete. Se trata de una única frase escrita en title-case, por lo que exceptuando artículos, preposiciones y similar la primera letra de las palabras debe ser mayúscula. No debe terminar en punto, ya que de lo contrario se nos avisará al subirlo al repositorio de CRAN. Tampoco conviene poner “en R” ya que los revisores lo consideran error por reiterativo.
- *Maintainer* indica la persona que lleva a cabo tareas de mantenimiento de la librería. Es importante que la información sea correcta ya que se usará para comunicar novedades sobre el estado del paquete durante el proceso de revisión y subida. Sigue el formato nombre apellido <correo electrónico>
- *Authors@R* indica las personas involucradas en el desarrollo del paquete. Anteriormente se usaba un formato similar al de *Maintainer*, pero ahora se aconseja usar la clase *person* [10]. Dispone de múltiples roles para asignar a cada persona en función del trabajo realizado.

- *Description* amplía la información del título sobre el uso de la librería. Hay que ser moderadamente específicos con lo que realiza ya que si usas términos como *diferentes funciones o varios métodos* los revisores pueden exigir elaborar sobre dichos métodos.
- *Version* indica la versión de desarrollo de la librería. Por defecto se sitúa en 0.1.0. Es importante que cada vez que subamos una actualización del paquete al repositorio modifiquemos manualmente este valor ya que de lo contrario se detectará una coincidencia y la subida no será efectiva.

Una vez tengamos el archivo *DESCRIPTION* completo, rellenaremos el formulario haciendo que el nombre y correo electrónico coincidan con los del *Maintainer* y subiremos un archivo con extensión *.tar.gz* creado previamente con la función *build()*. Pasados unos minutos después del envío, llegará al correo del maintainer un mensaje de confirmación con un enlace donde se deben aceptar los términos.

Si nos fijamos en el segundo de estos términos, confirmamos haber realizado una comprobación de tipo *check()* y nos adjunta una página web. Esto se debe a que según las directrices de CRAN es necesario realizar comprobaciones como las de la función *check()*, pero siguiendo las normas de la próxima versión de R planeada para ser desplegada.

Existen dos formas de realizar estas comprobaciones. La primera es instalar R-devel, es decir la siguiente versión de R en tu ordenador para realizar ahí las comprobaciones. Este método puede resultar pesado, por lo que si entramos al enlace mencionado anteriormente veremos un lugar donde realizar comprobaciones para la versión actual, anteriores o R-devel. Solo hace falta subir el archivo con extensión *.tar.gz* e indicar el correo donde se quiera recibir el resultado y en cuestión de minutos llegará un enlace a un archivo tipo log donde veremos los resultados de la misma manera que realizando *check()* en local. En el mejor caso recibiremos solo una nota encima del maintainer, lo cual es normal ya que se trata de un recordatorio para que los maintainer comprueben en los ficheros log que la actualización ha sido solicitada por él y no otra persona, según Uwe Ligges, maintainer de CRAN.

```

* using log directory 'd:/RCompile/CRANguest/R-devel/samplingR.Rcheck'
* using R Under development (unstable) (2023-04-30 r84357 ucrt)
* using platform: x86_64-w64-mingw32 (64-bit)
* R was compiled by
  gcc.exe (GCC) 12.2.0
  GNU Fortran (GCC) 12.2.0
* running under: Windows Server 2022 x64 (build 20348)
* using session charset: UTF-8
* checking for file 'samplingR/DESCRIPTION' ... OK
* checking extension type ... Package
* this is package 'samplingR' version '0.1.4'
* package encoding: UTF-8
* checking CRAN incoming feasibility ... [11s] NOTE
Maintainer: 'Javier Estévez <javier.estase@gmail.com>'

The Date field is over a month old.
* checking package namespace information ... OK
* checking package dependencies ... OK
* checking if this is a source package ... OK
* checking if there is a namespace ... OK
* checking for hidden files and directories ... OK
* checking for portable file names ... OK
* checking serialization versions ... OK
* checking whether package 'samplingR' can be installed ... OK
* checking installed package size ... OK
* checking package directory ... OK
* checking for future file timestamps ... OK
* checking DESCRIPTION meta-information ... OK
* checking top-level files ... OK
* checking for left-over files ... OK
* checking index information ... OK
* checking package subdirectories ... OK
* checking R files for non-ASCII characters ... OK
* checking R files for syntax errors ... OK
* checking whether the package can be loaded ... [2s] OK
* checking whether the package can be loaded with stated dependencies ... [1s] OK
* checking whether the package can be unloaded cleanly ... [1s] OK
* checking whether the namespace can be loaded with stated dependencies ... [1s] OK
* checking whether the namespace can be unloaded cleanly ... [1s] OK
* checking loading without being on the library search path ... [2s] OK
* checking startup messages can be suppressed ... [1s] OK
* checking use of S3 registration ... OK
* checking dependencies in R code ... OK
* checking S3 generic/method consistency ... OK
* checking replacement functions ... OK
* checking foreign function calls ... OK
* checking R code for possible problems ... [10s] OK
* checking Rd files ... [1s] OK
* checking Rd metadata ... OK
* checking Rd line widths ... OK
* checking Rd cross-references ... OK
* checking for missing documentation entries ... OK
* checking for code/documentation mismatches ... OK
* checking Rd \usage sections ... OK
* checking Rd contents ... OK
* checking for unstated dependencies in examples ... OK
* checking examples ... [2s] OK
* checking PDF version of manual ... [28s] OK
* checking HTML version of manual ... [2s] OK
* checking for detritus in the temp directory ... OK
* DONE
Status: 1 NOTE

```

Figura 2.3: Ejemplo de resultados usando la herramienta Winbuilder.

Una vez confirmados los términos llegará un segundo correo con los resultados de los tests automatizados. En cualquier caso, se notificará del resultado y se tendrá acceso al registro completo en un archivo tipo log. Una confusión generalizada al ver este archivo es ver una nota en el nombre del maintainer. En caso de pasar los tests se añadirá que el paquete queda pendiente de una última revisión manual. Esta última revisión será la que más se demore en la entrega de sus resultados y es aquí donde pueden sugerir cambios que no muestre la función `check()` como los antes mencionados del archivo `DESCRIPTION`, usar `TRUE` y `FALSE` en lugar de sus homónimos `T` y `F` en el código para mejorar su comprensión o reformular descripciones de funciones que contienen “esta función...” por redundantes.

Tras recibir el visto bueno por parte de los moderadores, su respuesta será una confirmación de que el paquete se ha creado correctamente y se encuentra de camino a CRAN. Generalmente se estima un periodo de 24 horas antes de poder ser descargado por los usuarios.

Capítulo 3

Muestreo aleatorio simple

El muestreo aleatorio simple es una de las técnicas más sencillas de tomar una muestra de una población, ya que se basa en la selección de unidades de forma individual y aleatoriamente, según un modelo uniforme discreto, por lo que todos los individuos de la población tienen la misma probabilidad de ser elegidos.

En la práctica, para seleccionar las muestras mediante este tipo de muestreo se requiere la existencia de un marco que identifique las unidades de muestreo, que son precisamente los elementos de la población. En España, para las estadísticas oficiales dirigidas a la población, este marco de muestreo se basa principalmente en los datos obtenidos del padrón municipal de habitantes, del cual se pueden extraer tanto individuos como hogares facilitando de esta manera una información sumamente valiosa de cara a diseñar diferentes tipos de muestreo. Para el caso de muestras dirigidas a empresas, el INE utiliza el Directorio Central de Empresas (DIRCE) [11], el cual reúne en un sistema de información único a todas las empresas española y a sus unidades locales ubicadas en el territorio nacional. Su objetivo básico es hacer posible la realización de encuestas económicas por muestreo.

Existen dos variantes de este tipo de muestreo según las características de sus muestras. El muestreo sin reposición impide la posibilidad de incluir individuos repetidos en la muestra a diferencia del muestreo con reposición, lo que repercute en la estimación de parámetros.

Tanto en el muestreo sin reemplazamiento como con reemplazamiento la probabilidad de inclusión en la muestra para todos los individuos de la población es:

$$\pi_i = \frac{n}{N} \quad (3.1)$$

donde n es el tamaño de la muestra deseada y N el tamaño poblacional.

3.1. Selección de muestras

La selección de la muestra sobre la que estudiar la variable de interés debe seguir un procedimiento que asegure la aleatoriedad del proceso. Generalmente se distin-

guen dos procedimientos con sus posibles aplicaciones en el campo de la informática.

El método de la urna se basa en introducir todos los elementos de la población en una urna, convenientemente removida, para sacar uno a uno elementos hasta llegar al tamaño de muestra deseado. Al no poder ver el interior de la urna se supone aleatoriedad.

En el muestreo con reposición, tras seleccionar un individuo, éste debe ser devuelto a la urna para permitir que pueda salir elegido de nuevo. En el caso de muestreo sin reposición, cada elemento seleccionado no se devuelve a la urna.

Para lograr un método de selección similar en un programa informático, se deben enumerar las unidades poblacionales del 1 a N , siendo N el tamaño poblacional. Después utilizando un método aleatorio se obtiene un número aleatorio entre 1 y N , obteniendo el índice del individuo seleccionado. El procedimiento se repite tantas veces como muestras se requieran. En el muestreo sin reposición tras la selección de cada individuo éste se retiraría de la lista y sería necesario tener un control extra sobre el número aleatorio para que no se repita. Un ejemplo sería seleccionar uno nuevo si coincide con el índice de algún individuo seleccionado previamente.

El segundo método consiste en usar tablas de números aleatorios, donde se crea una secuencia de números entre 1 y N . Se elige un punto de partida y se seleccionan tantos elementos de forma secuencial como requiera la muestra. Para ello es preciso numerar los individuos al igual que en el método anterior.

20	17	42	01	72	33	94	55	89	65	58	60
72	49	04	27	56	49	11	63	77	79	23	00
94	70	49	05	74	64	00	26	07	23	60	31
22	15	78	49	74	37	50	94	13	90	08	14
93	29	12	20	26	22	66	98	37	53	82	62
45	04	77	48	87	77	33	58	12	08	91	12
16	23	91	95	97	98	52	49	99	78	30	37
04	50	65	37	99	57	74	98	93	99	78	30
03	64	59	55	85	63	49	46	61	89	33	79
62	49	00	67	28	96	19	65	13	44	78	39
89	03	90	40	10	18	43	37	68	97	28	19

Figura 3.1: Ejemplo de tabla de números aleatorios.

Tomando como ejemplo la figura 3.1 para la selección de una muestra en una población de tamaño 100, se parte desde el primer número, y se seleccionan los números necesarios en orden, bien siguiendo las filas o las columnas, ya que no afecta a la aleatoriedad de la selección. En el caso de querer una muestra de tamaño 5

tomaremos los 5 primeros números de izquierda a derecha o de arriba hacia abajo, por lo que nuestra muestra constará de los individuos de la población numerados con los índices 20, 17, 42, 1 y 72.

Para su aplicación en software informático se indexan los individuos y se utiliza un generador aleatorio de números para obtener tantos números entre 0 y 1 como individuos existen en la población. De entre esos números se seleccionan tantos valores como se requieran para la muestra siguiendo una regla como por ejemplo, utilizando los valores más grandes o los más pequeños, etc. Haciendo coincidir el índice de dichos números con la numeración de la población obtenemos los individuos que forman la muestra aleatoria.

En éste trabajo se ha utilizado el método de la urna para la selección de muestras con reemplazamiento y las tablas de números aleatorios para el caso sin reemplazamiento.

3.2. Estimadores lineales insesgados

3.2.1. Muestreo aleatorio simple sin reemplazamiento

Tomando el estimador de Horwitz-Thompson como el estimador lineal insesgado general podemos obtener los estimadores lineales insesgados de los parámetros que precisamos. En concreto, tenemos:

$$\hat{\theta}_{HT} = \sum_{i=1}^n \frac{Y_i}{\pi_i} \quad (3.2)$$

Para obtener el estimador del total poblacional realizamos el siguiente desarrollo:

$$\theta = X = \sum_{i=1}^N X_i \Rightarrow Y_i = X_i \Rightarrow \hat{\theta} = \hat{X} = \sum_{i=1}^n \frac{X_i}{\pi_i} = \sum_{i=1}^n \frac{X_i}{\frac{n}{N}} = N \frac{1}{n} \sum_{i=1}^n X_i = N\bar{x} \quad (3.3)$$

Sustituyendo de forma similar para los parámetros media, proporción y total de clase tenemos:

$$\hat{\theta} = \hat{\bar{X}} = \sum_{i=1}^n \frac{\frac{X_i}{N}}{\pi_i} = \sum_{i=1}^n \frac{\frac{X_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{x} \quad (3.4)$$

$$\hat{\theta} = \hat{P} = \sum_{i=1}^n \frac{A_i}{\pi_i} = \sum_{i=1}^n \frac{\frac{A_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^n A_i = P \quad (3.5)$$

$$\hat{\theta} = \hat{A} = \sum_{i=1}^n \frac{A_i}{\pi_i} = \sum_{i=1}^n \frac{A_i}{\frac{n}{N}} = N \frac{1}{n} \sum_{i=1}^n A_i = NP \quad (3.6)$$

Donde A_i son observaciones de una variable dicotómica con valores 0 ó 1.

