ANÁLISIS ENCUESTA DE CONDICIONES DE VIDA DE MADRID

Guillermo Palomo y Daniel Aramburu

8/10/2021

Table of Contents

# 1. Introducción

Nuestro grupo se centrará en el estudio de la población de la Comunidad Autónoma de Madrid. En las diferentes encuestas realizadas a dicha población, se han extraído diferentes variables que proporcionan información útil acerca de las condiciones de vida durante un trimestre para el estudio que queremos realizar.

Este estudio se centra en la selección de muestras para la estimación de parámetros de la población a estudiar, en este caso Madrid, y tiene como objetivo que dicha estimación sea similar a la de los parámetros de la población total. Por otro lado, el segundo objetivo es la estimación correcta de los parámetros de una población a partir de una muestra.

La motivación que nos lleva a realizar este estudio es verificar la importancia que tiene una estimación adecuada de los parámetros cuando seleccionamos diferentes muestras dentro de la misma población(Madrid), así como la diferencia frente a la estimación de los mismos parámetros cuando partimos de que los datos son de la población total y no de una muestra.

En cuanto al conjunto de datos que se va a analizar en las encuestas están presentes la nacionalidad del encuestado, su edad y género, sus ingresos y las horas de trabajo realizadas, etc. Todo ello proporciona información acerca de las condiciones de vida de la población madrileña.

# 2. Descripción de variables

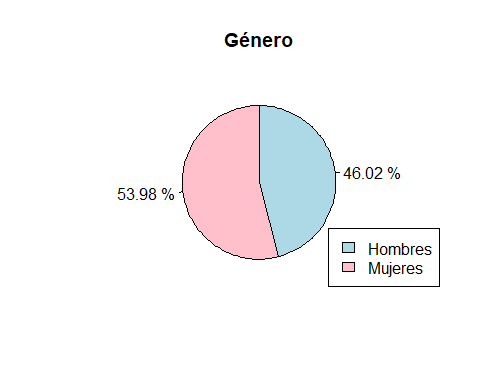
A lo largo de la encuesta podemos observar datos de 7 variables diferentes que hemos estudiado, mencionadas a continuación. La provincia (prov), variable cualitativa, nominal y discreta, en nuestro caso, la Comunidad Autónoma de Madrid. El género (gen) es una variable cualitativa, nominal y discreta y puede ser hombre (1) o mujer (2). La edad, variable cualitativa, ordinal y discreta, presenta datos distribuidos en cuatro intervalos, el primero de 16 a 24 años (1), el segundo de 25 a 49 años (2), el tercero de 50 a 64 años (3) y el último de mayores de 65 años (4). En cuanto a la nacionalidad (nac), se puede obtener nacionalidad española (1) u otras (2), siendo esta una variable cualitativa, nominal y discreta. El nivel de educación (neduc), es una variable cualitativa, ordinal y discreta, y se distingue en: primaria (1), secundaria (2) y postsecundaria (3). La situación laboral de los encuestados (sitlab), variable cualitativa, nominal y discreta, se centra en si la persona encuestada trabaja (1), es desempleado (2) o inactivo (3). Los ingresos normalizados (ingnorm) es una variable cuantitativa, nominal y continua. Las horas de trabajo a la semana (horas) es una variable cuantitativa, nominal y continua. Y el peso muestral (factorel) es una variable cualitativa, nominal y continua.

## Provincia

En nuestro caso solo estudiamos las condiciones de vida de la Comunidad Autónoma de Madrid, por lo que la variable no tiene un uso muy significativo.

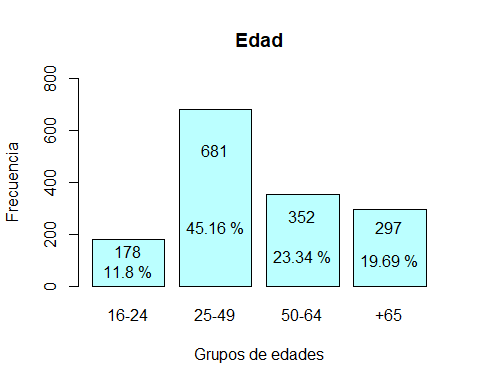
## Género

En cuanto a la variable género, podemos observar que el porcentaje de hombres y mujeres no es muy diferente, las mujeres forman un 53.98% de la población de Madrid mientras que los hombres algo menos, un 46.02% del total.



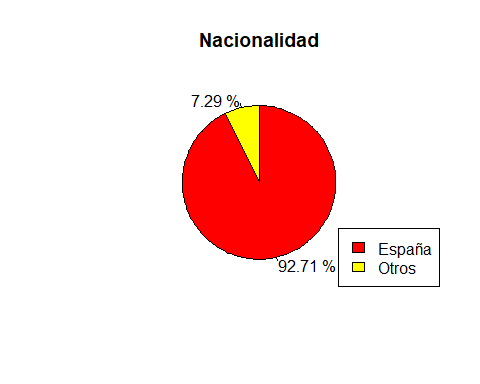
## Edad

Como ya hemos mencionado anteriormente, hemos dividido las edades en cuatro intervalos. La mayor concentración de personas encuestadas se encuentra entre 25 y 49 años, y corresponde con un 45.16%. También podemos observar que el segundo grupo más encuestado es el de 50 y 64 años, con casi la mitad, un 23.34%, seguido el intervalo de +65 años, con una cantidad algo menor, un 19.69%. Por último, tenemos el grupo de 16 a 24 años, con un 11.8% tan solo.



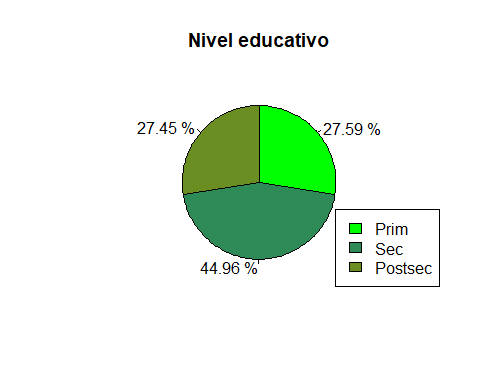
## Nacionalidad

En este gráfico podemos observar que la gran mayoría de personas encuestadas son españolas, y lo forman el 92.71%, siendo por tanto el 7.29% restante de personas de una nacionalidad diferente.



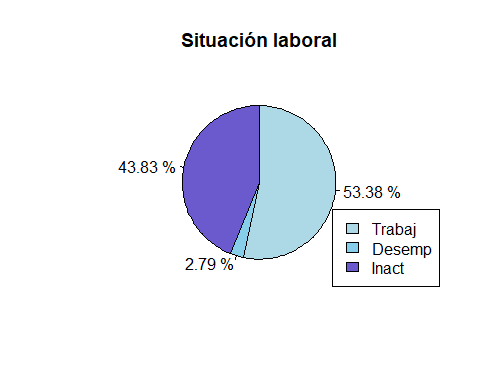
## Nivel de educación

En primer lugar, el mayor porcentaje del gráfico pertenece a personas graduadas en secundaria, un 44.96%. Por otro lado, hay que destacar que tanto el porcentaje de personas graduadas en primaria como en postsecundaria es prácticamente igual, un 27.45% y 27.59% respectivamente.



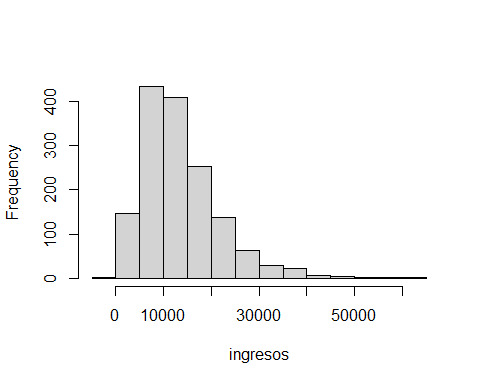
## Situación laboral

La mayoría de personas encuestadas, un 53.38% tienen un trabajo, y por tanto son trabajadores. Por otro lado, un 43.83% del total son inactivos, y por tanto, tan solo el 2.79% restante son personas desempleadas.



## Ingresos normalizados

La mayor concentración de ingresos se encuentra entre 0 y 25000, como se ve en el histograma, por lo que hemos decidido hacer intervalos de 6000 en 6000 para que no tengan frecuencias muy diferentes.

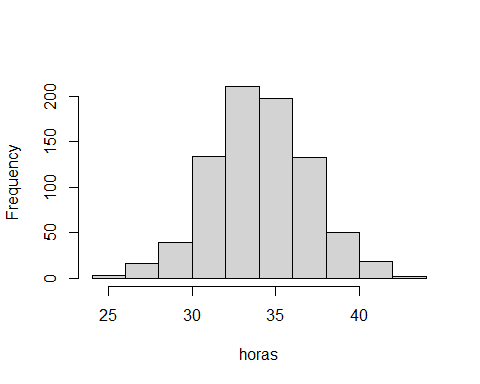


Hemos decidido hacer 5 intervalos de 6000 en 6000.

|  |  |
| --- | --- |
| intervalos\_salarios | Freq |
| -6000 | 215 |
| 6000-12000 | 563 |
| 12000-18000 | 384 |
| 18000-24000 | 200 |
| +24000 | 146 |

## Horas de trabajo a la semana

En cuanto a las horas de trabajo, estamos tratando con aquellas personas encuestadas que tienen un trabajo, es decir, con el 53.38% de las personas. Hay que destacar que la mayor concentración está localizada entre 30 y 38 horas aproximadamente, siendo la media de 34.08 horas. Además, se puede observar cómo dicho gráfico sigue una distribución normal a simple vista.



# 3. Ingresos normalizados según nivel educativo

Vemos como la gran mayoría cobra entre 6000-18000 euros.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Postsec | Prim | Sec |
| -6000 | 13 | 100 | 102 |
| +24000 | 98 | 12 | 36 |
| 12000-18000 | 116 | 83 | 185 |
| 18000-24000 | 85 | 30 | 85 |
| 6000-12000 | 102 | 191 | 270 |

# 4. Medidas

## 4.1 Medidas de la variable ingresos

### Centralidad

Existen tres medidas de centralidad, la media, mediana y moda, las cuales vamos a estudiar. La media tiene un valor de 13518.77, y la mediana de 11692.3. Con estos datos podemos concluir que al saber que la mediana es el valor central, y la media tiene un valor superior en este caso, existen numerosos datos atípicos mayores que la mediana, y que hacen que la media sea mayor. Por último, hay que mencionar que para la variable ingresos no se puede obtener un valor de la moda, al ser esta cuantitativa, nominal y continua.

