Universitat Oberta de Catalunya

Patricia Lázaro Tello
plazarotello@uoc.edu

A2 - Analítica descriptiva e inferencial

Patricia Lázaro Tello

Índice general

Patricia Lázaro Tello
plazarotello@uoc.edu

Apéndice: Funciones utilizadas

Mostrar área de confianza

```
confidence.plot <- function(type){
  ic.x <- seq(-4, 4, by=0.01)
  ic.y <- dnorm(ic.x, mean=0, sd=1)

  ic.xmin <- -2.0; ic.xmax <- 2.0
  ic.ymin <- ic.y[which(ic.x == ic.xmin)]
  ic.ymax <- ic.y[which(ic.x == ic.xmax)]
  ic.factual.x <- ifelse(switch(type, left = ic.x>ic.xmin,
                          two.sided = ic.x>ic.xmin & ic.x<ic.xmax,
                          right = ic.x<ic.xmax), ic.x, NA)

  ic.plot <- ggplot(mapping=aes(x=ic.x, y=ic.y)) +
    geom_area(mapping = aes(x = ic.factual.x), fill = default.color.main,
              alpha=0.75) + geom_line(size=1.05) +
    annotate('text', label='Área de \nconfianza', x=0,
             y=ic.y[which(ic.x==0)]/2, size=5, color='black') +
    no.axis.y + ylab('') + xlab('') + xlim(c(-3.5, 3.5))

  if(type %in% c('two.sided', 'left')){
    ic.plot <- ic.plot + geom_segment(aes(x=ic.xmin, y=0, xend=ic.xmin,
                                          yend=ic.ymin), size=1.05) +
      annotate('text', x=ic.xmin, y=ic.ymin + 0.05, parse=TRUE, size=4,
               label=switch(type, two.sided='frac(~alpha,2)', left='~alpha'))
  }
  if(type %in% c('two.sided', 'right')){
    ic.plot <- ic.plot + geom_segment(aes(x=ic.xmax, y=0, xend=ic.xmax,
                                          yend=ic.ymax), size=1.05) +
      annotate('text', x=ic.xmax, y=ic.ymax + 0.05, parse=TRUE, size=4,
               label=switch(type, two.sided='1-frac(~alpha,2)',
                            right='1-~alpha'))
  }
  return(ic.plot)
}
```

Patricia Lázaro Tello
plazarotello@uoc.edu

Cálculo de intervalos de confianza

```
confidence.interval.mean <- function(nc, dist_mean, dist_sd, dist_n, type){
  alpha <- 1.0 - nc
  se <- dist_sd / sqrt(dist_n)
  val <- switch(type, two.sided=alpha/2, right=1-alpha, left=alpha)
  z <- qt(val, df=dist_n-1, lower.tail=FALSE)

  ic.1 <- dist_mean - z*se
  ic.2 <- dist_mean + z*se
  if(ic.1 > ic.2){ tmp.ic <- ic.2; ic.2 <- ic.1; ic.1 <- tmp.ic; }

  return(switch(type, two.sided=c(ic.1, ic.2),
                right=c(ic.1, Inf), left=c(-Inf, ic.2)))
}
```

```
confidence.interval.2.means <- function(nc, mean1, mean2, se1, se2, v, type){
  alpha <- 1.0 - nc
  val <- switch(type, two.sided=alpha/2, right=1-alpha, left=alpha)
  z <- qt(val, df=v, lower.tail=FALSE)

  ic.1 <- (mean1-mean2) - z*sqrt(se1+se2)
  ic.2 <- (mean1-mean2) + z*sqrt(se1+se2)
  if(ic.1 > ic.2){ tmp.ic <- ic.2; ic.2 <- ic.1; ic.1 <- tmp.ic; }

  return(switch(type, two.sided=c(ic.1, ic.2),
                right=c(ic.1, Inf), left=c(-Inf, ic.2)))
}
```

```
confidence.interval.2.props <- function(nc, p1, p2, n1, n2, type){
  alpha <- 1-nc
  z <- qnorm(switch(type, two.sided=alpha/2, right=1-alpha, left=alpha),
             lower.tail=FALSE)
  ic.1 <- (p1-p2) - z * sqrt((p1*(1-p1)/n1) + (p2*(1-p2)/n2))
  ic.2 <- (p1-p2) + z * sqrt((p1*(1-p1)/n1) + (p2*(1-p2)/n2))

  if(ic.1 > ic.2){ tmp.ic <- ic.2; ic.2 <- ic.1; ic.1 <- tmp.ic; }

  return(switch(type, two.sided=c(ic.1, ic.2),
                right=c(ic.1, Inf), left=c(-Inf, ic.2)))
}
```