3.2.2. Muestreo aleatorio simple con reemplazamiento

Como ya comentamos al principio de la sección, la probabilidad de inclusión en la muestra para este tipo de muestreo es igual que para el muestreo sin reemplazamiento, por lo que sus estimadores lineales insesgados serán equivalentes a los de las fórmulas 3.3 a 3.6.

3.3. Estimadores de la varianza

3.3.1. Muestreo aleatorio simple sin reemplazamiento

Partiendo de la expresión general del estimador insesgado para el cálculo de la varianza en el muestreo sin reposición:

$$V(\widehat{\theta_{HH}}) = \sum_{i=1}^n \frac{Y_i^2}{\pi_i^2} (1 - \pi_i) + \sum_{i < j}^n \frac{Y_i Y_j}{\pi_i \pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \quad (3.7)$$

y sustituyendo para los estimadores tenemos:

$$\hat{V}(\hat{X}) = N^2(1 - f) \frac{\hat{S}^2}{n} \quad (3.8)$$

$$\hat{V}(\hat{X}) = (1 - f) \frac{\hat{S}^2}{n} \quad (3.9)$$

$$\hat{V}(\hat{P}) = (1 - f) \frac{\hat{P}\hat{Q}}{n - 1} \quad (3.10)$$

$$\hat{V}(\hat{A}) = N^2(1 - f) \frac{\hat{P}\hat{Q}}{n - 1} \quad (3.11)$$

donde $f = \frac{n}{N}$ y la cuasivarianza muestral es un estimador insesgado de la cuasivarianza poblacional, y queda definida por:

$$\hat{S}^2 = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{x})^2 \quad (3.12)$$

3.3.2. Muestreo aleatorio simple con reemplazamiento

La expresión general del estimador insesgado de la varianza en el muestreo aleatorio con reemplazamiento es:

$$V(\widehat{\theta}_{HH}) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{Y_i}{P_i} - \hat{Y}_{HH} \right)^2 \quad (3.13)$$

que al aplicarla al estimador del total, media, proporción y total de clase obtenemos:

$$\hat{V}(\hat{X}) = N^2 \frac{\hat{S}^2}{n} \quad (3.14)$$

$$\hat{V}(\hat{X}) = \frac{\hat{S}^2}{n} \quad (3.15)$$

$$\hat{V}(\hat{P}) = \frac{\hat{P}\hat{Q}}{n-1} \quad (3.16)$$

$$\hat{V}(\hat{A}) = N^2 \frac{\hat{P}\hat{Q}}{n-1} \quad (3.17)$$

donde la cuasivarianza muestral \hat{S}^2 es un estimador insesgado de la varianza poblacional.

Como puede comprobarse estas expresiones son equivalentes a las expresiones 3.8 a 3.11 del muestreo sin reposición excepto el factor $(1-f)$.

3.4. Tamaño de la muestra

Cuando se realiza el estudio de una variable de interés sobre una población mediante técnicas de muestreo, es deseable que al realizar la estimación se aseguren unas tolerancias en el error.

Desarrollando la expresión del error de muestreo $e = \sqrt{\hat{V}(\hat{\theta})}$ y despejando n se puede obtener el tamaño de muestra que debería tomarse dependiendo del estimador y el tipo de muestreo realizado para lograr los resultados deseados.

Las diferentes situaciones que se han usado en este trabajo, son las siguientes:

3.4.1. Error de muestreo dado

Es el escenario más simple, en el que determinamos el tamaño de la muestra en base al error de muestreo absoluto a cometer. Tomamos el ejemplo de la estimación

para la media en el muestreo aleatorio sin reemplazamiento:

$$e = \sqrt{\hat{V}(\hat{\theta})} = \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} \Rightarrow n = \frac{NS^2}{Ne^2 + S^2} \quad (3.18)$$

Al tomar muestras con este tamaño es posible que el error de muestreo sea ligeramente mayor que el especificado, pero siempre será cercano a su valor.

3.4.2. Error de muestreo y coeficiente de confianza dados

En el caso en el que se desea obtener un error de muestreo menor o igual al especificado de forma más consistente que en el anterior escenario es necesario introducir un coeficiente de confianza α entre 0 y 1 de forma que con probabilidad $1 - \alpha$ nos encontremos en la situación deseada.

Para ello tomamos λ_α como el cuartil de la distribución normal estándar para el valor de α dado. Desarrollando la expresión de la media:

$$e_\alpha = \lambda_\alpha e = \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} \Rightarrow n = \frac{\frac{\lambda_\alpha^2 S^2}{e_\alpha^2}}{1 + \frac{\lambda_\alpha^2 S^2}{e_\alpha^2}} = \frac{n_\alpha}{1 + \frac{n_\alpha}{N}} \quad (3.19)$$

Al introducir el coeficiente de confianza el valor de n siempre será mayor que en el apartado 3.4.1

3.4.3. Error relativo de muestreo dado

El error relativo representa otra medida de precisión distinta al error de muestreo y se define como el coeficiente del error de muestreo sobre el parámetro y el valor del parámetro sobre la población, o lo que es lo mismo su valor real.

$$e_r = \frac{\sqrt{\widehat{V}(\bar{X})}}{E(\bar{X})} = \frac{\sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}}{\bar{X}} \Rightarrow n = \frac{N\left(\frac{S}{\bar{X}}\right)^2}{e_r^2 + \frac{\left(\frac{S}{\bar{X}}\right)^2}{N}} = \frac{NC^2}{e_r^2 + \frac{C^2}{N}} \quad (3.20)$$

3.4.4. Error relativo de muestreo y coeficiente de confianza dados

Siguiendo la propuesta del apartado 3.4.2 y la expresión del error relativo del apartado 3.4.3 tenemos:

$$e_{r\alpha} = \lambda_\alpha \frac{\sqrt{\hat{V}(\bar{X})}}{E(\bar{X})} \Rightarrow n = \frac{\lambda_\alpha^2 C^2}{e_{r\alpha}^2 + \lambda_\alpha^2 \frac{C^2}{N}} \quad (3.21)$$

Desarrollando para el resto de parámetros y en ambos tipos de muestreo finalmente tenemos:

3.4.5. Muestreo aleatorio simple sin reemplazamiento

Se incluye en este apartado un resumen de las fórmulas para estimar los tamaños de muestra en el muestreo irrestricto aleatorio, agrupadas en la tabla 3.1.

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{N^2 S^2}{e^2 + NS^2}$	$\frac{NC^2}{Ne_r^2 + C^2}$	$\frac{\lambda_\alpha^2 N^2 S^2}{e^2 + \lambda_\alpha^2 NS^2}$	$\frac{\lambda_\alpha^2 NC^2}{Ne_{r\alpha}^2 + \lambda_\alpha^2 C^2}$
Media	$\frac{NS^2}{Ne^2 + S^2}$	$\frac{NC^2}{Ne_r^2 + C^2}$	$\frac{\lambda_\alpha^2 NS^2}{Ne^2 + \lambda_\alpha^2 S^2}$	$\frac{\lambda_\alpha^2 NC^2}{Ne_{r\alpha}^2 + \lambda_\alpha^2 C^2}$
Proporción	$\frac{NPQ}{e^2(N-1) + PQ}$	$\frac{NQ}{P(N-1)e_r^2 + Q}$	$\frac{\lambda_\alpha^2 NPQ}{e^2(N-1) + \lambda_\alpha^2 PQ}$	$\frac{NQ\lambda_\alpha^2}{e_{r\alpha}^2(N-1)P + \lambda_\alpha^2 Q}$
Total de clase	$\frac{N^3 PQ}{e^2(N-1) + N^2 PQ}$	$\frac{NQ}{P(N-1)e_r^2 + Q}$	$\frac{\lambda_\alpha^2 N^3 PQ}{e^2(N-1) + \lambda_\alpha^2 N^2 PQ}$	$\frac{NQ\lambda_\alpha^2}{e_{r\alpha}^2(N-1)P + \lambda_\alpha^2 Q}$

Tabla 3.1: Tamaños de muestra en el muestreo aleatorio simple sin reemplazamiento.

3.4.6. Muestreo aleatorio simple con reemplazamiento

En la tabla 3.2 se resumen las fórmulas de estimación del tamaño poblacional en el caso del muestreo aleatorio simple con reemplazamiento.

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{N^2 \sigma^2}{e^2}$	$\frac{C^2}{e_r^2}$	$\frac{\lambda_\alpha^2 N^2 \sigma^2}{e^2}$	$\frac{\lambda_\alpha^2 C^2}{e_{r\alpha}^2}$
Media	$\frac{\sigma^2}{e^2}$	$\frac{C^2}{e_r^2}$	$\frac{\lambda_\alpha^2 \sigma^2}{e^2}$	$\frac{\lambda_\alpha^2 C^2}{e_{r\alpha}^2}$
Proporción	$\frac{PQ}{e^2}$	$\frac{Q}{Pe_r^2}$	$\frac{\lambda_\alpha^2 PQ}{e^2}$	$\frac{\lambda_\alpha^2 Q}{Pe_{r\alpha}^2}$
Total de clase	$\frac{N^2 PQ}{e^2}$	$\frac{Q}{Pe_r^2}$	$\frac{N^2 \lambda_\alpha^2 PQ}{e^2}$	$\frac{\lambda_\alpha^2 Q}{Pe_{r\alpha}^2}$

Tabla 3.2: Tamaños de muestra en el muestreo aleatorio simple con reemplazamiento.

3.5. Dominios de estudio

Es posible que dentro de la población los individuos estén divididos en subpoblaciones o dominios de los cuales interesa estudiar la variable de interés sobre una de ellas. El problema reside en que no siempre el marco de estudio permite muestrear la subpoblación deseada hasta que no se evalúan los individuos de un marco más general.

Una vez tomada una muestra y etiquetados sus individuos en dominios, es posible estudiar la variable de interés para dicho dominio. Para el caso de la media de un dominio tendríamos:

$$\hat{Y}_j = \sum_{k=1}^{n_j} Y_{jk} \quad (3.22)$$

Donde $Y_{jk}, k = 1 \dots n_j$ son las medidas de la variable de interés para cada individuo del dominio y n_j el tamaño del dominio dentro de la muestra.

Para la estimación de la varianza usaríamos

$$\widehat{V(\hat{Y}_j)} = (1 - \frac{n_j}{N_j}) \frac{\hat{S}_j^2}{n_j} \quad (3.23)$$

Donde \hat{S}_j^2 es la cuasivarianza muestral del dominio y N_j el tamaño del dominio en la población. En caso de no conocer éste último parámetro se puede realizar la aproximación $\frac{n_j}{N_j} = \frac{n}{N}$.

3.6. Aplicaciones en la librería `samplingR`

Las funciones desarrolladas para la aplicación de los conceptos teóricos mostrados anteriormente utilizarán el prefijo *srs* como abreviación de *simple random sampling*, y son las siguientes.

- `srs.sample`: Dado un tamaño poblacional N y un tamaño de muestra n devuelve una muestra aleatoria simple. Dicha muestra puede ser tomada con o sin reemplazamiento, dependiendo del valor del parámetro *replace*. En caso de aportar un conjunto de datos devolverá los individuos de los datos que conforman la muestra. Al ser esto opcional, si no se aporta un conjunto de datos devolverá los índices de los individuos que conforman la muestra.

Realiza las funciones explicadas en el apartado 3.1.

- `srs.estimator`: Dada una muestra de datos obtiene el estimador poblacional del parámetro especificado, entre *total poblacional*, *media*, *proporción* y *total de clase*. También calcula su varianza estimada, error de muestreo y opcionalmente su error de estimación y un intervalo de confianza si se especifica el coeficiente de confianza en el parámetro *alpha*.

Se permite especificar si el muestreo se realiza con o sin reemplazamiento para realizar estimaciones más precisas.

Realiza las funciones explicadas en el apartado 3.2 y 3.3.

- `srs.domainestimator`: Realiza las mismas estimaciones de la función anterior, pero sobre una subpoblación o dominio de los datos muestrales.

Realiza las funciones explicadas en el apartado 3.5.

- `srs.samplesize`: Calcula el tamaño de muestra necesario para cometer un error de muestreo menor del especificado. Dicho error puede ser absoluto o relativo,

según el parámetro *relative*, y se permite la relajación de su estimación si se especifica un coeficiente de confianza.

