A2 - Analítica descriptiva e inferencial

Leonardo Segovia Vilchez

Noviembre 2021

Contents

Introducción

El conjunto de datos claim.csv se inspira (ha sido modificado por motivos académicos) en la base de datos disponible en la plataforma Kaggle: https://www.kaggle.com/c/actuarial-loss-estimation.

Este conjunto de datos contiene información de una muestra de indemnizaciones otorgadas por una compañía de seguros por el tiempo que ha estado de baja laboral el trabajador.

Las variables del fichero de datos (train3.csv) son:

- ClaimNumber: Identificador de la póliza.
- DateTimeOfAccident: Fecha del accidente.
- DateReported: Fecha que se comunica a la compañía y se abre el expediente.
- Age: Edad del trabajador.
- Gender: Sexo.
- MaritalStatus: Estado civil, (M)arried, (S)ingle, (U)nknown, (W)idowed, (D)ivorced.
- DependentChildren: Número de hijos dependientes.
- DependentsOther: Número de dependientes excluyendo hijos.
- WeeklyWages: Salario semanal (en EUR).
- PartTimeFullTime: Jornada laboral, Part time (P) o Full time(F).
- HoursWeek:: Número horas por semana.
- DaysWeek: Número de días por semana.
- ClaimDescription: Descripción siniestros.
- IniCost: Estimación inicial del coste realizado por la compañía.
- UltCost: Coste total pagado por siniestro.
- Time: Tiempo desde que se apertura a cierra el siniestro.

```r
# Librerias
# Paquetes y librerías..
#install.packages("tidyr")
#install.packages("dplyr")
#install.packages(gridExtra)
library(tidyr)
library(tibble)
library(dplyr)
library(ggplot2)
library(scales)
library(tidyr)
library(stringr)
library(grid)
library(gridExtra)
library(nortest)
library(BSDA)

# Cargamos el fichero de datos.
claim <- read.csv('train_clean2.csv',stringsAsFactors = FALSE)

# Mostramos si el dataset se ha cargado correctamente.
head(claim,2)
```

```
##   X ClaimNumber   DateTimeOfAccident         DateReported Age Gender
## 1 1   WC8285054 2002-04-09T07:00:00Z 2002-07-05T00:00:00Z  48      M
## 2 2   WC6982224 1999-01-07T11:00:00Z 1999-01-20T00:00:00Z  43      F
##   MaritalStatus DependentChildren DependentsOther WeeklyWages PartTimeFullTime
## 1             M                 0               0      500.00                F
## 2             M                 0               0      509.34                F
##   HoursWeek DaysWeek                                           ClaimDescription
```

```
## 1      38.0        5          LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
## 2      37.5        5 STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
##   IniCost UltCost Time
## 1    1500    4303   87
## 2    5500    6106   13
```

Examinamos la interpretación que hace R de cada una de las variables.

```r
# Ejecutamos la función *class* a cada variable del dataset
sapply(claim, class)
```

```
##                  X        ClaimNumber DateTimeOfAccident       DateReported
##          "integer"        "character"        "character"        "character"
##                Age             Gender      MaritalStatus  DependentChildren
##          "integer"        "character"        "character"          "integer"
##    DependentsOther        WeeklyWages    PartTimeFullTime          HoursWeek
##          "integer"          "numeric"        "character"          "numeric"
##           DaysWeek   ClaimDescription            IniCost            UltCost
##          "integer"        "character"          "integer"          "integer"
##               Time
##          "integer"
```

Lectura del fichero y preparación de los datos

Leed el fichero claim.csv y guardad los datos en un objeto con identificador denominado claim. A continuación, verificad que los datos se han cargado correctamente.

```r
# Valores resumenes.
summary(claim)
```

```
##        X          ClaimNumber        DateTimeOfAccident DateReported
##  Min.   :    1   Length:50526       Length:50526       Length:50526
##  1st Qu.:13515   Class :character   Class :character   Class :character
##  Median :27006   Mode  :character   Mode  :character   Mode  :character
##  Mean   :27004
##  3rd Qu.:40480
##  Max.   :54000
##       Age           Gender          MaritalStatus      DependentChildren
##  Min.   :13.00   Length:50526       Length:50526       Min.   :0.000
##  1st Qu.:24.00   Class :character   Class :character   1st Qu.:0.000
##  Median :32.00   Mode  :character   Mode  :character   Median :0.000
##  Mean   :34.04                                         Mean   :0.123
##  3rd Qu.:43.00                                         3rd Qu.:0.000
##  Max.   :81.00                                         Max.   :8.000
##  DependentsOther     WeeklyWages      PartTimeFullTime     HoursWeek
##  Min.   :0.00000   Min.   :   6.81   Length:50526       Min.   : 1.00
##  1st Qu.:0.00000   1st Qu.: 238.71   Class :character   1st Qu.:38.00
##  Median :0.00000   Median : 408.50   Mode  :character   Median :38.00
##  Mean   :0.01043   Mean   : 433.78                      Mean   :37.36
##  3rd Qu.:0.00000   3rd Qu.: 513.00                      3rd Qu.:40.00
##  Max.   :5.00000   Max.   :7497.00                      Max.   :93.00
##     DaysWeek     ClaimDescription      IniCost            UltCost
##  Min.   :1.000   Length:50526       Min.   :      1   Min.   :     6
##  1st Qu.:5.000   Class :character   1st Qu.:    735   1st Qu.:  1183
##  Median :5.000   Mode  :character   Median :   2000   Median :  3291
##  Mean   :4.906                      Mean   :   7988   Mean   : 10148
##  3rd Qu.:5.000                      3rd Qu.:   9500   3rd Qu.:  9226
##  Max.   :7.000                      Max.   :2000000   Max.   :492515
##       Time
##  Min.   :   0.00
##  1st Qu.:  14.00
##  Median :  22.00
##  Mean   :  39.05
##  3rd Qu.:  41.00
##  Max.   :1095.00
```