### Variabilidad

En cuanto a las medidas de variabilidad tenemos la varianza, desviación típica y rango intercuartílico.

En este caso, la varianza tiene un valor de 65216871.65, la desviación típica, al ser la raíz de la varianza, tiene un valor de 8075.7, y el valor del rango intercuartílico es de 9670.93. El valor de la varianza es un indicador de la dispersión de los datos frente a la media, mientras que el valor de la desviación indica la dispersión de los datos entre sí. Para finalizar, el valor del rango intercuartílico nos indica la diferencia entre el tercer cuartil y el primero.

### Forma

Por otro lado, tenemos las medidas de forma, que nos indican si una distribución de frecuencias es simétrica o asimétrica; cómo están concentrados los datos; y la curtosis o medida de apuntamiento. Todo ello permite que la clasifiquemos en un tipo de distribución concreta.

Tenemos por tanto un coeficiente de asimetría de 1.377, y un coeficiente de curtosis de 6.04. El coeficiente de asimetría nos indica que es asimétrica por la derecha, al ser un resultado positivo, y el de curtosis indica que es una distribución leptocúrtica (>3), es decir, más puntiaguda que una distribución normal.

## 4.2 Medidas de la variable ingresos según el nivel educativo

### Centralidad

Las medidas de centralidad para los ingresos en función del nivel de estudios son las siguientes: para primaria, la media es 10384.97 y la mediana 9494.6; para secundaria, la media es 12399.98, y la mediana 11262.6; y, por último, para postsecundaria, la media es 18499.91 y la mediana 16762.25.

Por ello podemos afirmar que, al incrementar nivel de educación, aumentan también los ingresos esperados.

### Variabilidad

En primer lugar, la varianza, la desviación y el rango intercuartílico de ingresos para primaria son 37203096.69, 6099.43 y 7280 respectivamente. En segundo lugar, la varianza, la desviación y el rango intercuartílico de ingresos para secundaria son 45495804.54, 6745.06 y 8119.95 respectivamente. Por último, en tercer lugar, los datos para estas tres medidas de variabilidad para postsecundaria son 89190721.28, 9444.08 y 11936.45 respectivamente.

### Forma

En cuanto a las medidas de forma para el nivel de educación primaria, el coeficiente de asimetría es 1.246, por lo que es asimétrica por la derecha; y el coeficiente de curtosis es 5.3 por lo que es leptocúrtica.

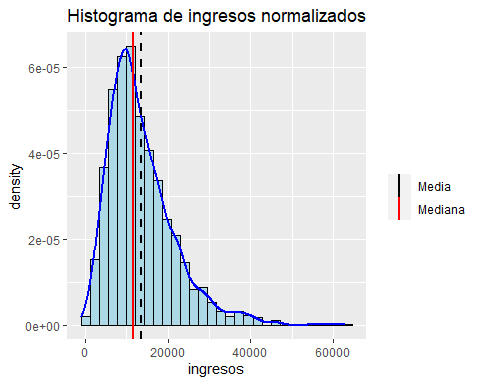
Para secundaria, el coeficiente de asimetría es 1.106, por lo que es asimétrica por la derecha; y el coeficiente de curtosis es 4.99 por lo que también es leptocúrtica.

Por último, para postsecundaria, el coeficiente de asimetría es 1.17, por lo que es asimétrica por la derecha; y el coeficiente de curtosis es 4.94, y como las dos anteriores, también es leptocúrtica.

# 5. Gráficos

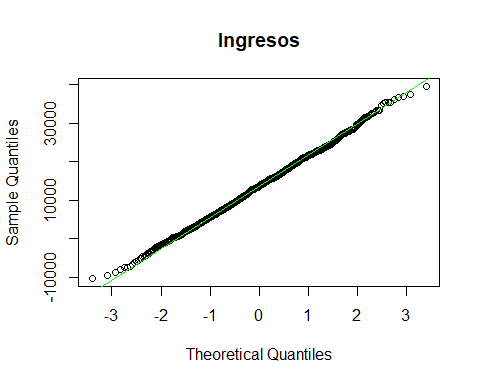
## 5.1 Gráficos de ingresos

### Histograma



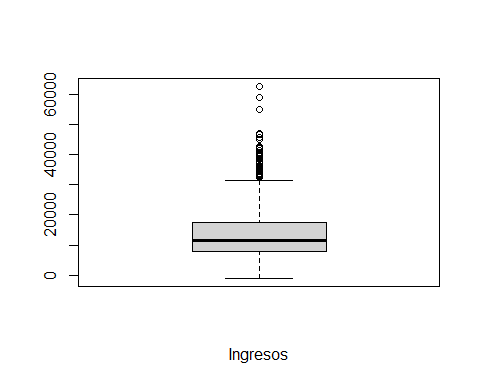
### Diagrama de probabilidad normal

Vemos como los datos se adecúan mayormente a una distribución normal excepto los extremos. Para ajustarlo más se podrían eliminar atípicos.



### Diagrama de caja

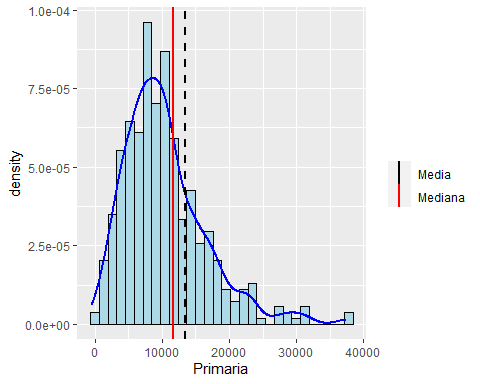
Vemos como hay algunos atípicos que empujan la media hacia arriba.

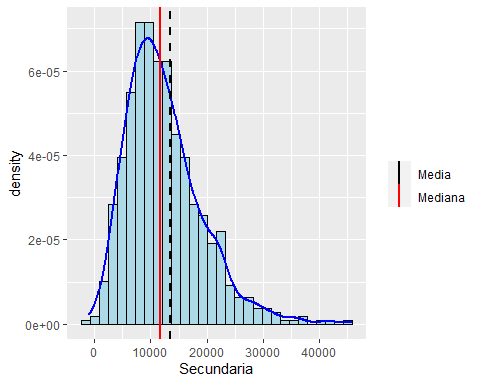


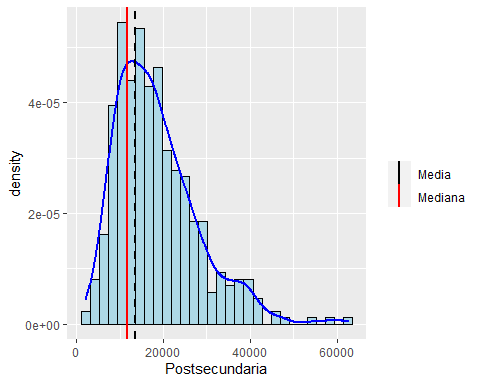
## 5.2 Gráficos de ingresos según el nivel educativo

Todos estos gráficos concuerdan con las explicaciones dadas sobre las diferentes medidas y sobre los gráficos anteriores.

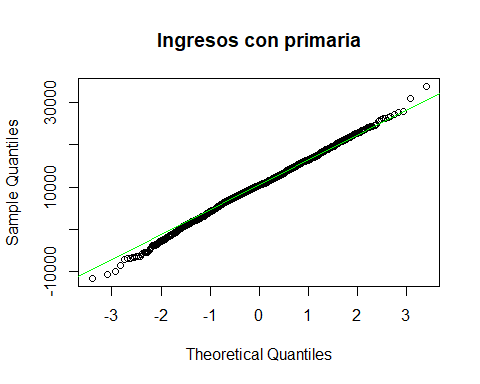
### Histogramas

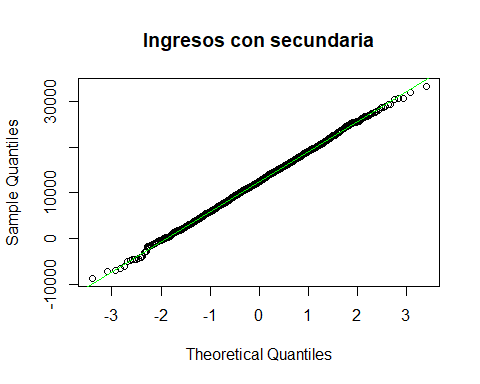


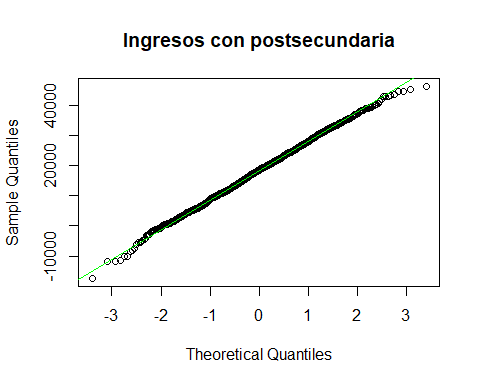




### Diagramas de probabilidad normal

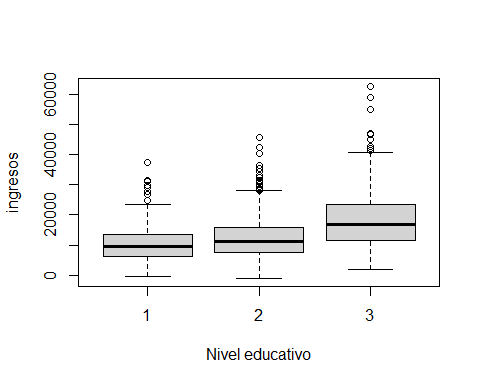






### Diagrama de caja

Como ya hemos observado anteriormente, pero que con este gráfico se ve notablemente, los ingresos crecen proporcionalmente al nivel educativo.