Para la estimación no se pide aportar los datos poblacionales por la posibilidad de no disponer de ellos, si no que se deben aportar medidas estadísticas tales como la estimación de la varianza y el tamaño poblacional. También se puede especificar el tipo de muestreo que se va a utilizar para realizar estimaciones más precisas.

Aplica los conceptos explicados en el apartado 3.4.

Capítulo 4

Muestreo estratificado

El objetivo del muestreo estratificado consiste en dividir una población heterogénea en subpoblaciones no solapadas lo más homogéneas posibles llamadas estratos, los cuales pueden venir determinados por factores demográficos, geográficos, socioeconómicos u otras variables relevantes para el estudio.

Gracias a esta estrategia es posible representar de manera más precisa la información de los distintos estratos, teniendo en cuenta la heterogeneidad que hay entre ellos. Además podemos realizar inferencias sobre cada uno de los estratos estudiados. Otra ventaja de esta división es la de poder destinar más recursos en la recogida de una muestra en aquellos estratos con mayor variabilidad para así ganar precisión reduciendo el error de muestreo.

Este tipo de muestreo es un método poderoso y flexible que es ampliamente usado en la práctica. En las encuestas económicas (son las dirigidas a las empresas) es el principal tipo de diseño utilizado. Por ejemplo, se utiliza por el INE en los cálculos del *Índice de Comercio al por Menor* o la *Estadística Estructural de Empresas*.

4.1. Selección de la muestra

Para formar una muestra estratificada basta con tomar una muestra aleatoria (existen métodos para utilizar diferentes tipos de muestreo que además no debe de ser el mismo en cada estrato) para cada uno de los estratos en los que se divide la población. De este procedimiento surge una nueva cuestión: ¿Cómo realizamos el reparto del tamaño muestral entre los diferentes estratos?

La *afijación* de la muestra es el nombre con el que se denomina esta adjudicación de tamaños de submuestra para cada uno de los estratos. Entre las más destacadas y utilizadas en la práctica, se encuentran la afijación uniforme, proporcional, de mínima varianza y óptima, que vemos a continuación.

4.1.1. Afijación uniforme

Se trata de la estrategia más sencilla, consistente en asignar el mismo número de individuos a la muestra de cada uno de los estratos, teniendo entonces que $n_h = \frac{n}{L}$ $\forall h = 1, \dots, L$ donde L es el número de estratos. En el caso de que $\frac{n}{L}$ no sea un número entero n_h se redondea según sea conveniente para que al final nos acerquemos lo máximo posible al tamaño muestral deseado.

4.1.2. Afijación proporcional

Consiste en asignar individuos de la muestra de forma proporcional al tamaño poblacional de cada estrato, asegurando así la equiprobabilidad de que un individuo de la población pertenezca a la muestra, con una probabilidad de inclusión de $\pi_{hi} = \frac{n_h}{N_h}$.

Con este tipo de afijación, el tamaño de la muestra para cada estrato queda definido por:

$$n_h = N_h \frac{n}{N} \quad (4.1)$$

4.1.3. Afijación de mínima varianza

También conocida como afijación de Neyman, pretende calcular los tamaños de muestra para cada estrato de forma que la varianza de los estimadores sea mínima. Se trata por lo tanto de resolver un problema de optimización en el que queremos minimizar la función objetivo $V(\theta)$ bajo la restricción $\sum_{h=1}^L n_h = n$.

El resultado de este problema da la siguiente expresión para el tamaño muestral del estrato h :

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^L N_h S_h} \quad (4.2)$$

Que coincide para el cálculo del estimador del total y de la media.

4.1.4. Afijación óptima

Este tipo de afijación busca obtener la mínima varianza posible de los estimadores para un coste fijado C . Este coste será el resultado de la suma de los costes de entrevistar individuos de los diferentes estratos para la muestra, cada uno con un coste determinado.

Esto conlleva un problema de optimización con la misma función objetivo que la afijación de Neyman y la restricción $\sum_{h=1}^L c_h n_h = C$, donde c_h representa el coste de entrevistar un individuo del estrato h para la muestra.

Resolviendo el problema llegamos a la expresión:

$$n_h = n \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \frac{N_h S_h}{\sqrt{c_h}}} \quad (4.3)$$

que también coincide para todos los estimadores.

Una variante de este enfoque implica añadir al problema anterior un coste máximo de estudio C y opcionalmente un coste base C_{ini} , llegando a una expresión similar dada por:

$$n_h = (C - C_{ini}) \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L N_h S_h \sqrt{c_h}} \quad (4.4)$$

4.2. Estimadores lineales insesgados

El cálculo de los estimadores lineales en el muestreo estratificado requiere de realizar la suma de los estimadores de todos los estratos, por lo que a partir de la fórmula general del estimador Horwitz-Thompson 3.2 llegamos a:

$$\hat{\theta}_{HH} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{X_{hi}}{\pi_{hi}} \quad (4.5)$$

Extendiendo este procedimiento a las fórmulas 3.3 a 3.6 del muestreo aleatorio simple obtenemos:

$$\hat{\theta} = \hat{X} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{X_{hi}}{\pi_{hi}} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{X_i}{\frac{n_h}{N_h}} = \sum_{h=1}^L N_h \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^L N_h \bar{x}_h \quad (4.6)$$

$$\hat{\theta} = \hat{X} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{\frac{X_{hi}}{N_h}}{\pi_{hi}} = \sum_{h=1}^L \frac{1}{N} \sum_{i=1}^{n_h} \frac{X_{hi}}{\frac{n_h}{N_h}} = \sum_{h=1}^L \frac{N_h}{N} \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^L W_h \bar{x}_h \quad (4.7)$$

$$\hat{\theta} = \hat{P} = \sum_{h=1}^L \frac{1}{N} \sum_{i=1}^{n_h} \frac{A_{hi}}{\pi_{hi}} = \sum_{h=1}^L \frac{1}{N} \sum_{i=1}^{n_h} \frac{A_{hi}}{\frac{n_h}{N_h}} = \sum_{h=1}^L \frac{N_h}{N} \frac{1}{n_h} \sum_{i=1}^{n_h} A_{hi} = \sum_{h=1}^L W_h P_h \quad (4.8)$$

$$\hat{\theta} = \hat{A} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{A_{hi}}{\pi_{hi}} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{A_{hi}}{\frac{n_h}{N_h}} = \sum_{h=1}^L N_h \frac{1}{n_h} \sum_{i=1}^{n_h} A_{hi} = \sum_{h=1}^L N_h P_h \quad (4.9)$$

los cuales coinciden con los estimadores del muestreo estratificado sin reemplazamiento.

4.3. Estimación de varianzas

La estimación de la varianza en cualquier proceso de investigación por muestreo es de suma importancia, ya que gracias a este valor podremos estimar la calidad del estudio realizado. A continuación se procede a relatar cómo se obtiene este valor distinguiendo si el método muestral utilizado es con o sin reemplazamiento.

4.3.1. Muestreo sin reemplazamiento

La estimación de la varianza para el estimador θ es igual a la suma de las varianzas de θ en cada uno de los estratos al ser las muestras independientes. Sustituyendo θ por cada uno de los estimadores obtenemos:

$$\hat{V}(\hat{X}) = \sum_{h=1}^L N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h} \quad (4.10)$$

$$\hat{V}(\hat{\hat{X}}) = \sum_{h=1}^L W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h} \quad (4.11)$$

$$\hat{V}(\hat{P}) = \sum_{h=1}^L W_h^2 (1 - f_h) \frac{\hat{P}_h \hat{Q}_h}{n_h - 1} \quad (4.12)$$

$$\hat{V}(\hat{A}) = \sum_{h=1}^L N_h^2 (1 - f_h) \frac{\hat{P}_h \hat{Q}_h}{n_h - 1} \quad (4.13)$$

4.3.2. Muestreo con reemplazamiento

Aplicando el mismo fundamento con las varianzas en el muestreo estratificado con reemplazamiento obtenemos:

$$\hat{V}(\hat{X}) = \sum_{h=1}^L N_h^2 \frac{\hat{S}_h^2}{n_h} \quad (4.14)$$

$$\hat{V}(\hat{\hat{X}}) = \sum_{h=1}^L W_h^2 \frac{\hat{S}_h^2}{n_h} \quad (4.15)$$

$$\hat{V}(\hat{P}) = \sum_{h=1}^L W_h^2 \frac{\hat{P}_h \hat{Q}_h}{n_h - 1} \quad (4.16)$$

$$\hat{V}(\hat{A}) = \sum_{h=1}^L N_h^2 \frac{\hat{P}_h \hat{Q}_h}{n_h - 1} \quad (4.17)$$

que de forma similar a como ocurre en el apartado 3.3.2 de estimación de varianzas en el muestreo aleatorio simple, se diferencian del muestreo sin reemplazamiento en el factor $(1 - f_h)$, denominado *factor de corrección para poblaciones finitas*. //

4.4. Tamaño de la muestra

Al igual que en el muestreo aleatorio simple, es posible determinar el tamaño de muestra necesario para realizar estimaciones cometiendo un error de muestreo menor que el determinado.

En el caso del muestreo estratificado, tras obtener el tamaño de muestra deben asignarse los recursos entre los distintos estratos de población, por lo que la afijación seleccionada también afectará al error cometido y por lo tanto a la estimación del tamaño muestral.

4.4.1. Afijación proporcional

Para estimar el tamaño de muestra con afijación proporcional dado un error de muestreo para el estimador de la media, usaremos el procedimiento de la sección 3.4.1 utilizando la fórmula de la varianza en la afijación proporcional.

$$e^2 = \hat{V}(\hat{X}) = \frac{1 - \frac{n}{N}}{n} \sum_{h=1}^L W_h S_h^2 \Rightarrow n = \frac{\sum_{h=1}^L W_h S_h^2}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2} \quad (4.18)$$

4.4.2. Afijación de mínima varianza

De forma similar utilizando el valor de varianza mínima del estimador de la media en la afijación de Neyman:

$$e^2 = \hat{V}(\hat{X}) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^L W_h S_h \right)^2}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2} \quad (4.19)$$

4.4.3. Afijación óptima

Tener en cuenta los costes por unidad de muestreo en cada estrato a la hora de obtener la varianza de los estimadores modifica la expresión del tamaño muestral:

$$e^2 = \hat{V}(\hat{X}) = \frac{1}{n} \left(\sum_{h=1}^L \frac{W_h S_h}{\sqrt{c_h}} \right) \left(\sum_{h=1}^L W_h S_h c_h \right) - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^L \frac{W_h S_h}{\sqrt{c_h}} \right) \left(\sum_{h=1}^L W_h S_h c_h \right)}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2} \quad (4.20)$$

Aplicando a las fórmulas de estimación con error absoluto y relativo, tanto sin coeficiente de confianza como con él, obtenemos los resúmenes de las siguientes tablas:

4.4.4. Muestreo sin reemplazamiento

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{N \sum_{h=1}^L N_h S_h^2}{e^2 + \sum_{h=1}^L N_h S_h^2}$	$\frac{N \sum_{h=1}^L N_h S_h^2}{N^2 \bar{X}^2 e^2 + \sum_{h=1}^L N_h S_h^2}$	$\frac{N \sum_{h=1}^L N_h S_h^2}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h S_h^2}$	$\frac{N \sum_{h=1}^L N_h S_h^2}{\frac{N^2 \bar{X}^2 e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h S_h^2}$
Media	$\frac{\sum_{h=1}^L W_h S_h^2}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{\sum_{h=1}^L W_h S_h^2}{\bar{X}^2 e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{\sum_{h=1}^L W_h S_h^2}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{\sum_{h=1}^L W_h S_h^2}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$
Proporción	$\frac{\sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{\sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}{P^2 e^2 + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{\sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{\sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}{\frac{P^2 e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$
Total de clase	$\frac{N \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}{e^2 + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{N \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}{N^2 P^2 e^2 + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{N \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{N \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}{\frac{N^2 P^2 e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$

Tabla 4.1: Tamaños de muestra con afijación proporcional en el muestreo sin reemplazamiento