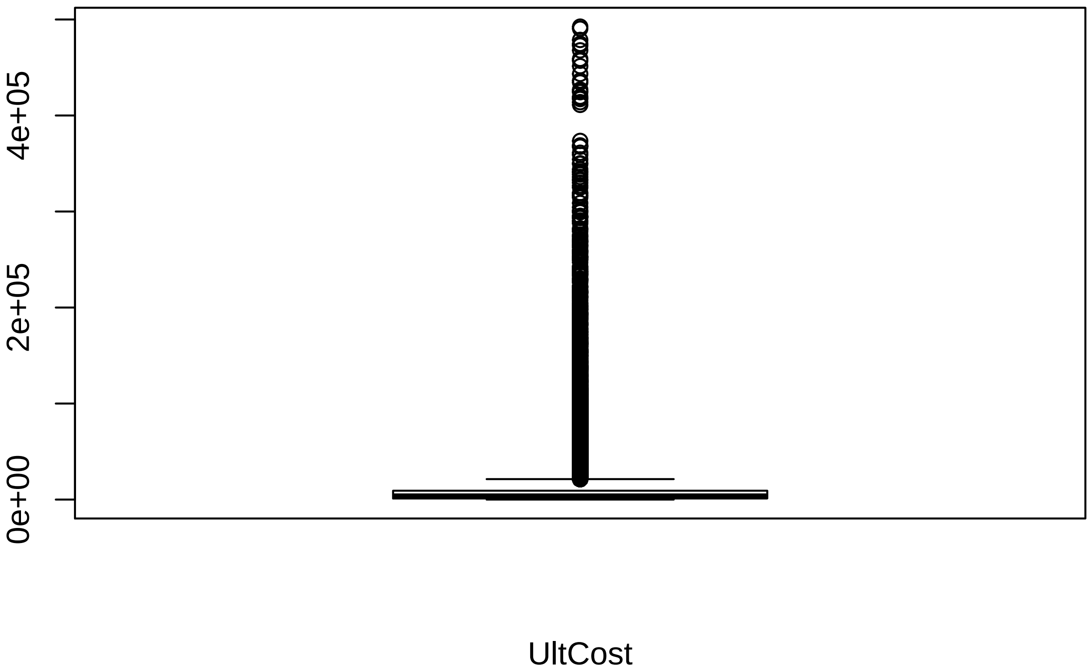
Coste de los siniestros y final de los siniestros

La compañía de seguros está interesada en investigar los valores que toma la variable coste de los siniestros en la población. Para ello, realizad un primer análisis visual de esta variable (UltCost) a partir de la muestra. Posteriormente, realizad un análisis de normalidad y calculad el intervalo de confianza de la variable UltCost de los siniestros. Seguid los pasos que se indican a continuación.

Análisis visual

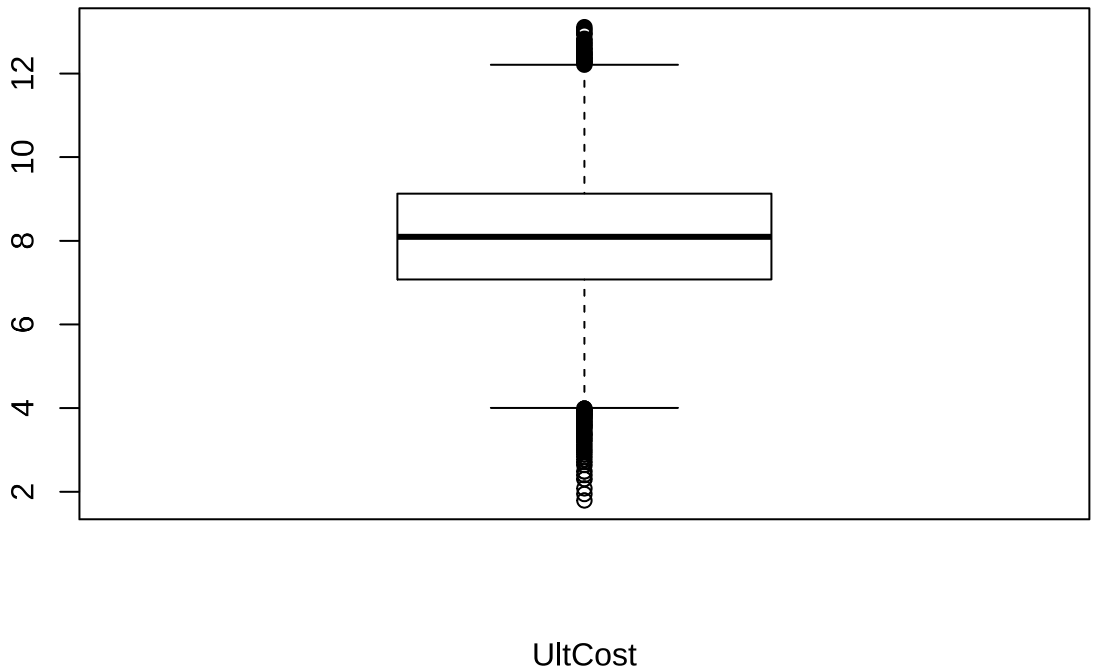
1. Mostrad con un diagrama de caja la distribución de la variable 'UltCost'.

```r
boxplot(claim$UltCost, xlab = "UltCost")
```



2. Transformad la variable 'UltCost' a escala logarítmica y mostrad el diagrama de caja.

```r
boxplot(log(claim$UltCost), xlab = "UltCost")
```



