Departamento de Ingeniería Civil Industrial
Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile
IN3242-1 Estadística

PROBLEM SET 1

Tarea nº1

Integrantes: Adolfo Rojas
Profesor: Matías R. Labbé
Auxiliar: Diego Guevara
Fabiana Torres
Ayudantes: Antonia González
Javiera Donoso T.

Fecha de entrega: 05 de abril
Santiago de Chile

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Índice de Códigos

1. Estadística Matemática Básica

1.1. Ítem 1

Dada la distribución normal $\tau \sim N(\mu, 9\sigma^2)$ se obtiene la variable aleatoria estandarizada $Z = \frac{\tau-\mu}{3\sigma}$. Luego para calcular la probabilidad

$$\begin{aligned}\mathbb{P}(\mu-\sigma<\tau\leq\mu+\sigma) &= F_\tau(\mu+\sigma)-F_\tau(\mu-\sigma)\\ &= \Phi(\frac{\mu+\sigma-\mu}{3\sigma})-\Phi(\frac{\mu-\sigma-\mu}{3\sigma})\\ &= \Phi(\frac{1}{3})-\Phi(-\frac{1}{3}) = 2\Phi(\frac{1}{3})-1\\ &\approx 0.261\end{aligned}$$

1.2. Ítem 2

Calculamos por definición la esperanza de la v.a continua X.

$$\begin{aligned}E[X] &= \int_0^\infty \lambda x e^{-\lambda x}dx\\ &= [-xe^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x}dx\\ &= -\frac{1}{\lambda}[e^{-\lambda x}]_0^\infty = \frac{1}{\lambda}\end{aligned}$$

Luego para calcular el segundo momento de X procedemos como sigue

$$\begin{aligned}Var[X] &= E[X^2]-E[X]^2\\ &= \int_0^\infty \lambda x^2 e^{-\lambda x}dx - E[X]^2\\ &= [-x^2e^{-\lambda x}]_0^\infty + 2\int_0^\infty xe^{-\lambda x}dx - E[X]^2 = \frac{2}{\lambda}E[X]-E[X]^2\\ &= \frac{2}{\lambda^2}-\frac{1}{\lambda^2}\\ &= \frac{1}{\lambda^2}\end{aligned}$$

1.3. Ítem 3

Primero recordando que la esperanza es un operador lineal por linealidad de la integral se sigue que $E[X+\alpha] = \mu_X + \alpha$

$$\begin{aligned} Var[X+\alpha] &= E[((X+\alpha) - (\mu_X+\alpha))^2] \\ &= E[(X-\mu_X)^2] := \sigma_X^2 \end{aligned} \tag{1}$$

De misma forma aplicamos la definición de varianza

$$\begin{aligned} Var[X+\alpha Y] &= E[((X+\alpha Y) - E[X+\alpha Y])^2] \\ &= E[(X-\mu_X+\alpha(Y-\mu_Y))^2] \\ &= E[(X-\mu_X)^2 + 2\alpha(X-\mu_X)(Y-\mu_Y) + \alpha^2(Y-\mu_Y)^2] \\ &= \sigma_X^2 + 2\alpha\sigma_{XY}^2 + \alpha^2\sigma_Y^2 \end{aligned} \tag{2}$$

Usamos la definición de covarianza

$$\begin{aligned} Cov(\alpha X, \beta Y) &= E[(\alpha X - E[\alpha X])(\beta Y - E[\beta Y])] \\ &= \alpha\beta E[(X-\mu_X)(Y-\mu_Y)] := \alpha\beta\sigma_{XY} \end{aligned} \tag{3}$$

1.4. Ítem 4

Usando la parte b del ítem anterior con $\alpha = 1, \sigma_{XY} \neq 0$ se tiene que $Var[X+Y] = \sigma_X^2 + 2\sigma_{XY}^2 + \sigma_Y^2$ luego tomando $\alpha = -1$ vemos que $Var[X-Y] = \sigma_X^2 - 2\sigma_{XY}^2 + \sigma_Y^2$ lo cual denota que en efecto no se tiene la igualdad en el caso en que no exista relación entre variables aleatorias

