Facultad de Ciencias Sociales y Jurídicas de Elche

ESTADÍSTICA EMPRESARIAL

Trabajo Fin de Grado

2017-2018

Manual del buen uso del muestreo estadístico.

Alumna: Elena Ruano Martínez.

Tutora: Mercedes Landete Ruiz.

ÍNDICE

1. INTRODUCCIÓN

El muestreo es una herramienta de la investigación científica que se utiliza para extraer conclusiones de una determinada población de interés a partir de la selección de una parte de los elementos de la misma. El muestreo estadístico es una disciplina que permite obtener conclusiones fiables para la población que se estudia.

Los primeros trabajos utilizando muestreo aparecen publicados a finales del siglo XIX **(1)**, hasta entonces los científicos descartaban la idea del muestreo, confiando más en el censo, procedimiento mediante el cual se examinan todos los elementos de la población. La necesidad de estudiar grandes colectivos sociales justifica la incorporación del muestreo en poblaciones finitas como técnica eficaz para reducir el coste y el tiempo empleado en el estudio, de esta manera, se podrían realizar un mayor número de estudios con los mismos recursos. El muestreo estadístico es una ciencia multidisciplinar que aparece en muchos campos de aplicación como por ejemplo:

- Biología: Recogida de datos del peso de cierto número de adolescentes de un determinado país para estudiar la tasa de desnutrición.
- Economía: Recolección de datos sobre la renta media de algunas de las familias de cierta población para conocer el desarrollo de la comunidad a la que pertenece.
- Agricultura: Muestra sobre el número plantaciones de árboles frutales en cierta región.
- Comercio: Muestreo acerca de las tiendas en un país dedicadas a la venta de calzado.
- Transporte de mercancías: Muestreo de graneles en buques.

Dado que existen tantos tipos de estudios, es lógico pensar que existen diferentes diseños muestrales con diversas características y que por tanto, según la población, se deberá elegir aquel diseño que sea más apropiado para conseguir, a partir de la muestra, las mejores estimaciones poblacionales. Además se podrán elegir diferentes estimadores, es decir, diferentes estadísticos para inferir el valor de un parámetro muestral. Como diseños muestrales básicos podemos destacar: muestreo aleatorio simple, muestreo aleatorio con reemplazamiento, muestreo sistemático, muestreo estratificado con MAS en cada estrato…

Las etapas principales para realizar un muestreo son:

i. Planteamiento de objetivos.
ii. Especificación de los elementos del muestreo, es decir, población, variables, parámetros y nivel de precisión.

iii. Plan de muestreo que conlleva escoger el tipo de muestreo, estimadores y errores de muestreo (inferencia estadística) y decidir tamaños muestrales según el presupuesto o la eficiencia del estudio.
iv. Organización del trabajo de campo basado en el entrenamiento del personal, calidad de los datos y modelos para el tratamiento de la no respuesta.
v. Resumen y análisis de los datos.

En muestreo en poblaciones finitas el espacio muestral es el conjunto de todas las muestras, éstas tienen carácter probabilístico pues el diseño muestral es una distribución de probabilidad sobre el conjunto de las muestras, donde cualquier subconjunto es un suceso que tiene asignada una probabilidad de ocurrencia.

1.1 Objetivos.

Uno de los objetivos de este TFG es destacar la importancia que tiene que la persona encargada de analizar cierta base de datos sea consciente del tipo de muestreo que se ha llevado a cabo para conseguirla. Esto es de vital importancia para que las conclusiones que se puedan extraer del análisis sean acertadas.

En segundo lugar se recalcará la importancia de estudiar teoría acerca de las técnicas de muestreo, conocer los tipos de muestreo y sus características, ventajas y desventajas para así, según la población objeto de estudio, seleccionar el tipo de muestreo más adecuado a ésta y, tras el análisis, conseguir mayor precisión en los resultados. Con un ejemplo demostraremos que, a pesar de que el análisis de los datos esté bien realizado y de que todas las conclusiones sean ciertas, la precisión y exactitud cambia mucho según si el tipo de muestreo es más o menos adecuado para la población de interés, ya que siempre existe un tipo de muestro más idóneo que otro para cada población.

Es importante distinguir entre los errores muestrales y los no muestrales, pues cada tipo de error pertenece a una sección de este trabajo.

Los errores muestrales son aquellos que se cometen por el hecho de utilizar muestreo en lugar de estudiar la población al completo. Un error muestral básico es aplicar un tipo de muestreo cuyas características no son apropiadas para la población de estudio, lo cual provoca menor precisión en los resultados y, por lo tanto, que el intervalo en el que se encuentra el parámetro de interés sea más grande. Además debemos saber que todas las muestras tienen este tipo de error, en mayor o menor medida.

En cambio, los errores no muestrales son mucho más variados y tienen un carácter eminentemente práctico. Un error no muestral son todos los que concierne errores en la encuesta (error de no respuesta, error de anotación de respuesta, errores de medición por

subjetividad…). Otros errores no muestrales son aquellos errores cometidos al codificar los datos o errores de desconocimiento del tipo de muestreo utilizado para obtener la base de datos que se está analizando.

En la primera sección abordamos errores no muestrales y en la segunda sección los errores muestrales.

1.2 Contexto.

La asignatura base sobre la cual he realizado este trabajo ha sido "Auditoría: Técnicas de muestreo" cursada en el último año de la titulación Estadística empresarial. Gracias al estudio de esta asignatura entendí lo importante que era estudiar los diferentes tipos diseños muestrales para así poder seleccionar el más adecuado para la población de interés. Estudié la posibilidad de que los estimadores estuvieran sesgados, estudié, de forma práctica y teórica, la forma el elegir el tamaño muestral teniendo en cuenta la precisión que podía conseguir y el coste que este tamaño supondría, etc.

Haciendo este trabajo de fin de grado he aprendido a utilizar el paquete *TeachingSampling* del programa RStudio, cuya funcionalidad se basa en la selección de muestras y estimación de parámetros. Este paquete es un descubrimiento muy útil para realizar dichas funciones de forma segura y sin errores.

Por otra parte, tras realizar el estudio, también he aprendido que no solo es importante seleccionar bien la técnica de muestreo para una población, si no que, además, es muy importante conocer qué técnica se ha utilizado para recoger la base de datos que estamos analizando ya que los resultados cambian y por tanto las conclusiones a las que llega el estadístico serán diferentes. Y que esto pasa incluso habiendo hecho bien el análisis, lo cual recalca aún más dicha importancia y hace que el estudio del muestreo tenga más sentido, se entienda su utilidad y por tanto, sea bonito estudiarlo.

1.3 Base de datos.

La base de datos utilizada (COCHES) tiene 384 observaciones y 8 variables en total. Es una base de datos original del programa estadístico SPSS.

Para realizar el estudio cogemos una variable cualitativa, que será "CILINDROS", corresponde a las cilindradas de cada coche. Esta variable la convertimos en factor para poder así distinguir sus categorías, lo cual nos será muy útil para obtener resultados menos generales, es decir, estimaciones específicas para cada categoría.

Las variables cuantitativas seleccionadas serán "CONSUMO", en cuanto a litros de gasolina por cada 100 kilómetros, y "ACELERACIÓN", donde se plasman el número de segundos que tarda cada coche en pasar de 0 a 100 kilómetros por hora.

Los datos pertenecen a coches relativamente antiguos, aproximadamente son coches de hace 25 años, por este motivo el consumo de litros de gasolina es mayor que el actual, y los segundos de aceleración son, también, superiores a los actuales.

1.4 Software.

Para seleccionar muestras y estimar parámetros podemos utilizar el programa RStudio, un programa de software libre para el análisis estadístico de datos que permite al usuario crear funciones o utilizar las ya creadas para satisfacer sus necesidades acerca del estudio que esté realizando. En concreto, utilizaremos el paquete "TeachingSampling" **(1) y (5).** Un paquete es un conjunto de funciones que ya están creadas y pueden ser utilizadas por cualquier usuario de R. Éste paquete lleva incorporado una serie de funciones útiles para realizar diferentes tipos de muestreo en poblaciones finitas.

En primer lugar debemos instalar el paquete con el siguiente comando install.packages("TeachingSampling") y cargarlo library("TeachingSampling").

Dependiendo del tipo de muestreo que queramos asumir, se usarán unas funciones u otras.

2. ERRORES NO MUESTRALES.

En este apartado veremos los errores no muestrales que se siguen del uso de distintos diseños muestrales en la etapa de selección y en la etapa de estimación. El objetivo, como ya comentamos anteriormente, es destacar la importancia de que la persona que analiza una base de datos conozca qué tipo de muestreo se ha utilizado, ya que, según cual crea que ha sido el tipo de muestreo podrá concluir de forma más o menos acertada.

Por ejemplo, si estamos estudiando la edad a la que los jóvenes empiezan a fumar para así saber cuándo deberían impartirse charlas en los institutos. Cambia mucho si las conclusiones del análisis, realizado por un estadístico que cree que se ha recogido la muestra con un tipo de muestreo A, dicen que la edad es entre los 15 y 17 años o si, por el contrario, el estadístico cree que los datos han sido recogidos por un tipo de muestreo B y los resultados de éste análisis dicen que son entre los 14 y 16 años. Sería un grave

error ya que es de vital importancia impartir la charla antes de que la mayoría de jóvenes propensos a fumar haya empezado a hacerlo.

Con los siguientes "Casos" pretendo ilustrar las discrepancias que provoca en los resultados este desconocimiento.

En cada "Caso" se realiza una hipótesis acerca del tipo de muestreo que se ha utilizado para recoger la muestra, y por tanto, se aplican las funciones correspondientes a ese determinado tipo de muestreo para analizar "Coches".