3. Interpretad los gráficos brevemente.

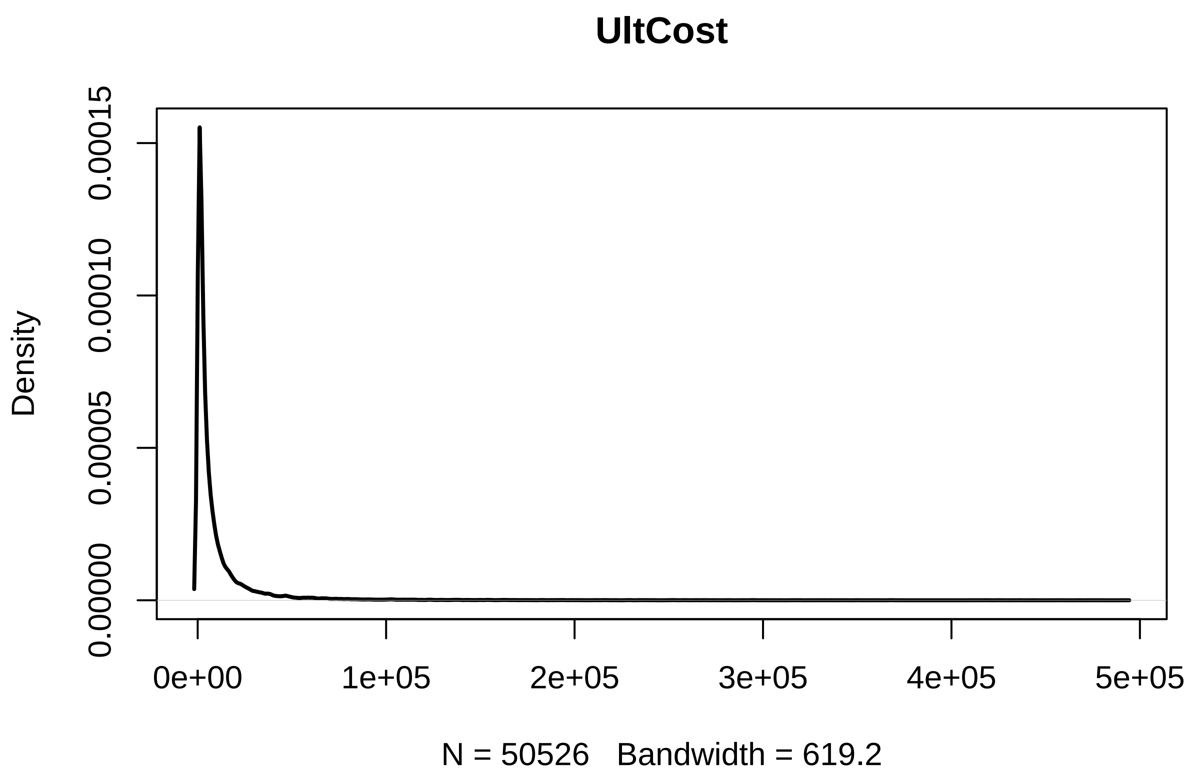
La función logarítmica cambia la escala de medida de los datos. Esta escala reduce la división entre muestras sobretodo los valores más altos (outlier), por este motivo la representación se hace mucho más visual.

Comprobación de normalidad

¿Podemos asumir que la variable UltCost tiene una distribución normal? Debéis justificar la respuesta a partir de métodos visuales y contrastes.

- Realizad inspección visual de normalidad en base a los gráficos que consideréis oportunos.

```
plot(density(claim$UltCost), lwd = 2,  main = "UltCost")
```



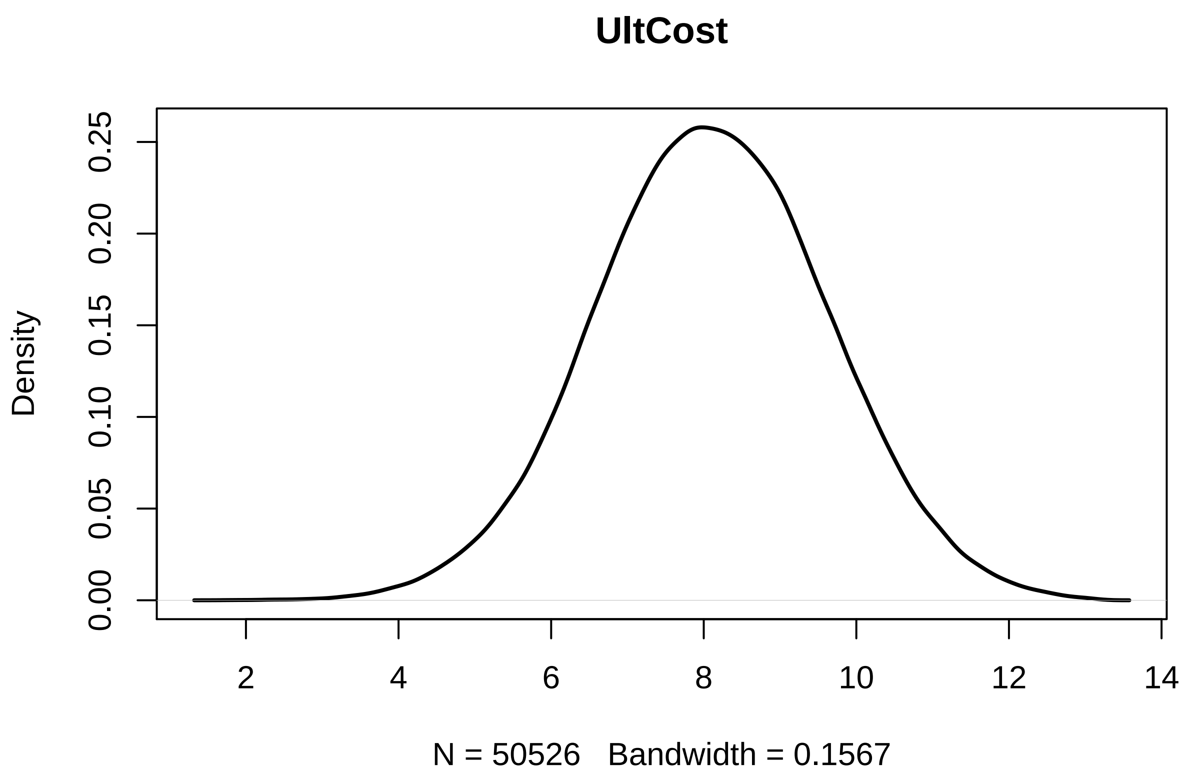
- Realizad contraste de normalidad de Lilliefors (p.ej. con función lillie.test de la libreria nortest).*

```
lillie.test(claim$UltCost)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  claim$UltCost
## D = 0.33597, p-value < 2.2e-16
```

- Realizad inspección visual y contraste de normalidad a la variable UltCost en escala logarítmica.

```
plot(density(log(claim$UltCost)), lwd = 2,  main = "UltCost")
```




```r
lillie.test(log(claim$UltCost))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  log(claim$UltCost)
## D = 0.0029142, p-value = 0.375
```

Podemos asumir que los datos escalados mediante una función logarítmica de la variable UltCost sige una distribucion normal. Esto queda demostrado visualmente mediante su distribución y los tests realizados, en los que no se puede rechazar la hipótesis nula para el escenario dodnde los datos han sido escalados.