2. Encuesta de Ocupación y Desocupación

2.1. Diciembre - 2022

2.1.1. Ítem 1

Abriendo el archivo *Diciembre2022.dta* de la base de datos podemos ver que este corresponde a una encuesta de Ocupación y Desocupación de diciembre del 2022 realizada por el Departamento de Economía de la Universidad de Chile, a través del Centro de Microdatos donde cada variable es de tipo entero/NA y que representan entre muchas cosas la información de un hogar, la situación ocupacional de jefa o jefe de hogar e integrantes, sus historiales bancarios, etc. En particular la variable **ident** representa el identificador del hogar la cual no es útil en el análisis estadístico pero sirve para diferenciar los hogares encuestados, **numper** la cantidad de personas en el hogar, importante para conocer la cantidad promedio de personas en un hogar y agrupar las mismas, **pcoh** el vinculo con el jefe del hogar (1 si el encuestador es la/el jefa/jefe de hogar), el cual serviría para sacar datos acerca del ingreso por hogar (como se verá en el ítem 5) entre otras utilidades para distintos análisis estadísticos que escapan este trabajo.

2.1.2. Ítem 2

Identificamos la variable de ingreso neto por trabajador **ingTrab** y filtramos para generar un histograma que solo comprenda los ingresos de la población activa

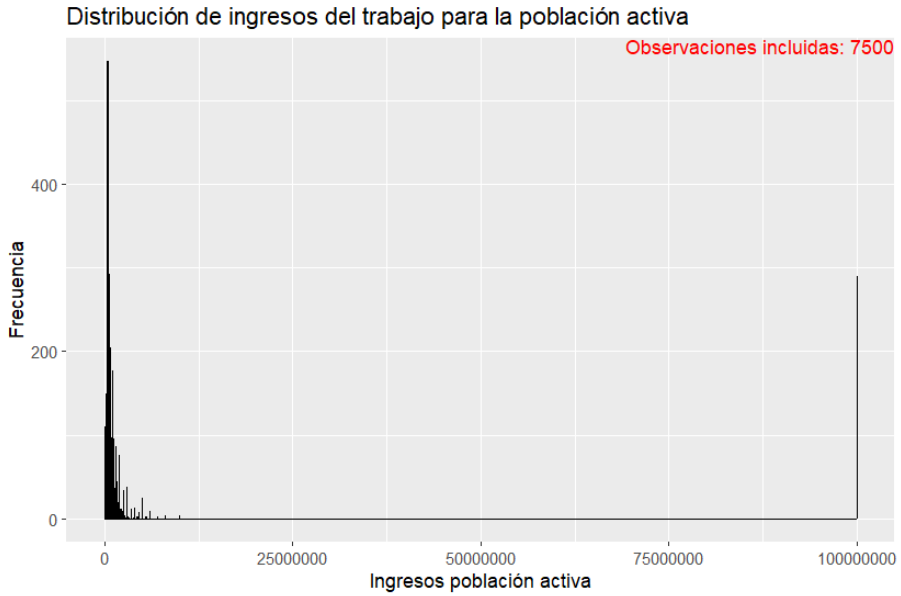

Figura 1: Gráfico con distribución de ingresos para la población muestral activa

Uno de los problemas que arruinan las dimensiones y el análisis son los ingresos *desconocidos* por los encuestados que fueron asignados con el valor 99999999 además de los valores NA que son aquellos que dejaron en blanco

2.1.3. Ítem 3

Una vez eliminado todos los ingresos desconocidos se obtiene el siguiente histograma y tabla con estadísticos básicos