El análisis consiste en obtener la estimación de la media para las 2 variables cualitativas de interés, así como la proporción de cada nivel de la variable cualitativa "Cilindros". Además calculamos tantos intervalos de confianza (IC) como variables y niveles tenemos. Con el IC **(3)** pretendemos determinar un rango de valores entre los que, con una determinada probabilidad, estará el valor del parámetro poblacional que buscamos. El nivel de confianza de los IC será del 95%, es decir, la probabilidad de que hayamos acertado al decir que el IC contiene el parámetro buscado es del 95%, siendo el 5% restante el nivel de significación o probabilidad de errar. Además, comentaremos también el efecto (DEFF) de cada tipo de muestreo con respecto al muestreo aleatorio simple (MAS). El efecto compara la eficacia de un determinado tipo de muestreo con la eficacia que se hubiera conseguido si en lugar de utilizar ese tipo, hubiéramos utilizado MAS. Cuando el valor resultante es un número superior a la unidad, podremos decir que el MAS es más eficaz, y si por el contrario, si el valor es menor a la unidad afirmaremos la mayor eficacia de ese tipo de muestreo que estamos comparando, con respecto al MAS.

Al final de este apartado, se compararán los resultados obtenidos y verificaremos la importancia de no hacer hipótesis acerca del tipo de muestreo que se ha utilizado para conseguir la muestra, si no de realmente conocerlo.

2.1 Casos.

Tomamos como suposición básica que, para todos los "Casos", son 9000 los coches que pertenecen a la población completa (N) y que de éstos, solo 384 (n) han sido seleccionados por el muestreo correspondiente a cada caso, justo los 384 elementos que encontramos en la base de datos "Coches", asumiendo ésta como la muestra a analizar. En la figura 1 vemos el código correspondiente.

```
N <- 9000
n <- dim(Coches)[1]
n
```

Figura 1: Código N y n común en errores no muestrales.

Tras preparar la base de datos para el análisis, aplicamos las funciones correspondientes a cada tipo de muestreo, que conseguimos gracias al paquete "*TeachingSampling*" del programa R-Studio.

La primera función de este paquete, que es vital para que el resto funcionen, es "*Domains*", ésta sirve para separar los niveles de la variable cualitativa de interés Cilindros;

CILINDROS <- Domains (CILINDROS)

Las variables de interés son Cilindros, Consumo y Aceleración;

var_interes <- data.frame (CONSUMO, ACELERACION, CILINDROS)

El resto de funciones son específicas de cada caso y las veremos a continuación.

- CASO 1: MUESTREO ALEATORIO SIMPLE (MAS).

El MAS consiste en seleccionar una muestra no ordenada de tamaño n, donde se efectúan n extracciones sucesivas sin reponer la unidad que sale cada vez.

Gracias a la siguiente línea de código que escribiremos en R-Studio obtenemos la tabla 1, que comentaremos a continuación, donde podemos observar la estimación del total de la población de cada variable de interés y por cada nivel de la variable "Cilindros", su error estándar estimado, su coeficiente de variación estimado y el efecto del diseño (DEFF).

Tabla_MAS <- E.SI (N, n, var_interes)

	N	CONSUMO	ACELERACIÓN	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	9000	101179.69	139800	4664.06	1945.31	2390.62
Error estándar		1759.62	1237.73	224.83	185.21	198.62
Coeficiente de variación		1.74	0.89	4.82	9.52	9.31
DEFF			1			

Tabla 1: Estimaciones según MAS.

En la primera columna y fila de datos verificamos que los datos se han calculado asumiendo 9000 como población total.

Según los cálculos, asumiendo que la base de datos se ha recogido a través de un MAS, 101.179 litros de gasolina por cada 100 km es la estimación del total de consumo de

todos los coches de la población, siendo su error estándar de 1759,62 litros/100km y con un coeficiente de variación de 1.74.

En la tercera columna de datos pertenece a la estimación de la variable "Aceleración". Son 139.800 los segundos que tardan el total de coches en pasar de 0 a 100 km/h, con desviación estándar de 1237,73 segundos y coeficiente de variación 0.89.

En la cuarta, quinta y sexta columna de datos tenemos las estimaciones de cada uno de los niveles de la variable cualitativa "Cilindros". Podemos concluir, mediante la estimación de cada del total partida 9000, que el 51,8% de los coches tienen 4 cilindros, el 21,6% tienen 6 cilindros y el 26,6% restante tienen 8 cilindros.

La utilidad de DEFF en este caso es nula, ya que, por defecto, compara el MAS con el tipo de muestreo utilizado en la tabla, que también es MAS, y por tanto, el valor es 1, ni mejor ni peor, si no, la misma efectividad.

Como hemos comentado con anterioridad, calcular los intervalos de confianza sería una buena forma de comparar resultados entre los obtenidos dependiendo de si se ha asumido un tipo de muestreo u otro. Los IC han sido calculados mediante los datos obtenidos en la tabla anterior 1, pero no en términos de *total* si no en términos de *medias*. El código utilizado para obtener los IC ha sido el que vemos en la figura 1.1 y las valores correspondientes se muestran en la tabla 2, una tabla resumen con los datos más importantes. Además vemos valores correspondientes a medias y proporciones, estos valores se han conseguido gracias a la división de cada estimación entre el total.

```
#para sacar el IC: Estimación_total / N +- 1.96 Error/N
N_MAS<-Tabla_MAS[1,1]

#PARA CONSUMO
estimacion_con<-Tabla_MAS[1,2]
error_con<-Tabla_MAS[2,2]
IC_sup_con<-((estimacion_con/N_MAS) + (1.96*error_con/N_MAS))
IC_inf_con<-((estimacion_con/N_MAS) - (1.96*error_con/N_MAS))
Efecto_con<-Tabla_MAS[4,2]
```

Figura 1.1: Código para calcular IC con datos de las funciones del MAS.

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,24	10,86	11,63	
ACELERACIÓN	15,53	15,26	15,8	
4 CILINDROS	0,52	0,47	0,57	1
6 CILINDROS	0,22	0,18	0,26	
8 CILINDROS	0,26	0,22	0,31	

Tabla 2: Tabla resumen con los resultados del MAS.

- Con un 95% de confianza, el consumo de litros de gasolina por cada 100 kilómetros será entre 10,86 l/100km y 11,63 l/100km, con media de 11,24 l/100km.
- Con un 95% de confianza, los coches se pondrán de 0 a 100 kilómetros por hora entre 15,26 segundos y 15,8 segundos. De media, tardan 15,53 segundos.
- Con un 95% de confianza, la proporción de coches de 4 cilindros suele estar entre el 0,47 y 0,57, es decir, entre el 47% y el 57%. El valor medio es de un 52% de coches que tienen 4 cilindros.
- La proporción de coches de 6 cilindradas suele estar entre el 0,18 y 0,26, al 5% de significación, es decir, la proporción real estará entre ese rango de valores con un nivel de confianza del 95%, siendo su valor medio 0.22.
- La proporción de coches con 8 cilindros, al 95% de confianza, está entre 0,22 y 0,31. En términos de media, el 26% de los coches tienen 8 cilindros.

❖ CASO 2: MUESTREO ALETORIO SIMPLE CON REEMPLAZAMIENTO (MACR).

MACR consiste en seleccionar una muestra ordenada de tamaño n, donde se efectúan n extracciones sucesivas reponiendo la unidad que ha salido en la última extracción.

La función correspondiente a MACR es "*E.WR*", la cual nos devuelve la tabla x con datos estimados.

Tabla_MAS_CONREEM <- E.WR (N, n, var_interes)

	N	CONSUMO	ACELERACIÓN	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	9000	101179,7	139800	4664,06	1945,31	2390,63
Error estándar		1798,408	1265,013	229,79	189,3	203,11
Coeficiente de variación		1,77	0.9	4,93	9,73	8,5
DEFF			1,04			

Tabla 3: Estimaciones según MACR.

Tras utilizar los comandos correspondientes a este tipo de muestreo y utilizar las salidas que nos proporcionan estos comandos (tabla 3), podemos, de nuevo, calcular IC, comentar e interpretar los rangos entre los que se mueve el parámetro de interés.

Se presenta la tabla 4 donde se muestra, además de los IC, la estimación de la media o proporción según se trate de una variable cuantitativa o cualitativa. Además observamos cual es el *efecto* de MACR con respecto a MAS.

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,24	10,85	11,63	
ACELERACIÓN	15,53	15,26	15,8	
4 CILINDROS	0,52	0,47	0,57	1,04
6 CILINDROS	0,22	0,17	0,26	
8 CILINDROS	0,26	0,22	0,3	

Tabla 4: Tabla resumen con los resultados del MACR.

Dado que nuestro objetivo es, más que interpretar resultados, comparar resultados, debemos destacar el gran parecido entre los valores resultantes de MAS y los valores resultantes de MACR plasmados en las tablas 1 y 3 (estimaciones según MAS o MACR), como en las tablas 2 y 4 (tablas resumen).

Esto es debido a la fracción de muestreo (f), es decir, debido a la proporción muestral con respecto a la población. Cuando éste valor es pequeño, las diferencias entre MAS y MACR son mínimas. Bajo nuestra suposición del tamaño poblacional (N) de 9000 coches y teniendo un tamaño muestral (n) de 384 coches, la fracción muestral (f = n / N) es de 0,0426. Dado que la varianza utiliza el valor (1-f), en nuestro caso (1 - 0.0426 = 0,9573) es un valor muy cercano a la unidad, y por tanto, el hecho de que la muestra sea recogida con o sin reemplazamiento no provoca grandes variaciones en los resultados, puesto que es poco probable que en el muestreo con reemplazamiento se seleccione una unidad que ya haya sido seleccionada con anterioridad.

Aun así, al fijarnos en el valor DEFF (efecto) nos damos cuenta de que el muestreo aleatorio simple sin reemplazamiento es más eficaz, en nuestra población de interés, que el muestreo aleatorio con reemplazamiento, dado que DEFF es 1,04 y por tanto, mayor que la unidad (1,04 > 1).

- CASO 3: MUESTREO SISTEMÁTICO LINEAL DE 1 EN K (MSL).