Por lo que podriámos trabajar con los datos como si tuvieran una distribución normal, recordando que trabajamos con datos escalados

IC de la media poblacional de la variable UltCost

- Calculad manualmente el intervalo de confianza al 95% de la media poblacional de la variable 'UltCost' en escala original (No se pueden utilizar funciones como t.test o z.test para el cálculo). Sí se pueden usar funciones como 'mean', 'sd', 'qnorm', 'pnorm', 'qt' y 'pt'.

```r
# Función propia de calculo de IC.
tInterval <- function(d, alfa=0.05){

  sd <- sd(d)
  n <- length(d)
  SE <- sd / sqrt(n)
  t <- qt( 1- alfa/2, df=n-1, lower.tail=FALSE )
  L <- mean(d) - t*SE
  U <- mean(d) + t*SE
  return(round( c(U,L), 2))
```

```r
}

# Intervalos de Sales.
tInterval(claim$UltCost, alfa=0.05)
```

```
## [1]  9938.86 10356.48
```

```r
# Comprobación del intervalo.
t.test(claim$UltCost, sigma.x=sd(claim$UltCost))
```

```
##
##  One Sample t-test
##
## data:  claim$UltCost
## t = 95.251, df = 50525, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9938.855 10356.479
## sample estimates:
## mean of x
##  10147.67
```

- ¿Podemos asumir la hipótesis de normalidad para el cálculo del intervalo de confianza sobre la media muestral del coste en escala original? Argumentar la respuesta.

Por el teorema del límite central, podemos asumir normalidad, puesto que tenemos una muestra de tamaño grande n>30 y se desea realizar un test sobre la media.

- A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

El IC del 95% de la media poblacional de UltCost es (9938.86, 10356.47).Por los que si se sacarán diferentes muestras de la población, el 95% de los IC contendría el valor de la media poblacional.

Coste inicial y final de los siniestros

La compañía de seguros está interesada en investigar si la estimación inicial del coste que hace de los siniestros (IniCost) en promedio es suficiente para cubrir el coste total pagado (UltCost). Para eso, nos plantean la pregunta siguiente:

¿Podemos aceptar que no hay diferencias entre IniCost y UltCost?

Responded a la pregunta utilizando un nivel de confianza del 95%.

Seguid los pasos que se detallan a continuación.

Justificación del test a aplicar

Explicad qué tipo de contraste se puede aplicar en este caso. Es decir, explicad si se trata de un contraste de una muestra o dos muestras, sobre la media/varianza/proporción, si es bilateral o unilateral, etcétera.

Asumimos normalidad gracias al teorema del límite central, puesto que tenemos una muestra de tamaño grande y se desea realizar un test sobre la media. Aplicaremos un test de hipótesis (bilateral) de dos muestras independientes sobre la media y dado que no se conoce la varianza de la población utilizamo la distribución t.

Escribid la hipótesis nula y la alternativa

H0 : μ IniCost = μ UltCost

H1 : μ IniCost != μ UltCost

Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%. Estos cálculos deben ser acordes con el método (contraste) elegido.

Nota: se deben realizar los cálculos manualmente. No se pueden usar funciones de R que calculen directamente el contraste como t.test o similar. Sí se pueden usar funciones como mean, sd, qnorm, pnorm, qt y pt.

Comprobarmos si las varianzas son iguales para poder elegir las formulas adecuadas

```
var.test(claim$UltCost, claim$IniCost )
```

```
##
##  F test to compare two variances
##
## data:  claim$UltCost and claim$IniCost
## F = 1.3434, num df = 50525, denom df = 50525, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.320217 1.367077
## sample estimates:
## ratio of variances
##           1.343443
```

Comprobamos que el valor p-value es inferior al valor de significanza 0.05 por lo que podemos rechazar la hipótesis nula, igualdad de varianzas, y queda demostrado que las varianzas dos los data set son diferentes

```
function_bilateral <- function( dt1, dt2, C_Interval=0.95, var_equal=TRUE){
    mean1<-mean(dt1);    n1<-length(dt1);    sd1<-sd(dt1);
    mean2<-mean(dt2);    n2<-length(dt2);    sd2<-sd(dt2)
        # Elección de la fórmulas de la varianza.
    if (var_equal==TRUE){
      s <- sqrt( ((n1-1)*sd1^2 + (n2-1)*sd2^2 )/(n1+n2-2) )
```

```r
      t_denominator <- s*sqrt(1/n1 + 1/n2)
      df <- n1+n2-2
    }
    else{
      t_denominator <- sqrt( sd1^2/n1 + sd2^2/n2 )
      df <- ( (sd1^2/n1 + sd2^2/n2)^2 ) / ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1))
    }
    # Cálculos de valores a devolver.
    alfa <- (1-C_Interval);     t <- (mean1-mean2) / t_denominator
    t_critical <- qt( alfa/2, df, lower.tail=FALSE )
    p_value <- pt( abs(t), df, lower.tail=FALSE )*2
    # Par una mejor visualización lo introducimos en un dataframe.
    df_resutl <- data.frame(t, df, p_value,t_critical)

    return(df_resutl)
}
```

Ejecutamos la función.

```r
function_bilateral( claim$UltCost, claim$IniCost, var_equal=FALSE)
```

```
##          t       df      p_value t_critical
## 1 15.34648 98925.26 4.306053e-53   1.959988
```

##Conclusión

A partir de los valores obtenidos, explicad si podemos aceptar o rechazar la hipótesis planteada. También debéis responder la pregunta de investigación formulada.

