Caso EasySHARE

MDSF – Examen Final 2019 - 2020

Mayra Goicochea Neyra

05/02/2020

- Asignatura: Fundamentos para el Análisis de Datos y la Investigación
- Profesor: Sonia De la Paz Cobo

1 INTRODUCCIÓN

La encuesta SHARE realiza continuamente estudios sobre atributos y criterios importantes (como Salud, Bienestar económico, Trabajo, Estudios entre otros) del colectivo de personas mayores de 50 años, asi como su entorno social y familiar, en los países europeos e Israel.

Contiene la información de alrededor de 380000 encuestados. Es de mucha importancia para estudios académicos y gubernamentales con la finalidad de estimar la calidad de vida y consideraciones futuras para la mejora del bienestar de ésta población.

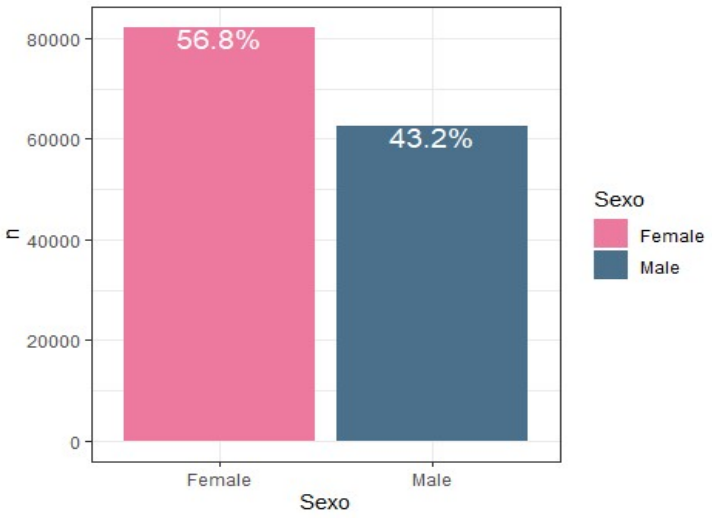
En el siguiente informe, se realiza un estudio sobre la paridad de Género con respecto a la Calidad de Vida, Trabajo y Situación Económica, para tal objetivo se utilizó el conjunto de datos de SHARE de fines académicos (EasySHARE) que contiene los datos de las 7 encuestas realizadas a las fechas, de las cuales sólo se consideraron las dos últimas (Wave 6 y 7).

2 FUENTE DE DATOS

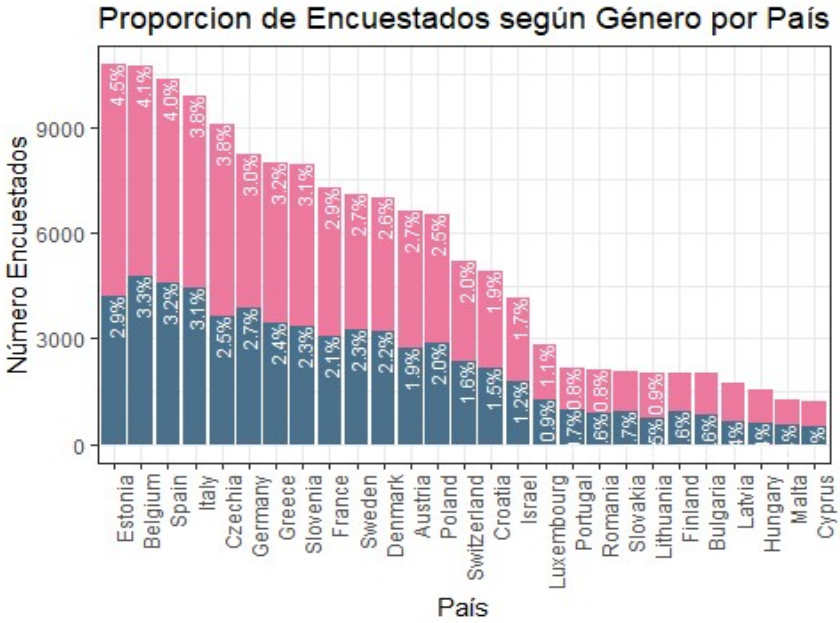
El conjunto de datos de las encuestas 6 y 7 contienen 144708 registros con información demográfica, sobre conexión social, sobre condiciones de infancia, de salud, riesgos de comportamiento, laboral y sobre la situación económica de cada encuestado. Con la finalidad de revisar a detalle las diferencias entre la población de mujeres y hombres sobre cómo perciben su Calidad de Vida, y sesgando hacia indicadores más resaltantes como es la situación económica, se consideraron sólo los atributos de Situación económica del Hogar, Empleo y la variable de la encuesta *casp*, que es un estimador del encuestado en alusión a su bienestar físico, material, social, emocional y de desarrollo.