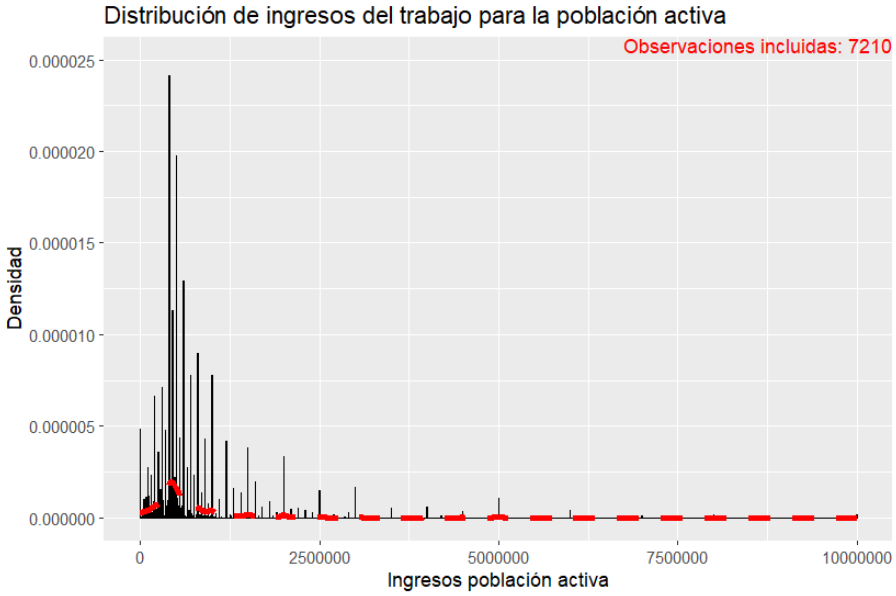

Figura 2: Gráfico con distribución de ingresos para la población muestral activa que sí registró sus ingresos

Observaciones	7210
Promedio	729912.3
Desviacion Estándar	812919.5
Minimo	0
Maximo	10000000
Mediana	500000
Percentil_1	0
Percentil_99	4500000

Tabla 1: Estadísticas básicas de los ingresos del trabajo

2.1.4. Ítem 4

A simple vista es posible decir que los ingresos de nuestra población no distribuyen normal debido a que los mismos se encuentran demasiado cargados a la media y existe un extremo donde se ganan cifras muy disparejas respecto a la población muestral, lo cual es confirmado con el uso del método gráfico Q-Q implementado en R tal como se muestra a continuación y nos permite concluir que la distribución no es de tipo normal.

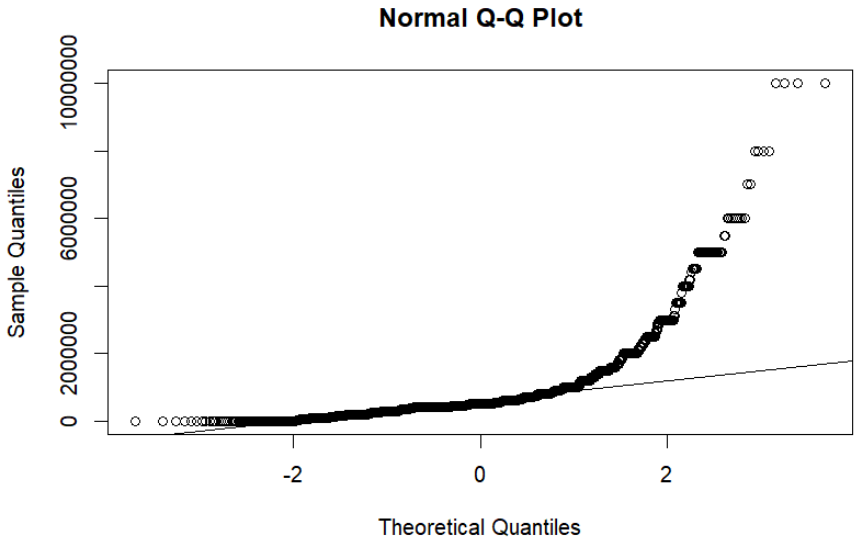

Figura 3: Gráfico con distribución de ingresos para la población muestral activa que sí registró sus ingresos