# 6. Tablas de contingencia

Las tablas de contingencia nos permiten llevar a cabo el análisis y estudio de la relación entre dos o más variables.

## Edad - Situación laboral

En cuanto a la relación entre la edad y la situación laboral cabe destacar la gran proporción de población que trabaja entre los 25 y 49 años, así como la creciente tasa de inactividad a medida que pasan los años hasta la jubilación.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Desemp | Inact | Trabaja |
| +65 | 0.0000 | 0.1950 | 0.0020 |
| 15-64 | 0.0046 | 0.1034 | 0.1253 |
| 16-24 | 0.0099 | 0.0623 | 0.0458 |
| 25-49 | 0.0133 | 0.0776 | 0.3607 |

## Género - Situación laboral

También, en cuanto a la relación del género y la situación laboral, podemos observar que la proporción de personas que trabajan y los desempleados es bastante similar en ambos géneros. Por otro lado, es sorprendente la gran diferencia de proporción en cuanto a las personas inactivas, siendo bastante superior el género femenino.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Desemp | Inact | Trabaja |
| Fem | 0.0192 | 0.2798 | 0.2407 |
| Masc | 0.0086 | 0.1585 | 0.2931 |

## Nacionalidad - nivel educativo

Por último, tenemos la relación entre la nacionalidad y el nivel educativo. En la que se ve como la mayoría de la población española y extranjera tienen un nivel de estudios de secundaria.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Postsec | Prim | Sec |
| Españoles | 0.2527 | 0.2533 | 0.4211 |
| Otros | 0.0219 | 0.0225 | 0.0285 |

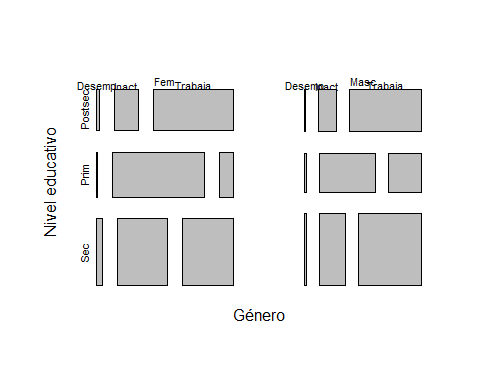
# 7. Gráfico de mosaico

Los gráficos de mosaico nos permiten el estudio y análisis de dos o más variables cualitativas. Estos gráficos representan mediante cuadrados o rectángulos el número de personas(en nuestro caso) que corresponden a un determinado emparejamiento entre dos variables. Dicha cuadrícula aumenta su tamaño cuando el número de personas sea mayor y lo reduce cuando sea menor.

## Género - Nivel educativo - Situación laboral

Nosotros hemos querido estudiar la relación entre el género, nivel educativo, y la situación laboral.

En este estudio, cabe destacar la gran superioridad de mujeres inactivas con primaria respecto a los hombres que estudian lo mismo. Por otro lado, también queremos destacar la superioridad tanto de mujeres inactivas como desempleadas que tienen secundaria respecto a los hombres.



# 8. Estimación de ingresos mediante diseño muestral

## 8.1 Estimación de ingresos completa

### Estimador Horvitz-Thompson

est <- (1/N)\*(sum(ingresos\*(factel)))  
est

## [1] 13571.36

media\_ing

## [1] 13518.77

Este estimador nos da una estimación del ingreso medio anual de Madrid. Se estima que es de 13571.36€, lo que se acerca bastante a los 13518.77€ que nos da la media. Mensualmente, el salario medio sería de 1166.30€.

### Error de muestreo

estvar\_ing <- sum(((1-(1/factel))/((1/factel)^2))\*(ingresos/N)^2)  
estdesv\_ing <- sqrt(estvar\_ing)  
estdesv\_ing

## [1] 505.1769

Para sacar los pesos hemos invertido la variable factor de elevación, para luego sacar la estimación del error de muestreo a través de la desviación típica con el estimador Horvitz-Thompson. El cálculo da 505.18€, siendo esta una cantidad no demasiado alta teniendo en cuenta la estimación de ingresos medios y la percepción que teníamos previa al trabajo siendo habitantes de la comunidad.

### Coeficiente de variación

estcoef <- estdesv\_ing/est  
estcoef

## [1] 0.03722374

EL coeficiente de variación es una medida adimensional que nos proporciona una dispersión del 3.72% reafirmando el apartado anterior.

### Intervalo de confianza al 95% con la desigualdad de Tchebychev

D <- estdesv\_ing/sqrt(1-0.95)  
IntTcheb <- c(est - D, est + D)  
IntTcheb

## [1] 11312.14 15830.58

Este intervalo nos señala que el 95% de los ingresos de Madrid se encuentran entre los 11312.14€ y los 15830.58€.

### Intervalo de confianza al 95% aproximado a la Normal

t <- qt(0.975, length(ingresos))  
IntAp <- c(est - estdesv\_ing \* t, est + estdesv\_ing \* t)  
IntAp

## [1] 12580.44 14562.29

Como se puede apreciar este intervalo es más preciso que el de Tchebychev puesto que dice que un 95% de los ingresos anuales se encuentran entre los 12580.44€ y los 14562.29€.

## 8.2 Estimación de ingresos por género

### Estimadores Horvitz-Thompson

estH <- 1/NH\*(sum(ingresos[genero == 1]\*(factel[genero == 1])))  
estM <- 1/NM\*(sum(ingresos[genero == 2]\*(factel[genero == 2])))  
estH

## [1] 14069.52

estM

## [1] 13120.39

Separando las estimaciones por género vemos como los hombres anualmente ganan cerca de 1000 € más que las mujeres en media. Siendo los ingresos de hombres y mujeres de 14069.52€ y 13120.39€.

### Errores de muestreo

estvar\_ingH <- sum(((1-(1/factel[genero == 1]))/(1/factel[genero == 1])^2)\*((1/NH)\*ingresos[genero == 1])^2)  
estvar\_ingM <- sum(((1-(1/factel[genero == 2]))/(1/factel[genero == 2])^2)\*((1/NM)\*ingresos[genero == 2])^2)  
estdesv\_ingH <- sqrt(estvar\_ingH)  
estdesv\_ingM <- sqrt(estvar\_ingM)  
estdesv\_ingH

## [1] 750.8079

estdesv\_ingM

## [1] 681.4926

La diferencia en los errores no es muy grande, siendo mayor el de los hombres. Esto podría signficar que la diferencia en ingresos anuales entre hombres y mujeres podría ser mayor.

### Coeficientes de variación

estcoefH <- estdesv\_ingH/estH  
estcoefM <- estdesv\_ingM/estM  
estcoefH

## [1] 0.05336415

estcoefM

## [1] 0.05194148

En los coeficientes de variación se confirman las diferencias de los errores viendo como el salario de los hombres varía más.

### Intervalos de confianza al 95% con la desigualdad de Tchebychev

DH <- estdesv\_ingH/sqrt(1-0.95)  
DM <- estdesv\_ingM/sqrt(1-0.95)  
IntTchebH <- c(estH - DH, estH + DH)  
IntTchebM <- c(estM - DM, estM + DM)  
IntTchebH

## [1] 10711.80 17427.23

IntTchebM

## [1] 10072.66 16168.12

Se ve como el intervalo de los hombres es mayor que el de las mujeres por lo que los salarios más altos serán mayormente de hombres.

### Intervalos de confianza al 95% aproximado a la Normal

t <- qt(0.975, length(ingresos))  
IntApH <- c(estH - estdesv\_ingH \* t, estH + estdesv\_ingH \* t)  
IntApM <- c(estM - estdesv\_ingM \* t, estM + estdesv\_ingM \* t)  
IntApH

## [1] 12596.78 15542.26

IntApM

## [1] 11783.62 14457.17

Este intervalo es más preciso que el anterior y sigue confirmando las suposiciones anteriores.

### Conclusiones

mean(as.numeric(datos$horas[genero == 1]))

## [1] 21.7868

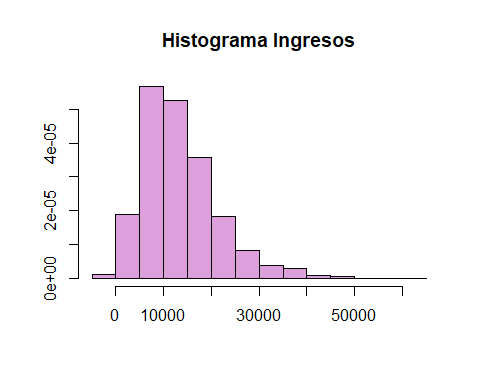
mean(as.numeric(datos$horas[genero == 2]))

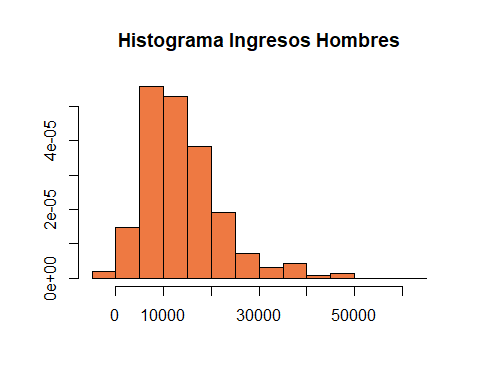
## [1] 15.12748

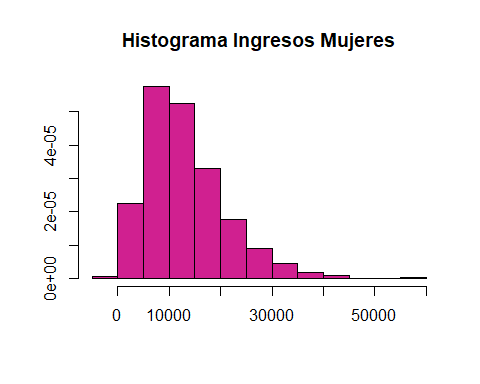
Como hemos visto, los hombres parecen ocupar los puestos con ingresos más altos en la Comunidad de Madrid, pero si observamos la media de horas trabajadas vemos que los hombres trabajan de media alrededor de 6 horas más por semana lo que podría explicar también los salarios más altos y por tanto poner en duda una hipotética brecha salarial.