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{(\sum_{h=1}^L N_h S_h)^2}{e^2 + \sum_{h=1}^L N_h S_h^2}$	$\frac{(\sum_{h=1}^L N_h S_h)^2}{N^2 \bar{X}^2 e^2 + \sum_{h=1}^L N_h S_h^2}$	$\frac{(\sum_{h=1}^L N_h S_h)^2}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h S_h^2}$	$\frac{(\sum_{h=1}^L N_h S_h)^2}{\frac{N^2 \bar{X}^2 e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h S_h^2}$
Media	$\frac{(\sum_{h=1}^L W_h S_h)^2}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{(\sum_{h=1}^L W_h S_h)^2}{\bar{X}^2 e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{(\sum_{h=1}^L W_h S_h)^2}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{(\sum_{h=1}^L W_h S_h)^2}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$
Proporción	$\frac{(\sum_{h=1}^L W_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{(\sum_{h=1}^L W_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{P^2 e^2 + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{(\sum_{h=1}^L W_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{(\sum_{h=1}^L W_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{\frac{P^2 e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h \frac{N_h}{N_{h-1}} P_h Q_h}$
Total de clase	$\frac{(\sum_{h=1}^L N_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{e^2 + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{(\sum_{h=1}^L N_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{N^2 P^2 e^2 + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{(\sum_{h=1}^L N_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$	$\frac{(\sum_{h=1}^L N_h \sqrt{\frac{N_h}{N_{h-1}}} P_h Q_h)^2}{\frac{N^2 P^2 e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h \frac{N_h}{N_{h-1}} P_h Q_h}$

Tabla 4.2: Tamaños de muestra con afijación de mínima varianza en el muestreo sin reemplazamiento

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{e^2 + \sum_{h=1}^L N_h S_h^2}$	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{\bar{X}^2 e^2 + \sum_{h=1}^L N_h S_h^2}$	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h S_h^2}$	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2} + \sum_{h=1}^L N_h S_h^2}$
Media	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{\bar{X}^2 e^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$

Tabla 4.3: Tamaños de muestra con afijación óptima en el muestreo sin reemplazamiento

Para obtener los tamaños para la proporción y el total de clase se debe sustituir S_h por $\sqrt{\frac{N_h}{N_h-1}} P_h(1 - P_h)$

4.4.5. Muestreo con reemplazamiento

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{N \sum_{h=1}^L N_h \sigma_h^2}{e^2}$	$\frac{N \sum_{h=1}^L N_h \sigma_h^2}{N \bar{X}^2 e^2}$	$\frac{N \sum_{h=1}^L N_h \sigma_h^2}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{N \sum_{h=1}^L N_h \sigma_h^2}{\frac{N \bar{X}^2 e^2}{\lambda_\alpha^2}}$
Media	$\frac{\sum_{h=1}^L W_h \sigma_h^2}{e^2}$	$\frac{\sum_{h=1}^L W_h \sigma_h^2}{\bar{X}^2 e^2}$	$\frac{\sum_{h=1}^L W_h \sigma_h^2}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{\sum_{h=1}^L W_h \sigma_h^2}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2}}$
Proporción	$\frac{\sum_{h=1}^L W_h P_h Q_h}{e^2}$	$\frac{\sum_{h=1}^L W_h P_h Q_h}{P^2 e^2}$	$\frac{\sum_{h=1}^L W_h P_h Q_h}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{\sum_{h=1}^L W_h P_h Q_h}{\frac{P^2 e^2}{\lambda_\alpha^2}}$
Total de clase	$\frac{N \sum_{h=1}^L N_h P_h Q_h}{e^2}$	$\frac{N \sum_{h=1}^L N_h P_h Q_h}{N P^2 e^2}$	$\frac{N \sum_{h=1}^L N_h P_h Q_h}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{N \sum_{h=1}^L N_h P_h Q_h}{\frac{N P^2 e^2}{\lambda_\alpha^2}}$

Tabla 4.4: Tamaños de muestra con afijación proporcional en el muestreo con reemplazamiento

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{(\sum_{h=1}^L N_h \sigma_h)^2}{e^2}$	$\frac{(\sum_{h=1}^L N_h \sigma_h)^2}{N^2 \bar{X}^2 e^2}$	$\frac{(\sum_{h=1}^L N_h \sigma_h)^2}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{(\sum_{h=1}^L N_h \sigma_h)^2}{\frac{N^2 \bar{X}^2 e^2}{\lambda_\alpha^2}}$
Media	$\frac{(\sum_{h=1}^L W_h \sigma_h)^2}{e^2}$	$\frac{(\sum_{h=1}^L W_h \sigma_h)^2}{\bar{X}^2 e^2}$	$\frac{(\sum_{h=1}^L W_h \sigma_h)^2}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{(\sum_{h=1}^L W_h \sigma_h)^2}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2}}$
Proporción	$\frac{(\sum_{h=1}^L W_h \sqrt{P_h Q_h})^2}{e^2}$	$\frac{(\sum_{h=1}^L W_h \sqrt{P_h Q_h})^2}{P^2 e^2}$	$\frac{(\sum_{h=1}^L W_h \sqrt{P_h Q_h})^2}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{(\sum_{h=1}^L W_h \sqrt{P_h Q_h})^2}{\frac{P^2 e^2}{\lambda_\alpha^2}}$
Total de clase	$\frac{(\sum_{h=1}^L N_h \sqrt{P_h Q_h})^2}{e^2}$	$\frac{(\sum_{h=1}^L N_h \sqrt{P_h Q_h})^2}{N^2 P^2 e^2}$	$\frac{(\sum_{h=1}^L N_h \sqrt{P_h Q_h})^2}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{(\sum_{h=1}^L N_h \sqrt{P_h Q_h})^2}{\frac{N^2 P^2 e^2}{\lambda_\alpha^2}}$

Tabla 4.5: Tamaños de muestra con afijación de mínima varianza en el muestreo con reemplazamiento

Tipo de error Parametro	Absoluto	Relativo	Absoluto y coeficiente de confianza adicional	Relativo y coeficiente de confianza adicional
Total	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{e^2}$	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{\bar{X}^2 e^2}$	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{(\sum_{h=1}^L N_h S_h / \sqrt{c_h})(\sum_{h=1}^L N_h S_h c_h)}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2}}$
Media	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{e^2}$	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{\bar{X}^2 e^2}$	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{\frac{e^2}{\lambda_\alpha^2}}$	$\frac{(\sum_{h=1}^L W_h S_h / \sqrt{c_h})(\sum_{h=1}^L W_h S_h c_h)}{\frac{\bar{X}^2 e^2}{\lambda_\alpha^2}}$

Tabla 4.6: Tamaños de muestra con afijación óptima en el muestreo con reemplazamiento

Para obtener los tamaños para la proporción y el total de clase se debe sustituir S_h por $\sqrt{\frac{N_h}{N_h-1}} P_h (1 - P_h)$

4.5. Tamaño de muestra dado un presupuesto

En ocasiones, el estudio está limitado en cuanto a la cantidad de información que es capaz de recoger debido a un presupuesto máximo establecido. Otro factor a tener en cuenta es el coste que supone poner en marcha el estudio, antes incluso de empezar a recopilar información. Por lo tanto, tenemos un presupuesto para tomar la muestra de $C - C_{ini}$, donde C es el presupuesto del estudio y C_{ini} el coste de iniciar el proyecto.

Tomando afijación uniforme, el tamaño de muestra sería:

$$n = \frac{(C - C_{ini})L}{\sum_{h=1}^L c_h} \quad (4.21)$$

Con afijación proporcional:

$$n = \frac{(C - C_{ini})}{\sum_{h=1}^L W_h c_h} \quad (4.22)$$

Con afijación de mínima varianza, la cual coincide con la afijación óptima al optimizar con la restricción de costes:

$$n = \frac{(C - C_{ini}) \frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L W_h S_h \sqrt{c_h}} \quad (4.23)$$

4.6. Aplicaciones para el muestreo estratificado de la librería `samplingR`

Las funciones desarrolladas para la aplicación de los conceptos teóricos mostrados en este capítulo utilizarán el prefijo *strata* como abreviación de *stratified sampling*, y son las siguientes.

- `strata.sample`: Dado un tamaño poblacional N devuelve una muestra estratificada con tantos individuos de cada estrato como se especifique en el vector de tamaños de muestra n . Dicha muestra puede ser tomada con o sin reemplazamiento, dependiendo del valor del parámetro *replace*.

Realiza las funciones explicadas en el apartado 4.1.

- `strata.allocation`: Realiza la división del tamaño de muestra general para cada uno de los estratos, dependiendo del tipo de afijación especificado en el parámetro *allocation*.

En el caso de especificar la afijación óptima, se debe incluir un vector de costes ch , en el cual opcionalmente se puede declarar un coste máximo de estudio y un coste de inicio del proyecto.

Realiza las funciones explicadas en los subapartados 4.1.1 a 4.1.4.

- `strata.estimator`: Dada una muestra de datos obtiene el estimador poblacional del parámetro especificado. También calcula su varianza estimada, error de muestreo y opcionalmente su error de estimación y un intervalo de confianza si se especifica el coeficiente de confianza en el parámetro *alpha*.

Se permite indicar si el muestreo se realiza con o sin reemplazamiento para realizar estimaciones más precisas.

Realiza las funciones explicadas en el apartado 4.2 y 4.3.

- `strata.samplesize`: Calcula el tamaño de muestra necesario para cometer un error de muestreo menor del especificado. Dicho error puede ser absoluto o relativo, según el parámetro *relative*, y se permite la relajación de su estimación si se especifica un coeficiente de confianza.

Para el cálculo no se pide aportar los datos poblacionales por la posibilidad de no disponer de ellos, si no que se deben aportar medidas estadísticas tales como la estimación de la varianza y el tamaño poblacional. También se puede especificar el tipo de muestreo y la afijación que se va a utilizar para realizar estimaciones más precisas.

Aplica los conceptos explicados en el apartado 4.4.

- `strata.samplesize.cost`: Calcula el tamaño de muestra dada la afijación a utilizar y un vector de costes, incluyendo la posibilidad de especificar el coste máximo del estudio y el coste de inicio del proyecto.

Realiza las funciones explicadas en el apartado 4.5.

Capítulo 5

Muestreo sistemático

Otra de las técnicas muestrales muy utilizadas consiste en el denominado muestreo sistemático. Este método consiste en seleccionar aleatoriamente y con la misma probabilidad una unidad entre los primeros “ k ” elementos del marco poblacional. Este entero positivo k se fija previamente y tiene como denominación intervalo muestral. El resto de elementos de la muestra se eligen de una manera sistemática, dejando una separación de k elementos entre todas las unidades seleccionadas. De esta forma, obtenemos las denominadas *muestras 1 en k* .

El muestreo sistemático ofrece varias ventajas prácticas, en particular su simplicidad de ejecución. El hecho de solo realizar una única selección aleatoria es una gran ventaja. Es fácil, por ejemplo, para un entrevistador seleccionar una muestra sistemática mientras está en campo.

En el muestreo sistemático, podemos decir que existe un efecto que podemos llamar de extensión o “estratificación”, si cada grupo de k elementos consecutivos a partir del primero se considera como un estrato. No obstante, debe tenerse en cuenta que en el muestreo estratificado la selección se realiza de forma independiente en cada estrato, mientras que en el muestreo sistemático todos los elementos seleccionados ocupan el mismo lugar dentro de cada grupo de k elementos.

Entre las ventajas de este tipo de muestreo, se pueden destacar:

- Extiende la muestra a toda la población.
- Recoge el posible efecto de estratificación debido al orden en que figuran las unidades en la población.
- Su aplicación y comprobación son fáciles de realizar.

También presenta algunos inconvenientes:

- La posibilidad de aumento de la varianza si existe periodicidad.
- El problema teórico que se presenta en la estimación de varianzas.

5.1. Selección de la muestra

Como ya se ha comentado anteriormente, la mayor virtud de este muestreo es su simplicidad. Al seleccionar el tamaño de muestra n , se divide la población en n filas, llamadas *zonas sistemáticas* de tamaño $\frac{N}{n} = k$, ordenando a la población en una matriz como la de la figura 5.1, lo cual facilita su comprensión y ejecución.

$i \backslash j$	1	2	3	...	j	...	k
1	u_1	u_2	u_3	...	u_j	...	u_k
2	u_{k+1}	u_{k+2}	u_{k+3}	...	u_{k+j}	...	u_{k+k}
3	u_{2k+1}	u_{2k+2}	u_{2k+3}	...	u_{2k+j}		u_{2k+k}
\vdots	\vdots	\vdots		\vdots			\vdots
i	$u_{(i-1)k+1}$	$u_{(i-1)k+2}$	$u_{(i-1)k+3}$...	$u_{(i-1)k+j}$...	$u_{(i-1)k+k}$
\vdots	\vdots	\vdots	\vdots			\vdots	\vdots
n	$u_{(n-1)k+1}$	$u_{(n-1)k+2}$	$u_{(n-1)k+3}$...	$u_{(n-1)k+j}$...	$u_{(n-1)k+k}$

Figura 5.1: Selección de una muestra sistemática 1 en k .

Una vez ordenada ya solo resta seleccionar un número aleatoriamente entre 1 y k , el cual determinará la columna del cuadro anterior de la cual obtendremos los elementos de nuestra muestra.