2.1.5. Ítem 5

Trabajando con los mismos datos ya limpios generamos el siguiente gráfico que muestra el porcentaje de hogares con deuda educacional por quintil

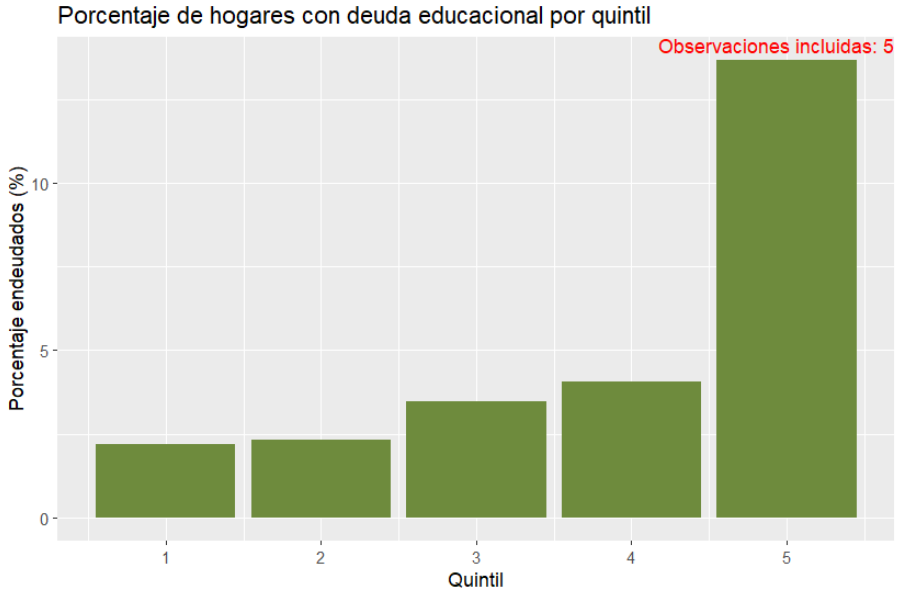

Figura 4: Gráfico con distribución de ingresos para la población muestral activa que sí registró sus ingresos

De este gráfico podemos ver cómo a mayor cantidad de ingresos, mayor es la cantidad de hogares endeudados por educación (mas no es posible inferir montos pues se trata solo de una variable dicotómica).
Repitiendo el proceso por deciles se obtiene el siguiente gráfico

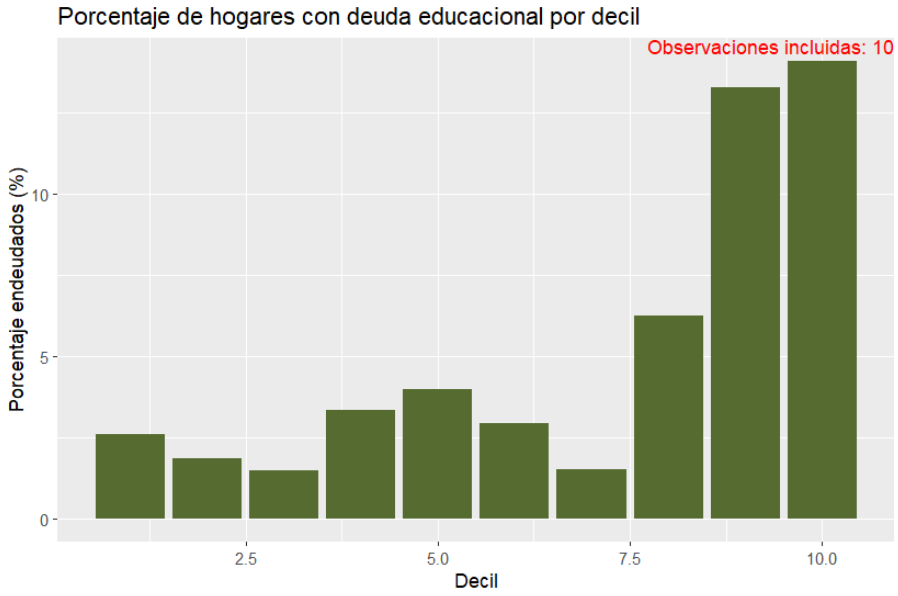

Figura 5: Gráfico con distribución de ingresos para la población muestral activa que sí registró sus ingresos