Para un nivel confianza del 95% obtenemos un valor critico del 1.959988 y un valor t igual a 15.34648, por lo que nos entontramos en la zona de rechazo de la hipótesis nula. También queda demostrado a través del valor p siendo este inferior a 0.05 (4.306053e-53)

H1 : μ IniCost != μ UltCost

##Comprobación

Comprobar si los valores obtenidos coinciden con los de la función de R t.test.

```r
result <- t.test(claim$UltCost, claim$IniCost, var.equal=FALSE, alternative = 'two.sided')
result$p.value;
```

```
## [1] 4.306053e-53
```

```r
result
```

```
##
##  Welch Two Sample t-test
##
## data:  claim$UltCost and claim$IniCost
## t = 15.346, df = 98925, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1883.569 2435.136
## sample estimates:
## mean of x mean of y
## 10147.667  7988.315
```

Diferencia de salario según género

Existe una opinión generalizada que las mujeres cobran menos que los hombres. Vamos a comprobar qué dicen los datos al respecto. Nos preguntamos si las mujeres reciben un menor salario (WeeklyWages) que los hombres. Para ello, debéis obtener dos muestras. La primera muestra contiene todas las mujeres (Gender igual a F). La segunda muestra contiene todos los hombres

Análisis visual

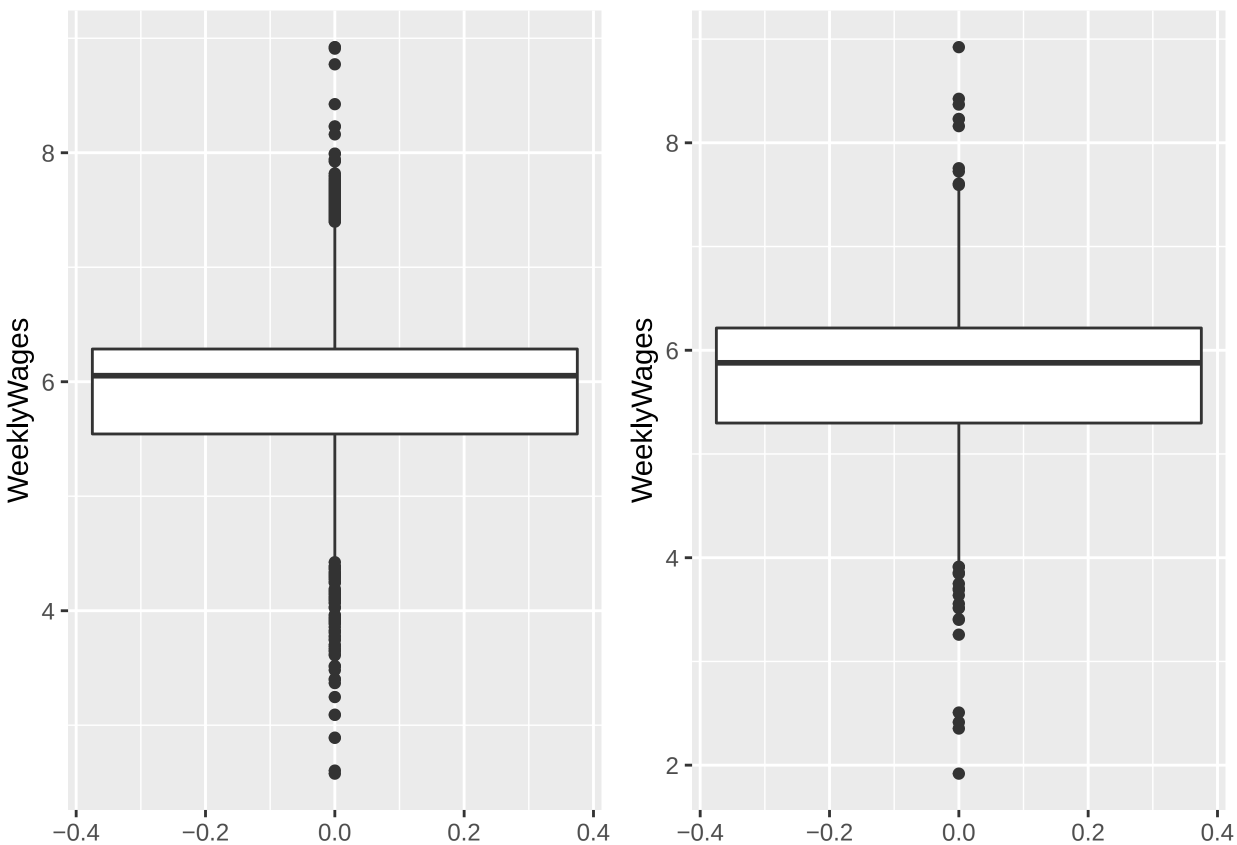
En primer lugar, mostrad en un diagrama de caja la distribución de la variable WeeklyWages (en logaritmos) según el género.

```r
# Filtrado por sexo.
df_hombres  <- claim[claim$Gender=="M",]
df_mujeres  <- claim[claim$Gender=="F",]

# Applicamos el logaritmo sín modificar el data ser original.
df_hombres$WeeklyWages  <- log(df_hombres$WeeklyWages)
df_mujeres$WeeklyWages  <- log(df_mujeres$WeeklyWages)

# plots diagrama de caja.
p1 <- ggplot( df_hombres, aes(y=WeeklyWages)) + geom_boxplot()
p2 <- ggplot( df_mujeres, aes(y=WeeklyWages)) + geom_boxplot()

grid.arrange(p1,p2, nrow=1)
```



Interpretación

Interpretad cualitativamente el gráfico mostrado en el apartado anterior, explicando si se pueden observar diferencias (visualmente) entre el salario de mujeres y hombres.

Se observa una pequeña diferencia en la tendencia entre el conjunto de muestras por genero, siendo sutilmente más elevada en el caso de los hombre. Pero no tenemos la certeza de que la diferencia observada entre ambos grupos sea a causa del genero y no, simplemente, de la aleatoriedad de las muestras (random chance). Ya que, el azar podría ser el causante de la diferencia observada. Necesitaremos una prueba que asegure que la diferencia de los grupos observada es más extrema que la misma que el azar podría producir

Escribid la hipótesis nula y la alternativa

Escribid las hipótesis nula y alternativa para la pregunta de investigación siguiente:

¿Podemos aceptar que los hombres cobran más que las mujeres en promedio a la semana?

Responded a la pregunta utilizando un nivel de confianza del 95%, usando la variable WeeklyWages en sus unidades originales (sin usar logaritmo).

H0 : μ *Salarios Hombres* = μ *Salarios Mujeres*

H1 : μ *Salarios Hombres* > μ *Salarios Mujeres*

Seguid los pasos que se detallan a continuación.