En MSL se elige la muestra a partir de los elementos de una lista, los n elementos serán seleccionados según un cierto orden fijado con anterioridad, o bien, sin orden pero recorriendo la lista a partir de un número aleatorio determinado. **(1)**

La función correspondiente a MSL es "*E.SY*", que nos solicita tres argumentos; el tamaño población (*N*), el vector o matriz de datos pertenecientes a las variables de interés (*var_interes*), y (*k*), siendo este valor la parte entera por debajo de N/n**,** en nuestro caso k=23. El valor k sirve **(1)** para crear el algoritmo de este tipo de diseño, donde se elige un elemento r al azar en {1,…,k}, y se completa la muestra con los elementos r+k, r+2*k,…. Por este motivo el N que vemos en la tabla x(la de abajo) de estimaciones según MSL no es exactamente 9000 como era hasta ahora, si no algo menor.

Tabla_SISTEMATICO <- E.SY (N, k, var_interes)

La función "*E.SY*" nos devuelve la tabla 5 con datos estimados acerca de la población. Si comparamos esta tabla de estimaciones con las tablas de estimaciones de MAS y MACR somos conscientes de que tanto el error estándar como el coeficiente de variación es menor en todas las variables y niveles de las estimaciones si consideramos que la muestra ha sido seleccionada según un MSL, por lo que podríamos decir que tendríamos mejores estimaciones o resultados.

	N	CONSUMO	ACELERACIÓN	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	8832	99291	137190,4	4577	1909	2346
Error estándar		1742,38	1225,605	222,63	183,4	196,79
Coeficiente de variación		1,75	0,89	4,86	9,6	8,39
DEFF			1			

Tabla 5: Estimaciones según MSL.

Tras utilizar la función correspondiente a este tipo de muestreo así como los datos de la tabla 5, se calculan los IC. Éstos se presentan en tabla 6 donde se muestra, además de los IC, la estimación de la media o proporción según se trate de una variable cuantitativa o cualitativa y el *efecto* de MSL de 1 en k con respecto a MAS.

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,24	10,85	11,63	
ACELERACIÓN	15,53	15,26	15,8	
4 CILINDROS	0,52	0,47	0,57	1
6 CILINDROS	0,22	0,18	0,26	
8 CILINDROS	0,26	0,22	0,31	

Tabla 6: Tabla resumen con los resultados del MSL.

Como se puede ver, el efecto toma el valor de 1, y el resto de datos son los mismos que en MAS. Hemos de saber que el MSL de 1 en k es más preciso que el MAS cuando la variabilidad dentro de las muestras es superior a la variabilidad dentro de las unidades de la población, además, MSL es más eficiente cuando el marco presenta una tendencia lineal. El efecto, en este caso nos informa de que ambos diseños son equivalentes en cuanto a error de muestreo.

❖ CASO 4: MUESTREO ESTRATIFICADO CON MAS (MAE(n)).

En este tipo de muestreo la población se divide en L clases, estratos o grupos, y todos ellos son examinados parcialmente. *Nh* es el número de unidades del estrato h-ésimo, seleccionadas a través de un MAS **(2)**. Es ideal para poblaciones que, de forma natural, están divididas según cierto criterio elegido, criterio que para nosotros es la variable de estudio.

En primer lugar selecciono esa variable mediante la cual voy a dividir la población en grupos. Tomaremos la variable cualitativa "Origen", que tiene tres niveles, éstos indican la procedencia de cada coche (Estados Unidos, Europa o Japón).

Es interesante estudiar las diferencias que pueden haber entre los resultados dependiendo de si los datos recogidos como muestra de cada grupo son proporcionales al tamaño poblacional de su grupo, o si por el contrario, la cantidad de datos de cada grupo de la muestra no tiene relación con la cantidad de individuos de la población. Por este motivo vamos a diferencia en "Caso 4 A" y "Caso 4 B".

Dado que el tamaño población es un dato que se está asumiendo, también asumiremos los tamaños de cada grupo de esa población y se asignarán valores *Nh* cuyo sumatorio sea igual a 9000, para así ser coherentes con el resto de diseños y poder comparar este diseño muestral estratificado con el resto. Los *nh* (tamaños de los estratos en la muestra) son comunes en los dos casos de estudios (ver figura 2).


```
#Definir el tamaño de los estratos de mi muestra
n1<-summary(ORIGEN)[[1]]
n2<-summary(ORIGEN)[[2]]
n3<-summary(ORIGEN)[[3]]
nh<-c(n1,n2,n3)
```

Figura 2: Código de tamaños de los estratos en la muestra Caso 4.

La función que se utiliza en el tipo de muestro estratificado es "*E.STSI*", esta función requiere la variable por la que segmentaremos la población, así como el vector de tamaños poblacionales y el vector de tamaños muestrales, además de las variables de interés.

➢ **Caso 4 A.**

En este caso asumiremos tamaños poblacionales proporcionales al tamaño muestral que tenemos en la base de datos. Dado que en ésta hay 244 coches de EEUU que corresponde a un 63,54% de la muestra, hay 65 coches de Europa que corresponde al 16,93% y hay 75 coches de Japón que corresponde al 19,53% de la

muestra restante, los valores poblaciones proporcionales son 5746 coches en EEUU, 1524 en Europa y 1730 coches en Japón, respectivamente. Véase figura 3 donde aparece la codificación;

```
# Caso 4 A #

#Asumir tamaños poblacionales de cada estrato
N1<-5746
N2<-1524
N3<-1730
Nh <- c(N1,N2,N3)
```

Figura 3: Codificación Nh caso 4 A.

Los parámetros a introducir en la función "*E.STSI*" son los siguientes;

Tabla_estratosA <- E.STSI (ORIGEN , Nh ,nh , var_interes)

Las tablas resultantes de esta función aparecen a continuación en la figura 4, donde constan 5 tablas, las dos primeras corresponden a las variables cuantitativas "Consumo" y "Aceleración, y las 3 siguientes corresponden a cada uno de los niveles de "Cilindros". En cada una de estas tablas vemos las estimaciones poblacionales de los parámetros de interés.

CONSUMO	EEUU	EUROPA	JAPÓN	POBLACIONAL
Nh	5746	1524	1730	9000
Estimación	74062,17	13465,66	13609,33	101317,2
Error estándar	1369,99	394,29	324,94	1462,163
Coeficiente de variación	1,85	2,89	2,39	1,44
DEFF	1	1	1	0,69
ACELERACIÓN	**EEUU**	**EUROPA**	**JAPÓN**	**POBLACIONAL**
Nh	5746	1524	1730	9000
Estimación	86051,06	25464,87	28247,44	139763,4
Error estándar	983,52	574,26	367,2	1196,63
Coeficiente de variación	1,14	2,26	1,3	0,86
DEFF	1	1	1	0,93
4 CILINDROS	**EEUU**	**EUROPA**	**JAPÓN**	**POBLACIONAL**
Nh	5746	1524	1730	9000
Estimación	1624,9	1430,22	1591,6	4646,71
Error estándar	162,4	44,79	53,36	176,75
Coeficiente de variación	9,99	3,13	3,35	3,8
DEFF	1	1	1	0,62
6 CILINDROS	**EEUU**	**EUROPA**	**JAPÓN**	**POBLACIONAL**
Nh	5746	1524	1730	9000
Estimación	1719,09	93,78	138,4	1951,27
Error estándar	165,16	44,79	53,36	179,26
Coeficiente de variación	9,6	47,76	38,56	9,19
DEFF	1	1	1	0,94
8 CILINDROS	**EEUU**	**EUROPA**	**JAPÓN**	**POBLACIONAL**
Nh	5746	1524	1730	9000
Estimación	2402,02	0	0	2402,02
Error estándar	177,91	0	0	177,91
Coeficiente de variación	7,4			7,41
DEFF	1			0,8

Figura 4: Estimaciones según Muestro estratificado caso 4 A.

A pesar de que el coste de este tipo de muestreo sea, generalmente, más elevado que el MAS, gracias al muestreo estratificado somos capaces de tener una estimación por cada grupo de la población, y por ello podemos ver fácilmente que en Europa y Japón no hay coches de 8 cilindros, además si nos fijamos en el error estándar que hay en las estimaciones de EEUU, Europa o Japón podremos ver que siempre es menor al error estándar de la última columna de datos, es decir, el error en la estimación poblacional de cada variable. Otra ventaja es que dado que los problemas de muestreo pueden darse en distintas partes de la población, si se ha usado muestreo estratificado, seremos conscientes de esto y quizás puedan ser corregidos con mayor facilidad. El hecho de que el muestreo estratificado se efectúa realizando de forma independiente en cada estrato un MAS(nh), el *efecto* de cada nivel de la variable *Origen* es 1.

El código que se debe de crear para obtener los intervalos de confianza y las medias o proporciones es mucho más largo que en el resto de diseños muestrales, ya que, en este caso podemos tener el triple intervalos y estimaciones. El trozo de código que se ve en la figura 5 que aparece a continuación es únicamente de la variable

"Consumo" y, únicamente, de EEUU, por lo que se realiza otro código para Europa, Japón y Poblacional, y de nuevo, otros códigos para el resto de variables con cada nivel de "*Origen*". Dado que debemos coger los datos de la figura 4 y ésta tiene diferentes tablas, la forma de recoger un dato específico es teniendo en cuenta el orden en el que indicarlo (fila x columna x tabla), por ejemplo [1,1,2] hace referencia a la fila 1 de la columna 1 de la tabla 1, es decir, la estimación del consumo en EEUU.

```
#CONSUMO

#EEUU
estimacion_con_EEUU<-Tabla_estratosA[1,1,2]
error_con_EEUU<-Tabla_estratosA[2,1,2]
IC_sup_con_EEUU<-((estimacion_con_EEUU/N1) + (1.96*error_con_EEUU/N1))
IC_inf_con_EEUU<-((estimacion_con_EEUU/N1) - (1.96*error_con_EEUU/N1))
Efecto_con_EEUU<-Tabla_estratosA[4,1,2]
```

Figura 5: Código IC para muestreo estratificado.

Gracias al código creado conseguiremos los datos que se muestran en la figura 6, datos que nos servirán para comparar los diseños muestrales.