Patricia Lázaro Tello
plazarotello@uoc.edu

Contraste de hipótesis de igualdad de varianzas

```
variance.equals.2.samples <- function(nc, dist1, dist2){
  alpha <- 1-nc

  n1 <- length(dist1); n2 <- length(dist2)
  sd1 <- sd(dist1); sd2 <- sd(dist2)
  fobs <- sd1^2 / sd2^2

  fcrit.lower <- qf(alpha/2, df1=n1-1, df2=n2-1)
  fcrit.upper <- qf(1-alpha/2, df1=n1-1, df2=n2-1)

  pvalue <- min(pf(fobs, df1=n1-1, df2=n2-1, lower.tail=FALSE),
                pf(fobs, df1=n1-1, df2=n2-1, lower.tail=TRUE))*2

  return(list(fobs=fobs, fcrit=c(fcrit.lower, fcrit.upper), pvalue=pvalue))
}
```

Contraste de hipótesis de dos muestras independientes sobre la media con varianzas desconocidas diferentes

```
means.contrast.2.samples.vars.unknown.dif <- function(x1, x2, dif, nc, type){
  x1.media <- mean(x1); x1.n <- length(x1); x1.sd <- sd(x1);
  x1.se <- x1.sd^2/x1.n

  x2.media <- mean(x2); x2.n <- length(x2); x2.sd <- sd(x2);
  x2.se <- x2.sd^2/x2.n

  alfa <- 1-nc
  t <- (x1.media - x2.media - dif) / sqrt(x1.se + x2.se)
  v <- ceiling((x1.se + x2.se)^2 / ((x1.se)^2/(x1.n-1) + (x2.se)^2/(x2.n-1)))
  pvalue <- pt(t, df=v, lower.tail=switch(type, two.sided=FALSE,
                                          right=FALSE, left=TRUE))
  tcrit <- qt(switch(type, two.sided=alfa/2, right=1-alfa, left=alfa), v)
  ic <- confidence.interval.2.means(nc, x1.media, x2.media, x1.se, x2.se,
                                    v, type)

  return(list(t=t, v=v, p=pvalue, tcrit=tcrit, ic=ic))
}
```

Patricia Lázaro Tello
plazarotello@uoc.edu

1 Carga del archivo

Se procede a la carga del archivo `train_clean2.csv` y a la visualización de sus datos.

```
claim <- read.csv(file = "train_clean2.csv", header = TRUE)

claim.rows <- dim(claim)[1]
claim.cols <- dim(claim)[2]

head(claim, n = 3L)
```

```
##   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
## 3 3   WC5481426 1996-03-25T00:00:00Z 1996-04-14T00:00:00Z  30      M
##   MaritalStatus DependentChildren DependentsOther WeeklyWages PartTimeFullTime
## 1             M                 0               0      500.00                F
## 2             M                 0               0      509.34                F
## 3             M                 0               0      709.10                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
## 3      38.0        5                          CUT ON SHARP EDGE CUT LEFT THUMB
##   IniCost UltCost Time
## 1    1500    4303   87
## 2    5500    6106   13
## 3    1700    2099   20
```

El fichero contiene 50526 registros con 17 atributos. En este documento se van a analizar las variables `IniCost` (coste estimado del siniestro), `UltCost` (coste real del siniestro), `WeeklyWages` (salario semanal) y `PartTimeFullTime` (tipo de jornada: jornada parcial o completa).