En su traducción a un lenguaje informático, el método es muy similar. Los datos serán ordenados como una matriz, pero sin utilizar los objetos *matrix*. Esto se debe a que es posible que no haya un número exacto para formar una matriz con las filas y columnas deseadas por el usuario. Este problema viene dado cuando $\frac{N}{n} = k$ no es un número entero. Para solucionar este problema, k será redondeado hacia abajo y los individuos que no completarían esa última fila de la matriz no serán seleccionados.

En algunos casos de la vida real, este problema llega a tener una relevancia importante, como ocurre en el método de selección de miembros a jurado popular. Para obtener estos candidatos, lo que se hace realmente es utilizar un muestreo sistemático con arranque aleatorio. Pero al realizar este tipo de selección, los electores del final de la lista tendrían probabilidad cero de ser elegidos.

Para evitar esta deficiencia se publicó el Real Decreto 1398/1995 artículo 3 [12], según el cual se mantienen hasta 5 decimales del valor de k , obteniendo por lo tanto una sucesión de índices no exactos. Estos índices serán redondeados para seleccionar los individuos de una muestra que sí abarque toda la población.

Posteriormente elegimos un número al azar entre 1 y k al que denominaremos m , y seleccionaremos los individuos con índice $m + ik, \forall i = 0, \dots, n - 1$. De ésta forma obtenemos los individuos de la columna m siguiendo el ejemplo de la figura 5.1.

5.2. Análisis de la varianza y correlación

En el muestreo sistemático resulta de gran utilidad utilizar una descomposición de la información en un formato análogo al que se utiliza para el análisis de la varianza (Análisis of Variance o ANOVA) ya que nos permite obtener medidas de precisión que nos ayuden a determinar el método de estimación más preciso para nuestros datos de entre los posibles.

La tabla ANOVA poblacional para el análisis sistemático contiene los siguientes valores:

Fuente de variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios
Entre muestras	$k - 1$	$\sum_i^n \sum_j^k (\bar{x}_j - \bar{X})^2$	S_{bs}^2
Intramuestras	$N - k$	$\sum_i^n \sum_j^k (X_{ij} - \bar{x}_j)^2$	S_{ws}^2
Total	$N - 1$	$\sum_i^n \sum_j^k (X_{ij} - \bar{X})^2$	S^2

Tabla 5.1: Tabla ANOVA

Donde S_{bs}^2 es la cuasivarianza intermuestral definida por:

$$\frac{\sum_i^n \sum_j^k (\bar{x}_j - \bar{X})^2}{k - 1} \quad (5.1)$$

Y S_{ws}^2 la cuasivarianza intramuestral:

$$\frac{\sum_i^n \sum_j^k (X_{ij} - \bar{x}_j)^2}{N - k} \quad (5.2)$$

Pudiendo escribir la siguiente igualdad:

$$(N - 1)S^2 = (N - k)S_{ws}^2 + (k - 1)S_{bs}^2 \quad (5.3)$$

La comparación de la cuasivarianza intermuestral e intramuestral con la cuasivarianza poblacional nos permite realizar una comparación de los distintos tipos de muestreo vistos hasta ahora.

Si $S_{bs}^2 > S^2$ entonces el muestreo sistemático será más preciso que el muestreo estratificado. De la misma forma si $S_{ws}^2 > S^2$ el muestreo sistemático será más preciso que el aleatorio simple. A igualdad de cuasivarianzas la precisión será la misma

en ambos tipos de muestreo.

Otras medidas de comparación son los coeficientes de correlación intramuestral e intermuestral. Se define el primero como:

$$\rho_w = \frac{2 \sum_j^k \sum_{i < z}^n (X_{ij} - \bar{X})(X_{zj} - \bar{X})}{N(n-1)\sigma^2} \quad (5.4)$$

que obtiene la precisión mínima con valor 1 y precisión máxima con valor $\frac{-1}{n-1}$.

Con $\rho_w = 0$ la precisión del muestreo sistemático coincide con el aleatorio simple con reposición. Entre este valor y 1 el muestreo aleatorio simple será más preciso que el sistemático. Valores menores que 0 hasta el valor de varianza mínima suponen el caso contrario.

El coeficiente de correlación intermuestral viene dado por la expresión:

$$\rho_{wst} = \frac{2 \sum_j^k \sum_{i < z}^n (X_{ij} - \bar{X}_i)(X_{zj} - \bar{X}_z)}{n(n-1)(k-1)S_{wst}^2} \quad (5.5)$$

que obtiene sus valores de precisión mínima y máxima de la misma forma que con la correlación intramuestral y con la misma interpretación de sus valores permite comparar la precisión entre el muestreo sistemático y el muestreo estratificado.

5.3. Estimadores lineales insesgados

Sustituyendo en la fórmula general para una muestra sistemática, obtenemos los estimadores lineales insesgados en el muestreo sistemático:

$$\hat{\theta} = \hat{X} = \sum_{i=1}^n \sum_{j=1}^1 \frac{X_{ij}}{\frac{1}{k}} = \sum_{i=1}^n k X_{ij} = N \frac{1}{n} \sum_{i=1}^n X_{ij} = N \bar{x}_j \quad (5.6)$$

$$\hat{\theta} = \hat{X} = \sum_{i=1}^n \sum_{j=1}^1 \frac{X_{ij}}{\frac{1}{k}} = \frac{1}{n} \sum_{i=1}^n X_{ij} = \bar{x}_j \quad (5.7)$$

$$\hat{\theta} = \hat{P} = \sum_{i=1}^n \sum_{j=1}^1 \frac{A_{ij}}{\frac{1}{k}} = \frac{1}{n} \sum_{i=1}^n A_{ij} = \bar{P}_j \quad (5.8)$$

$$\hat{\theta} = \hat{A} = \sum_{i=1}^n \sum_{j=1}^1 \frac{A_{ij}}{\frac{1}{k}} = N \frac{1}{n} \sum_{i=1}^n A_{ij} = N \bar{P}_j \quad (5.9)$$

5.4. Estimadores de la varianza

La estimación de la varianza es uno de los mayores problemas del muestreo sistemático. Al no tener un método directo, disponemos de 3 métodos de estimación diferentes con distintos resultados. En el apartado 6.2 vimos cómo, a partir de los valores de cuasivarianzas y correlaciones, se puede saber qué tipo de muestreo es más preciso. Diferenciamos 3 casos:

- ρ_w próximo a 0

En el caso de ρ_w positivo y próximo a 0 podemos suponer que la población está ordenada de forma aleatoria, por lo que podremos estimar la varianza con las fórmulas de muestreo aleatorio simple sobre nuestra muestra sistemática.

- ρ_{wst} próximo a 0

Para valores positivos próximos a 0 de ρ_{wst} se puede considerar la falta de aleatoriedad en la selección de la muestra sobre cada zona sistemática (fila) de nuestra matriz de población. Por ello la varianza se estima tomando cada fila como un estrato y realizando las estimaciones sobre una muestra estratificada con una unidad por estrato.

En la práctica se supone un estrato cada 2 zonas sistemáticas, teniendo en la muestra $\frac{n}{2}$ estratos de 2 individuos por estrato. En los casos en los que $\frac{n}{2}$ no sea un número entero se repite un individuo al azar de la muestra al final para cuadrar los resultados.

- Ni ρ_w ni ρ_{wst} próximos a 0

En este caso utilizaremos el método de las muestras interpenetrantes, que se utiliza para formar un estimador cuando tenemos una o más muestras elegidas con el mismo esquema de muestreo y proporcionen por sí mismas un estimador válido del parámetro con el mismo error de muestreo.

Para ello se indica un parámetro extra t que representa el número de arranques o submuestras que utilizaremos para la estimación. En lugar de tomar la muestra tomada originalmente de tamaño n tomaremos t submuestras sistemáticas de tamaño $\frac{n}{t}$.

A partir de las submuestras podemos formar estimadores de la media y el total como:

$$\bar{x}_c = \frac{1}{t} \sum_1^t \bar{x}_i \quad (5.10)$$

$$x_c = \frac{1}{t} N \sum_1^t \bar{x}_i \quad (5.11)$$

Y sus estimadores de varianzas:

$$\widehat{V(\bar{x}_c)} = \frac{1}{t(1-t)} \sum_1^t (\bar{x}_i^2 - \bar{x}_c^2) \quad (5.12)$$

$$\widehat{V(x_c)} = \frac{1}{t(1-t)} \sum_1^t ((N\bar{x}_i)^2 - x_c^2) \quad (5.13)$$

5.5. Aplicaciones para el muestreo sistemático de la librería `samplingR`

Las funciones desarrolladas para la aplicación de los conceptos teóricos mostrados en este capítulo utilizarán el prefijo *syst* como abreviación de *systematic sampling*, y son las siguientes.

- `syst.sample`: Dado un tamaño poblacional N y un tamaño de muestra n devuelve una muestra sistemática, comunmente denominada 1 en k . En caso de aportar un conjunto de datos devolverá los individuos de los datos que conforman la muestra. Al ser esto opcional, si no se aporta un conjunto de datos devolverá los índices de los individuos que conforman la muestra.

Realiza las funciones explicadas en el apartado 5.1.

- `syst.all.samples`: Función que devuelve todas las posibles muestras sistemáticas dato un tamaño de muestra n y un tamaño poblacional N .
- `syst.anova`: Obtiene una tabla de análisis de la varianza sobre los datos de la población, siguiendo el esquema comentado en la tabla 5.1.
- `syst.intercorr`: Realiza el cálculo del coeficiente de correlación interclase de los datos.
- `syst.intracorr`: Realiza el cálculo del coeficiente de correlación intraclase de los datos.

Esto permite realizar una comparación de la precisión de la estimación entre el muestreo sistemático y el muestreo aleatorio simple.

Estas tres funciones abarcan los conceptos explicados en el apartado 5.2.

- `syst.estimator`: Dada una muestra de datos obtiene el estimador poblacional del parámetro especificado. También calcula su varianza estimada, error de muestreo y opcionalmente su error de estimación y un intervalo de confianza si se especifica el coeficiente de confianza en el parámetro *alpha*.

Realiza las funciones explicadas en el apartado 5.3 y 5.4.

Capítulo 6

Muestreo por conglomerados

El muestreo por conglomerados considera que los M individuos de la población están a su vez agrupados en N unidades llamadas conglomerados, de forma que sean lo más heterogéneas posibles, pero lo más homogéneas posibles entre ellas, y de forma que no haya solapamiento. Se trata por tanto del caso opuesto al del muestreo estratificado

Por lo tanto, en este tipo de muestreo, la unidad primaria de muestreo pasa a ser el conglomerado en lugar del individuo y por lo tanto una muestra primaria de tamaño n en el muestreo por conglomerados está formada por n conglomerados compuestos por la totalidad de sus individuos (esto en el supuesto de una muestra monoetápica). Esta forma de selección de la muestra ofrece las siguientes ventajas:

- No requiere de un marco muy específico y detallado ya que no necesita la información individual de los individuos de la población.
- Al seleccionar aleatoriamente conglomerados en lugar de elementos individuales, se reduce la cantidad de recursos necesarios para la recopilación de datos, con el consiguiente ahorro de costes.
- En poblaciones grandes y dispersas, el muestreo por conglomerados puede ser más factible logísticamente, ya que por el tipo de selección facilita la organización y la ejecución del estudio.

Pero también es causa de algunas de sus principales desventajas:

- En los casos en los que la variación dentro de los conglomerados sea mayor que la variación entre los conglomerados puede resultar en una mayor variabilidad y una reducción de la precisión de las estimaciones en comparación a otras técnicas de muestreo.
- Su eficiencia disminuye a medida que aumenta el tamaño de los conglomerados, que es el caso en el que este tipo de muestreo es más útil.
- En el caso de no seleccionar unidades de muestreo correctamente se corre el riesgo de no representar correctamente a la población e introducir sesgos en la estimación.

En este capítulo se abarcará la estimación en el caso de que los estratos sean de igual tamaño o de tamaños similares. En este último caso, lo que se suele hacer es calcular el tamaño promedio de conglomerado como $\bar{M} = \frac{\sum_i^n M_i}{n}$

6.1. Selección de la muestra

Para la selección de la muestra y la estimación de parámetros se han diferenciado los dos casos que a continuación se indica:

- Cuando los datos de la variable a investigar están dados a nivel de conglomerados (por ejemplo total de clase, media del conglomerado, etc.).
- Cuando se disponen de los datos de las unidades poblacionales de los conglomerados en la muestra.

Para obtener una muestra en el primer caso basta con tomar una de tipo aleatoria simple de los datos. Al estar agrupados por conglomerado cada instancia seleccionada para la muestra se trata de un conglomerado completo.