## 8.3 Histogramas y gráficos de caja

### Histogramas

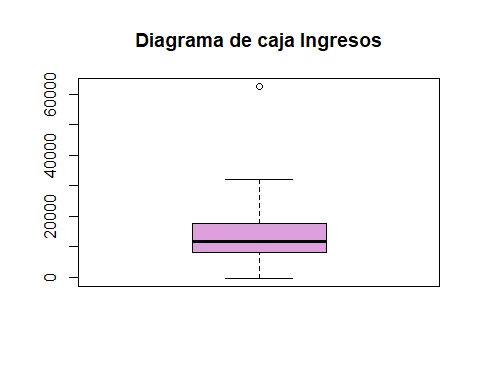


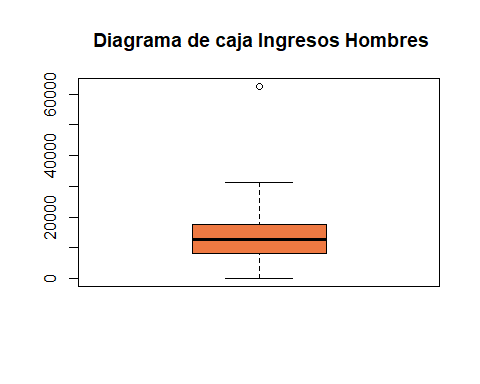


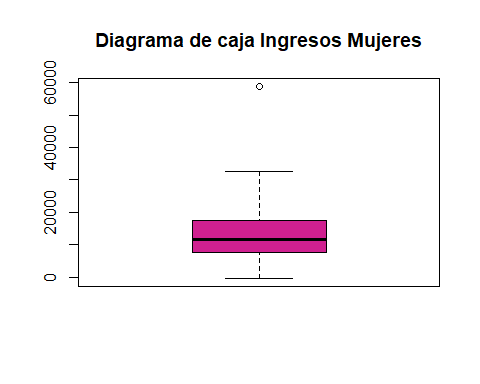


En los histogramas vemos como los hombres tienen más representación en los dos extremos haciendo más grandes los intervalos y más alta la variabilidad, mientras que las mujeres están más concentradas en el medio.

### Gráficos de caja







Se puede apreciar con estos gráficos la existencia de lo que se podría llamar una “clase alta” muy reducida con unos salarios muy por encima de lo común, siendo datos atípicos por completo.

# 9. Estimación de la tasa de individuos en riesgo de pobreza

Hemos creado en nuestros datos una nueva columna representando el riesgo de pobreza para continuar nuestro estudio.

## 9.1 Estimación de la tasa de individuos en riesgo de pobreza completa

### Estimador Horvitz-Thompson

estRP <- (1/N)\*sum(riesgo\_pobr\*factel)  
estRP

## [1] 0.1658606

Se estima que alrededor del 16.58% de la poblacion de Madrid se encuentra en riesgo de pobreza.

### Error de muestreo

var\_RP <- sum(((1-(1/factel))/(1/factel)^2) \* ((1/N) \* riesgo\_pobr)^2)  
desv\_RP <- sqrt(var\_RP)  
desv\_RP

## [1] 0.01293631

La desviación de los datos es del 1.29%.

### Coeficiente de variación

estRPcoef <- desv\_RP/estRP  
estRPcoef

## [1] 0.07799508

El coeficiente de variación para la proporción de individuos es del 7.79%. Podríamos decir que es algo alto teniendo en cuenta el valor de nuestro estimador.

### Intervalo de confianza al 95% con la desigualdad de Tchebychev

DRP <- desv\_RP/sqrt(0.05)  
IntTchebRP <- c(estRP - DRP, estRP + DRP)  
IntTchebRP

## [1] 0.1080076 0.2237135

Según el intervalo, hay entre un 10.8% y un 22.4% de personas por debajo del umbral de pobreza en Madrid, lo cual es un nivel alarmante.

### Intervalo de confianza al 95% aproximado a la Normal

IntApRP <- c(estRP - desv\_RP\*t, estRP + desv\_RP\*t)  
IntApRP

## [1] 0.1404855 0.1912356

Según el intervalo aproximado a la normal(más preciso), el número de gente en riesgo de pobreza está entre el 14% y el 19.1%.

### Estimación del total de individuos en riesgo de pobreza

TotalRP <- N\*estRP  
TotalRP

## [1] 784896.1

La estimación del total de individuos en riesgo de pobreza en Madrid es de 784896 personas.

### Error de muestreo del total

varTotalRP <- sum(((1-(1/factel))/(1/factel)^2)\*(riesgo\_pobr^2))  
desv\_TotalRP <- sqrt(varTotalRP)  
desv\_TotalRP

## [1] 61218.03

El error de muestreo es de 61218 personas, el cual es moderadamente alto teniendo en cuenta el total de personas en riesgo de pobreza.

### Coeficiente de variación del total

estcoefTotalRP <- desv\_TotalRP/TotalRP  
estcoefTotalRP

## [1] 0.07799508

El coeficiente de variación del total es el mismo que el de la proporción como cabría esperar.

## 9.2 Estimación de la tasa de individuos en riesgo de pobreza por género

### Estimadores Horvitz-Thompson

estRPH <- (1/NH)\*sum(riesgo\_pobr[genero == 1]\*factel[genero == 1])  
estRPM <- (1/NM)\*sum(riesgo\_pobr[genero == 2]\*factel[genero == 2])  
estRPH

## [1] 0.1392226

estRPM

## [1] 0.1899756

El estimador para hombres es del 13.92% mientras que para mujeres es del 18.99%, siendo este alrededro de un 6% mayor, lo que es una distancia bastante significativa.

### Errores de muestreo

var\_RPH <- sum(((1-(1/factel[genero == 1]))/(1/factel[genero == 1])^2) \* ((1/NH) \* riesgo\_pobr[genero == 1])^2)  
var\_RPM <- sum(((1-(1/factel[genero == 2]))/(1/factel[genero == 2])^2) \* ((1/NM) \* riesgo\_pobr[genero == 2])^2)  
desv\_RPH <- sqrt(var\_RPH)  
desv\_RPM <- sqrt(var\_RPM)  
desv\_RPH

## [1] 0.01885248

desv\_RPM

## [1] 0.0177824

Los errores para hombres y mujeres son del 1.88% y 1.77% respectivamente. Igualmente viendo el estimador anterior, podemos concluir que los hombres tienen mucho menor riesgo de caer en la pobreza.

### Coeficientes de variación

estRPcoefH <- desv\_RPH/estRPH  
estRPcoefM <- desv\_RPM/estRPM  
estRPcoefH

## [1] 0.1354125

estRPcoefM

## [1] 0.09360363

Confirmando los errores anteriores las diferencias en coeficientes de variación son más notables aún, difiriendo entre ellos en un 4% siendo superior la variabilidad en hombres.

### Intervalos de confianza al 95% con la desigualdad de Tchebychev

DRPH <- desv\_RPH/sqrt(0.05)  
DRPM <- desv\_RPM/sqrt(0.05)  
IntTchebRPH <- c(estRPH - DRPH, estRPH + DRPH)  
IntTchebRPM <- c(estRPM - DRPM, estRPM + DRPM)  
IntTchebRPH

## [1] 0.05491178 0.22353351

IntTchebRPM

## [1] 0.1104503 0.2695009

Aquí se confirma que los hombres tienen muchas menos probabilidades de caer en pobreza como vemos comparando los extremos izquierdos de los intervalos. El intervalo de las mujeres está mucho más a la derecha y es más pequeño que el de los hombres.

### Intervalos de confianza al 95% aproximado a la Normal

IntApRPH <- c(estRPH - desv\_RPH\*t, estRPH + desv\_RPH\*t)  
IntApRPM <- c(estRPM - desv\_RPM\*t, estRPM + desv\_RPM\*t)  
IntApRPH

## [1] 0.1022428 0.1762025

IntApRPM

## [1] 0.1550947 0.2248565

Como en el apartado anterior el intervalo de las mujeres está más desplazado hacia la derecha, teniendo mayor riesgo de entrar en la pobreza, siedo este de entre 15.5% y 22.48%.

### Estimaciones del total de individuos en riesgo de pobreza

TotalRPH <- NH\*estRPH  
TotalRPM <- NM\*estRPM  
TotalRPH

## [1] 313044.1

TotalRPM

## [1] 471852

Como cabía esperar por las proporciones anteriores, hay un mayor número de mujeres bajo el umbral de pobreza con 471852 mujeres y 313044 hombres.

### Errores de muestreo del total

varTotalRPH <- sum(((1-(1/factel[genero == 1]))/(1/factel[genero == 1])^2)\*(riesgo\_pobr[genero == 1]^2))  
varTotalRPM <- sum(((1-(1/factel[genero == 2]))/(1/factel[genero == 2])^2)\*(riesgo\_pobr[genero == 2]^2))  
desv\_TotalRPH <- sqrt(varTotalRPH)  
desv\_TotalRPM <- sqrt(varTotalRPM)  
desv\_TotalRPH

## [1] 42390.08

desv\_TotalRPM

## [1] 44167.06

Como en la proporción, los errores son muy proximos teniendo en cuenta los totales. El error en los hombres podría parecer menor que el de las mujeres pero al tener también menos hombres en total la estimación parece coincidir con la de la proporción.

### Coeficientes de variación del total

estcoefTotalRPH <- desv\_TotalRPH/TotalRPH  
estcoefTotalRPM <- desv\_TotalRPM/TotalRPM  
estcoefTotalRPH

## [1] 0.1354125

estcoefTotalRPM

## [1] 0.09360363

Los coeficientes siguen siendo iguales que los de la proporción.