3 MUESTRAS POBLACIONALES

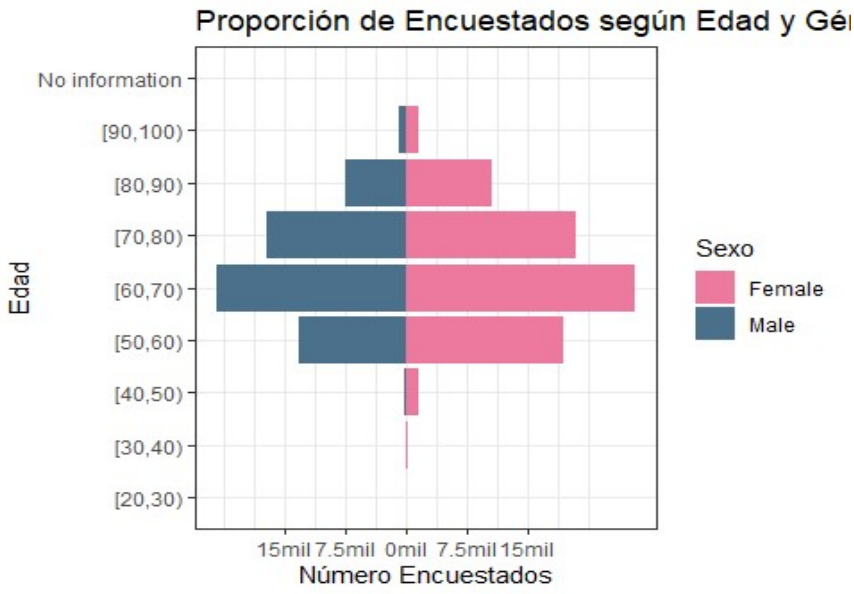
Se tiene dos muestras poblacionales, las cuales se identifican mediante la variable *female*, que son el colectivo de Mujeres (con un 56.8% de encuestados) y el colectivo de Hombres (con un 43.2%). Ambas muestras son independientes porque se generaron de encuestas realizadas a individuos distintos, esta premisa es considerada en el análisis de los atributos relacionados.



La encuesta se realizó en 27 países europeos e Israel, la proporción de encuestados según género se muestra en el siguiente gráfico. Estonia es el país con mayor tasa de encuestados. También se puede observar que las tasas de ambas poblaciones se mantienen entre los países, esto rasgo es muy importante porque nos indica que no hay sesgo por país en cuanto a una población determinada.



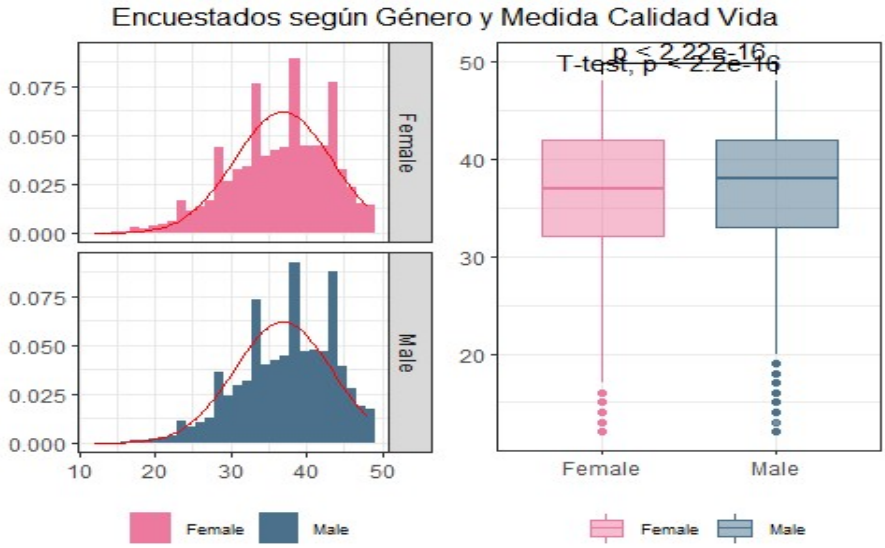
En cuanto al rango de Edad, el rango mayoritario es el de 60 a 70 años. El siguiente gráfico muestra que la distribución de los colectivos se mantiene, lo que permite que el análisis sea adecuado porque no hay sesgo hacia determinada población.



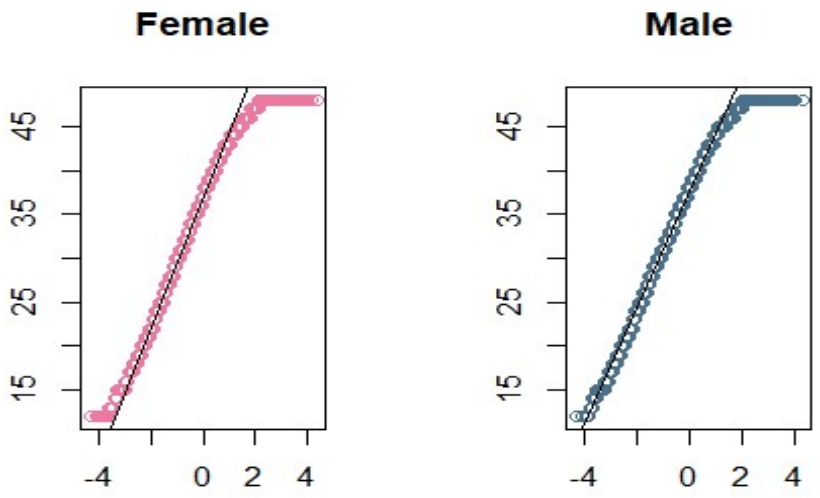
3.1 ÍNDICE DE CALIDAD DE VIDA

La variable *casp* mide la calidad de vida y se basa en cuatro subes calas de control, autonomía, placer y auto correlación. Con fines de hacer un análisis comparativo según el Género, los valores negativos que significan que no se tiene información verificada de los encuestados se apartaran para revisar los casos de los encuestados que si brindaron el estimador (corresponde al 0.06% casos con valores ausentes, no es muy significativo, además que para las pruebas de inferencia se tiene suficiente número de observaciones).

Como se observa en el histograma, la distribución de ésta variable en ambos colectivos no parece una Normal, es asimétrica con cola a la izquierda, sin embargo, se contrastará mediante las pruebas correspondientes. El diagrama de bloques nos muestra en detalle, que los rangos, sin considerar los valores atípicos, que las mujeres tienen un valor mínimo ligeramente menor al de los hombres, aunque no es lo suficiente para concluir que el grupo femenino estime su calidad de vida menor al de los varones.