Con los datos no agrupados se crea una lista con todos los conglomerados existentes en la población. Se extrae una muestra aleatoria simple de la lista de conglomerados y se toman todos los individuos cuyo conglomerado coincida con alguno de los conglomerados de la muestra.

6.2. Estimadores lineales insesgados

Si al extraer la muestra se utilizara una variable de apoyo e_i que toma valores 1 si el conglomerado i pertenece a la muestra con probabilidad $p = \frac{n}{N}$ y 0 en caso contrario con probabilidad $1-p$, tenemos que $E(e_i) = \pi_i = \frac{n}{N}$, por lo que aplicando el estimador lineal general de Horwitz-Thompson podemos estimar los parámetros para el muestreo por conglomerados sin reemplazamiento:

$$\hat{X} = \frac{\sum_i^n \sum_j^{\bar{M}} X_{ij}}{\pi_i} = \frac{\sum_i^n \sum_j^{\bar{M}} X_{ij}}{\frac{N}{n}} = \frac{N}{n} \sum_i^n \sum_j^{\bar{M}} X_{ij} = \frac{N}{n} \sum_i^n X_i \quad (6.1)$$

$$\hat{X} = \frac{1}{n} \sum_i^n \frac{1}{\bar{M}} \sum_j^{\bar{M}} X_{ij} = \frac{1}{n} \sum_i^n \frac{1}{\bar{M}} X_i \quad (6.2)$$

$$\hat{P} = \frac{1}{n} \sum_i^n \frac{1}{\bar{M}} \sum_j^{\bar{M}} A_{ij} = \frac{1}{n} \sum_i^n \frac{1}{\bar{M}} A_i \quad (6.3)$$

$$\hat{A} = \frac{N}{n} \sum_i^n \sum_j^{\bar{M}} A_{ij} = \frac{N}{n} \sum_i^n A_i \quad (6.4)$$

Las probabilidades de inclusión en la muestra coinciden en el muestreo con reemplazamiento por lo que sus estimadores insesgados también.

6.3. Estimación de varianzas

Para expresar la estimación de varianzas es preciso conocer la fórmula de la varianza entre conglomerados:

$$\hat{S}_b^2 = \frac{\sum_i^n \sum_j^{\bar{M}} (\bar{X}_i - \bar{X})^2}{n - 1} \quad (6.5)$$

con lo que la estimación de la varianza para el muestreo por conglomerados sin reposición es:

$$\widehat{V(X)} = N^2 \bar{M}^2 (1 - f) \frac{\hat{S}_b^2}{n \bar{M}} \quad (6.6)$$

$$\widehat{V(\bar{X})} = (1 - f) \frac{\hat{S}_b^2}{n \bar{M}} \quad (6.7)$$

$$\widehat{V(P)} = (1 - f) \frac{\hat{S}_b^2}{n \bar{M}} \quad (6.8)$$

$$\widehat{V(A)} = N^2 \bar{M}^2 (1 - f) \frac{\hat{S}_b^2}{n \bar{M}} \quad (6.9)$$

Para el muestreo con reposición las fórmulas se mantienen salvo por el factor (1-f).

6.4. Tamaño de la muestra

A la hora de su aplicación práctica y llevar a cabo la investigación estadística en campo, el diseño de muestreo por conglomerados generalmente conlleva una función de costes de muestreo en su realización.

El cálculo del tamaño muestral implica por lo tanto un problema de optimización con la finalidad de minimizar la varianza del estimador, con restricciones impuestas por dicha función de costes, cuya resolución es compleja y suele realizarse mediante algoritmos algebraicos de optimización iterativos, por lo que quedaría fuera del alcance de los objetivos marcados en este trabajo.

Si eliminamos la restricción de la función de costes podemos obtener las siguientes estimaciones para un error absoluto de muestreo dado en el muestreo sin reemplazamiento:

$$e = \sqrt{\widehat{V(\hat{X})}} = \sqrt{(1-f) \frac{\hat{S}_b^2}{n\bar{M}}} \Rightarrow n = \frac{(1-f)\hat{S}_b^2}{e^2\bar{M}} \quad (6.10)$$

$$e = \sqrt{\widehat{V(\hat{X})}} = \sqrt{(1-f) \frac{N^2\bar{M}\hat{S}_b^2}{n}} \Rightarrow n = \frac{(1-f)N^2\bar{M}\hat{S}_b^2}{e^2} \quad (6.11)$$

Si añadimos un coeficiente de confianza para relajar el cálculo tenemos:

$$e = \lambda_\alpha \sqrt{\widehat{V(\hat{X})}} = \sqrt{(1-f) \frac{\hat{S}_b^2}{n\bar{M}}} \Rightarrow n = \lambda_\alpha^2 \frac{(1-f)\hat{S}_b^2}{e^2\bar{M}} \quad (6.12)$$

$$e = \lambda_\alpha \sqrt{\widehat{V(\hat{X})}} = \sqrt{(1-f) \frac{N^2\bar{M}\hat{S}_b^2}{n}} \Rightarrow n = \frac{\lambda_\alpha^2(1-f)N^2\bar{M}\hat{S}_b^2}{e^2} \quad (6.13)$$

Al igual que dijimos en la estimación de varianzas, cuando realicemos el muestreo con reemplazamiento será necesario eliminar el factor (1-f) de las igualdades anteriores para realizar las estimaciones correctas.

6.5. Aplicaciones para el muestreo por conglomerados de la librería `samplingR`

Las funciones desarrolladas para la aplicación de los conceptos teóricos mostrados en este capítulo utilizarán el prefijo *cluster* como abreviación de *cluster sampling*, y son las siguientes.

- `cluster.sample`: Devuelve una muestra monoetápica de conglomerados para datos no agrupados, es decir, en los que se tiene los datos de cada individuo para cada conglomerado. Dicha muestra puede ser tomada con o sin reemplazamiento, dependiendo del valor del parámetro *replace*.

Realiza las funciones explicadas en el apartado 6.1.

- `cluster.estimator`: Dada una muestra de datos, ya sean agrupados o sin agrupar, obtiene el estimador poblacional del parámetro especificado. También calcula su varianza estimada, error de muestreo y opcionalmente su error de estimación y un intervalo de confianza si se especifica el coeficiente de confianza en el parámetro *alpha*.

Se permite indicar si el muestreo se realiza con o sin reemplazamiento para realizar estimaciones más precisas.

Realiza las funciones explicadas en el apartado 6.2 y 6.3.

- `cluster.samplesize`: Calcula el tamaño de muestra necesario para cometer un error de muestreo absoluto menor del especificado. Se permite la relajación de su estimación si se especifica un coeficiente de confianza en el parámetro *alpha*.

Realiza las funciones explicadas en el apartado 6.4.

Capítulo 7

Conclusiones y líneas futuras

7.1. Conclusiones

En los apartados anteriores, se ha mostrado un resumen del aparato teórico que rodea a los métodos de muestreo normalmente utilizados para poblaciones finitas.

Como ha podido comprobarse, esos métodos se basan en una cantidad ingente de fórmulas difíciles de memorizar para poder proceder a obtener resultados de manera rápida como hoy en día demanda nuestra sociedad. Es por ello que para poder conseguir este objetivo se hace preciso ocupar un espacio hoy en día poco cubierto con herramientas informáticas.

En consecuencia y dada la capacidad que ofrece R para implementar técnicas estadísticas, se ha utilizado esta herramienta para implementar las técnicas que de una forma teórica se han expuesto en los apartados anteriores y además se ha hecho de una manera que sea lo más eficiente posible, tanto en la reducción del número de funciones necesarias como desde el punto de vista de la optimización del número de parámetros requeridos en esas funciones para poder conseguir el mayor número posible de estimaciones derivados de la aplicación de las mismas.

En resumen, desde un punto de vista informático, se ha creado una librería compuesta por 17 funciones que permiten al usuario obtener muestras, realizar estimación de parámetros sobre una muestra poblacional, calcular el tamaño muestral necesario para cometer como máximo un error de muestreo dado, realizar tablas de análisis de la varianza sobre muestras sistemáticas y obtener medidas de interés. En resumen se ha pretendido facilitar al máximo la aplicación práctica de las técnicas muestrales mostradas en apartados anteriores de este TFG

Estas funciones han sido creadas y diseñadas con el objetivo de minimizar el número de nombres de función que el usuario debe recordar, procurando que para la realización de una misma tarea para un tipo de muestreo determinado se deba usar siempre la misma denominación de función variando los parámetros de control de la misma.

La nomenclatura de las funciones está definida para poder acceder a ella de manera sencilla, empezando siempre por el tipo de muestreo con el que vayamos a trabajar (*srs*, *strata*, *sys* y *cluster* para muestreo aleatorio simple, estratificado, sistemático y de conglomerados respectivamente) como si de un prefijo se tratara, separado por un punto del resto del nombre. Así por ejemplo tenemos que la función *srs.sample* es la indicada para tomar una muestra en el muestreo aleatorio simple.

Además de las funciones de uso público, también constan en la librería funciones auxiliares como aquellas destinadas a evitar la repetición de código u otras como la que muestra el mensaje de bienvenida cuando se utiliza la función *library(samplingR)* para poder empezar a hacer uso de sus funciones.

Dada la estructura que se dispone en la creación de los paquetes de R, se ha procedido a implementar líneas ayuda durante la creación de las funciones que constituyen el paquete creado en este TFG, consiguiendo de esta manera poder obtener de forma automática un pequeño manual de uso de esta librería, el cual se puede consultar en la web proporcionada por CRAN [5], documento PDF que igualmente se puede ver en el anexo que figura al final de este TFG.

Igualmente, a continuación del manual mencionado anteriormente, también se puede ver una *vignette* creada en este TFG con la cual se pretende que con la ayuda de un ejemplo se puedan asimilar de forma eficiente los resultados de ciertas funciones que se han desarrollado dentro del paquete creado y diseñado en ese TFG.

Se incluye en la bibliografía un enlace a un repositorio GitHub donde se encuentra el código creado durante este trabajo para su revisión y uso, así como este documento [6].

7.2. Líneas futuras

Como se ha podido comprobar y debido a la extensión máxima establecida y a los objetivos marcados en este TFG, se ha abarcado en el mismo hasta la técnica muestral de conglomerados monoetápica y considerando además que los conglomerados son de tamaño igual o similar, por lo que aquí se deja un campo abierto a posibles futuras líneas de actuación.

En consecuencia y como posibles líneas de trabajo futuras están la inclusión del muestreo por conglomerados para conglomerados de distintos tamaños, así como el muestreo bietápico por conglomerados.

Otra posible modificación de interés sería la implantación del método actualizado de selección de muestras en el muestreo sistemático para valores de k no enteros. [12]

Igualmente, cabría la posibilidad de desarrollar fórmulas de R que faciliten los cálculos para la obtención de diversos procedimientos indirectos de estimación, como pueden los denominados métodos de la razón, o de regresión.

Bibliografía

- [1] *The Comprehensive R Archive Network*. URL: <https://cran.r-project.org/>.
- [2] Yves Tillé y Alina Matei. *sampling: Survey Sampling*. R package version 2.9. 2021. URL: <https://CRAN.R-project.org/package=sampling>.
- [3] Hugo Andres Gutierrez Rojas. *TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population*. R package version 4.1.1. 2020. URL: <https://CRAN.R-project.org/package=TeachingSampling>.
- [4] César Pérez López. *Técnicas de muestreo estadístico: teoría, práctica y aplicaciones informáticas*. 519.52 P4. 2000.
- [5] Javier Estévez. *samplingR: Sampling and Estimation Methods*. R package version 1.0.1. 2023. URL: <https://CRAN.R-project.org/package=samplingR>.
- [6] *Proyecto en GitHub*. URL: <https://github.com/javeste/samplingR>.
- [7] Hadley Wickham et al. *roxygen2: In-Line Documentation for R*. R package version 7.2.3. 2022. URL: <https://CRAN.R-project.org/package=roxygen2>.
- [8] Hadley Wickham et al. *devtools: Tools to Make Developing R Packages Easier*. R package version 2.4.5. 2022. URL: <https://CRAN.R-project.org/package=devtools>.
- [9] *World Clock API*. URL: <http://worldclockapi.com/>.
- [10] *Person function - RDocumentation*. URL: <https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/person>.
- [11] *Directorio Central de Empresas*. URL: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550.
- [12] *Real Decreto 1398/1995, de 4 de agosto, por el que se regula el sorteo para la formación de las listas de candidatos a jurados*. URL: <https://www.boe.es/buscar/pdf/1995/BOE-A-1995-18778-consolidado.pdf>.
- [13] *samplingR: Sampling and Estimation Methods*. URL: <https://cran.r-project.org/web/packages/samplingR/index.html>.