Justificación del test a aplicar

Explicad qué tipo de contraste se puede aplicar en este caso.

Asumimos normalidad gracias al teorema del límite central, puesto que tenemos una muestra de tamaño grande y se desea realizar un test sobre la media. Aplicaremos un test de hipótesis (unilateral por la derecha) de dos muestras independientes sobre la media y dado que no se conoce la varianza de la población utilizamos la distribución t.

Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%. Nota: se deben realizar los cálculos manualmente. Para el cálculo del contraste, implementad una función que os permita utilizarla en los siguientes apartados.

Comprobamos si las varianzas son iguales para poder elegir las formulas adecuadas adecuadas

```r
var.test( claim$WeeklyWages[claim$Gender=="M"], claim$WeeklyWages[claim$Gender=="F"])
```

```
##
##  F test to compare two variances
##
## data:  claim$WeeklyWages[claim$Gender == "M"] and claim$WeeklyWages[claim$Gender == "F"]
## F = 1.4084, num df = 38903, denom df = 11619, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.367605 1.450156
## sample estimates:
## ratio of variances
##           1.408441
```

Comprobamos que el valor p-value es inferior al valor de significanza 0.05 por lo que podemos rechazar la hipótesis nula, igualdad de varianzas, y queda demostrado que las varianzas dos los data set son diferentes

```r
function_greater <- function( dt1, dt2, C_Interval=0.95, var_equal=TRUE){
    mean1<-mean(dt1);    n1<-length(dt1);    sd1<-sd(dt1);
    mean2<-mean(dt2);    n2<-length(dt2);    sd2<-sd(dt2)

    # Elección de la fórmulas de la varianza.
    if (var_equal==TRUE){
      s <- sqrt( ((n1-1)*sd1^2 + (n2-1)*sd2^2 )/(n1+n2-2) )
      t_denominator <- s*sqrt(1/n1 + 1/n2)
      df <- n1+n2-2
    }
    else{
      t_denominator <- sqrt( sd1^2/n1 + sd2^2/n2 )
      df <- ( (sd1^2/n1 + sd2^2/n2)^2 ) / ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1))
    }

    # Cálculos de valores a devolver.
    alfa <- (1-C_Interval);     t <- (mean1-mean2) / t_denominator
    t_critical <- qt( alfa, df, lower.tail=FALSE )
    p_value<-pt( t, df, lower.tail=FALSE )

    # Par una mejor visualización lo introducimos en un dataframe.
    df_resutl <- data.frame(t, df, p_value,t_critical)

    return(df_resutl)
}
```

```r
function_greater(claim$WeeklyWages[claim$Gender=="M"], claim$WeeklyWages[claim$Gender=="F"], var_equal=F
```

```
##         t       df       p_value t_critical
## 1 28.8127 22273.68 1.437086e-179   1.644922
```

Conclusión

A partir de los valores obtenidos, debéis concluir si podemos aceptar o rechazar la hipótesis. Asimismo, responded la pregunta formulada.

Para un nivel confianza del 95% obtenemos un valor critico del 1.644922 y un valor t (valor observado) igual a 28.8127, por lo que nos entontramos en la zona de rechazo de la hipotesis nula. También queda demostrado a través del valor p siendo este inferior a 0.05 (1.437086e-179)

H1 : μ Salarios Hombres > μ Salarios Mujeres

Comprobación

Comprobar si los valores obtenidos coinciden con los de la función de R t.test.

```r
df_hombres  <- claim[claim$Gender=="M",]
df_mujeres  <- claim[claim$Gender=="F",]

resutl <- t.test( df_hombres$WeeklyWages,df_mujeres$WeeklyWages, var.equal=FALSE, alternative = "greate
resutl$p.value
```

```
## [1] 1.437086e-179
```

```r
resutl
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  df_hombres$WeeklyWages and df_mujeres$WeeklyWages
## t = 28.813, df = 22274, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  64.18029      Inf
## sample estimates:
## mean of x mean of y
##  449.4311  381.3649
```

Salario semanal (II)

En este apartado, seguimos interesados en investigar si los hombres tienen un salario mayor a las mujeres, y si éste es mayor en una cantidad de 50 euros como mínimo. Por tanto, la pregunta que realizamos es:

¿Podemos aceptar que los hombres cobran al menos 50 euros más que las mujeres en promedio a la semana?

Seguid los pasos que se indican a continuación.

Escribid la hipótesis nula y la alternativa

H0 : μ *Salarios Hombres* = μ *Salarios Mujeres*

H1 : μ *Salarios Hombres* > μ *Salarios Mujeres*

Justificación del test a aplicar

Justificad qué tipo de contraste podemos aplicar en este caso.

Aplicamos el mismo test del apartado anterior.

Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%. Se recomienda usar las funciones desarrolladas en apartados anteriores, si éstas son útiles para este contraste.

Comprobarmos si las varianzas son iguales para poder elegir las formulas adecuadas adecuadas, aunque ya quedaron demostradas en el apartado anterior.

```r
df_hombres  <- claim[claim$Gender=="M",]
df_mujeres  <- claim[claim$Gender=="F",]

#mean(df_hombres$WeeklyWages)
df_hombres$WeeklyWages <- df_hombres$WeeklyWages - 50
# mean(df_hombres$WeeklyWages) # Media desplazada 50 Euros

var.test(df_hombres$WeeklyWages, df_mujeres$WeeklyWages)
```

```
##
##  F test to compare two variances
##
## data:  df_hombres$WeeklyWages and df_mujeres$WeeklyWages
## F = 1.4084, num df = 38903, denom df = 11619, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.367605 1.450156
## sample estimates:
## ratio of variances
##           1.408441
```

```r
function_greater( df_hombres$WeeklyWages, df_mujeres$WeeklyWages, var_equal=FALSE)
```

```
##          t       df      p_value t_critical
## 1 7.647497 22273.68 1.066063e-14   1.644922
```

Conclusión

A partir de los valores obtenidos, responded si podemos aceptar o rechazar la hipótesis y responded la pregunta formulada.