Mediante las gráficas qqnorm muestran asimetría hacia la derecha y los test de shapiro wilk encuentran evidencias significativas de que los datos no proceden de poblaciones con distribución normal. Sin embargo, dado que el tamaño del grupo es grande se puede considerar que el t-test sigue siendo suficientemente robusto.




```
##
##  Shapiro-Wilk normality test
##
## data:  muestraFemale
## W = 0.97665, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  muestraMale
## W = 0.97298, p-value < 2.2e-16
```

A continuación, mediante pruebas de contraste de igualdad de varianza muestral se estima si es necesario ajustar la prueba t-test con corrección de Welch. Dado que las muestras no cumplen el criterio de normalidad, es recomendable usar el test Leven o el test no paramétrico de Fligner-Killen (ambos basados en la mediana).

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  casp by Sexo
## Fligner-Killeen:med chi-squared = 162.56, df = 1, p-value < 2.2e-16

## Levene's Test for Homogeneity of Variance (center = "median")
##           Df F value    Pr(>F)
## group      1  149.08 < 2.2e-16 ***
##       135734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los tests muestran que no se puede aceptar la hipótesis nula “Varianzas homogéneas”, es por ello que la prueba t-test se debe realizar con corrección de Welch.

```
##
##  Welch Two Sample t-test
##
## data:  data.casp[data.casp$Sexo == "Female", "casp"] and data.casp[data.casp$Sexo
== "Male", "casp"]
## t = -24.416, df = 128688, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9204728 -0.7836750
## sample estimates:
## mean of x mean of y
##  36.53339  37.38547

##
## Cohen's d
##
## d estimate: -0.1329838 (negligible)
## 95 percent confidence interval:
##      lower      upper
## -0.1437390 -0.1222286
```

Con un nivel de confianza del 95%, la prueba de T-test no acepta la hipótesis nula con un intervalo de confianza de <-0.92: -0.78>. Se puede concluir que hay evidencias para considerar que existen diferencias entre el valor medio de casp en las mujeres y en los hombres. Aunque, mediante la prueba de d-Cohen, estima que el efecto es muy pequeño (-0.13).

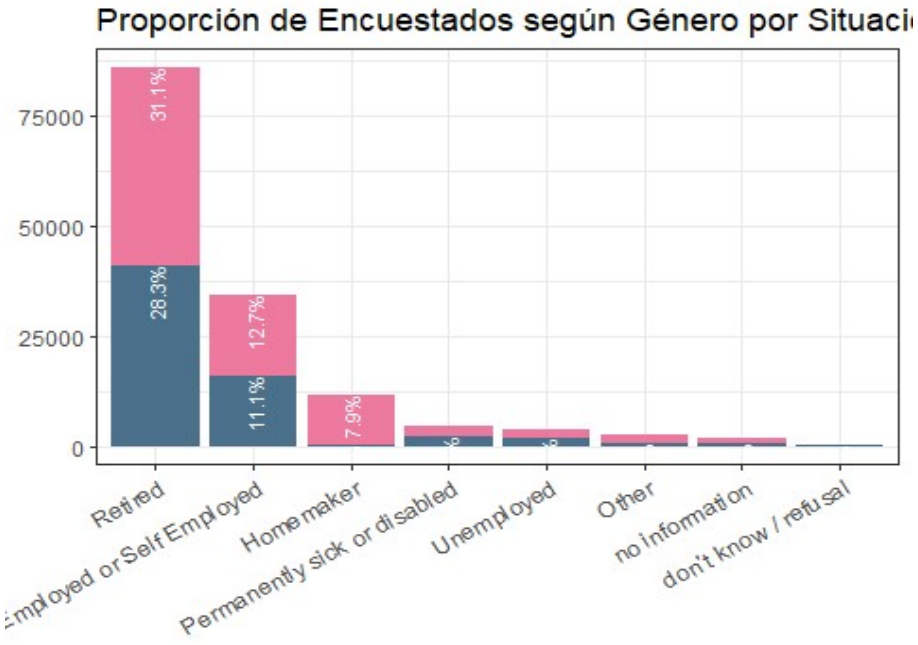
3.2 MÓDULO EMPLEO

La población objetivo de esta encuesta es las personas mayores de 50 años, por lo que atributos relacionados a la situación laboral del encuestado es muy relevante para el estudio de su calidad de vida. Se sabe que una persona puede sentirse con más confianza y motivación, si realiza actividades que le gusten o también si reciben incentivos, o cuando recibe cargos de responsabilidad. Con la finalidad de revisar estas hipótesis, consideró que las variables *ep005*, situación actual laboral, y *ep011_mod*, régimen laboral, mostrarán resultados que lleven a conclusiones sobre esta población y las muestras según género.