Anexos

En esta sección se incluyen contenidos de interés disponibles en la página web proporcionada por CRAN sobre la librería creada en este trabajo [13], compuestos por el manual de usuario con información detallada sobre las funciones desarrolladas y una viñeta creada para ejemplificar el uso de las funciones del muestreo estratificado de cálculo del tamaño de muestra y afijación, utilizando para ello herramientas creadas en la librería generada en este TFG.

Package ‘samplingR’

June 20, 2023

Type Package

Title Sampling and Estimation Methods

Version 1.0.1

Date 2023-05-20

Maintainer Javier Estévez <javier.estase@gmail.com>

Description Functions to take samples of data, sample size estimation and getting useful estimators such as total, mean, proportion about its population using simple random, stratified, systematic and cluster sampling.

Imports dplyr, methods

License GPL-2

Encoding UTF-8

RoxygenNote 7.2.3

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

cluster.estimator	2
cluster.sample	3
cluster.samplesize	4
srs.domainestimator	5
srs.estimator	6
srs.sample	7
srs.samplesize	8
strata.allocation	9
strata.estimator	10
strata.sample	11
strata.samplesize	11
strata.samplesize.cost	13
syst.all.samples	14
syst.anova	14
syst.estimator	15
syst.intercorr	16
syst.intracorr	17
syst.sample	17

Index**19**

cluster.estimator	<i>Parameter estimation for cluster samples</i>
-------------------	---

Description

Estimates parameters with optional confidence interval for clustered data of similar cluster size.

Usage

```
cluster.estimator(
  N,
  data,
  estimator = c("total", "mean", "proportion", "class total"),
  replace = FALSE,
  alpha
)
```

Arguments

N	Number of clusters for the population
data	Cluster sample
estimator	Estimator to compute. Can be one of "total", "mean", "proportion", "class total". Default is "total".
replace	Whether the sample to be taken can have repeated instances or not.
alpha	Optional value to calculate estimation error and build 1-alpha

Details

This function admits both grouped and non-grouped by cluster data.

Non-grouped data must have interest variable data in the first column and cluster name each individual belongs to in the last column.

Grouped by cluster data must have interest variable data in the first column, cluster size in the second and the cluster name in the last column. Interest values of grouped data must reflect the total value of each cluster.

Value

A list containing different interest values:

- estimator
- variance
- sampling.error
- estimation.error
- confint

Examples

```
d<-cbind(rnorm(500, 50, 20), rep(c(1:50),10)) #Non-grouped data
sample<-cluster.sample(d, n=10) #Non-grouped sample
sampleg<-aggregate(sample[,1], by=list(Category=sample[,2]), FUN=sum)
sampleg<-cbind(sampleg[,2], rep(10,10), sampleg[,1]) #Same sample but with grouped data
sum(d[,1])
cluster.estimator(N=50, data=sample, estimator="total", alpha=0.05)
cluster.estimator(N=50, data=sampleg, estimator="total", alpha=0.05)
```

cluster.sample	<i>Cluster sample for non-grouped data.</i>
----------------	---

Description

Retrieves a sample of clusters for data which interest variable values are not grouped by cluster.

Usage

```
cluster.sample(data, n, replace = FALSE)
```

Arguments

data	Matrix or data.frame containing the population data in first column and the cluster it belongs to in the last column.
n	Number of clusters of the returning sample.
replace	Whether the sample to be taken can have repeated clusters or not.

Details

#' If your data is grouped by cluster use [srs.sample](#) function to retrieve your sample. Remember grouped by cluster data must have interest variable data in the first column, cluster size in the second and the cluster name in the last column. Interest values of grouped data must reflect the total value of each cluster.

Value

Data frame of a clustered sample

Examples

```
data<-cbind(rnorm(500, 50, 20), rep(c(1:50),10))
sample<-cluster.sample(data, 10);sample
```

cluster.samplesize	<i>Sample size estimation on cluster sampling</i>
--------------------	---

Description

Calculates the required sample size in order to achieve an absolute sampling error less or equal to the specified for an specific estimator and an optional confidence interval in cluster sampling.

Usage

```
cluster.samplesize(
  N,
  data,
  error,
  alpha,
  estimator = c("total", "mean", "proportion", "class total"),
  replace = FALSE
)
```

Arguments

N	Number of clusters in the population.
data	Dataset.
error	Sampling error.
alpha	Significance level to obtain confidence intervals.
estimator	The estimator to be estimated. Default is "total".
replace	Whether the samples to be taken can have repeated instances or not.

Details

This function admits both grouped and non-grouped by cluster data.

Non-grouped data must have interest variable data in the first column and cluster name each individual belongs to in the last column.

Grouped by cluster data must have interest variable data in the first column, cluster size in the second and the cluster name in the last column. Interest values of grouped data must reflect the total value of each cluster.

Value

Number of clusters to be taken.

Examples

```
d<-cbind(rnorm(500, 50, 20), rep(c(1:50),10)) #Non-grouped data
sample<-cluster.sample(d, n=10) #Non-grouped sample
sampleg<-aggregate(sample[,1], by=list(Category=sample[,2]), FUN=sum)
sampleg<-cbind(sampleg[,2], rep(10,10), sampleg[,1]) #Sample sample with grouped data

#Cluster size to be taken for estimation
cluster.samplesize(N=50, data=sample, error=500, estimator="total", replace=TRUE)
```

```

newsample<-cluster.sample(d, n=26) #New sample for estimation
sum(d[,1])
cluster.estimator(N=50, data=newsample, estimator="total", alpha=0.05, replace=TRUE)
cluster.estimator(N=50, data=samplg, estimator="total", alpha=0.05)

```

srs.domainestimator *Simple Random Sample parameter estimation of domains.*

Description

Function to make estimations of diferent parameters on a given domain based on a Simple Random Sample.

Usage

```

srs.domainestimator(
  Nh,
  data,
  estimator = c("total", "mean", "proportion", "class total"),
  domain,
  replace = FALSE,
  alpha
)

```

Arguments

Nh	Number of instances of the data set domain.
data	Sample of the data. It must constain a column with the data to estimate and a second column with the domain of each instance.
estimator	One of "total", "mean". Default is "total".
domain	Domain of the sample from which parameter estimation will be done.
replace	Whether the sample to be taken can have repeated instances or not.
alpha	Optional value to calculate estimation error and build 1-alpha confidence interval.

Details

Data columns must be arranged with interest values on the first column and domain values on the last column.

Domain parameter can be either numeric or character and must be equal to one of the values of the domain column of data.

Value

A list containing different interest values:

- estimator
- variance
- sampling.error
- estimation.error
- confint

References

Pérez, C. (1999) Técnicas de muestreo estadístico. Teoría, práctica y aplicaciones informáticas. 193-195

Examples

```
data<-cbind(rnorm(500, 50, 20), rep(c(1:2),250))
sample<-data[srs.sample(500, 100),]
sum(data[which(data[,1]==1),1])
srs.domainestimator(Nh = 250, data = sample, estimator="total", domain=1)
```

srs.estimator

Simple Random Sampling parameter estimation.

Description

Function to make estimations of different parameters based on a Simple Random Sample.

Usage

```
srs.estimator(
  N,
  data,
  estimator = c("total", "mean", "proportion", "class total"),
  replace = FALSE,
  alpha
)
```

Arguments

N	Number of instances of the data set.
data	Sample of the data. It must only contain a single column of the data to estimate.
estimator	Estimator to compute. Can be one of "total", "mean", "proportion", "class total". Default is "total".
replace	Whether the sample has been taken with replacement or not.
alpha	Optional value to calculate estimation error and build 1-alpha confidence interval.

Value

A list containing different interest values:

- estimator
- variance
- sampling.error
- estimation.error
- confint

Examples

```
data<-rnorm(200, 100, 20)
sample<-data[srs.sample(200, 50)]
tau<-sum(data);tau
srs.estimator(200, sample, "total", alpha=0.05)
```

```
mu<-mean(data);mu
srs.estimator(200, sample, "mean", alpha=0.05)
```

srs.sample	<i>Simple Random Sample</i>
------------	-----------------------------

Description

With this function you receive a simple random sample consisting on a list of the instances index

Usage

```
srs.sample(N, n, replace = FALSE, data)
```

Arguments

N	Number of instances of the data set.
n	Number of instances of the returning sample.
replace	Whether the sample to be taken can have repeated instances or not.
data	Optional matrix or data.frame containing the population data. If specified an object of same class as data will be returned with sample instances.

Value

List of size n with numbers from 1 to N indicating the index of the data set's instances to be taken.

Examples

```
srs.sample(10,3)
```

```
data<-matrix(data=c(1:24), nrow=8)
N<-dim(data)[1]
sample<-srs.sample(N, 3, data = data)
sample
```

srs.samplesize

*Simple Random Sample size.***Description**

Calculates the required sample size in order to achieve a relative or absolute sampling error less or equal to the specified for an specific estimator and an optional confidence interval in simple random sampling.

Usage

```
srs.samplesize(
  N,
  var,
  error,
  alpha,
  estimator = c("total", "mean", "proportion", "class total"),
  p,
  mean,
  replace = FALSE,
  relative = FALSE
)
```

Arguments

N	Number of instances of the data set.
var	Estimated quasivariance.
error	Sampling error
alpha	Significance level to obtain confidence intervals.
estimator	One of "total", "proportion", "mean", "class total". Default is "total"
p	Estimated proportion. If estimator is not "proportion" or "class total" it will be ignored.
mean	Estimated mean. If relative=FALSE it will be ignored.
replace	Whether the sample to be taken can have repeated instances or not.
relative	Whether the specified error is relative or not.

Details

If the sample size result is not a whole number the number returned is the next whole number so `srs.samplesize>=n` is satisfied.

To estimate sample size of estimators "total" and "mean" estimated quasivariance must be provided. If the error is relative then estimated mean must also be provided.

To estimate sample size of estimator "proportion" and "class total" estimated proportion must be provided. If p is not specified sample size will be estimated based on worst-case scenario of $p=0.5$. N must be always be provided for calculations.

Value

Number of instances of the sample to be taken.

Examples

```
data<-rnorm(200, 100, 20)
n<-srs.samplesize(200, var(data), estimator="total", error=400, alpha=0.05);n
sample<-data[srs.sample(200, n)]
srs.estimator(200, sample, "total", alpha=0.05)$sampling.error
```

strata.allocation	<i>Strata allocation given a sample size</i>
-------------------	--

Description

Function to allocate the number of samples to be taken for each strata given the total sample size and the strata.allocation method. The number of allocations returned will be equal to the length of the parameters.

Usage

```
strata.allocation(
  Nh,
  n,
  var,
  alloc = c("unif", "prop", "min", "optim"),
  C,
  cini,
  ch
)
```

Arguments

Nh	Vector of population strata sizes.
n	Sample size
var	Vector of strata variances.
alloc	The allocation method to be used. Default is "unif".
C	Total study cost.
cini	Overhead study cost.
ch	Vector of costs to take an individual from a strata for the sample.

Details

alloc="optim" is the only that requires cost function data. Total study and overhead study costs are optional. If given allocation will be done so total study cost is not surpassed.

Value

Vector of strata sample sizes.

Examples

```
strata.allocation(Nh=rep(125,4), n=100, alloc="unif") #25, 25, 25, 25
strata.allocation(Nh=c(100, 50, 25), n=100, alloc="prop")
```

strata.estimator	<i>Parameter estimation of stratified data</i>
------------------	--

Description

Function to make estimations of different parameters based on a stratified sample.

Usage

```
strata.estimator(  
  N,  
  Nh,  
  data,  
  estimator = c("total", "mean", "proportion", "class total"),  
  replace = FALSE,  
  alpha  
)
```

Arguments

N	Population size.
Nh	Size of each population strata.
data	Stratified sample.
estimator	Estimator to compute. Can be one of "total", "mean", "proportion", "class total". Default is "total".
replace	Whether the sample to be taken can have repeated instances or not.
alpha	Optional value to calculate estimation error and build 1-alpha

Details

Nh length must be equal to number of strata in data.
data is meant to be a returned object of [strata.sample](#) function.

Value

A list containing different interest values:

- estimator
- variance
- sampling.error
- estimation.error
- confint

strata.sample	<i>Stratified sample</i>
---------------	--------------------------

Description

With this function you receive a sample of each strata within your data with specified size for each strata.

Usage

```
strata.sample(data, n, replace = FALSE)
```

Arguments

data	Population data consisting of a number of columns of data and a last column specifying the strata each instance belongs to.
n	Numeric array of sample sizes for each strata to be taken.
replace	Whether the sample to be taken can have repeated instances or not.

Details

n length must be equal to number of strata in data.

On return list each strata sample can be accessed calling object\$strataname where strataname are values of the last column of the original data.

Value

A list containing one strata sample per index.