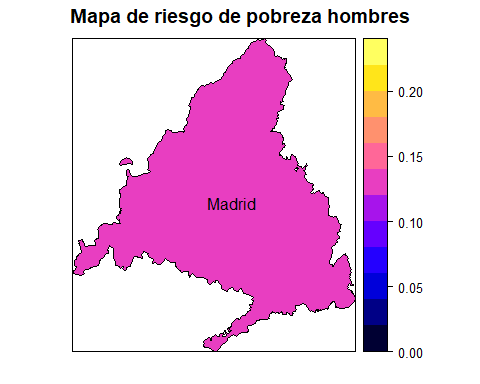
### Conclusiones

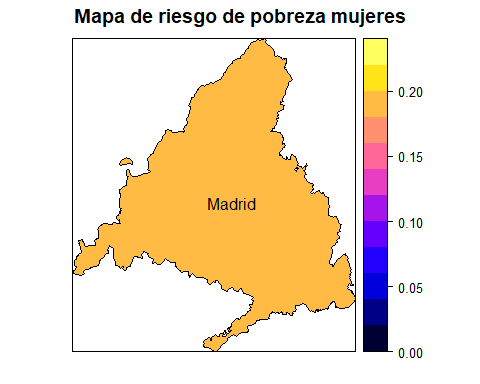
A la luz de los intervalos de la proporción y los totales al estar muy claramente más a la derecha el intervalo de las mujeres podemos confirmar la existencia de una brecha de género en los salarios más bajos poniendo a más mujeres bajo el umbral de pobreza.

Como estudiado con los ingresos podría tratarse también de la influencia de las horas trabajadas, haciendo que más mujeres no lleguen a las horas suficientes de trabajo para cobrar lo suficiente para salir del umbral.

### Mapas por género

Como la Comunidad de Madrid solo tiene una provincia vamos a hacer una separación por género para la proporción del riesgo de pobreza.





Como podemos observar, los mapas corresponden en color con las tasas de riesgo de pobreza obtenidas en anteriores apartados, coincidiendo con que las mujeres tienen mayor tasa y, por tanto, color más anaranjado.

# 10. Muestreo estratificado

## 10.1 Tipos de afijación muestral

1. Afijación uniforme: una afijación uniforme se da cuando se asigna el mismo tamaño de la muestra a cada estrato, es decir, el mismo número de unidades, sin importar el peso que dichos estratos tienen en la población. Es por ello que al dar la misma importancia a todos los estratos, va a favorecer a los estratos con un tamaño menor, perjudicando a su vez a aquellos con un tamaño mayor. Es obvio por tanto que este tipo de afijación es recomendable llevarla a cabo cuando los estratos son de tamaños parecidos.
2. Afijación proporcional: se da cuando dividimos una población en varios estratos que son de diferentes tamaños. Hay que tener en cuenta que para que se de una afijación proporcional, la muestra de los estratos tiene que tener la misma proporción que tenía en la población original.
3. Afijación óptima: se lleva a cabo cuando la variabilidad de la variable que estudiamos en un determinado estrato es alta.Cuanto mayor sea dicha variabilidad, mayor será el número de unidades que se deben tomar de dicho estrato. En este caso, el tamaño de cada estrato de nuestra muestra no guarda la proporcionalidad con la población, al contrario que en la afijación proporcional, siempre buscando de esta manera reducir el error basándonos en las cuasivarianzas. Por otro lado, Ch es el coste de muestrear una unidad en el estrato h, el coste de selección de las unidades muestrales en ese estrato sería Ch\*nh. Para obtener entonces el coste total, sumamos los costes anteriores para todos los estratos y a todo ello le añadimos el coste fijo del muestreo.
4. Afijación de Neyman: en este caso, no existen diferencias en los costes entre los diferentes estratos, y por ello se emplea este método. Se lleva a cabo que: C = C0 + nc con c = Ch. El coste total solo depende de n y no de cual sea la afijación, y no se impone ninguna restricción en cuanto a quienes sean los nh. Este es un caso particular de la afijación óptima.

## 10.2 Muestra aleatoria estratificada con una fracción muestral total igual a 1/10

Seleccionamos una muestra aleatoria estratificada con una fracción muestral total igual a 1/10, según la variable que referencia el nivel de estudios “neduc” como estratos, realizando muestreo aleatorio simple dentro de cada estrato. Dicha variable tiene 3 niveles, el nivel 1 representa primaria, el 2, secundaria, y el 3 post-secundaria. Con la función dada en el tema 4 sacamos una tabla con los siguientes datos:

## media varmues conf\_int Nhest ybar yvar  
## Primaria 13518.77 3660.741 121.0081 416 10384.97 37203097  
## Secundaria 13518.77 3660.741 121.0081 678 12399.98 45495805  
## Postsec 13518.77 3660.741 121.0081 414 18499.91 89190721

Vemos como cuanto más crece el nivel de estudios mayores son los ingresos medios estimados. De aquí sacamos el tamaño poblacional de los estratos que usaremos más tarde con 416, 678 y 414.

## 10.3 Muestras por proporcional, Neyman y óptima

En este apartado vamos a sacar los distintos tamaños muestrales de cada estrato y comparar.

### Afijación proporcional

n1 <- length(neduc[neduc == 1])  
n2 <- length(neduc[neduc == 2])  
n3 <- length(neduc[neduc == 3])  
  
t1 <- round(1/10\*n1,0)  
t2 <- round(1/10\*n2,0)  
t3 <- round(1/10\*n3,0)  
  
t1

## [1] 42

t2

## [1] 68

t3

## [1] 41

Para esta, los tamaños son 42 para primaria, 68 para secundaria y 41 para post-secundaria.

### Afijación de Neyman

nneducs <- c(length(neduc[neduc == 1]), length(neduc[neduc == 2]), length(neduc[neduc == 3]))  
cuasvars <- c(var(ingresos[neduc == 1]),var(ingresos[neduc == 2]),var(ingresos[neduc == 3]))  
  
n1ney <- (length(ingresos)/10) \* ((length(neduc[neduc == 1]) \* var(ingresos[neduc == 1]))/(sum(nneducs\*cuasvars)))  
n2ney <- (length(ingresos)/10) \* ((length(neduc[neduc == 2]) \* var(ingresos[neduc == 2]))/(sum(nneducs\*cuasvars)))  
n3ney <- (length(ingresos)/10) \* ((length(neduc[neduc == 3]) \* var(ingresos[neduc == 3]))/(sum(nneducs\*cuasvars)))  
n1ney

## [1] 28.03509

n2ney

## [1] 55.87669

n3ney

## [1] 66.88822

Con Neyman al ser el de mayor varianza el estrato de post-secudaria, su tamaño muestral crece a 67 mientras que el de primaria se queda en 28 y el de secundaria en 56.

### Afijción óptima

nestr <- length(neduc)/10  
Ch <- c(5,10,15)  
  
nopt1 <- nestr \* (((length(neduc[neduc == 1]) \* var(ingresos[neduc == 1]))/sqrt(Ch[1])))/(sum(nneducs \* cuasvars / sqrt(Ch)))  
nopt2 <- nestr \* (((length(neduc[neduc == 2]) \* var(ingresos[neduc == 2]))/sqrt(Ch[2])))/(sum(nneducs \* cuasvars / sqrt(Ch)))  
nopt3 <- nestr \* (((length(neduc[neduc == 3]) \* var(ingresos[neduc == 3]))/sqrt(Ch[3])))/(sum(nneducs \* cuasvars / sqrt(Ch)))  
nopt1

## [1] 39.82235

nopt2

## [1] 56.12295

nopt3

## [1] 54.8547

Para la afijación óptima, hemos puesto unos costes variables de muestreo inventados de 5 para primaria, 10 para secundari y 15 para post-secundaria. Esta vez la variabilidad del estrato de post-secundaria se contrarresta con el alto coste de muestreo, por lo que no es tan grande el tamaño como con Neyman, siendo este de 55. Secundaria tiene un tamaño de 56 y primaria uno de 40.

### Conclusiones

wh <- c(n1/nestr,n2/nestr,n3/nestr)  
varprop <- ((1-(n1/length(neduc)))/nestr) \* sum(cuasvars \* wh)  
varnney <- (1/nestr) \* ((sum(wh\*sqrt(cuasvars)))^2) - (1/length(neduc)) \* sum(wh \* cuasvars)  
varprop

## [1] 2650882

varnney

## [1] 35048979

Nuestros datos son medianamente homogéneos y desconocemos los costes de muestrear cada estrato por lo que no pensamos utilizar la afijación óptima. Entre Neyman y proporcional nos hemos decantado finalmente por la afijación proporcional ya que la varianza con esta es menor.

## 10.4 Estimación de ingresos medios

### Estimación ingresos y error en la muestra estratificada

svymean(~as.numeric(ingnorm), disestrat)

## mean SE  
## as.numeric(ingnorm) 14049 661.42

coefestestrat <- 661.42/14049  
coefestestrat

## [1] 0.04707951

Con la muestra estratificada nos sale una media de ingresos de 14049€ que comparada con la media poblacional de 13518€ es una aproximación decente. La desviación es de 661.42€ y el coeficiente de variación es de 4.7%, el cual no es muy alto tratando esta variable.

### Estimación por estratos

svyby(~as.numeric(ingnorm), ~neduc, disestrat, svymean)

## neduc as.numeric(ingnorm) se  
## 1 1 11146.91 1037.561  
## 2 2 12717.60 1144.165  
## 3 3 19145.94 1098.388

coefestestrat1 <- 1037.56/11146.91  
coefestestrat2 <- 1144.17/12717.6  
coefestestrat3 <-1098.39/19145.94  
coefestestrat1

## [1] 0.0930805

coefestestrat2

## [1] 0.08996745

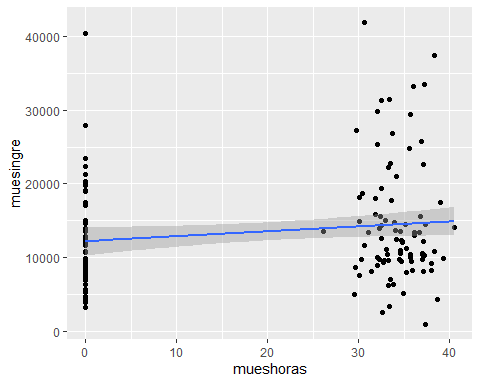
coefestestrat3

## [1] 0.05736934

Como vemos en la muestra por estratos se confirma que la gente que ha cursado post-secundaria llega en su futuro a cobrar bastante más en media. La desviaciones se mantienen bastante similares. Para analizar mejor la variación, miramos los coeficientes, siendo esta vez post-secundaria la más alejada hacia abajo.