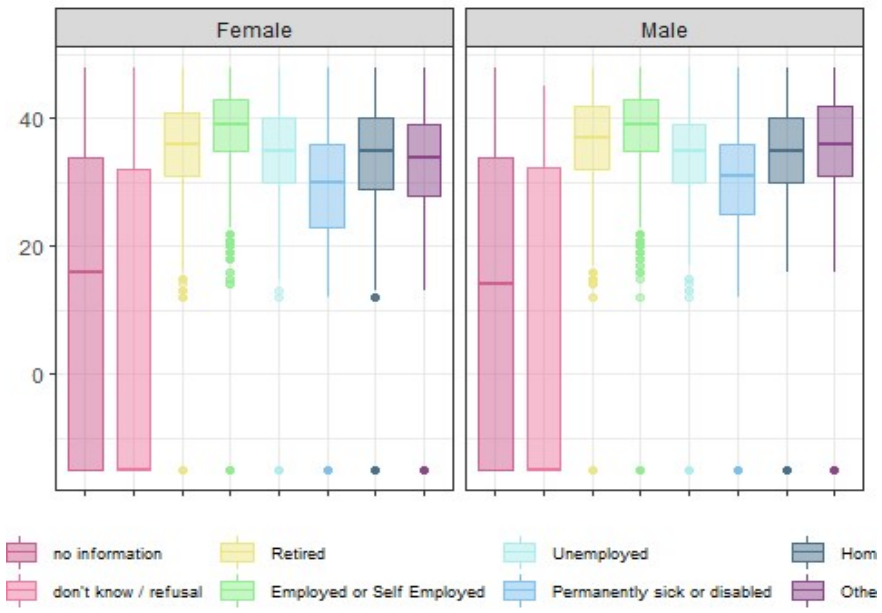
Con respecto a la “Situación Laboral Actual”, se tiene que 59.40% de los encuestados son retirados y el 23.82% son empleados de alguna empresa o autónomos.

```
##                                  Count  Total %
## Var1
## no information                 1907.00     1.32
## don’t know / refusal             90.00     0.06
## Retired                       85958.00    59.40
## Employed or Self Employed     34466.00    23.82
## Unemployed                     3557.00     2.46
## Permanently sick or disabled   4470.00     3.09
## Homemaker                     11746.00     8.12
## Other                          2514.00     1.74
```

La distribución según género muestra una proporción homogénea en todas las categorias de la variable. A excepción, de la categoría *Homemaker* que representa el 7.9% de mujeres encuestadas.



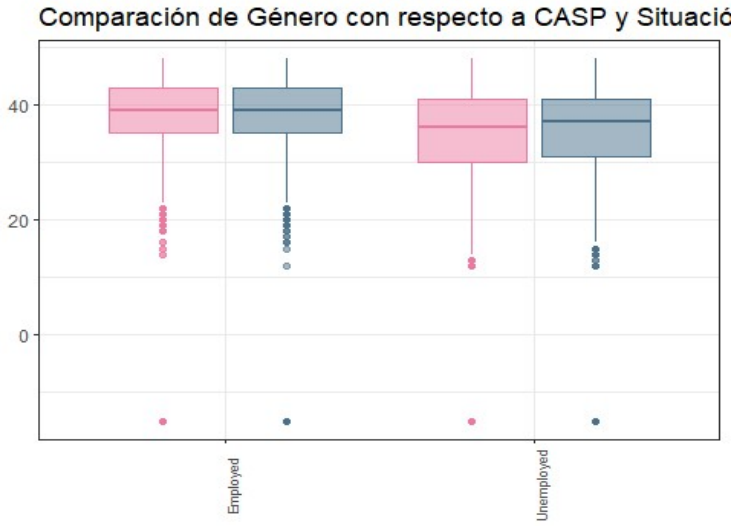
El siguiente gráfico de cajas muestra que las categorías se distribuyen de forma similar en ambas muestras poblacionales. También se observa que la caja de "Empleados" tiene rangos mayores en la variable *casp* en ambas muestras.



Existen información ausente en algunos encuestados, que, para fines de análisis más detallado, y considerando que se tiene un gran conjunto de observaciones, se extraen los casos "No information" y "Don't Know / Refusal" por no tener información verificada. Asimismo, se agruparon las categorías en "Empleados" y "No Empleados"

(donde se incluyen los retirados) para estudiar a detalle las características relevantes de cada muestra (Hombres y Mujeres).

El diagrama de cajas resultante muestra que la categoría Empleados tiene mayor rango de casp con respecto a la caja de No Empleados.



Se tienen 18413 mujeres empleadas que corresponden al 22.7% de la muestra de mujeres encuestadas. Y 16053 hombres que son empleados que corresponde al 25.97%.

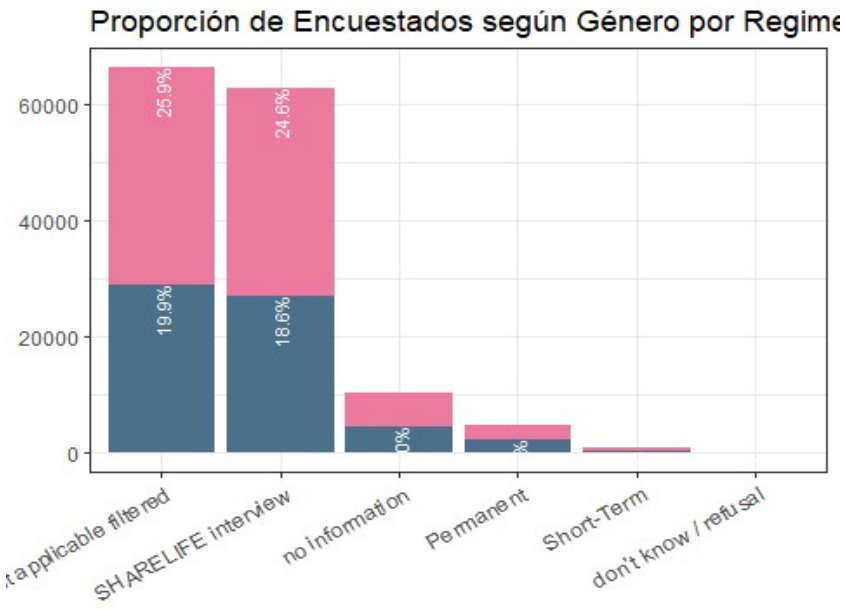
```
##         Employed Unemployed
##
## Female     18413      62485
## Male       16053      45760
```

La prueba de contraste de igualdad de proporciones nos permite afirmar, con un 95% de confianza, que hay una tasa entre 2.759% y 3.66% mayor de incidencia de empleos en hombres que mujeres. Además, que se confirma con un p-valor menor a 0.05, que la tasa de empleo es diferente entre mujeres y hombres.