Examples

```
data<-cbind(rnorm(500, 50, 20), rep(c("clase 1", "clase 2","clase 3","clase4"),125))
strata.sample(data=data, n=c(10,20,30,40))
```

strata.samplesize	<i>Sample size estimation on stratified sampling</i>
-------------------	--

Description

Calculates the required sample size in order to achieve an absolute or relative sampling error less or equal to the specified for an specific estimator and an optional confidence interval in stratified sampling.

Usage

```
strata.samplesize(
  Nh,
  var,
  error,
  alpha,
  estimator = c("total", "mean", "proportion", "class total"),
  alloc = c("prop", "min", "optim"),
  ch,
  p,
  mean,
  replace = FALSE,
  relative = FALSE
)
```

Arguments

Nh	Vector of population strata sizes.
var	Vector of estimated strata variances.
error	Sampling error.
alpha	Significance level to obtain confidence intervals.
estimator	The estimator to be estimated. Default is "total".
alloc	The allocation to be used when taking samples. Default is "prop".
ch	Vector of cost per strata to select an individual for the sample.
p	Estimated population proportion. If estimator is not "proportion" or "class total" it will be ignored.
mean	Estimated population mean. If relative=FALSE it will be ignored.
replace	Whether the samples to be taken can have repeated instances or not.
relative	Whether the specified error is relative or not.

Details

With "proportion" and "class total" estimators variance vector must contain `var` return values equal to $\frac{N_h}{(N_h-1)}p * (1 - p)$ values.

Value

Number of instances of the sample to be taken.

Examples

```
strata.samplesize(c(120,100,110,50), c(458, 313,407,364), error=5, alpha=0.05, "mean", "prop")
```

`strata.samplesize.cost`*Strata sample size by costs function*

Description

This function returns the total sample size given a costs function consisting on the fixed total study cost, overhead study cost and a vector of costs by strata. can be given so the allocation is calculated to not exceed the total study cost.

Usage

```
strata.samplesize.cost(  
  Nh,  
  var,  
  C,  
  cini,  
  ch,  
  alloc = c("unif", "prop", "optim")  
)
```

Arguments

Nh	Vector of population strata sizes.
var	Vector of strata variance values.
C	Total study cost.
cini	Overhead study cost.
ch	Vector of costs to take an individual from a strata for the sample.
alloc	The allocation method to be used. Default is "unif".

Details

Strata variance values are only necessary for optim allocation.

Value

Sample size.

Examples

```
strata.samplesize.cost(Nh=c(100,500,200), C=1000, cini=70, ch=c(9,5,12), alloc="prop")
```

syst.all.samples	<i>Systematic samples</i>
------------------	---------------------------

Description

Returns all possible systematic samples of size n

Usage

```
syst.all.samples(data, n)
```

Arguments

data	Population data
n	Sample size

Value

List with a sample per entrance

Examples

```
data<-c(1,3,5,2,4,6,2,7,3)
syst.all.samples(data, 3)
```

syst.anova	<i>Analysis of variance of population data</i>
------------	--

Description

Analysis of variance of population data

Usage

```
syst.anova(data, n)
```

Arguments

data	Population data
n	Sample size

Value

Summary

Examples

```
data<-c(1,3,5,2,4,6,2,7,3)
syst.anova(data,3)
```

syst.estimator	<i>Parameter estimation on a systematic sample</i>
----------------	--

Description

Parameter estimation on a systematic sample

Usage

```
syst.estimator(
  N,
  sample,
  estimator = c("total", "mean", "proportion", "class total"),
  method = c("srs", "strata", "syst"),
  alpha,
  data,
  t
)
```

Arguments

N	Population size
sample	Vector containing the systematic sample
estimator	Estimator to compute. Can be one of "total", "mean", "proportion", "class total". Default is "total".
method	Method of variance estimation. Can be one of "srs", "strata", "syst".
alpha	Optional value to calculate estimation error and build 1-alpha confidence interval.
data	Population data.
t	Number of systematic samples to take with interpenetrating samples method.

Details

Variance estimation has no direct formula in systematic sampling, thus estimation method must be done. Refer to [syst.intracorr](#) and [syst.intercorr](#) functions details for more information.

"syst" method uses interpenetrating samples method in which t systematic samples of size= $\frac{n}{t}$ are taken to estimate. $\frac{n}{t}$ must be even.

By choosing the start at random for all the samples they can be considered random taken. With this method population data and t must be given.

Value

A list containing different interest values:

- estimator
- variance
- sampling.error
- estimation.error
- confint

Examples

```
data<-c(1,3,5,2,4,6,2,7,3)
sample<-syst.sample(9, 3, data)
syst.estimator(N=9, sample, "mean", "srs", 0.05)
```

syst.intercorr	<i>Stratified correlation coefficient</i>
----------------	---

Description

Stratified correlation coefficient

Usage

```
syst.intercorr(N, n, data)
```

Arguments

N	Population size
n	Sample size
data	Population data

Details

This value serves as a comparison between systematic and stratified sampling precision. At value=1 the systematic precision is minimum. At value=0 both sampling methods precision are equal. At value= $\frac{-1}{n-1}$ systematic precision is maximum. Summarising at values between 1 and 0 stratified sampling estimation has more precision than systematic, so method="strata" should be set at [syst.estimator](#). The other way method="syst" of interpenetrating samples method is better.

Value

Correlation coefficient

Examples

```
data<-c(1,3,5,2,4,6,2,7,3)
syst.intercorr(9,3,data) #0.09022556
```

syst.intracorr	<i>Intraclass correlation coefficient</i>
----------------	---

Description

Intraclass correlation coefficient

Usage

```
syst.intracorr(N, n, data)
```

Arguments

N	Population size
n	Sample size
data	Population data

Details

This value serves as a comparison between systematic and simple random sampling precision. At value=1 the systematic precision is minimum. At value=0 both sampling methods precision are equal. At value= $\frac{-1}{n-1}$ systematic precision is maximum. Summarising at values between 1 and 0 simple random sampling estimation has more precision than systematic, so method="srs" should be set at [syst.estimator](#). The other way method="syst" of interpenetrating samples method is better.

Value

Intraclass correlation

Examples

```
data<-c(1,3,5,2,4,6,2,7,3)
syst.intracorr(9, 3, data) #0.34375 example 1
```

syst.sample	<i>Systematic sampling sample</i>
-------------	-----------------------------------

Description

Retrieves a $\frac{N}{n}$ systematic sample

Usage

```
syst.sample(N, n, data)
```

Arguments

N	Population size.
n	Sample size
data	Optional data of the population.

Details

If $\frac{N}{n}$ is not an even number a 1 in $\text{floor}(\frac{N}{n})$ sample will be taken.

Value

Vector of size n with numbers from 1 to N indicating the index samples to be taken. If data is provided then the instances will be returned.

Examples

```
data<-runif(40)
syst.sample(40,8, data)
```

Index

cluster.estimator, 2
cluster.sample, 3
cluster.samplesize, 4

srs.domainestimator, 5
srs.estimator, 6
srs.sample, 3, 7
srs.samplesize, 8
strata.allocation, 9
strata.estimator, 10
strata.sample, 10, 11
strata.samplesize, 11
strata.samplesize.cost, 13
syst.all.samples, 14
syst.anova, 14
syst.estimator, 15, 16, 17
syst.intercorr, 15, 16
syst.intracorr, 15, 17
syst.sample, 17

var, 12

Ejemplo de viñeta. Determinar el calculo de tamaño de muestra utilizando distintas afijaciones, considerando costes y sin considerar costes

```
library(samplingR)
#>
#>
#>
#> / _ _ / _ _ \ _ _ \ _ _ \ _ _ \ _ _ \ _ _ \ _ _ \
#> \ _ _ \ _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ /
#> / _ _ \ _ _ \ _ _ \ _ _ \ _ _ \ _ _ \ _ _ \ _ _ \
#> \ _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ /
#>
#>
#>
#> Type 'citation("samplingR")' for citing this R package in publications.
```

Preparamos los datos

La población está compuesta por 2 estratos generados sintéticamente a partir de dos distribuciones normales.

Primer estrato: pensionistas en CyL

```
N1<-585479
pen<-rnorm(N1, 750, 100)
```

Segundo estrato: asalariados en CyL

```
N2<-932992
ass<-rnorm(N2, 1500, 500)
```

```
datos<-cbind(c(pen, ass), c(rep("pensionista", N1), rep("asalariado", N2)))
N<-N1+N2
```

Ejercicio 1. Consideremos una muestra global de $n=800$ individuos y afijacion uniforme

a) Reparto de la muestra

```
Nh<-c(N1, N2)
n<-800
strata.allocation(Nh=Nh, n=n, alloc="unif")
#> [1] 400 400
```

b) Suponiendo una función de coste

```
C<-12000
Cini<-5000
ch<-c(45, 20)
size<-strata.samplesize.cost(Nh=Nh, C=C, cini=Cini, ch=ch, alloc="unif")
paste("Tamaño de muestra", size)
#> [1] "Tamaño de muestra 215.384615384615"
nh.unif<-strata.allocation(Nh=Nh, n=size, alloc="unif")

paste(c("Estrato 1:", "Estrato 2"), nh.unif)
#> [1] "Estrato 1: 107.692307692308" "Estrato 2 107.692307692308"
paste("Coste:", Cini+sum(ch*nh.unif))
#> [1] "Coste: 12000"
```

Como los tamaños de muestra deben ser enteros redondeamos al entero inferior.

```
nh.unif<-floor(nh.unif)
paste(c("Estrato 1:", "Estrato 2"), nh.unif)
#> [1] "Estrato 1: 107" "Estrato 2 107"
paste("Coste:", Cini+sum(ch*nh.unif))
#> [1] "Coste: 11955"
```

Ejercicio 2. Afijación proporcional

a) Reparto de la muestra

```
strata.allocation(Nh=Nh, n=n, alloc="prop")
#> [1] 308.4571 491.5429
```

b) Con función de coste igual a la anterior

```
size<-strata.samplesize.cost(Nh=Nh, C=C, cini=Cini, ch=ch, alloc="prop")
paste("Tamaño de muestra", size)
#> [1] "Tamaño de muestra 236.173037187271"
nh.unif<-floor(strata.allocation(Nh=Nh, n=size, alloc="prop"))
paste(c("Estrato 1:", "Estrato 2"), nh.unif)
#> [1] "Estrato 1: 91" "Estrato 2 145"
paste("Coste:", Cini+sum(ch*nh.unif))
#> [1] "Coste: 11995"
```

Ejercicio 3. Afijación de mínima varianza

La afijación de Neyman depende de las cuasivarianzas de los estratos, por lo que se deben estimar.

a) Reparto de la muestra

Opción 1: usar las varianzas reales como estimadores (solución teórica)

```
var<-c(var(pen), var(ass))
strata.allocation(Nh=Nh, n=n, var=var, alloc="min")
#> [1] 89.15727 710.84273
```

Opción 2: tomar una muestra previa para estimar las cuasivarianzas de los estratos.

```
sample<-strata.sample(data=datos, n=c(20, 20))
var<-c(var(sample[which(sample[,2]=="asalariado"),1]), var(sample[which(sample[,2]=="pensionista"),1]))
strata.allocation(Nh=Nh, n=n, var=var, alloc="min")
#> [1] 612.9519 187.0481
```

Opción 3: estimación más conservadora. Suponemos cuasivarianza máxima en todos los estratos = $\frac{N_h}{N_h-1}p(1-p)$ con $p=0.5$

```
var<-c(Nh/(Nh-1)*0.5*(1-0.5))
strata.allocation(Nh=Nh, n=n, var=var, alloc="min")
#> [1] 308.4572 491.5428
```

Si no se fija la varianza en la función se mostrará un warning y la varianza declarada en la función será la máxima para cada estrato.

```
strata.allocation(Nh=Nh, n=n, alloc="min")
#> Warning in strata.allocation(Nh = Nh, n = n, alloc = "min"):
#> Necessary var argument missing, will be set to worst case scenario value for each strata.
#> [1] 308.4572 491.5428
```

La estimación conservadora coincide con el cálculo de afijación proporcional.

b) Con función de coste igual a la anterior y solución teórica

En la afijación de mínima varianza optimizando con una función de costes es equivalente a utilizar la afijación óptima.

```
size<-strata.samplesize.cost(Nh=Nh, var=var, C=C, cini=Cini, ch=ch, alloc="optim")
paste("Tamaño de muestra", size)
#> [1] "Tamaño de muestra 319.21044160007"
nh.optim<-floor(strata.allocation(Nh=Nh, n=size, var=var, alloc="optim", C=C, cini=Cini, ch=ch))
paste(c("Estrato 1:", "Estrato 2"), nh.optim)
#> [1] "Estrato 1: 24" "Estrato 2 294"
paste("Coste:", Cini+sum(ch*nh.optim))
#> [1] "Coste: 11960"
```