De aquí podemos ver como la tendencia de endeudamiento educacional en verdad es como una "normal.ª excepción del último decil el cual rompe esta tendencia con una cantidad de hogares endeudados muy grande.
Finalmente con los datos filtrados creamos dicha variable dicotómica para testear si existe una diferencia significativa en la proporción de hogares con deudas educativas entre los quintiles más bajos y el quintil más alto, los resultados obtenidos del test t realizado es el siguiente De aquí concluimos

Estadístico	Valor
Promedio x	0.0303901437371663259
Promedio y	0.1368760064412238298
IC inferior	-0.1344356802243264204
IC superior	-0.0785360451837885803
t	-7.4801796036478451768
Grados de libertad	700.5633997941017696576
p-value	0.0000000000002227931

Tabla 2: Resultados del test t

que la hipótesis en que estas medias fuesen iguales resultó ser rechazada, es decir, el quintil más alto se endeuda con creces (en cantidad de hogares, pues desconocemos la suma de endeudamiento) en comparación.

2.2. Una visión de largo plazo

2.2.1. Ítem 1

A continuación el snippet de código R en el cual creamos la seudo data frame con la cual se trabajará en los siguientes ítems

Código 1: Snippet con ciclo iterativo para la creación de dataframe con todos los años

```r
setwd("C:/Users/adolf/Desktop/CC/Progra/Estadística/Problem_Set_1_-_Data/mix_encuestas")
# Ítem 1

# Crear listas para almacenar los datos de cada trimestre
trimestre1 <- vector(mode = "list", length = 26)
trimestre2 <- vector(mode = "list", length = 26)
trimestre3 <- vector(mode = "list", length = 26)
trimestre4 <- vector(mode = "list", length = 26)

# Loop para cargar los datos de cada trimestre desde 1997 hasta 2022 (codigo puede ser optimizable)
for (x in 1997:2022){
  # Cargar datos de cada trimestre
  marzo <- read_dta(paste0("Marzo", x, ".dta"))
  junio <- read_dta(paste0("Junio", x, ".dta"))
  septiembre <- read_dta(paste0("Septiembre", x, ".dta"))
  diciembre <- read_dta(paste0("Diciembre", x, ".dta"))

  # Asignar trimestre y año a los datos y almacenarlos en las listas correspondientes
  trimestre1[[x-1996]] <- marzo
  trimestre1[[x-1996]]$trimestre <- 1
  trimestre1[[x-1996]]$año <- x

  trimestre2[[x-1996]] <- junio
  trimestre2[[x-1996]]$trimestre <- 2
  trimestre2[[x-1996]]$año <- x

  trimestre3[[x-1996]] <- septiembre
  trimestre3[[x-1996]]$trimestre <- 3
  trimestre3[[x-1996]]$año <- x

  trimestre4[[x-1996]] <- diciembre
  trimestre4[[x-1996]]$trimestre <- 4
  trimestre4[[x-1996]]$año <- x
}

# Combinar los datos de cada trimestre en un solo objeto
datos_agrupados <- c(trimestre1, trimestre2, trimestre3, trimestre4)
# Añadir las variables de identificación de filas faltantes y seleccionar las variables especificadas
for (i in 1:length(datos_agrupados)){
  colf <- setdiff(c("año", "trimestre", "ident", "orden", "numper", "pcoh", "sexo", "edad", "horas", "
    ↪ ingfam", "ingtrab", "arriendo", "arrienm", "sitocup1", "sitocup2"), names(datos_agrupados[[i
    ↪ ]]))
```

```r
  for (columna in colf) {
    datos_agrupados[[i]][[columna]] <- NA
  }
  datos_agrupados[[i]] <- datos_agrupados[[i]][, c("año", "trimestre", "ident", "orden", "numper", "
    ↪ pcoh", "sexo", "edad", "horas", "ingfam", "ingtrab", "arriendo", "arrienm", "sitocup1", "
    ↪ sitocup2")]
}
# Combinar todos los datos en un solo dataframe
datos_combinados <- do.call(bind_rows, datos_agrupados)
```