CONSUMO	Media o proporción	IC límite inferior	IC límite superior	Efecto
EEUU	12,89	12,42	13,36	1
EUROPA	8,95	8,44	9,46	1
JAPÓN	7,87	7,5	8,23	1
POBLACIONAL	11,26	10,94	11,58	0,69
ACELERACIÓN	**Media o proporción**	**IC límite inferior**	**IC límite superior**	**Efecto**
EEUU	14,98	14,64	15,31	1
EUROPA	16,71	15,97	17,45	1
JAPÓN	16,33	15,91	16,74	1
POBLACIONAL	15,53	15,27	15,89	0,93
4 CILINDROS	**Media o proporción**	**IC límite inferior**	**IC límite superior**	**Efecto**
EEUU	0,28	0,22	0,34	1
EUROPA	0,94	0,88	0,99	1
JAPÓN	0,92	0,86	0,98	1
POBLACIONAL	0,52	0,48	0,55	0,62
6 CILINDROS	**Media o proporción**	**IC límite inferior**	**IC límite superior**	**Efecto**
EEUU	0,3	0,24	0,36	1
EUROPA	0,06	0	0,12	1
JAPÓN	0,08	0,02	0,14	1
POBLACIONAL	0,22	0,18	0,26	0,94
8 CILINDROS	**Media o proporción**	**IC límite inferior**	**IC límite superior**	**Efecto**
EEUU	0,41	0,36	0,48	1
EUROPA				
JAPÓN				
POBLACIONAL	0,27	0,23	0,31	0,8

Figura 6: Tablas resumen muestro estratificado caso 4 A.

Dado que en el resto de diseños muestrales sólo tenemos estimaciones acerca de la población total y no según el origen de los coches, la forma de compararlos también debe de ser genérica, y por tanto debemos basarnos en la última fila de datos de cada variable, que tenemos en la figura 6.

Como podemos ver, los *efectos* a nivel *poblacional* de cada variable son todos valores inferiores a la unidad, lo que quiere decir que para nuestra población es más eficaz realizar muestreo por estratos que un MAS, que un MACR o que un MSL.

En cuanto a los intervalos de confianza y las medias y proporciones caben destacar los siguientes aspectos;

- Según el muestreo estratificado, el consumo de litros de gasolina por cada 100 kilómetros es ligeramente mayor al indicado con el resto de muestreos, la media sube de 11,24 l/100km a 11,26 l/100km, teniendo cambios destacables en los intervalos de confianza proporcionales, pues lo que se dice en el resto de casos es que el consumo está entre 10,86 l/100km y 11,63 l/100km, y ahora ambos intervalos son bastante superiores (pues debemos tener en cuenta que hablamos de consumo), el consumo está entre 10,94 l/100km y 11,58 l/100km, con un 95% de confianza.
- La media respecto a los segundos que tardan los coches en ponerse de 0 a 100 kilómetros por hora sigue siendo de 15,53 segundos, pero los intervalos de confianza son algo inferiores. En el muestreo estratificado se puede decir que con un 95% de confianza, los coches tardan entre 15,27 segundos y 15,9 segundos.
- Con un 95% de confianza, la proporción de coches de 4 cilindros suele estar entre el 0,48 y 0,55, valores diferentes a los 0,47 y 0,57 que veíamos en el resto de diseños. En todos los muestreos el valor medio de coches que tienen 4 cilindros es de un 52%.

- **Caso 4 B**:

Como se ha comentado en la presentación del Caso 4 en general, la diferencia entre los "Casos 4" es la asignación de tamaños poblaciones que se asumen, en este caso B tomamos tamaños poblaciones de forma aleatoria. El código de la asignación de tamaños aleatoria lo vemos en la figura 7.

```
# CASO 4 B #

# Asumir tamaños poblacionales de cada estrato
N1b<-1500
N2b<-3000
N3b<-4500
Nhb <- c(N1b,N2b,N3b)
```

Figura 7: Codificación Nh caso 4 B.

La línea de código en la que se usa la función correspondiente y se introducen los parámetros adecuados al caso es;

Tabla_estratosB <- E.STSI (ORIGEN, Nhb, nh, var_interes)

Aunque nos devuelve la estimación de cada variable y por cada origen del coche, en la tabla 7 sólo se muestran las estimaciones en términos *poblacional* ya que son los datos que utilizamos para comparar los diseños muestrales.

POBLACIONAL	N	CONSUMO	ACELERACIÓ	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	9000	81595,55	146067,4	7379,56	993,39	627,05
Error estándar		1209,046	1516,998	171,24	171,39	43,43
Coeficiente de variación		1,48	1,04	2,32	17,25	6,93
DEFF		0,47	1,5	0,58	0,86	0,05

Tabla 7: Estimaciones según Muestro estratificado caso 4 B, poblacional.

Como vemos en la tabla 7 el error estándar de todas las estimaciones es, excepto en "aceleración", menor en las estimaciones que se han conseguido bajo la hipótesis del muestreo estratificado que bajo la hipótesis de MAS, MACR o MSL. Además, los *efectos* de cada variable indican, de nuevo exceptuando la variable "aceleración" que la hipótesis de haber obtenido la muestra con un MAS provoca peores estimaciones que si pensáramos que ha sido recogida mediante un muestro estratificado.

De nuevo, tras crear el código para extraer los datos de las 5 tablas que devuelve la función "*E.STSI*", podemos obtener tantos intervalos de confianza como estimaciones de parámetros de interés, es decir, de cada variable y por cada origen. En la tabla 8 tenemos los datos poblacionales necesarios para seguir con la comparación entre hipótesis de muestreo.

POBLACIONAL	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	9,06	8,8	9,33	0,47
ACELERACIÓN	16,23	15,9	16,56	1,5
4 CILINDROS	0,82	0,78	0,86	0,58
6 CILINDROS	0,11	0,07	0,15	0,86
8 CILINDROS	0,07	0,08	0,06	0,05

Tabla 8: Tabla resumen muestro estratificado caso 4 B, poblacional.

Podemos observar que las estimaciones poblacionales cambian drásticamente con respecto al resto de muestreos. Por ejemplo, decir que un coche, de media, consume 9,06 litros cada 100km o decir que consume 11,24 litros cada 100km es un cambio importante en las conclusiones. Si nos fijamos en los segundos que tarda un coche en ponerse de 0 a 100km/h observamos que hay casi un segundo de diferencia, lo cual es mucha diferencia estando en el ámbito de coches. Además, hasta ahora, se decía que había mayor número de coches con 4 cilindros (52%) y que el menor número de coches correspondía a los que tenían 6 cilindros, en cambio, si asumimos que la muestra ha sido recogida con un muestro estratificado, concluiremos que el mayor número de coches es también de 4 cilindros pero estos representan no el 52%, si no el 82%, y el menor porcentaje de coches no es de 6 cilindros, si no de 8 cilindros.

- CASO 5: MUESTREO ALEATORIO CON PROBABILIDADES PROPORCINALES AL TAMAÑO (MAPPT).

También es conocido como muestreo con probabilidades de selección desiguales. Este diseño trata de **(1)** asignar una probabilidad a cada elemento de la población, seleccionar un elemento de la población bajo esas probabilidades y continuar este proceso de forma iterativa hasta que se consigan los *n* requeridos.

La función que se utiliza en este caso requiere de probabilidades de inclusión, para asignar dichas probabilidades a cada elemento de la población se usa la siguiente expresión: $Pi = (tiempo[i] * n / T)$ siendo *T* el total de años que tienen los coches de la población.

"*TIEMPOS*" es una variable creada específicamente para guardar los valores en años que hace desde que se fabricó cada coche de la base de datos, es decir, de la muestra. Esto se consigue gracias a la diferencia entre el valor 118 (ya que estamos en el año 2018) y cada dato de la variable original "*FECHA*" donde están guardados el año (los 2 últimos números del año) en el que se fabricó cada coche. Además, como estamos asumiendo que la población de coches es 9000, ahora también necesitamos asumir el valor correspondiente a la suma del total de años de todos los coches de la población. Para ello, se multiplica por 30 el total de años de todos los coches de la muestra, ya que la muestra es aproximadamente el 30% de la población. Ver figura 8 donde se muestra el código.

```
Coches$TIEMPOS <- 118 - Coches$FECHA
sumatorio<-sum(Coches$TIEMPOS)
T<-sumatorio*30

probabilidades<-c(rep(0,n))
 for(i in 1:n)
  {
    probabilidades[i]<-(Coches$TIEMPOS[i]*n/T)
 }
```

Figura 8: Código para asignar probabilidades en MAPPT.

La función asociada a este diseño muestral es "*E.piPS*", cuyos argumentos de entrada son las variables de interés y las probabilidades de inclusión asignadas a cada elemento.

Tabla_PROBA <- E.piPS (var_interes, probabilidades)

Gracias a esta función conseguimos la siguiente tabla 9 de estimaciones conseguidas si asumimos que la muestra se ha recogido mediante un MAPPT.

	N	CONSUMO	ACELERACIÓ	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	11609,82	128282	181143,1	6173,7	2503,51	2932,61
Error estándar	51,71731	1963,322	1971,135	301,6	240,49	245,7
Coeficiente de variación	0,445618	1,53	1,08	4,89	9,6	8,38
DEFF		0,74	1,51	1,07	1	0,91

Tabla 9: Estimaciones según MAPPT.

Si nos fijamos en la primera fila y columna de datos, vemos que 11.609 corresponde a N, este valor es superior a 9000 (valor asumido como población total) lo que quiere decir que por azar, se han seleccionado elementos muy representativos. En consecuencia las estimaciones del total de la población (fila 1) son mayores que en el resto de diseños muestrales. Se debe prestar atención al error estándar de las estimaciones, pues todos son valores superiores al resto de errores estándar de los otros tipos de muestreos. Por lo que podemos decir que si asumimos la hipótesis de que el diseño muestral que se ha escogido para seleccionar la muestra es un MAPPT tendremos un error estándar superior al resto de hipótesis.

Tras extraer los datos necesarios de la salida de la función "*E.piPS*" y realizar los cálculos necesarios, obtenemos la siguiente tabla 10:

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,04	10,72	11,38	0,74
ACELERACIÓN	15,6	15,27	15,94	1,51
4 CILINDROS	0,53	0,48	0,58	1,07
6 CILINDROS	0,22	0,18	0,26	1
8 CILINDROS	0,25	0,21	0,29	0,9

Tabla 10: Tabla resumen MAPPT.