Para un nivel confianza del 95% obtenemos un valor critico del 1.644922 y un valor t (valor observado) igual a 7.647497, por lo que nos encontramos en la zona de rechazo de la hipótesis nula. También queda demostrado a través del valor p siendo este inferior a 0.05 (1.066063e-14)

H1 : μ Salarios Hombres > μ Salarios Mujeres

Comprobación

Comprobar si los valores obtenidos coincicen con los de la función de R t.test.

```
resutl <- t.test( df_hombres$WeeklyWages, df_mujeres$WeeklyWages, var.equal=FALSE, alternative = "great
resutl$p.value
```

```
## [1] 1.066063e-14
```

```
resutl
```

```
##
##  Welch Two Sample t-test
##
## data:  df_hombres$WeeklyWages and df_mujeres$WeeklyWages
## t = 7.6475, df = 22274, p-value = 1.066e-14
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  14.18029      Inf
## sample estimates:
## mean of x mean of y
##  399.4311  381.3649
```

Diferencia de jornada según género

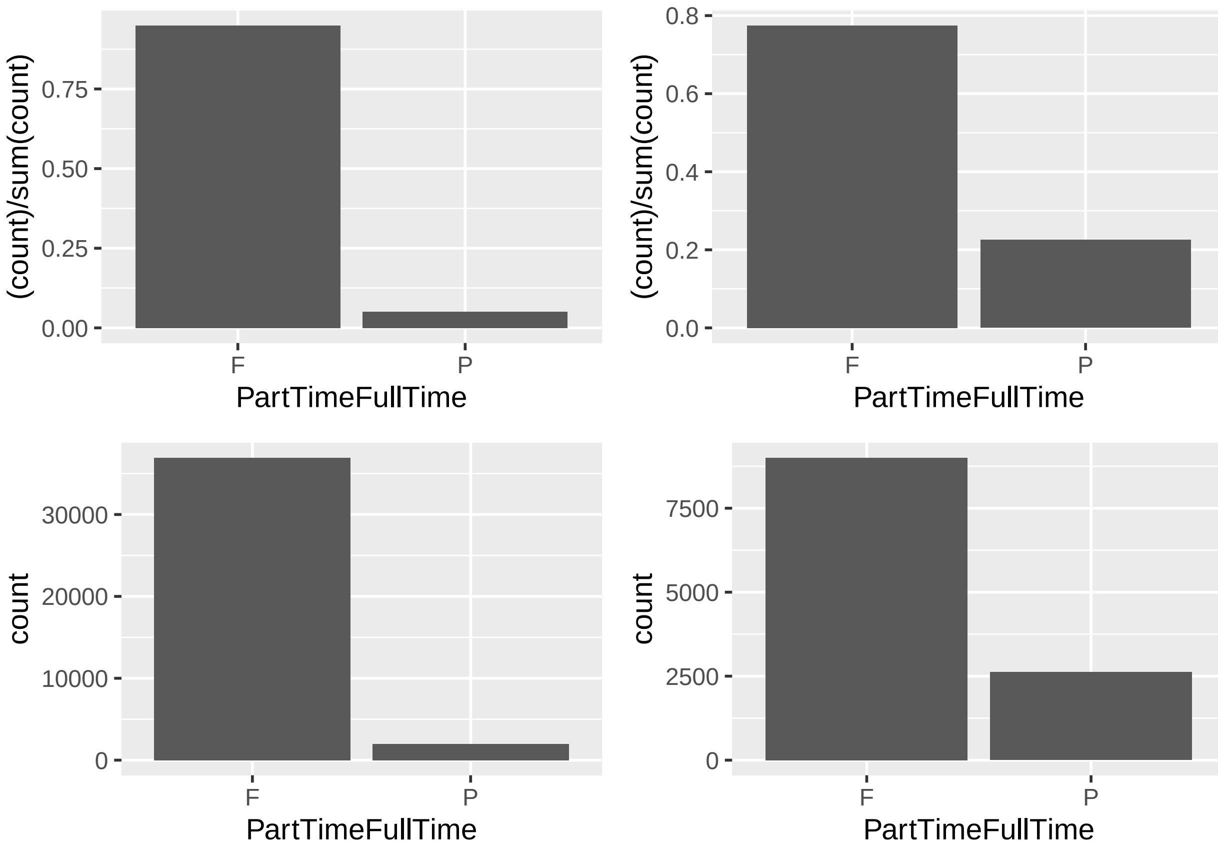
Existe una opinión generalizada que las mujeres tienden a utilizar más la jornada a tiempo parcial. Vamos a comprobar qué dicen los datos al respecto. Nos preguntamos si las mujeres realizan más frecuentemente una jornada a tiempo parcial PartTimeFullTime que los hombres.

Análisis visual

Mostrad un un diagrama de barras que muestre los porcentajes de cada categoría de la variable PartTimeFullTime según el género.

```
p1<-ggplot(claim[claim$Gender=="M",], aes(x=PartTimeFullTime)) + geom_bar(aes(y = (..count..)/sum(..cour
p2<-ggplot(claim[claim$Gender=="F",], aes(x=PartTimeFullTime)) + geom_bar(aes(y = (..count..)/sum(..cour
p3<-ggplot(claim[claim$Gender=="M",], aes(x=PartTimeFullTime)) + geom_bar()
p4<-ggplot(claim[claim$Gender=="F",], aes(x=PartTimeFullTime)) + geom_bar()

grid.arrange(p1,p2,p3,p4, nrow=2)
```



Interpretación

Interpretad los resultados y dad respuesta a la pregunta planteada.

Se puede observar que en proporción hay más mujeres que utilizan la jornada partida que la completa. Y en términos absolutos vemos que hay un número parecido de personas que utilizan la jornada partida.

Hipótesis nula y alternativa

La pregunta que realizamos sobre los datos es:

¿La proporción de personas que trabajan a tiempo completo es diferente para hombres que para mujeres?

Escribid la hipótesis nula y la alternativa teniendo en cuenta la pregunta formulada.

H0 : p JComp. H. = p JComp. M.

H1 : P JComp. H. != P JComp. M.

Tipo de test

Indicad qué tipo de test aplicaréis y justificadlo.