2.2.2. Ítem 2

Tomando ambas muestras realizamos un test t de student para refutar o no la hipótesis de que las horas promedio trabajadas en los años 1997 y 2017 son relativamente iguales. A continuación se presenta la tabla con los estadísticos relevantes del test Con lo que en efecto esta hipótesis es falsa

Variables	Valores
Media de x	48.4949427955563
Media de y	42.4955048409405
Intervalo de confianza inferior	5.721843
Intervalo de confianza superior	6.277033
t	42.3605686930211
df	37008.6772530479
valor-p	< 0.00000000000000022

Tabla 3: Estadísticas del test t

2.2.3. Ítem 3

Una vez filtradas las horas de trabajo generamos el siguiente gráfico con el valor esperado de horas trabajadas en cada trimestre (de cada año) y su intervalo de confianza con significancia del $\alpha = 0,05$

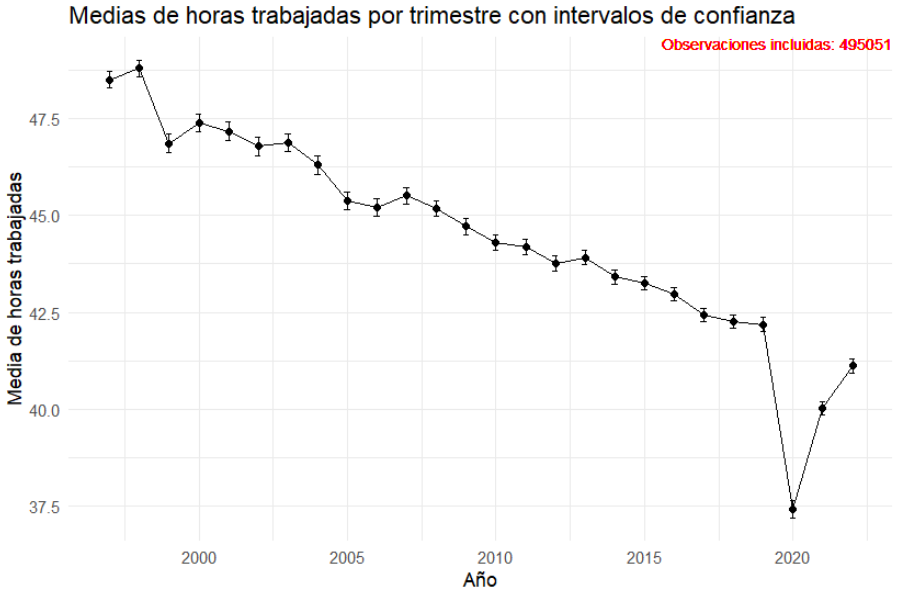

Figura 6: Gráfico con las horas trabajas esperadas

De aquí podemos ver que existe una tendencia a la disminución de horas trabajadas con el pasar de los años