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(16053, 18413) out of c(61813, 80898)
## X-squared = 197.04, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.02758993 0.03660017
## sample estimates:
##    prop 1    prop 2
## 0.2597027 0.2276076
```

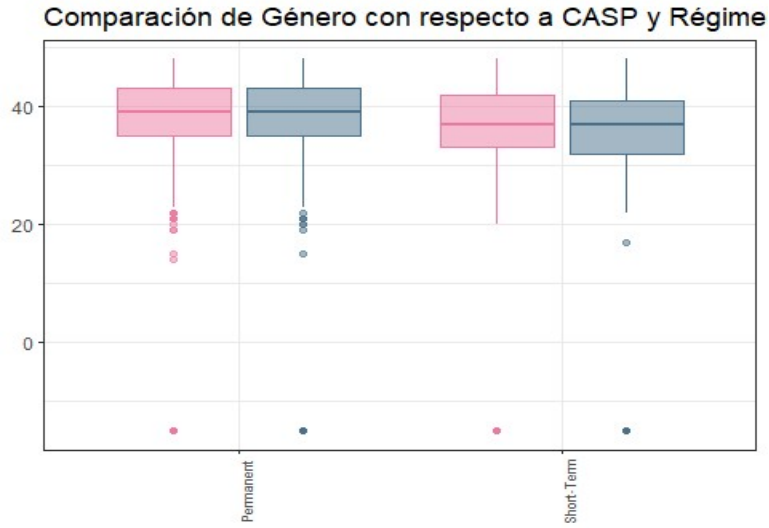
En cuanto al Régimen Laboral de los encuestados, se tiene que 3.33% de “Régimen Permanente” y el 0.53% representan a “Temporales”. Adicionalmente, se tiene mucha información no verificada que recae en las categorías “No information”, “don’t know / refusal”, “Sharelife interview” y “No applicable filtered”.

```
##                              Count  Total %
## Var1
## no information           10276.00     7.10
## don't know / refusal        22.00     0.02
## SHARELIFE interview      62561.00    43.23
## not applicable filtered  66262.00    45.79
## Short-Term                 763.00     0.53
## Permanent                 4824.00     3.33
```



De manera similar al indicador de Situación Laboral, se filtrarán solo los casos con régimen laboral para continuar con el estudio detallado de esta característica.

La variable *casp* es mayor en las observaciones con Régimen Permanente tanto para mujeres como para hombres. También se observa que en la categoría "Temporal", las mujeres presentan mayor valor en casp a diferencia de los hombres, pero esto puede deberse a que la mayoría de la categoría Temporal son casos de mujeres encuestadas.



El 87.4% de casos de hombres son de régimen Permanente, en el caso de las mujeres, es el 85.5%.

```
##         Permanent Short-Term
##
## Female       2683        456
## Male         2141        307
```

La prueba de contraste de igualdad de proporciones nos permite afirmar, con un 95% de confianza, que hay una tasa entre 0.1858% y 3.79% mayor de incidencia de que el empleo en hombres sea por régimen Permanente a comparación del que tiene las mujeres. Además, que se confirma con un p-valor menor a 0.05, que la tasa de empleo en régimen permanente es diferente entre mujeres y hombres.

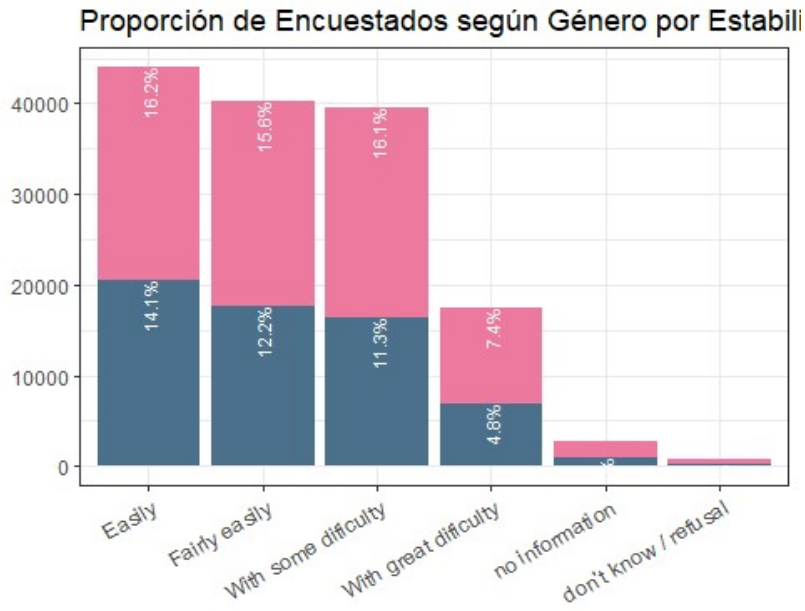
```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(2141, 2683) out of c(2448, 3139)
## X-squared = 4.6009, df = 1, p-value = 0.03196
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.001858842 0.037862552
## sample estimates:
##    prop 1    prop 2
## 0.8745915 0.8547308
```

3.3 MODULO INGRESO HOGAR

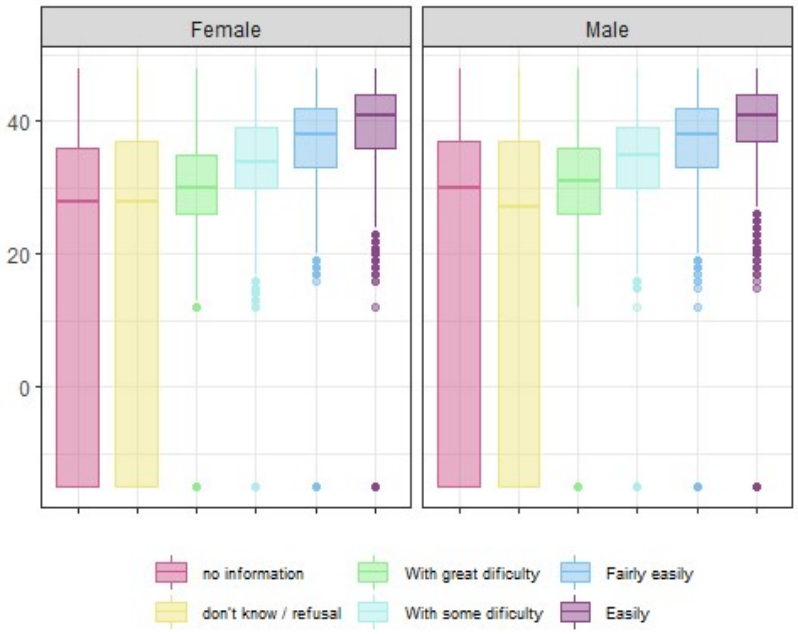
Otros aspectos importantes de interés para la población objetivo de esta encuesta, son los relacionados a la seguridad y situación económica del hogar, debido a que contribuye al bienestar del encuestado, mientras su hogar no tenga muchas dificultades puede vivir con tranquilidad, es por ello muy importante considerarlas para el estudio de paridad de género.

La variable *co007_* guarda las categorías de respuesta ante la pregunta si hay estabilidad económica en el hogar. Se observa que el 30.38% de los encuestados respondieron que su familia llega a fin de mes con tranquilidad, un 27.77% con tranquilidad más ajustada, mientras que el 39.48% con dificultad y el 2.37% de información ausente o no verificada.