Un cambio destacable con respecto al resto de hipótesis muestrales es la media de consumo de litros de gasolina, la cual es casi medio litro inferior si asumimos MAPPT. En la última columna de datos de la tabla 10 vemos que dependiendo de la variable es más o menos efectivo el MAS que el MAPPT. Sería recomendable trabajar bajo la hipótesis de MAPPT cuando queramos estimar el consumo de litros de gasolina, así como la proporción de coches de 8 cilindros. En el caso de querer estimar los coches de 6 cilindros será indiferente al asumir cualquiera de los dos tipos de muestreo.

- CASO 6: MUESTREO ALEATORIO CON PROBABILIDADES PROPORCINALES AL TAMAÑO CON REPOSICIÓN (MAPPTCR).

Este diseño muestral funciona de forma similar al anterior (MAPPT) pero con la diferencia de la reposición de cada elemento tras su extracción **(1)**. De igual manera se debe asignar una probabilidad a cada elemento de la población y seleccionar n elementos de ésta bajo esas probabilidades.

Para asignar probabilidades debemos de crear un código nuevo con respecto al creado en el caso anterior, puesto que antes la función específica de ese muestreo precisaba de las probabilidades de inclusión en lugar de las probabilidades de selección que precisa la función correspondiente a MAPPTCR. Para calcular estas probabilidades, que se asignan a cada elemento de la población, se usa la siguiente expresión: *Pi = (tiempo[i] / T)* siendo *T* el total de años que tienen los coches de la población y *tiempo* es una variable donde se guarda los años que hace que se fabricó el coche *i*. El código utilizado para calcular las probabilidades de selección se puede ver en la figura 9.

```
probabilidades_MAPPTCR<-c(rep(0,n))
 for(i in 1:n)
  {
    probabilidades_MAPPTCR[i]<-(Coches$TIEMPOS[i]/T)
 }
```

Figura 9: Código para asignar probabilidades en MAPPTCR.

Se usa la función "*E.PPS*" para obtener la tabla 11 de estimaciones según MAPPTCR.

TABLA_PROBA_REPO <- E.PPS(var_interes, probabilidades_MAPPTCR)

	N	CONSUMO	ACELERACIÓN	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	11609,82	128282	181143,1	6173,7	2503,51	2932,61
Error estándar	52,58763	1997,565	2004,821	306,68	244,55	249,99
Coeficiente de variación	0,4529582	1,56	1,11	4,97	9,77	8,52
DEFF		0,76	1,56	1,11	1,04	0,94

Tabla 11: Estimaciones según MAPPTCR.

De igual manera que pasa con el muestreo aleatorio simple sin reposición (MAS) y con reposición (MACR), el MAPPT y el MAPPTCR asignan los mismos valores a las estimaciones. Esto es debido a la fracción de muestreo (f), pues n / N es un valor cercano a 0 y esto se ve reflejado en la varianza muestral que depende de (1-f). El hecho de que la muestra sea recogida con o sin reemplazamiento no hace variar los resultados.

Aún así vemos que el error estándar de las estimaciones es superior comparado con el error estándar del resto de estimaciones correspondientes a cada diseño muestral. Por lo que podríamos decir que el resto de estimaciones son más apropiadas. Solo en el consumo y en la proporción de coches de 8 cilindros el *efecto* indica que MAPPTCR es mejor que MAS ya que el *efecto* es un valor inferior a la unidad.

En cuanto a la media del consumo y media de aceleración, así como la proporción de coches de 4,6 u 8 cilindros tenemos las mismas conclusiones que en MAPPT, donde el cambio más destacado es el medio litro superior de consumo medio que se indica tanto en el caso de con reposición como en el de sin reposición, podemos verificar esto con la siguiente tabla 12.

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,05	10,71	11,38	0,77
ACELERACIÓN	15,6	15,26	15,94	1,56
4 CILINDROS	0,53	0,48	0,58	1,11
6 CILINDROS	0,22	0,17	0,26	1,04
8 CILINDROS	0,25	0,21	0,29	0,94

Tabla 12: Tabla resumen MAPPTCR.

2.2 Conclusiones.

Tras hacer un repaso de todos los casos anteriores donde se comentan los diferentes tipos de funciones capaces de devolver datos según si asumimos que la muestra ha sido recogida a través de un diseño muestral u otro, y tras hacer pequeñas comparativas, llegamos a la idea de que efectivamente sí es importante conocer qué tipo de muestreo se ha utilizado para recoger una muestra que se está analizando, ya que las conclusiones a las que llegamos varían según que hipótesis se asume.

Observamos cambios tan drásticos como una diferencia en media de 2,18 litros consumidos cada 100 km entre una hipótesis y otra (caso MAS y caso MAE B), así como 0,7 segundos de diferencia en lo que tardan los vehículos de media en pasar de 0 a 100km/hora. Además, también cambia el porcentaje de coches un 4, 6 y 8 cilindros, incluso variando el orden porcentual, es decir, en unas hipótesis la minoría de coches corresponde a los que tienen 6 cilindros, y en otras se especifica que la minoría son los coches que tienen 8 cilindros.

3. ERRORES MUESTRALES.

En este último apartado se pretende recalcar la importancia de estudiar y conocer los diferentes tipos de muestreo para así poder elegir el más adecuado a la población de estudio que queramos analizar.

Con los siguientes "Casos" pretendo ilustrar las discrepancias en cuanto a precisión que se consigue tras aplicar un tipo de muestreo u otra a la población de interés. La base de datos ("Coches") y las variables de interés ("Cilindros", "Consumo" y "Aceleración") son las mismas que las utilizadas en el apartado 2 de este documento.

En cada "Caso" se aplica un tipo de muestreo mediante la utilización de ciertas funciones **(1)** del paquete "*TeachingSampling"* consiguiendo así la muestra que se va a utilizar. Seguidamente, se utilizan las funciones específicas de cada muestreo para obtener la estimación del total de la población de cada variable de interés y por cada nivel de la variable "Cilindros", el error estándar por cada estimación, el coeficiente de variación estimado y el efecto del diseño (DEFF). Por último, gracias a los datos que proporcionan las salidas de dichas funciones, tras unos cálculos se consigue obtener la estimación de la media, en el caso de las variables cuantitativas, y la estimación de la proporción en el caso de las cualitativas, así como los intervalos de confianza para cada parámetro. Intervalos que son de gran utilidad para estudiar y comparar la precisión de los diferentes tipos de muestreo.

3.1 Casos.

En cada uno de los siguientes casos se asume como población todas las observaciones (N=384) de la base de datos "COCHES", y en cada "Caso" se realiza un tipo de muestreo. Todos los diseños muestrales seleccionan 30 observaciones (n), y podremos ver que, a pesar de que en todos los casos se ha aplicado correctamente un tipo de muestreo y el tamaño muestral es el mismo para todos, la precisión cambia según si el tipo de muestreo es más o menos apropiado para nuestra población de interés. En la figura 10 vemos la asignación de tamaños muestrales.

```
N <- dim(Coches)[1]
N
n <- 30
```

Figura 10: Código N y n común en errores muestrales.

⇒ CASO 1: MUESTREO ALEATORIO SIMPLE (MAS).

El muestreo aleatorio simple es un tipo de muestreo muy sencillo, que se suele utilizar cuando se dispone de un listado de todos los elementos de la población y cuando la población de interés es desconocida, es decir, cuando no se sabe si la población está agrupada en conjuntos ni se conoce la respuesta de los individuos a una pregunta cuantitativa relacionada con la de interés.

Para seleccionar los 30 elementos de la muestra se utiliza, en el caso de MAS, la función "*S.SI*", función que necesita como argumentos de entrada el tamaño poblacional (N) y el tamaño muestral que queremos obtener (n). En la figura 11 vemos una parte de la salida proporcionada por dicha función, donde se aprecia que los elementos 25, 26 y 29 si han sido seleccionados.

muestra_MAS <- S.SI (N, n)

```
[22,]    0
[23,]    0
[24,]    0
[25,]   25
[26,]   26
[27,]    0
[28,]    0
[29,]   29
[30,]    0
```

Figura 11: Salida función S.SI (384,30)

Con la anterior función conseguimos los elementos de la lista que han sido seleccionados, por lo que se debe indicar qué es lo que queremos de esos elementos. En nuestro caso, queremos utilizar los datos correspondientes a las variables de interés de esos elementos seleccionados. La línea de código para conseguir esto es:

datos_MAS <- var_interes[muestra_MAS,]

Una vez se tiene la muestra con todos los datos que interesan, gracias a la función "E.SI" y la introducción de los parámetros que ésta requiere, conseguimos la tabla 13 de estimaciones que vemos a continuación.

Tabla_MAS <- E.SI(N, n, datos_MAS)

	N	CONSUMO	ACELERACIÓN	4 CILINDROS	6 CILINDROS	8 CILINDROS
Estimación	384	4428,8	6050,56	204,8	64	115,2
Error estándar		268,52	189,78	34,16	25,51	31,37
Coeficiente de variación		6,06	3,14	16,68	39,87	27,23
DEFF			1			

Tabla 13: Estimaciones según MAS(30).

La anterior tabla de estimaciones nos sirve no tanto para comentar los datos que nos ofrece, si no para utilizar estos de forma que podamos conseguir la media o proporción de cada variable o nivel, así como los intervalos de confianza. Estos datos son los que se muestran en la tabla 14:

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,53	10,16	12,9	
ACELERACIÓN	15,76	14,79	16,73	
4 CILINDROS	0,53	0,36	0,71	1
6 CILINDROS	0,17	0,03	0,3	
8 CILINDROS	0,3	0,14	0,46	

Tabla 14: Tabla resumen MAS(30).

Si comparamos esta tabla con la tabla que se obtuvo cuando se trabajó bajo la hipótesis de un MAS de n=384, en el apartado 2 de este documento, podemos ver que la amplitud de los 5 intervalos que se muestran es bastante superior en la tabla 14 que acabamos de obtener. Esto se debe simplemente al tamaño muestral con el que se trabaja, es lógico pensar que a mayor tamaño muestral mayor precisión en las estimaciones. La decisión acerca del tamaño muestral que se debe tomar es una

cuestión en la que se debe tener en cuenta no solo esa precisión si no el coste que se tendrá que asumir.