# 11 Muestreo de razón y regresión

## 11.1 Estimador de razón y regresión



## Con el estimador de la razón R = 675.3248 sacamos las estimaciones de la media = 23014.58 y del total = 34705980 con varianza = 974589.6

## Con el estimador de regresión sacamos para la media = 13145.5 y para el total = 19823408 con varianza = 334608

El estimador de razón en nuestro caso es algo deficiente, suponemos que por la variable horas, mientras que el de regresión estima bastante bien la media en comparación con el anterior.

## 11.2 Comparación con m.a.e

Con el muestreo aleatorio estratificado los resultados han sido de una media de 14049€ y una desviación típica de 661.42€ mientras que con nuestro estimador de regresión sacamos una media de 13145.5€ y una desviación de 578.45. En nuestro caso sabiendo que la media poblacional es de 13518.77€ el estimador que parece más preciso es el de regresión. Aun así, nos parece que al ser las horas de trabajo datos tan dispares conteniendo muchos “0”, las muestras escogidas podrían variar tanto como para que al final la estratificada acabara siendo mejor estimador.

# 12. Conclusiones

Como podemos observar tras el estudio de las medidas de centralidad para la variable ingresos según el nivel de estudio, al aumentar el nivel de estudios, aumentan los ingresos. Los menores ingresos son en primaria, y los mayores se dan en post-secundaria.

Por otro lado, hay que destacar la diferencia en los salarios entre hombres y mujeres, debido a que los hombres ganan de media 1000€ más al año que las mujeres en la Comunidad de Madrid. Los ingresos son de 14069.52€ para los hombres y 13120.39€ para las mujeres. Sin embargo, hay que tener en cuenta también las horas trabajadas de ambos géneros, en la que podemos observar que los hombres trabajan 6 horas más que las mujeres por semana, lo que podría explicar la diferencia salarial y poner en duda la brecha mencionada.

Otro dato e información impactante es que casi el 17% de la población en la Comunidad de Madrid está en riesgo de pobreza. Por ello, la estimación del total de individuos en riesgo de pobreza es de 784896 personas.

Consultando en la página de INE.es podemos comprobar que desde 2016 la tendencia de pobreza estaba disminuyendo notablemente, pero en los dos ultimos años parece que vuelve a aumentar en un 3%. En el último año registrado (2020),la Comunidad de Madrid tiene un 15.4% de pobreza, por lo que tan solo 5 provincias tienen un riesgo de pobreza menor a nuestra Comunidad. Las provincias con mayor porcentaje son Ceuta y Melilla con un 36% aproximadamente, mientras que las que tienen el menor porcentaje son Navarra y País Vasco con un 10%.

En cuanto al riesgo de pobreza por género, en los hombres es del 13.92%, mientras que para mujeres es del 18.99%, un 6% mayor, lo que nos ha sorprendido.

Nosotros haríamos una extensión del trabajo centrándonos en los ingresos por hora para determinar la brecha salarial en trabajadores actuales.

# Anexo

knitr::opts\_chunk$set(echo = TRUE)  
  
#Preparación de datos  
  
library(moments)  
library(ggplot2)  
library(ggpubr)  
library(survey)  
library(sp)  
library(sampling)  
  
datos <- read.table("datosECV\_ComMadrid.txt")  
names(datos) <- datos[1,]  
datos <- datos[-1,]  
rownames(datos) <- 1:nrow(datos)  
  
genero <- as.numeric(datos$datosECVmas16.gen)  
edad <- as.numeric(datos$edad)  
nac <- as.numeric(datos$nac)  
neduc <- as.numeric(datos$neduc)  
sitlab <- as.numeric(datos$sitlab)  
factel <- as.numeric(datos$factorel)  
  
genero1 <- genero  
neduc1 <- neduc  
sitlab1 <- sitlab  
edad1 <- edad  
nac1 <- nac  
  
genero1[genero1 == 1] <- "Masc"  
genero1[genero1 == 2] <- "Fem"  
neduc1[neduc1 == 1] <- "Prim"  
neduc1[neduc1 == 2] <- "Sec"  
neduc1[neduc1 == 3] <- "Postsec"  
sitlab1[sitlab1 == 1] <- "Trabaja"  
sitlab1[sitlab1 == 2] <- "Desemp"  
sitlab1[sitlab1 == 3] <- "Inact"  
edad1[edad1 == 1] <- "16-24"  
edad1[edad1 == 2] <- "25-49"  
edad1[edad1 == 3] <- "15-64"  
edad1[edad1 == 4] <- "+65"  
nac1[nac1 == 1] <- "Españoles"  
nac1[nac1 == 2] <- "Otros"  
  
N <- sum(factel)  
NH <- sum(factel[genero == 1])  
NM <- sum(factel[genero == 2])  
  
#Gráfico para el género  
  
genero\_grafico <- c(round((length(genero[genero == 1])/length(genero))\*100,2), round((length(genero[genero == 2])/length(genero))\*100,2))  
pie(genero\_grafico, main = "Género", clockwise = T, labels = paste(genero\_grafico,"%"), col = c("lightblue", "pink"))  
legend("bottomright", legend = c("Hombres","Mujeres"), fill = c("lightblue", "pink"))  
  
  
#Gráfico para la edad  
  
Graf\_edad <- barplot(table(edad), col = "paleturquoise1", ylim = c(0,800), main = "Edad", names.arg = c("16-24","25-49","50-64","+65"), ylab = "Frecuencia", xlab = "Grupos de edades")  
text(Graf\_edad, table(edad)/1.3, labels = table(edad))  
text(Graf\_edad, table(edad)/3, labels = paste(round(table(edad)/length(edad)\*100,2),"%"))  
  
#Gráfico para la nacionalidad  
  
nac\_grafico <- c(round((length(nac[nac == 1])/length(nac))\*100,2), round((length(nac[nac == 2])/length(nac))\*100,2))  
pie(nac\_grafico, main = "Nacionalidad", clockwise = T, labels = paste(nac\_grafico,"%"), col = c("red", "yellow"))  
legend("bottomright", legend = c("España","Otros"), fill = c("red", "yellow"))  
  
#Gráfico para el nivel de educación  
  
neduc\_grafico <- c(round((length(neduc[neduc == 1])/length(neduc))\*100,2), round((length(neduc[neduc == 2])/length(neduc))\*100,2), round((length(neduc[neduc == 3])/length(neduc))\*100,2))  
pie(neduc\_grafico, main = "Nivel educativo", clockwise = T, labels = paste(neduc\_grafico,"%"), col = c("green", "seagreen", "olivedrab"))  
legend("bottomright", legend = c("Prim","Sec","Postsec"), fill = c("green", "seagreen", "olivedrab"))  
  
#Gráfico para la situación laboral  
  
sitlab\_grafico <- c(round((length(sitlab[sitlab == 1])/length(sitlab))\*100,2), round((length(sitlab[sitlab == 2])/length(sitlab))\*100,2), round((length(sitlab[sitlab == 3])/length(sitlab))\*100,2))  
pie(sitlab\_grafico, main = "Situación laboral", clockwise = T, labels = paste(sitlab\_grafico,"%"), col = c("lightblue", "skyblue", "slateblue"))  
legend("bottomright", legend = c("Trabaj","Desemp","Inact"), fill = c("lightblue", "skyblue", "slateblue"))  
  
#Gráfico para los ingresos  
  
ingresos <- as.numeric(datos$ingnorm)  
hist(ingresos, main = NULL)  
  
#Tabla ingresos  
  
intervalos\_salarios <- c()  
  
for(i in 1:length(ingresos)){  
   
 if(ingresos[i] <= 6000) {   
 intervalos\_salarios <- c(intervalos\_salarios,"-6000") }  
 if(ingresos[i] <= 12000 & ingresos[i] > 6000) {   
 intervalos\_salarios <- c(intervalos\_salarios,"6000-12000") }  
 if(ingresos[i] <= 18000 & ingresos[i] > 12000) {   
 intervalos\_salarios <- c(intervalos\_salarios,"12000-18000") }  
 if(ingresos[i] <= 24000 & ingresos[i] > 18000) {   
 intervalos\_salarios <- c(intervalos\_salarios,"18000-24000")}  
 if(ingresos[i] > 24000) {   
 intervalos\_salarios <- c(intervalos\_salarios,"+24000")}  
   
}  
intervalos <- as.data.frame(table(intervalos\_salarios))  
int <- c("-6000","6000-12000","12000-18000","18000-24000","+24000")  
intervalos <- intervalos[match(int,intervalos$intervalos\_salarios),]  
rownames(intervalos) <- 1:nrow(intervalos)  
knitr::kable(intervalos)  
  
#Gráfico para las horas de trabajo  
  
horas <- as.numeric(datos$horas[datos$sitlab == 1])  
Media\_horas <- mean(horas)  
hist(horas, main = NULL)  
#Tabla ingresos según nivel  
  
IE <- round(table(intervalos\_salarios,neduc1),4)  
knitr::kable(IE)  
#Medidas  
  
media\_ing <- mean(ingresos)  
mediana\_ing <- median(ingresos)  
  
  
var\_ing <- var(ingresos)  
desv\_ing <- sqrt(var\_ing)  
IQR\_ing <- IQR(ingresos)  
  
  
asimetria\_ing <- skewness(ingresos)  
curtosis\_ing <- kurtosis(ingresos)  
  