```
##                          Count  Total %
## Var1
## no information         2718.00     1.88
## don't know / refusal    708.00     0.49
## With great dificulty  17531.00    12.11
## With some dificulty   39602.00    27.37
## Fairly easily         40188.00    27.77
## Easily                43961.00    30.38
```

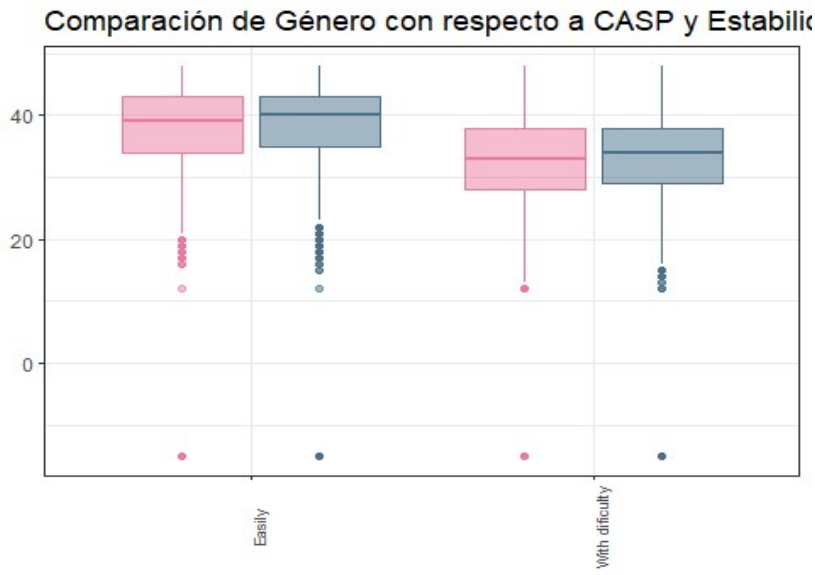


El siguiente gráfico de cajas muestra que las categorias se distribuyen de forma similar en ambas muestras poblacionales. También se observa que la caja de "Easily" tiene rangos mayores en la variable *casp* en ambas muestras.



De manera similar a los atributos del módulo Empleo, se filtrarán solo los casos con información verificada, excluyendo las categorías "no information" y "don't know / refusal". También, se agruparán las categorías "Easily" y "Fairly Easily" como "Easily" y las otras dos categorias como "With Dificulty" para realizar el estudio de diferencias de medias.

La variable *casp* es mayor en las observaciones con estabilidad económica en el hogar tanto para mujeres como para hombres, lo cual es razonable porque al tener menos preocupaciones el encuestado puede percibir una mejor calidad de vida.



El 57.5% de mujeres encuestadas han respondido que llevan sin problemas la situación económica en su hogar. Y el 62.2% de hombres encuestados han respondido lo mismo.

```
##         Easily With dificulty
##
## Female   46012          33923
## Male     38137          23210
```

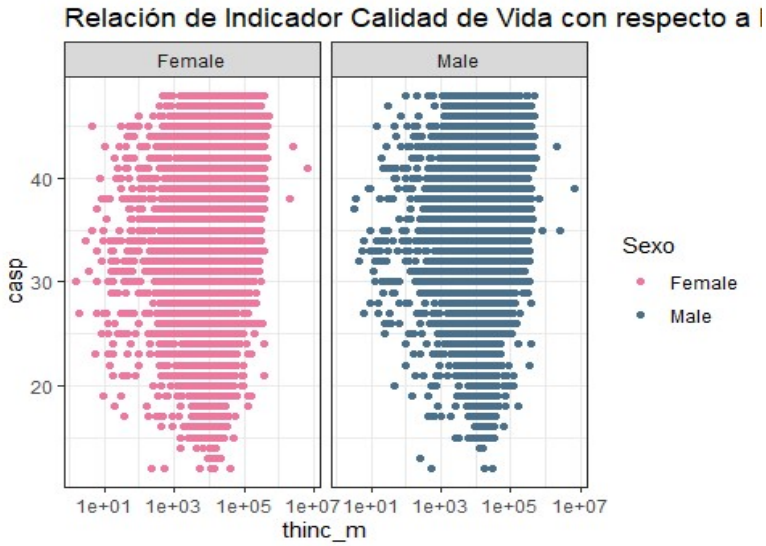
La prueba de contraste de igualdad de proporciones nos permite afirmar, con un 95% de confianza, que hay una tasa entre 4.1% y 5.1% mayor de incidencia de que haya estabilidad económica en el hogar de los hombres a comparación de las mujeres. Además, que se confirma con un p-valor menor a 0.05, que la tasa de estabilidad económica en el hogar es diferente entre mujeres y hombres.

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(38137, 46012) out of c(61347, 79935)
## X-squared = 305.49, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04089806 0.05118734
## sample estimates:
##    prop 1    prop 2
## 0.6216604 0.5756177
```