⇒ CASO 2: MUESTREO ALETORIO SIMPLE CON REEMPLAZAMIENTO (MACR).

El muestreo aleatorio simple con reemplazamiento se suele usar bajo las mismas condiciones poblacionales que se usa el MAS. Sin embargo, MACR es un n/(N–n)% peor que MAS, diferencia que no se aprecia cuando el tamaño poblacional (N) es muy grande o infinito **(1).** En nuestro caso, ese valor corresponde a un 8,47%, porcentaje que nos indica que sí hay diferencia en términos de calidad de muestra que podemos verificar a continuación.

Para seleccionar la muestra, se usa la siguiente función:

muestra_MAS_CON <- S.WR (N, n)

De nuevo, debemos no sólo quedarnos con la lista de elementos que nos proporciona la anterior función, si no seleccionar los datos de las variables de interés correspondientes a esos elementos.

datos_MAS_CON <- var_interes [muestra_MAS_CON,]

Con la siguiente línea de código obtenemos una tabla que proporciona la estimación del total de la población, error estándar, coeficiente de variación y efecto del diseño (DEFF) para cada variable de interés y por cada nivel. Datos cuya funcionalidad es ser utilizados para conseguir la tabla 15, que vemos a continuación.

Tabla_MAS_CONREEM <- E.WR (N, n, datos_MAS_CON)

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,13	9,76	12,51	
ACELERACIÓN	15,17	14,38	15,98	
4 CILINDROS	0,57	0,39	0,75	1,08
6 CILINDROS	0,27	0,11	0,43	
8 CILINDROS	0,17	0,03	0,3	

Tabla 15: Tabla resumen MACR(30).

Gracias al efecto, cuyo valor es mayor a la unidad (en 0.08), podemos decir que el MAS es más apropiado para "Coches", es decir, MAS consigue una muestra de

mayor calidad a la muestra que ha conseguido MACR en nuestra población de interés, pues tiene menor error de muestreo.

En cuanto a los intervalos de confianza, si comparamos MACR(30) con MAS(30) no observamos mucha diferencia en cuanto al rango de valores que abarca la amplitud de los intervalos. Por lo que, la precisión que se consigue en ambos diseños muestrales no cambia de forma llamativa.

⇒ CASO 3: MUESTREO SISTEMÁTICO LINEAL DE 1 EN K (MSL).

Este tipo de muestreo se recomienda aplicar cuando no existe una lista donde aparezcan todas las unidades de la población, es decir, cuando no existe *marco* o cuando si exista y los elementos estén ordenados. Se sabe que se pueden conseguir, con este diseño muestral, muestras muy representativas si las unidades de la población están numeradas de forma que cuanto más parecidas sean más cerca estarán en la lista de la población, aunque también podemos conseguir muestras representativas si las unidades están numeradas al azar. Además, se considera que la selección de la muestra es sencilla, tanto en el trabajo de oficina como en el de campo.

Como ya se explicó en el punto 2, necesitamos fijar un k que será la parte entera por debajo de N/n. Para seleccionar esta parte entera utilizamos la función "*trunc*". En nuestro caso k=12, parámetro de entrada, junto al tamaño poblacional, que necesita la función "*S.SY*" para seleccionar los elementos de la muestra.

k <- N / n

k <- trunc (k)

muestra_SISTEMATICO <- S.SY (N, k)

Esta función devuelve un vector de tamaño n, cada elemento de este vector indica la unidad que se seleccionó, por lo que falta agrupar los datos que nos interesan de cada unidad:

datos_SIST <- var_interes [muestra_SISTEMATICO,]

Gracias a la función "*E.SY*" obtendremos los datos necesarios para construir la siguiente tabla 16 para poder comparar los diseños muestrales.

Tabla_SISTEMATICO <- E.SY (N, k, datos_SIST)

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,91	10,45	13,37	
ACELERACIÓN	15,22	14,29	16,15	
4 CILINDROS	0,47	0,3	0,64	1
6 CILINDROS	0,22	0,08	0,36	
8 CILINDROS	0,31	0,16	0,47	

Tabla 16: Tabla resumen MSL.

MAS y MSL de 1 en k son, en cuanto a error de muestreo, equivalentes, ya que el *efecto* toma el valor de la unidad. Se dice que MSL es preferible a MAS si la variabilidad dentro de las muestras sistemáticas es mayor que la variabilidad global, o si las muestras sistemáticas son internamente heterogéneas, es decir, tienen correlación interna negativa.

Si nos fijamos en la amplitud de los intervalos de confianza, podemos ver que el rango de valores es relativamente mayor si se aplica un MSL a si se aplica un MAS. Por lo que en términos de precisión, nos decantaríamos por un MAS(30) para muestrear nuestra población de interés. Si restamos el límite superior al inferior de los intervalos de confianza de cada variable y en cada tipo de muestreo obtenemos las siguientes diferencias que podemos ver en la figura 12, en la última fila tenemos el sumatorio de estas diferencias.

Diseños	MAS	MSL
CONSUMO	2,74	2,92
ACELERACIÓN	1,94	1,86
4 CILINDROS	0,35	0,34
6 CILINDROS	0,27	0,28
8 CILINDROS	0,32	0,31
Sumatorio	5,62	5,71

Figura 12: Amplitud de IC, MAS y MSL.

⇒ CASO 4: MUESTREO ESTRATIFICADO CON MAS (MAE(n)).

El muestreo estatificado es aconsejable cuando la varianza dentro de los estratos es pequeña en relación con la varianza entre estratos, es decir, cuando los estratos construidos son grupos internamente homogéneos y externamente heterogéneos con respecto a la variable de interés. Así mismo, es interesante utilizar este diseño muestral cuando se busca estimaciones sobre cada conjunto de la población. La

estratificación permite obtener a menor coste estimadores con varianza más pequeña, por lo que al aumentar el número de estratos (hasta cierto límite) suele aumentar la precisión, pero también el coste, por lo que se busca un equilibrio entre precisión y coste.

La variable que vamos a usar para dividir la población en grupos o estratos es "ORIGEN", variable con 3 niveles, por lo que nos ceñiremos a crear 3 estratos. "ORIGEN" indica si los coches son de EEUU, de Europa o de Japón. En términos de porcentajes, a EEUU corresponden el 63.54% de la población, a Europa el 16.93% y a Japón el 19.53% restante. Como se ha indicado con anterioridad, la base de datos coches, en este apartado 3, se asume como población, por lo que asignaremos el número de coches de cada origen de la base de datos como tamaños poblacionales de los estratos. Véase la siguiente figura 13.

```
> summary(ORIGEN)
 EE.UU Europa  Japón
   244     65     75
> N1<-summary(ORIGEN)[[1]]
> N2<-summary(ORIGEN)[[2]]
> N3<-summary(ORIGEN)[[3]]
> Nh <- c(N1,N2,N3)
```

Figura 13: Código asignación tamaños poblacionales MAE(30).

De nuevo vamos a distinguir 2 casos en el muestreo estatificado, estos casos dependen de la asignación del tamaño muestral, esto es, de la forma de repartir el tamaño de la muestra entre los distintos estratos. El tamaño relativo de la muesrta correspondiente al *estrato h* sería "*wh=nh/n"*, donde el sumatorio de *wh* debe ser *1* o, lo que es lo mismo, el *sumatorio de los nh* debe ser *igual a n*, en nuestro caso, igual a 30, pues es el tamaño muestral que hemos decidido seleccionar en todos los anteriores casos.

- **Caso 4 A.**

La asignación proporcional consiste en asignar a cada estrato un tamaño muestral proporcional al tamaño de dicho estrato. El código correspondiente es el que se ve en la figura 14, donde los tamaños de los estratos son guardados en un vector *nha*.

```
## CASO A : ASIGNACIÓN PROPORCIONAL

n1<-16
n2<-5
n3<-6
nha<-c(n1,n2,n3)
```

Figura 14: Código asignación tamaños muestrales MAE(30) caso 4 A.

Para seleccionar los 30 elementos en total, de los cuales 16, 5 y 6 corresponden a EEUU, Europa y Japón respectivamente, utilizamos la función "*S.STSI*", cuyos parámetros de entrada son la variable a partir de la cual se realizan los grupos, y dos vectores, uno con los tamaños poblacionales y otro con los tamaños muestrales de cada estrato.

muestra_ESTRATI_A <- S.STSI (ORIGEN, Nh, nha)

La función utilizada devuelve un vector donde se indican las unidades seleccionadas, con el siguiente código conseguimos los datos que nos interesan correspondientes a eses unidades.

datos <- Coches[muestra_ESTRATI_A,]

attach(datos)

Gracias a la función "*E.STSI*" obtenemos una matriz compuesta por varias matrices que representan cada variable de interés. Las columnas de cada matriz corresponden a los parámetros estimados de las variables de interés en cada estrato y en toda la población. Mediante estas matrices o tablas podemos extraer los datos que nos interesan para realizar los cálculos necesarios para conseguir la tabla 17. Para indicar el dato que se quiere extraer se debe indicar en el orden "fila x columna x tabla".

Tabla_estratosA <- E.STSI (ORIGEN, Nh, nha, var_interes)

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	14,13	12,81	14,45	0,86
ACELERACIÓN	12,52	11,75	13,28	0,59
4 CILINDROS	0,24	0,1	0,37	0,56
6 CILINDROS	0,14	0,02	0,26	0,69
8 CILINDROS	0,62	0,49	0,75	0,64

Tabla 17: Tabla resumen MAE(30) caso A.

Todo indica que el muestreo estratificado con asignación proporcional del tamaño muestral es más idóneo para nuestra población de interés "COCHES" que el resto de muestreos utilizados con anterioridad. Respecto al MAS(30), confirmamos que tiene menor efectividad que MAE, fijándonos en la última columna de datos de la tabla

17, donde se indica que el *efecto* de cada variable con respecto al MAS, es menor a la unidad, por lo que descartamos de forma casi automática el muestreo aleatorio simple. Respecto al resto de muestreos podemos ver en la figura 15 que la amplitud de los intervalos de confianza son todos rangos mayores que los conseguidos en MAE(30) caso A, por lo que se consigue estimaciones bastante más precisas que el resto de diseños.