Aplicamos un test para la diferencia de dos proporciones independientes. Por un lado la proporción de casos en los que el hombre utiliza la jornada completa. Y lo mismo para genero femenino. Se obtienen dos proporciones p1 y p2 y se compara si la primera es significativamente diferente de la segunda (bilateral)

Cálculos

Realizad todos los cálculos con instrucciones propias. Calculad el valor observado, el valor crítico y el valor p. Mostrad los resultados.

```
df_hombres  <- claim[claim$Gender=="M",]
df_mujeres  <- claim[claim$Gender=="F",]
n1 <- nrow(claim[claim$Gender=="M",])
n2 <- nrow(claim[claim$Gender=="F",])
p1 <- sum(df_hombres$PartTimeFullTime == 'F') / n1
p2 <- sum(df_mujeres$PartTimeFullTime == 'F') / n2
p1; n1; p2; n2
```

```
## [1] 0.9489513

## [1] 38904

## [1] 0.7744406

## [1] 11620
```

```
p <- ((n1*p1 + n2*p2)/(n1+n2))
z <- ((p1-p2) / sqrt(p*(1-p)*(1/n1 + 1/n2)))
z_crit <- qnorm(0.025)
p_value <- pnorm(abs(z),lower.tail=FALSE)*2
z; z_crit; p_value
```

```
## [1] 57.3423

## [1] -1.959964

## [1] 0
```

Conclusión

A partir de los valores obtenidos, responded la pregunta formulada.

Debido a que p es menor alfa=0.05, estamos en la zona de rechazo de la hipótesis nula y podemos afirmar la diferencias de proporciones es significativamente diferente con un nivel de confianza del 95%.

Confirmamos que la jornada de trabajo a tiempo completo es diferente para hombres que para mujeres, dado el procedimiento de muestreo podemos

Comprobación

Comprobar si los valores obtenidos coinciden con los de la función de R prop.test.

```r
success <- c(p1*n1,p2*n2)
n <- c(n1,n2)
resutl <- prop.test( success, n, alternative="two.side", correct=FALSE)
resutl$p.value
```

```
## [1] 0
```

```r
resutl
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of n
## X-squared = 3288.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1666030 0.1824183
## sample estimates:
##    prop 1    prop 2
## 0.9489513 0.7744406
```

Salario por hora

Anteriormente hemos comparado el salario semanal entre hombres y mujeres. Ahora bien, la pregunta que nos hacemos ahora es:

¿Podemos afirmar que los hombres cobran más que las mujeres por hora trabajada?

Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa teniendo en cuenta la pregunta formulada.

H0 : p Exhr Hombres = p Exhr Mujeres

H1 : P Exhr Hombres > P Exhr Mujeres

Tipo de test

Indicad qué tipo de test aplicaréis y justificadlo.

Asumimos normalidad gracias al teorema del límite central, puesto que tenemos una muestra de tamaño grande y se desea realizar un test sobre la media. Aplicaremos un test de hipótesis (unilateral por la derecha) de dos muestras independientes sobre la media y dado que no se conoce la varianza de la población utilizamos la distribución t.

Cálculos

Calculad el estadístico de contraste, el valor crítico y el valor p con un nivel de confianza del 95%. Para realizar estos cálculos, usad la función que habéis implementado previamente.

```
df_hombres_hrWk  <- claim$WeeklyWages[claim$Gender=="M"]/claim$HoursWeek[claim$Gender=="M"]
df_mujeres_hrWk  <- claim$WeeklyWages[claim$Gender=="F"]/claim$HoursWeek[claim$Gender=="F"]

var.test(df_hombres_hrWk, df_mujeres_hrWk)
```

```
##
##  F test to compare two variances
##
## data:  df_hombres_hrWk and df_mujeres_hrWk
## F = 0.624, num df = 38903, denom df = 11619, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6059053 0.6424789
## sample estimates:
## ratio of variances
##          0.6239975
```

```
result <- function_greater(df_hombres_hrWk, df_mujeres_hrWk, var_equal=FALSE)
result
```

```
##           t       df  p_value t_critical
## 1 0.7098522 16185.74 0.238903   1.644948
```

Conclusión

A partir de los valores obtenidos, responded la pregunta formulada.

Para un nivel confianza del 95% obtenemos un valor critico del 1.644948 y un valor t (valor observado) igual a 0.7098522, por lo que nos entontramos en la zona de aceptación de la hipótesis nula. También queda demostrado a través del valor p siendo este superior a 0.05 (0.238903)

H0 : p Exhr Hombres = p Exhr Mujeres

Comprobación

Comprobar si los valores obtenidos coincicen con los de la función de R t.test.

```
resutl <- t.test( df_hombres_hrWk, df_mujeres_hrWk, var.equal=FALSE, alternative = "greater")
resutl$p.value
```

```
## [1] 0.238903
```

```
resutl
```

```
##
##  Welch Two Sample t-test
##
## data:  df_hombres_hrWk and df_mujeres_hrWk
## t = 0.70985, df = 16186, p-value = 0.2389
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.08790322         Inf
## sample estimates:
## mean of x mean of y
##  11.83031  11.76358
```

Resumen ejecutivo

Resumid las conclusiones principales del análisis. Para ello, podéis resumir las conclusiones de cada uno de los apartados.

3. ¿Podemos aceptar que no hay diferencias entre IniCost y UltCost?

Podemos afirmar que existe una diferencia entre el coste estimado por la empresa y valor del coste final con un nivel de confianza del 95%.

4. ¿Podemos aceptar que los hombres cobran más que las mujeres en promedio a la semana?

Con un nivel de confianza del 95%, se puede afirmar que los hombres cobran en promedio semanal más que las mujeres.

5. ¿Podemos aceptar que los hombres cobran al menos 50 euros más que las mujeres en promedio a la semana?

sí, también queda demostrado que la diferencia es como mínimo superior a 50 euros.

6. ¿las mujeres realizan más frecuentemente una jornada a tiempo parcial PartTimeFullTime que los hombres?

Existe una diferencia proporcional en el uso de la jornada a tiempo parcial entre las mujeres y los hombres con un nivel de confianza del 95%. De las gráficas también se observa que los hombres hacen un uso superior de la jornada a tiempo completo

7. ¿Podemos afirmar que los hombres cobran más que las mujeres por hora trabajada?

No se observan diferencias en la proporción de hombre que cobren más con respecto a las mujeres con un nivel de confianza del 95%.