2 Coste de los siniestros

El coste de los siniestros está representado en la variable `UltCost` en el fichero de datos. Sus características principales son:

```
summary(claim$UltCost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       6    1183    3291   10148    9226  492515
```

```
tail(sort(unique(claim$UltCost)), 10)
```

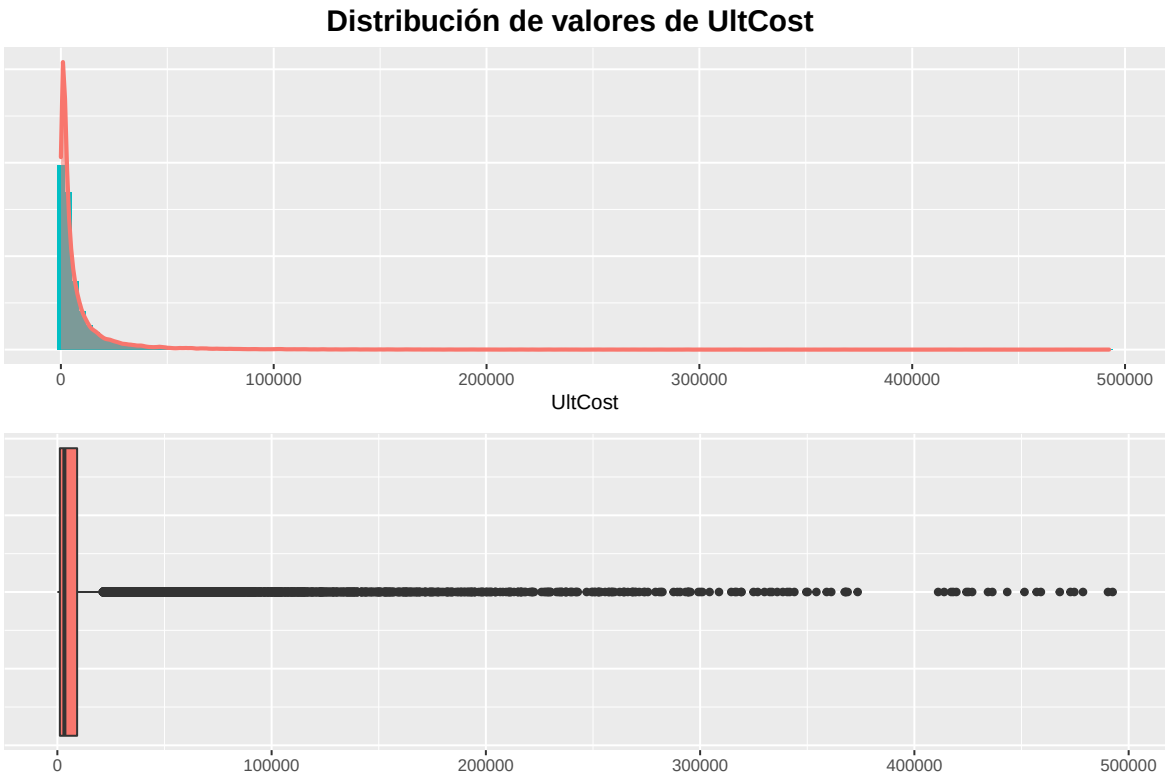
```
##  [1] 443334 451403 457061 459027 467861 472843 474644 478694 490368 492515
```

La mayoría de los valores se concentran en el rango (0, 10.000)€, con algunos valores por encima de los 100.000€. Se procede a realizar un análisis visual más exhaustivo para comprobar la largura de la cola derecha y observar en global la distribución de la variable.

2.1 Análisis visual

```
annotate_figure(ggarrange(
  ggplot(data=claim, mapping=aes(x=UltCost)) +
    geom_histogram(fill=default.color.secondary, bins=150,
                   mapping=aes(y=..density..)) +
    geom_density(alpha=0.5, size=1.05, color=default.color.main,
                 fill=default.color.main) + ylab('') + no.axis.y,
  ggplot(data=claim, mapping=aes(x=UltCost)) +
    geom_boxplot(fill=default.color.main) + no.axis.y + xlab(''),
  nrow=2, ncol=1, align='hv'),
  top=text_grob('Distribución de valores de UltCost', face='bold', size=16))
```

Patricia Lázaro Tello
plazarotello@uoc.edu



La variable `UltCost` presenta una asimetría a la derecha (o asimetría positiva, o sesgo a la derecha) pronunciado, con una cola muy larga.

El análisis visual ratifica las conclusiones obtenidas del resumen de características de la variable visto anteriormente: los valores se encuentran concentrados mayormente en un rango pequeño, con 5598 valores anormales de hasta 500.000€ (frente a los 44928 valores normales).