Diseños	MAS	MACR	MSL	MAE caso A
CONSUMO	2,74	2,75	2,92	1,64
ACELERACIÓN	1,94	1,6	1,86	1,53
4 CILINDROS	0,35	0,36	0,34	0,27
6 CILINDROS	0,27	0,32	0,28	0,24
8 CILINDROS	0,32	0,27	0,31	0,26
Sumatorio	5,62	5,3	5,71	3,94

Figura 15: Amplitud de IC con MAS, MACR, MSL y MAE caso A.

Es interesante destacar que, como podemos ver en la anterior figura 15 columna 1 y 2 de datos, en cuanto a amplitud de intervalos, MACR es mejor que MAS. Teóricamente, esto no puede suceder ya que, si se tiene la misma muestra, la fórmula del ECM (error cuadrático medio) del MACR da un valor mayor que la fórmula ECM del MAS, pero, en la práctica si es posible, ya que la muestra es diferente, son muestras aleatoriamente distintas y, al azar, ha salido mejor muestra a través del MACR que a través del MAS.

- **Caso 4 B.**

Seguimos en el caso 4 en el que se ejemplifica el diseño muestral estratificado pero, en este apartado B, tenemos la peculiaridad de utilizar lo que se conoce como "asignación igual", o sea, se asigna el mismo tamaño a todos los estratos de la muestra. Si el tamaño de muestra que queremos es 30 y tenemos 3 estratos, cada uno será de tamaño 10, como podemos ver en la figura 16 donde se muestral la asignación de tamaños y el vector donde éstos se guardan.

```
## CASO B : ASIGNACIÓN IGUAL

n1b<-10
n2b<-10
n3b<-10
nhb<-c(n1b,n2b,n3b)
```

Figura 16: Código asignación tamaños muestrales MAE(30) caso 4 B.

De nuevo usamos la función "*S.STSI"* para seleccionar los 30 elementos de la muestra, y las siguientes líneas de código para guardar los datos de interés que se piden como parámetros en la función "*E.STSI*", función que devuelve la tabla de datos que utilizaremos para conseguir la tabla 18 que aparece a continuación.

muestra_ESTRATI_B <- S.STSI(ORIGEN, Nh, nhb)

datos_B <- Coches[muestra_ESTRATI_B,]

attach(datos_B)

Tabla_ESTRATI_B<-E.STSI(ORIGEN ,Nh, nhb, var_interes)

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	14,59	13,56	15,61	0,58
ACELERACIÓN	11,74	11,08	12,4	0,49
4 CILINDROS	0,17	0,09	0,24	0,19
6 CILINDROS	0,07	0,01	0,13	0,17
8 CILINDROS	0,76	0,69	0,84	0,24

Tabla 18: Tabla resumen MAE(30) caso B.

Si nos paramos a examinar la tabla anterior 18 vemos que todo indica que el muestreo estratificado es más idóneo para nuestra población que el resto de muestreos. El *efecto* revela que el error de muestreo cometido es menor si se utiliza MAE que si se utiliza MAS. Y los intervalos de confianza abarcan un menor rango de valores que el resto de intervalos conseguidos con los otros diseños muestrales. Para poder compararlos de forma sencilla y numérica, podemos fijarnos en la figura 17, donde se ha calculado la diferencia entre el límite superior y el límite inferior de los intervalos, y, en la última fila, podemos ver la suma de estas diferencias. La menor diferencia, que corresponde a MAE, está coloreada en morado.

Diseños	MAS	MACR	MSL	MAE caso B
CONSUMO	2,74	2,75	2,92	2,05
ACELERACIÓN	1,94	1,6	1,86	1,32
4 CILINDROS	0,35	0,36	0,34	0,15
6 CILINDROS	0,27	0,32	0,28	0,12
8 CILINDROS	0,32	0,27	0,31	0,15
Sumatorio	5,62	5,3	5,71	3,79

Figura 17: Amplitud de IC con MAS, MACR, MSL y MAE caso B.

Ahora bien, podemos plantearnos qué tipo de estratificado es más acertado para "COCHES", en otros términos, qué tipo de asignación del tamaño muestral en los estratos es más adecuado. Con esto, nos damos cuenta, una vez más, de la importancia de conocer las características y la variedad de diseños muestrales que existen, pues aún puede cambiar la precisión dependiendo de si elegimos seleccionar la muestra con un MAE con asignación proporcional (caso A) o un MAE con asignación igual (caso B). Si nos fijamos en la figura 18, nos decantaremos por un MAE con asignación igual, ya que éste consigue una mayor precisión reflejada en la amplitud de sus intervalos de confianza.

Diseños	MAE caso A	MAE caso B
CONSUMO	1,64	2,05
ACELERACIÓN	1,53	1,32
4 CILINDROS	0,27	0,15
6 CILINDROS	0,24	0,12
8 CILINDROS	0,26	0,15
Sumatorio	3,94	3,79

Figura 18: Amplitud de IC con MAE caso A y MAE caso B.

⇒ CASO 5: MUESTREO ALEATORIO CON PROBABILIDADES PROPORCINALES AL TAMAÑO (MAPPT).

Este diseño muestral se suele utilizar cuando se conoce una variable cuantitativa auxiliar relacionada con la de interés para así conseguir muestras representativas de la población.

En la base de datos que asumimos como población, existe una variable a la que se le llama "FECHA", podemos pensar que las características de un coche, el consumo que tiene, la aceleración o la cantidad de cilindros que tiene, puede tener relación con la antigüedad de los coches. Motivo por el cual se ha decido, ya en el punto 2 de este documento, crear una variable "TIEMPOS" que indique los años que tiene cada coche.

La función "*S.piPS*" es la función que se debe de utilizar para seleccionar los elementos de la población y su probabilidad si utilizamos un MAPPT. Dicha función devuelve una matriz de 30 filas y dos columnas. Cada elemento de la primera columna indica la unidad que se seleccionó y cada elemento de la segunda columna indica la probabilidad de inclusión de esta unidad. En la figura 19 podemos ver una parte de esta selección.

```
> Muestra_PROBA <- S.piPS (n, Coches$TIEMPOS)
> Muestra_PROBA
      samp      Pik.s
[1,]   27 0.08923592
[2,]    4 0.08923592
[3,]    2 0.08923592
[4,]    1 0.08923592
[5,]   43 0.08737684
[6,]   37 0.08737684
```

Figura 19: Código y salida función S.piPS

Lo que llamaos "*datos_PROBA*", en las siguientes líneas de código, es uno de los parámetros, el cual guarda los datos correspondientes a las unidades seleccionadas por la anterior función, que pide la función "E.piPS". Función que proporciona datos que utilizaremos para crear la tabla 19 a partir de la cual se evalúa la precisión conseguida por el muestreo MAPPT. El segundo parámetro que pide la función "E.piPS" son las probabilidades de inclusión de los elementos seleccionados, por lo que, seleccionaremos la segunda columna de la salida "Muestra_PROBA".

```
probabilidades <- Muestra_PROBA[,2]
datos_PROBA  <-  var_interes[Muestra_PROBA, ]
Tabla_PROBA  <- E.piPS (datos_PROBA,  probabilidades)
```

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	11,84	10,64	13,04	0,68
ACELERACIÓN	15,17	13,85	16,49	1,8
4 CILINDROS	0,52	0,39	0,7	1,1
6 CILINDROS	0,14	0,02	0,26	1,1
8 CILINDROS	0,4	0,18	0,5	0,86

Tabla 19: Tabla resumen MAPPT.

El *efecto* denota la conveniencia de utilizar un MAS(30) antes que un MAPPT(30), ya que a excepción de "Consumo" y "8 Cilindros", el valor del *efecto* es superior a la unidad. En cuanto a precisión, podemos examinar la siguiente figura 20.

Diseños	MAS	MACR	MSL	MAE caso A	MAE caso B	MAPPT
CONSUMO	2,74	2,75	2,92	1,64	2,05	2,4
ACELERACIÓN	1,94	1,6	1,86	1,53	1,32	2,64
4 CILINDROS	0,35	0,36	0,34	0,27	0,15	0,31
6 CILINDROS	0,27	0,32	0,28	0,24	0,12	0,24
8 CILINDROS	0,32	0,27	0,31	0,26	0,15	0,32
Sumatorio	5,62	5,3	5,71	3,94	3,79	5,91

Figura 20: Amplitud de IC con MAS, MACR, MSL, MAE caso A, caso B y MAPPT.

Como podemos apreciar en la anterior figura, MAPPT es sin duda el tipo de muestreo menos acertado para aplicar a nuestra población de interés, ya que la precisión que consigue luce bastante peor que el resto (casilla roja).

⇒ CASO 6: MUESTREO ALEATORIO CON PROBABILIDADES PROPORCINALES AL TAMAÑO CON REPOSICIÓN (MAPPTCR).

En este diseño muestral, como en MAPPT, se necesita una variable auxiliar cuantitativa que esté relacionada con la de interés. Esta variable es "TIEMPOS". Los diseños MAPPT y MAPPTCR funcionan mejor que los anteriores cuando la variable auxiliar que usamos para definir las probabilidades que utilizamos está suficientemente correlacionada con la de interés.