#Medidas según nivel  
  
media\_ing1 <- mean(ingresos[neduc == 1])  
mediana\_ing1 <- median(ingresos[neduc == 1])  
  
media\_ing2 <- mean(ingresos[neduc == 2])  
mediana\_ing2 <- median(ingresos[neduc == 2])  
  
media\_ing3 <- mean(ingresos[neduc == 3])  
mediana\_ing3 <- median(ingresos[neduc == 3])  
  
  
var\_ing1 <- var(ingresos[neduc == 1])  
desv\_ing1 <- sqrt(var\_ing1)  
IQR\_ing1 <- IQR(ingresos[neduc == 1])  
  
var\_ing2 <- var(ingresos[neduc == 2])  
desv\_ing2 <- sqrt(var\_ing2)  
IQR\_ing2 <- IQR(ingresos[neduc == 2])  
  
var\_ing3 <- var(ingresos[neduc == 3])  
desv\_ing3 <- sqrt(var\_ing3)  
IQR\_ing3 <- IQR(ingresos[neduc == 3])  
  
  
asimetria\_ing1 <- skewness(ingresos[neduc == 1])  
curtosis\_ing1 <- kurtosis(ingresos[neduc == 1])  
  
asimetria\_ing2 <- skewness(ingresos[neduc == 2])  
curtosis\_ing2 <- kurtosis(ingresos[neduc == 2])  
  
asimetria\_ing3 <- skewness(ingresos[neduc == 3])  
curtosis\_ing3 <- kurtosis(ingresos[neduc == 3])  
  
#Gráficos ingresos  
  
ing\_dat <- as.data.frame(ingresos)  
  
Hist\_ing <- ggplot(data = ing\_dat, aes(x = ingresos)) +   
 geom\_histogram(aes(y = ..density..), fill = "lightblue", color = "black") +   
 geom\_density( size = 1, col = "blue") +  
 ggtitle("Histograma de ingresos normalizados") +   
 geom\_vline(aes(xintercept = mean(ingresos), col = "Media"), linetype = "dashed", size = 1) +  
 geom\_vline(aes(xintercept = median(ingresos), col = "Mediana"), size = 1) +   
 scale\_color\_manual(name = "", values = c("Media" = "black", "Mediana" = "red"))  
  
Hist\_ing  
  
prob\_graf = rnorm(length(ingresos),media\_ing,desv\_ing)  
qqnorm(prob\_graf, main = "Ingresos")   
qqline(prob\_graf, col = "green")  
  
  
  
boxplot(ingresos, xlab = "Ingresos")  
  
#Gráficos ingresos según nivel  
  
ing\_neduc\_dat <- data.frame(ingresos, neduc)  
  
Hist\_ing1 <- ggplot(data = as.data.frame(ing\_neduc\_dat$ingresos[neduc == 1]), aes(x = ingresos[neduc == 1])) +   
 geom\_histogram(aes(y = ..density..), fill = "lightblue", color = "black") +   
 geom\_density( size = 1, col = "blue") +  
 geom\_vline(aes(xintercept = mean(ingresos), col = "Media"), linetype = "dashed", size = 1) +  
 geom\_vline(aes(xintercept = median(ingresos), col = "Mediana"), size = 1) +   
 xlab("Primaria") +  
 scale\_color\_manual(name = "", values = c("Media" = "black", "Mediana" = "red"))  
  
Hist\_ing1  
  
  
Hist\_ing2 <- ggplot(data = as.data.frame(ing\_neduc\_dat$ingresos[neduc == 2]), aes(x = ingresos[neduc == 2])) +   
 geom\_histogram(aes(y = ..density..), fill = "lightblue", color = "black") +   
 geom\_density( size = 1, col = "blue") +  
 geom\_vline(aes(xintercept = mean(ingresos), col = "Media"), linetype = "dashed", size = 1) +  
 geom\_vline(aes(xintercept = median(ingresos), col = "Mediana"), size = 1) +   
 xlab("Secundaria") +  
 scale\_color\_manual(name = "", values = c("Media" = "black", "Mediana" = "red"))  
  
Hist\_ing2  
  
  
Hist\_ing3 <- ggplot(data = as.data.frame(ing\_neduc\_dat$ingresos[neduc == 3]), aes(x = ingresos[neduc == 3])) +   
 geom\_histogram(aes(y = ..density..), fill = "lightblue", color = "black") +   
 geom\_density( size = 1, col = "blue") +  
 geom\_vline(aes(xintercept = mean(ingresos), col = "Media"), linetype = "dashed", size = 1) +  
 geom\_vline(aes(xintercept = median(ingresos), col = "Mediana"), size = 1) +   
 xlab("Postsecundaria") +  
 scale\_color\_manual(name = "", values = c("Media" = "black", "Mediana" = "red"))  
  
Hist\_ing3  
  
  
prob\_graf1 = rnorm(length(ingresos),media\_ing1,desv\_ing1)  
qqnorm(prob\_graf1, main = "Ingresos con primaria")  
qqline(prob\_graf1, col = "green")  
  
  
prob\_graf2 = rnorm(length(ingresos),media\_ing2,desv\_ing2)  
qqnorm(prob\_graf2, main = "Ingresos con secundaria")  
qqline(prob\_graf2, col = "green")  
  
  
prob\_graf3 = rnorm(length(ingresos),media\_ing3,desv\_ing3)  
qqnorm(prob\_graf3, main = "Ingresos con postsecundaria")  
qqline(prob\_graf3, col = "green")  
  
  
boxplot(ingresos ~ neduc, xlab = "Nivel educativo")  
  
  
#Tablas de contingencia  
  
x <- round(prop.table(table(edad1,sitlab1)),4)  
knitr::kable(x)  
  
  
y <- round(prop.table(table(genero1,sitlab1)),4)  
knitr::kable(y)  
  
  
z <- round(prop.table(table(nac1,neduc1)),4)  
knitr::kable(z)  
  
  
#Gráfico de mosaico  
  
datos1 <- data.frame(genero1,neduc1,sitlab1)  
  
mosaicplot(~genero1 + neduc1 + sitlab1, main = NULL, xlab = "Género", ylab = "Nivel educativo", off = 22)  
  
  
est <- (1/N)\*(sum(ingresos\*(factel)))  
est  
media\_ing  
  
  
estvar\_ing <- sum(((1-(1/factel))/((1/factel)^2))\*(ingresos/N)^2)  
estdesv\_ing <- sqrt(estvar\_ing)  
estdesv\_ing  
  
  
estcoef <- estdesv\_ing/est  
estcoef  
  
  
D <- estdesv\_ing/sqrt(1-0.95)  
IntTcheb <- c(est - D, est + D)  
IntTcheb  
  
  
t <- qt(0.975, length(ingresos))  
IntAp <- c(est - estdesv\_ing \* t, est + estdesv\_ing \* t)  
IntAp  
  
  
estH <- 1/NH\*(sum(ingresos[genero == 1]\*(factel[genero == 1])))  
estM <- 1/NM\*(sum(ingresos[genero == 2]\*(factel[genero == 2])))  
estH  
estM  
  
  
estvar\_ingH <- sum(((1-(1/factel[genero == 1]))/(1/factel[genero == 1])^2)\*((1/NH)\*ingresos[genero == 1])^2)  
estvar\_ingM <- sum(((1-(1/factel[genero == 2]))/(1/factel[genero == 2])^2)\*((1/NM)\*ingresos[genero == 2])^2)  
estdesv\_ingH <- sqrt(estvar\_ingH)  
estdesv\_ingM <- sqrt(estvar\_ingM)  
estdesv\_ingH  
estdesv\_ingM  
  
  
estcoefH <- estdesv\_ingH/estH  
estcoefM <- estdesv\_ingM/estM  
estcoefH  
estcoefM  
  
  
DH <- estdesv\_ingH/sqrt(1-0.95)  
DM <- estdesv\_ingM/sqrt(1-0.95)  
IntTchebH <- c(estH - DH, estH + DH)  
IntTchebM <- c(estM - DM, estM + DM)  
IntTchebH  
IntTchebM  
  
  
t <- qt(0.975, length(ingresos))  
IntApH <- c(estH - estdesv\_ingH \* t, estH + estdesv\_ingH \* t)  
IntApM <- c(estM - estdesv\_ingM \* t, estM + estdesv\_ingM \* t)  
IntApH  
IntApM  
  
  
mean(as.numeric(datos$horas[genero == 1]))  
mean(as.numeric(datos$horas[genero == 2]))  
  
  
diseñoing = svydesign (id=~1, weights =~factel, data = datos )  
svyhist(~ingresos, diseñoing ,main = "Histograma Ingresos", col = "plum", xlab = "", ylab = "")  
  
  
diseñoH = svydesign (id=~1, weights =~factel[genero == 1], data = datos[genero ==1,])  
svyhist(~ingresos[genero == 1], diseñoH, main = 'Histograma Ingresos Hombres', col = "sienna2", xlab = "", ylab = "")  
  
  
diseñoM = svydesign (id=~1, weights =~factel[genero == 2], data = datos[genero == 2,])  
svyhist(~ingresos[genero == 2], diseñoM, main = 'Histograma Ingresos Mujeres', col = "violetred", xlab = "", ylab = "")  
  
  
svyboxplot(ingresos~1, diseñoing ,main = "Diagrama de caja Ingresos", col = "plum", xlab = "", ylab = "")  
  
  
svyboxplot(ingresos[genero == 1]~1, diseñoH, main = " Diagrama de caja Ingresos Hombres", col = "sienna2", xlab = "", ylab = "")  
  
  
svyboxplot(~ingresos[genero == 2]~1, diseñoM, main = "Diagrama de caja Ingresos Mujeres", col = "violetred", xlab = "", ylab = "")  
  
  
riesgo\_pobr <- c()  
  
for (i in 1:nrow(datos)){  
   
 if(ingresos[i] < 6600){  
   
 xa <- 1  
 riesgo\_pobr <- c(riesgo\_pobr, xa)  
   
 }  
 else{  
   
 xa <- 0  
 riesgo\_pobr <- c(riesgo\_pobr, xa)  
   
 }  
}  
  
datos <- cbind(datos, riesgo\_pobr)  
  
  
estRP <- (1/N)\*sum(riesgo\_pobr\*factel)  
estRP  
  
  
var\_RP <- sum(((1-(1/factel))/(1/factel)^2) \* ((1/N) \* riesgo\_pobr)^2)  
desv\_RP <- sqrt(var\_RP)  
desv\_RP  
  
  
estRPcoef <- desv\_RP/estRP  
estRPcoef  
  
  
DRP <- desv\_RP/sqrt(0.05)  
IntTchebRP <- c(estRP - DRP, estRP + DRP)  
IntTchebRP  
  