2.2.4. Ítem 4

Diferenciando las muestras por genero se obtiene el siguiente gráfico

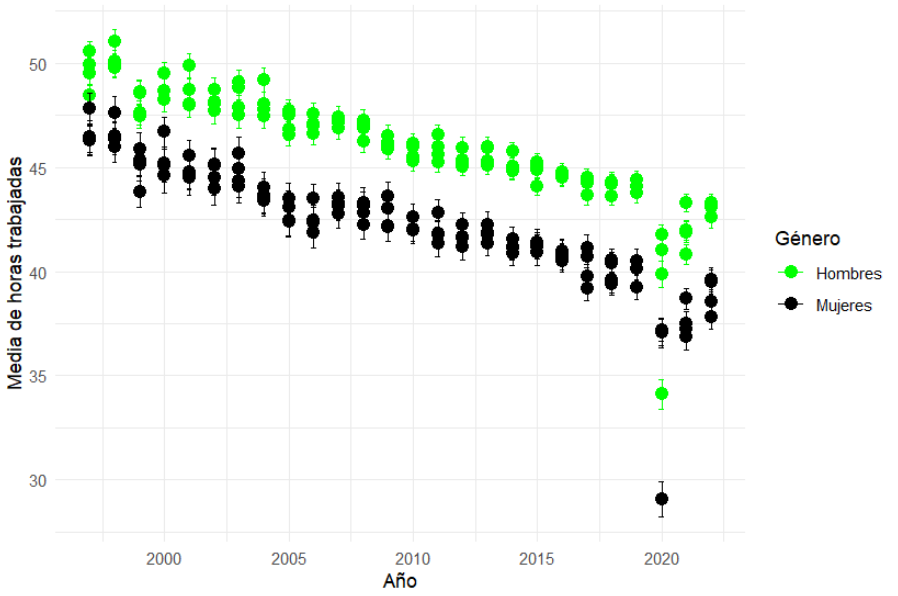

Figura 7: Gráfico con las horas trabajas esperadas por género

De aquí podemos ver que nuestra conclusión previa no distingue entre géneros y que la cantidad de horas trabajas por los hombres es mayor que el de las mujeres (por lo que se registra en la encuesta)

2.2.5. Ítem 5

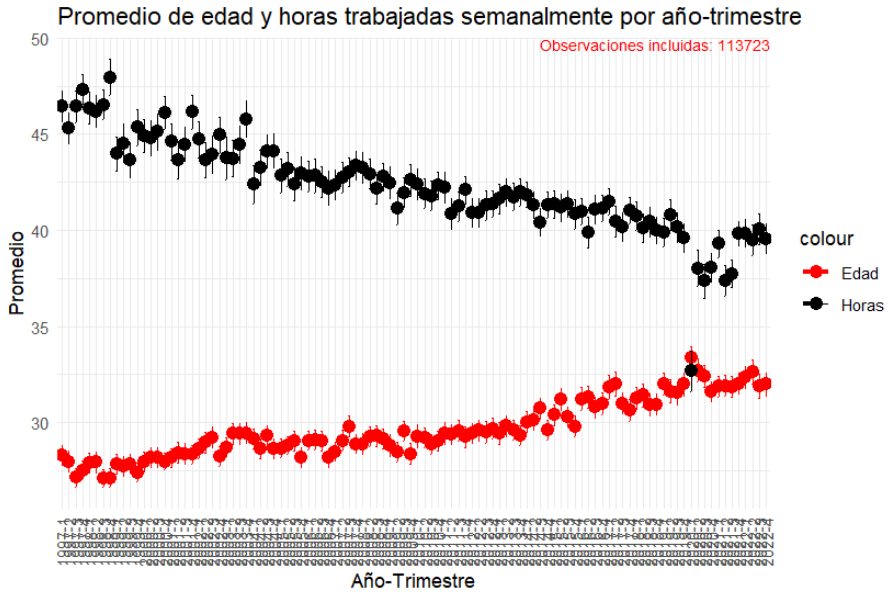

Figura 8: Gráficos superpuestos de los promedios muestrales de edad trabajador y horas de trabajo por semana

A medida que disminuyen las horas trabajadas se ha visto un aumento en el ingreso tardío al mercado laboral (puede ser coincidencia, dependencia entre variables o factores extra que se escapan del análisis por lo que no se puede inferir con seguridad)