La variable *thinc_m* es de carácter numérico y guarda el ingreso neto de la vivienda, como indica la metodología de SHARE se tiene valores imputados y algunos etiquetados con valores negativos que muestran que no se tiene la información verificada. De forma similar a la variable *casp*, y como se tiene una muestra grande, se consideraron solo los valores mayores a 0 para el análisis de inferencia.

¿El ingreso neto del Hogar tendrá relación con la Calidad de Vida?, dado que la Calidad de Vida o *casp* estima el bienestar en varios aspectos del encuestado entre ellos el de desarrollo (sobre cumplimiento de metas) y emocional, si debe guardar relación por que un individuo que tenga un ingreso de hogar menor podría pasar

situaciones difíciles y su bienestar verse afectado. Mediante un gráfico de dispersión, se puede observar que mientras el estimador de *casp* es mayor, también es mayor el ingreso neto.



En cuanto a las ratios de correlación, se puede observar que son variables ligeramente correlacionadas. Es así que podriamos concluir que no están muy relacionadas la variable *casp* y el ingreso neto del hogar para las muestras.

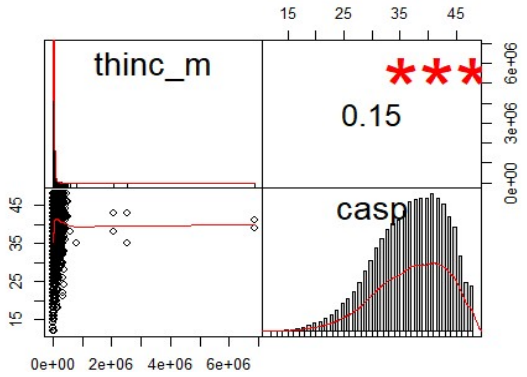


Ilustración 1: Correlación con toda la muestra

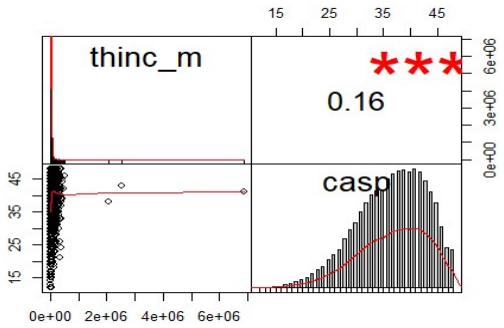


Ilustración 2: Correlación Muestra Mujeres

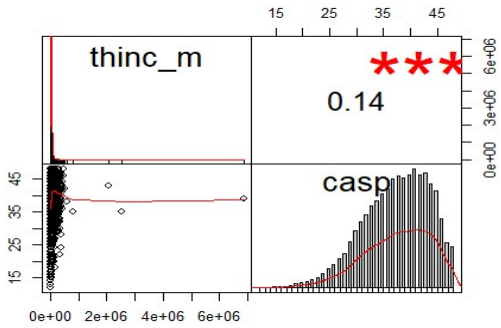
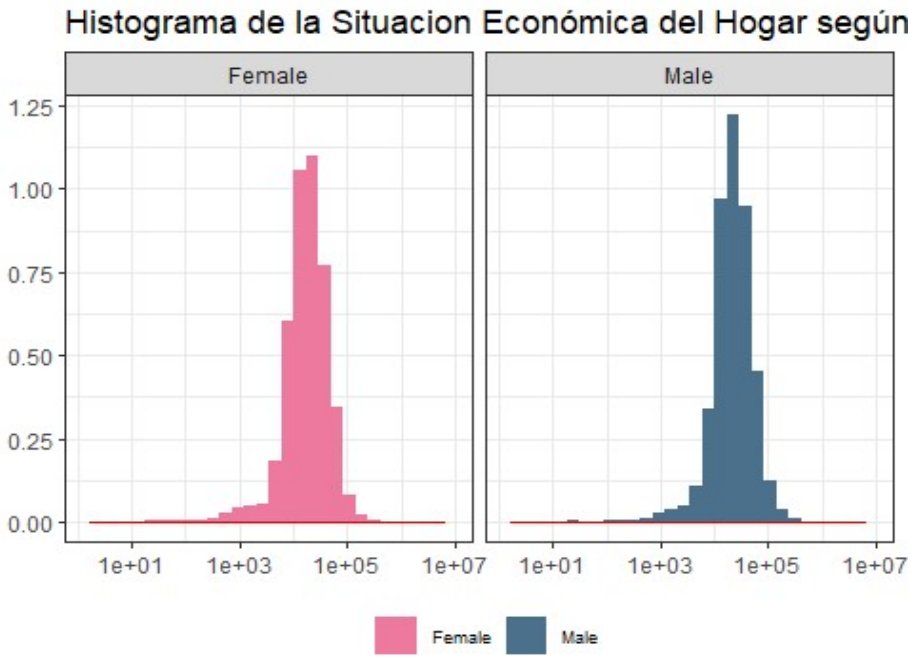
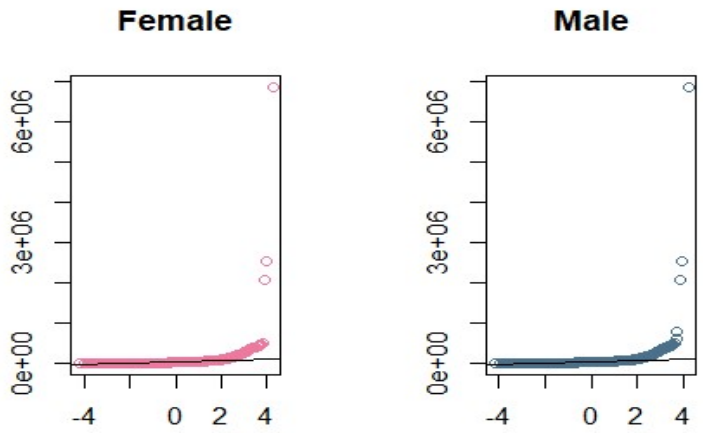