  
IntApRP <- c(estRP - desv\_RP\*t, estRP + desv\_RP\*t)  
IntApRP  
  
  
TotalRP <- N\*estRP  
TotalRP  
  
  
varTotalRP <- sum(((1-(1/factel))/(1/factel)^2)\*(riesgo\_pobr^2))  
desv\_TotalRP <- sqrt(varTotalRP)  
desv\_TotalRP  
  
  
estcoefTotalRP <- desv\_TotalRP/TotalRP  
estcoefTotalRP  
  
  
estRPH <- (1/NH)\*sum(riesgo\_pobr[genero == 1]\*factel[genero == 1])  
estRPM <- (1/NM)\*sum(riesgo\_pobr[genero == 2]\*factel[genero == 2])  
estRPH  
estRPM  
  
  
var\_RPH <- sum(((1-(1/factel[genero == 1]))/(1/factel[genero == 1])^2) \* ((1/NH) \* riesgo\_pobr[genero == 1])^2)  
var\_RPM <- sum(((1-(1/factel[genero == 2]))/(1/factel[genero == 2])^2) \* ((1/NM) \* riesgo\_pobr[genero == 2])^2)  
desv\_RPH <- sqrt(var\_RPH)  
desv\_RPM <- sqrt(var\_RPM)  
desv\_RPH  
desv\_RPM  
  
  
estRPcoefH <- desv\_RPH/estRPH  
estRPcoefM <- desv\_RPM/estRPM  
estRPcoefH  
estRPcoefM  
  
  
DRPH <- desv\_RPH/sqrt(0.05)  
DRPM <- desv\_RPM/sqrt(0.05)  
IntTchebRPH <- c(estRPH - DRPH, estRPH + DRPH)  
IntTchebRPM <- c(estRPM - DRPM, estRPM + DRPM)  
IntTchebRPH  
IntTchebRPM  
  
  
IntApRPH <- c(estRPH - desv\_RPH\*t, estRPH + desv\_RPH\*t)  
IntApRPM <- c(estRPM - desv\_RPM\*t, estRPM + desv\_RPM\*t)  
IntApRPH  
IntApRPM  
  
  
TotalRPH <- NH\*estRPH  
TotalRPM <- NM\*estRPM  
TotalRPH  
TotalRPM  
  
  
varTotalRPH <- sum(((1-(1/factel[genero == 1]))/(1/factel[genero == 1])^2)\*(riesgo\_pobr[genero == 1]^2))  
varTotalRPM <- sum(((1-(1/factel[genero == 2]))/(1/factel[genero == 2])^2)\*(riesgo\_pobr[genero == 2]^2))  
desv\_TotalRPH <- sqrt(varTotalRPH)  
desv\_TotalRPM <- sqrt(varTotalRPM)  
desv\_TotalRPH  
desv\_TotalRPM  
  
  
estcoefTotalRPH <- desv\_TotalRPH/TotalRPH  
estcoefTotalRPM <- desv\_TotalRPM/TotalRPM  
estcoefTotalRPH  
estcoefTotalRPM  
  
  
espana <- readRDS("gadm36\_ESP\_2\_sp.rds")  
madrid <- espana[espana$NAME\_1 == "Comunidad de Madrid",]  
madrid$cantidadH <- estRPH  
madrid$cantidadM <- estRPM  
texto <- "Madrid"   
coords <- coordinates(madrid)  
lista <- list("sp.text",coords, texto)  
spplot(madrid,"cantidadH", sp.layout = lista, at = seq(from = 0, to = .25, by = .02), main = "Mapa de riesgo de pobreza hombres")  
  
  
spplot(madrid,"cantidadM", sp.layout = lista, at = seq(from = 0, to = .25, by = .02), main = "Mapa de riesgo de pobreza mujeres")  
  
  
estringr1 <- ingresos[neduc == 1]  
estringr2 <- ingresos[neduc == 2]  
estringr3 <- ingresos[neduc == 3]  
  
Nh <- c(length(neduc[neduc == 1]), length(neduc[neduc == 2]), length(neduc[neduc == 3]))  
  
Xh <- list(Primaria = estringr1, Secundaria = estringr2, Postsec = estringr3)  
  
estrat <- function(Nhest, listmues, detalles = "no") {  
   
NTot = sum(Nhest)  
fpc = 1/10  
Nhest = unlist(lapply(listmues, length ))  
  
ybar = unlist(lapply(listmues, mean))  
yvar = unlist(lapply(listmues, var))  
  
media = sum(Nhest \* ybar) / NTot  
varmues = sum(Nhest^2 \* fpc \* yvar / Nhest) / (NTot^2)  
  
conf\_int = 2 \* sqrt(varmues)  
if(detalles == "no") {  
cbind(media, varmues, conf\_int) }  
else{cbind(media, varmues, conf\_int, Nhest, ybar, yvar)}  
  
}  
  
estrat(Nh, Xh, detalles = "si")  
  
  
n1 <- length(neduc[neduc == 1])  
n2 <- length(neduc[neduc == 2])  
n3 <- length(neduc[neduc == 3])  
  
t1 <- round(1/10\*n1,0)  
t2 <- round(1/10\*n2,0)  
t3 <- round(1/10\*n3,0)  
  
t1  
t2  
t3  
  
nneducs <- c(length(neduc[neduc == 1]), length(neduc[neduc == 2]), length(neduc[neduc == 3]))  
cuasvars <- c(var(ingresos[neduc == 1]),var(ingresos[neduc == 2]),var(ingresos[neduc == 3]))  
  
n1ney <- (length(ingresos)/10) \* ((length(neduc[neduc == 1]) \* var(ingresos[neduc == 1]))/(sum(nneducs\*cuasvars)))  
n2ney <- (length(ingresos)/10) \* ((length(neduc[neduc == 2]) \* var(ingresos[neduc == 2]))/(sum(nneducs\*cuasvars)))  
n3ney <- (length(ingresos)/10) \* ((length(neduc[neduc == 3]) \* var(ingresos[neduc == 3]))/(sum(nneducs\*cuasvars)))  
n1ney  
n2ney  
n3ney  
  
  
nestr <- length(neduc)/10  
Ch <- c(5,10,15)  
  
nopt1 <- nestr \* (((length(neduc[neduc == 1]) \* var(ingresos[neduc == 1]))/sqrt(Ch[1])))/(sum(nneducs \* cuasvars / sqrt(Ch)))  
nopt2 <- nestr \* (((length(neduc[neduc == 2]) \* var(ingresos[neduc == 2]))/sqrt(Ch[2])))/(sum(nneducs \* cuasvars / sqrt(Ch)))  
nopt3 <- nestr \* (((length(neduc[neduc == 3]) \* var(ingresos[neduc == 3]))/sqrt(Ch[3])))/(sum(nneducs \* cuasvars / sqrt(Ch)))  
nopt1  
nopt2  
nopt3  
  
  
wh <- c(n1/nestr,n2/nestr,n3/nestr)  
varprop <- ((1-(n1/length(neduc)))/nestr) \* sum(cuasvars \* wh)  
varnney <- (1/nestr) \* ((sum(wh\*sqrt(cuasvars)))^2) - (1/length(neduc)) \* sum(wh \* cuasvars)  
varprop  
varnney  
  
  
set.seed(123)  
s <- strata(datos, "neduc", size = c(t1, t2, t3), method = "srswor")  
  
strat\_pobl = getdata(datos, s)  
  
strat\_pobl$pobtam = with(strat\_pobl, ifelse(neduc == "1", n1, ifelse(neduc == "2", n2, n3)))  
  
strat\_pobl$mispesos = 1/strat\_pobl$Prob  
  
disestrat = svydesign(id=~1, weights=~mispesos, fpc=~pobtam, strat=~neduc, data=strat\_pobl)  
  
  
svymean(~as.numeric(ingnorm), disestrat)  
  
coefestestrat <- 661.42/14049  
coefestestrat  
  
  
svyby(~as.numeric(ingnorm), ~neduc, disestrat, svymean)  
  
coefestestrat1 <- 1037.56/11146.91  
coefestestrat2 <- 1144.17/12717.6  
coefestestrat3 <-1098.39/19145.94  
coefestestrat1  
coefestestrat2  
coefestestrat3  
  
set.seed(123)  
muestreomas <- sample(length(ingresos), length(ingresos)/10)  
muestramas <- datos[muestreomas,]  
muesingre <- as.numeric(muestramas$ingnorm)  
mueshoras <- as.numeric(muestramas$hora)  
qplot(mueshoras, muesingre) + geom\_point() + geom\_smooth(method=lm)  
  
  
razon <- function(x, y, medobs, N, n) {  
   
 r <- sum(y)/sum(x)  
 medR <- r \* medobs  
 totR <- medR \* N  
 varR <- ((1 - (n/N))/(n\*(n-1))) \* (sum(x^2) \* r^2 + sum(y^2) - 2 \* r \* sum(x \* y))  
   
 cat("Con el estimador de la razón R =", r, "sacamos las estimaciones de la media =", medR, " y del total =", totR,"con varianza =", varR, "\n")  
   
}  
  
regresion <- function(x, y, medobs, N, n){  
   
 datos <- data.frame(x, y, N)  
  
 dis = svydesign(id=~1, fpc=~N, data=datos)  
  
 regre <- svyglm(y~x, design = dis)  
   
 estregre <- coefficients(regre)[1] + coefficients(regre)[2] \* (medobs - mean(x))  
   
 esttotregre <- estregre \* N  
   
 varregre <- ((1 - (n/N))/(n\*(n-2))) \* (sum(y^2) - n \* mean(y)^2 - coefficients(regre)[2] \* (sum(x \* y) - n \* mean(x) \* mean(y)))  
   
 cat("Con el estimador de regresión sacamos para la media =", estregre, " y para el total =", esttotregre, " con varianza =", varregre, "\n")  
   
}  
  
razon(mueshoras,muesingre,mean(horas),length(ingresos), nestr)  
regresion(mueshoras,muesingre,mean(horas),length(ingresos), nestr)