La función utilizada para seleccionar los elementos de la población que van a pasar a formar parte de la muestra, así como para asignar probabilidades de selección es "*S.PPS*". Sus parámetros de entrada son el tamaño de muestra que queremos (n=30) así como la variable auxiliar. Esta función devuelve una matriz, vemos parte de ésta en la figura 21.

```
> Muestra_PROBA_REEM <- S.PPS (n,Coches$TIEMPOS)
> Muestra_PROBA_REEM
      sam
[1,] 135 0.002726653
[2,]  63 0.002850592
[3,] 336 0.002292867
[4,] 177 0.002664684
[5,] 233 0.002540745
[6,] 325 0.002354837
```

Figura 21: Código y salida función S.PPS

Para conseguir la media y proporción de las variables de interés, así como los intervalos de confianza, necesitamos los datos que proporciona la función "E.PPS", es decir, la estimación del total de la población, el error de estimación, etc. Dicha función exige dos parámetros de entrada, el primero son las probabilidades que proporciona la segunda columna de la matriz "Muestra_PROBA_REEM" que vemos en la figura 21, para guardarlas en un vector utilizamos la siguiente línea de código:

probabilidades_reem <- Muestra_PROBA_REEM[,2]

El segundo parámetro de entrada son los datos correspondientes a las variables de interés de las observaciones que han sido seleccionadas como muestra de la población (*datos_PROBA_REEM).*

datos_PROBA_REEM <- var_interes[Muestra_PROBA_REEM,]
Tabla_PROBA_REEM <-E.PPS(datos_PROBA_REEM, probabilidades_reem)

Gracias a los datos de esta "Tabla_PROBA_REEM" conseguimos calcular la media o proporción y los intervalos de confianza que se ven en la siguiente tabla 20.

	Media o proporción	IC límite inferior	IC límite superior	Efecto
CONSUMO	10,24	9,05	11,44	0,74
ACELERACIÓN	16,61	15,6	17,63	1,87
4 CILINDROS	0,68	0,5	0,85	1,14
6 CILINDROS	0,16	0,03	0,29	1,01
8 CILINDROS	0,16	0,03	0,29	1,01

Tabla 20: Tabla resumen MAPPTCR.

De nuevo, gracias a los valores correspondientes a la columna de *efecto* que vemos en la tabla 20, podemos decir que realizar un MAS es preferible a realizar un MAPPTCR en cuanto a error de muestreo se refiere. Pero si nos fijamos en la precisión, MAPPTCR consigue mejor precisión en sus estimaciones que muchos de los muestreos anteriores, consiguiendo el tercer mejor lugar en la lista de

opciones de tipos de muestreo idóneos para nuestra población de interés. Quizás esto sea consecuencia de la fuerte correlación de la variable "TIEMPO" con las variables de interés. Como vemos en la siguiente figura 22, el orden de mejor a peor (de izquierda a derecha) en cuanto a precisión es:

Diseños	MAE caso B	MAE caso A	MAPPTCR	MACR	MAS	MSL	MAPPT
CONSUMO	2,05	1,64	2,39	2,75	2,74	2,92	2,4
ACELERACIÓN	1,32	1,53	2,03	1,6	1,94	1,86	2,64
4 CILINDROS	0,15	0,27	0,35	0,36	0,35	0,34	0,31
6 CILINDROS	0,12	0,24	0,26	0,32	0,27	0,28	0,24
8 CILINDROS	0,15	0,26	0,26	0,27	0,32	0,31	0,32
Sumatorio	3,79	3,94	***5,29***	5,3	5,62	5,71	5,91

Figura 22: Amplitud de IC con MAS, MACR, MSL, MAE caso A, caso B y MAPPT, MAPPTCR.

Si examinamos las columnas 3 y 7 de datos de la figura anterior, vemos que MAPPTCR es bastante mejor que MAPPT en cuanto al total de precisión alcanzada. Que el diseño con reposición de valores sea mejor que el diseño sin reposición sí es matemáticamente posible por la selección distinta de muestras. En la práctica, para evitar esta paradoja, se han diseñado nuevos diseños aleatorios con probabilidades de inclusión proporcionales al tamaño, como por ejemplo el diseño de Midzuno corregido. Estos diseños no son tenidos en cuenta en el paquete de R que se está utilizando, por lo que no los hemos considerado.

Como comentario general a la figura 22 podría decirse que sería conveniente elegir un MAPPTCR en el caso en el que no se pudiera realizar un muestreo estratificado, pues éste sigue siendo la mejor opción con bastante ventaja respecto al resto de muestreos.

3.2 Conclusiones.

Si revisamos los casos que se han expuesto en el apartado 3.1 de este documento, donde se han realizado diferentes tipos de diseños muestrales asumiendo que en la base de datos "COCHES" teníamos todos los elementos de una población, podemos llegar a la conclusión de que si a una misma población se le aplican diferentes diseños muestrales con un mismo tamaño muestral, la precisión de los estimadores varía en función si el tipo de diseños muestral es más o menos adecuado a la población. Motivo por el cual es de vital importancia elegir el tipo de muestro habiendo estudiado y conociendo las particularidades de cada uno de ellos.

De forma gráfica podemos apreciar esta diferencia en precisión si nos fijamos en la siguiente figura 23, un gráfico de columnas donde aparece una columna por tipo de muestreo, la altura de ésta depende del sumatorio de las diferencias entre los límites superior e inferior de los intervalos que el muestreo ha conseguido, por lo que, cuanto menor sea este valor, mejor será la precisión del muestreo, pues menor es la amplitud de los intervalos de confianza.

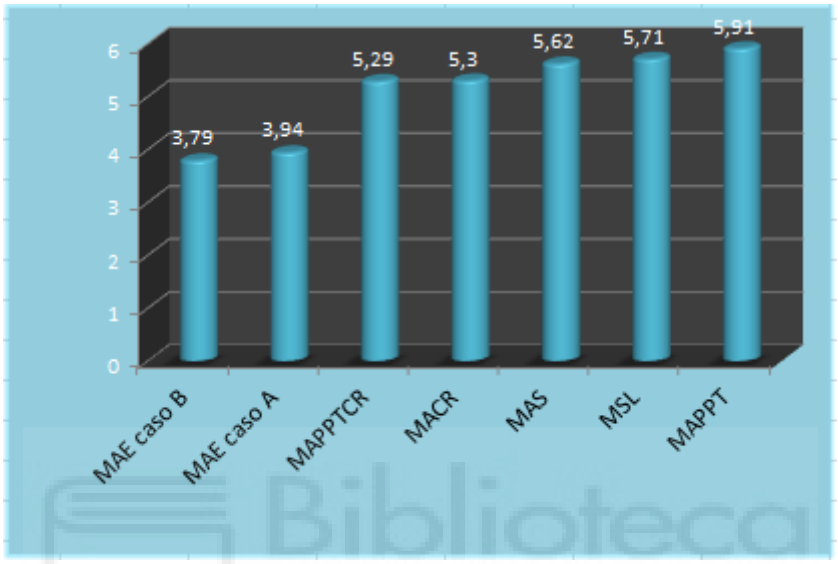

Figura 23: Gráfico de columnas con la precisión de cada tipo de muestreo.

Como podemos apreciar, la diferencia en precisión es bastante alta, concretamente hay una diferencia de hasta 2,12 valores. Siendo esto lo que se quería demostrar, podemos dar por finalizado este estudio.

Puede ser interesante comentar que los resultados también cambian, ya que si comparamos MAE caso B con MAPPT, somos conscientes de la diferencia en cuanto a resultados en medias pues son 2,75 litros de consumo más si se aplica MAE que si se aplica MAPTT. Adicionalmente, consta una diferencia de 3,43 segundos si hablamos en términos de aceleración de 0 a 100km/hora. Por otra parte según MAE caso B, la mayoría de coches tienen de 8 cilindros y la minoría tienen 6 cilindros, además la mayoría goza de una ventaja del 59%, en cambio, MAPPT indica que la mayoría de coches tienen 4 cilindros, en lugar de los 8 que indicaba MAE.

4. CONCLUSIONES.

Durante la realización de este trabajo he tenido claro que quería transmitir dos ideas. En primer lugar quería enjuciar el hecho de que una persona dedicada al análisis de datos, dedicada a la estadística, deba llevar a cabo un análisis sin conocer qué tipo de muestreo se ha utilizado para recoger los datos que va a estudiar. Para alegar la importancia que tiene el desconocimiento del diseño muestral utilizado he llevado a cabo, en el apartado 2, un análisis de una misma muestra bajo diferentes hipótesis. En cada hipótesis he realizado un tipo de muestreo consiguiendo resultados que, al compararlos, revelan los significativos cambios en los que me baso para recalcar dicha importancia. Estos cambios hacen que las conclusiones sean menos acertadas de lo que podrían ser, pues, como se puede leer en las conclusiones específicas del apartado "Errores no muestrales", no debería poder aceptarse que se llegue a la conclusión de que el consumo medio es de 11,24 L/100km cuando en realidad el consumo medio es de 9,06 L/100km, o viceversa. Estamos ante una diferencia notable para el cliente.

En segundo lugar, quería hacer énfasis y resaltar lo positivo de estudiar este aspecto básico y primer paso en todo análisis que es el muestreo, lo positivo que es dedicarle tiempo a estudiar qué tipos de diseños muestrales existen y cuáles son sus características. De forma práctica he subrayado que según el muestreo que apliquemos podemos conseguir resultados muy preciosos o todo lo contrario. Los diferentes casos que se muestran en el apartado 3 de este documento son diseños muestrales, del mismo tamaño (n), que se aplican a una misma población. De esta forma se aprecia que, aunque todos los resultados son acertados, la precisión de éstos cambia visiblemente. Los resultados de un análisis deben de ser precisos, deben de ser concretos y con un margen limitado, pues esa es la magia de la estadística. Por lo que cuanto menor sea la amplitud del intervalo de confianza, herramienta que he usado para hacer el estudio, mayor será la satisfacción de conseguir ciertos resultados. Podemos apreciar los cambios en cuánto a precisión conseguida por los diseños muestrales en las conclusiones específicas del apartado 3 "Errores muestrales".

5. REFERENCIAS

(1) Métodos cuantitativos para la toma de decisiones. Muestreo. – Mercedes Landete.

(2) Muestreo en poblaciones finitas – Domingo Morales González.

(3) CEU, Universidad Cardenal Herrera.
https://www.uv.es/~mamtnez/IECRC.pdf

(4) Monografías.
http://www.monografias.com/trabajos39/muestreo-estadistico/muestreo-estadistico.shtml

(5) Dialnet – Universidad de la Rioja.
https://dialnet.unirioja.es/servlet/articulo?codigo=4770371