Ilustración 3: Correlación Muestra Hombres

Por otro lado, los gráficos de histograma de la variable *Ingreso Neto del Hogar* con respecto a las poblaciones según Género muestran un comportamiento similar en ambas, aunque la muestra de mujeres muestra que su mediana se encuentra en los 18000 euros, mientras que el de los hombres es 22000.



Como siguiente paso, se realizaron las pruebas de inferencia para comprobar si ambas muestras presentan diferencias significativas. De manera similar a lo realizado con la variable *casp*, primero se evalúa si ambas muestras se distribuyen en forma de una distribución normal (gaussiana), aparentemente por lo visto en los histogramas no son simétricas. Los gráficos *qqnorm* muestran que tienen asimetrías hacia la derecha.



Mientras que las pruebas de Shapiro-Wilk encuentran evidencias significativas de que los datos no se distribuyen en forma normal (Contraste de Normalidad). Como se indicó en el análisis de la variable *casp*, el tamaño de la muestra en grande y la prueba de t-test se puede considerar suficientemente robusta.

```
##
##  Shapiro-Wilk normality test
##
## data:  muestraFemale
## W = 0.6519, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  muestraMale
## W = 0.58724, p-value < 2.2e-16
```

El siguiente paso, fue realizar pruebas de contrastes de igualdad de varianza, con la finalidad de revisar si se requiere ajustar la prueba t-test mediante la corrección de Welch. Las pruebas de Leven y Fligner-Killen, recomendables dado que no se cumple el criterio de normalidad por que se basan en el estudio de la mediana, mostraron que no se puede aceptar la hipótesis nula "Varianzas homogéneas" (con p-valor <0.05).

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  thinc_m by Sexo
## Fligner-Killeen:med chi-squared = 309.08, df = 1, p-value < 2.2e-16

## Levene's Test for Homogeneity of Variance (center = "median")
##          Df F value    Pr(>F)
## group     1  51.423 7.509e-13 ***
##       80740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La prueba T-test con corrección de Welch muestra que hay diferencias significativas entre el ingreso medio de las mujeres encuestadas y el ingreso medio de los hombres encuestados con un intervalo de confianza al 95% entre los valores -5639.431 a -4307.327 (Dado que el intervalo es negativo, el valor del ingreso medio de los hombres encuestados es mayor al de las mujeres). El tamaño de efecto medido por d-Cohen es muy pequeño (-0.11).

```
##
##  Welch Two Sample t-test
##
## data:  data.think[data.think$Sexo == "Female", "thinc_m"] and data.think[data.thin
k$Sexo == "Male", "thinc_m"]
## t = -14.635, df = 69364, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5639.431 -4307.327
## sample estimates:
## mean of x mean of y
##  25348.88  30322.26
```

```
##
## Cohen's d
##
## d estimate: -0.1058927 (negligible)
## 95 percent confidence interval:
##       lower       upper
## -0.11982003 -0.09196546
```

4 CONCLUSIONES

- La mayoria de encuestados es de sexo Femenino, que representa el 56.8% de las encuestas 6 y 7. El rango de edad mayoritario es entre los 60 a 70 años.
- La calidad de Vida o variable *casp* es una variable numérica, no presenta un comportamiento normal en ambas poblaciones, sin embargo, se puede afirmar que hay diferencias significativas entre hombres y mujeres encuestados. A un 95% de confianza, se considera que los hombres tienen un índice medio mayor que las mujeres.
- Sobre las variables de Empleo, se comprobó por pruebas de inferencias en las muestras, que la tasa de empleo en hombres es entre 2.759% a 3.66% mayor a las mujeres a un 95% de confianza. En cuanto a la tasa de incidencia en Régimen Permanente, los hombres presentan mayor incidencia.
- Sobre la estabilidad en el Hogar, de forma similar existe mayor incidencia en hombres a comparación de las mujeres. También, mediante la prueba t-test, se estimó que hay diferencias significativas entre ambas poblaciones.

5 REFERENCIAS BIBLIOGRÁFICAS

- Análisis de la Actividad Física y Satisfacción Vital en personas mayores de 60 años, Tesis Doctoral, Maria Antonia Parra Rizo, Universidad Miguel Hernández de Elche, Departamento de Psicología de la Salud.
- T-test de medias independientes y dependientes con R, por Joaquin Amat Rodrigo, *link*
- Inferencia para variables categóricas dicotómicas (proporciones). Intervalos de confianza y test de hipótesis, por Joaquin Amat Rodrigo, *link*