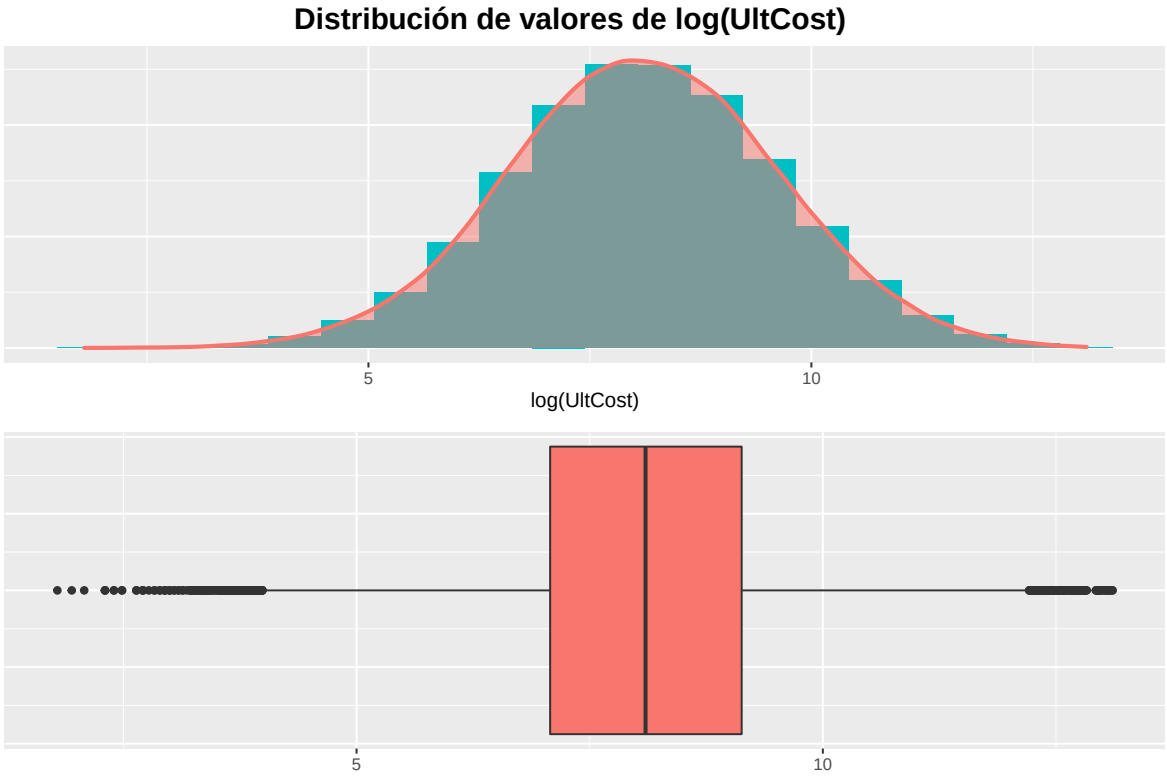
Para reducir el sesgo a la derecha de la variable, se plantea aplicar una transformación logarítmica. Esta transformación tiene, además, otro objetivo: conseguir una distribución de la variable transformada más cerca de la normal.

```
claim <- claim %>% dplyr::mutate(UltCost.log = log(UltCost))
summary(claim$UltCost.log)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.792   7.076   8.099   8.101   9.130  13.107
```

```
annotate_figure(ggarrange(
  ggplot(data=claim, mapping=aes(x=UltCost.log)) +
    geom_histogram(fill=default.color.secondary, bins=20,
```

```
                  mapping=aes(y=..density..)) +
  geom_density(alpha=0.5, size=1.05, color=default.color.main,
               fill=default.color.main) +
  ylab('') + no.axis.y + xlab('log(UltCost)'),
ggplot(data=claim, mapping=aes(x=UltCost.log)) +
  geom_boxplot(fill=default.color.main) + no.axis.y + xlab(''),
nrow=2, ncol=1, align='hv'),
top=text_grob('Distribución de valores de log(UltCost)', face='bold',
              size=16))
```



Se comprueba visualmente que la transformación logarítmica ha eliminado el sesgo a la derecha. A su vez, los datos se encuentran más centrados, con 50182 valores dentro del rango normal y 344 valores fuera del rango normal.

2.2 Comprobación de normalidad

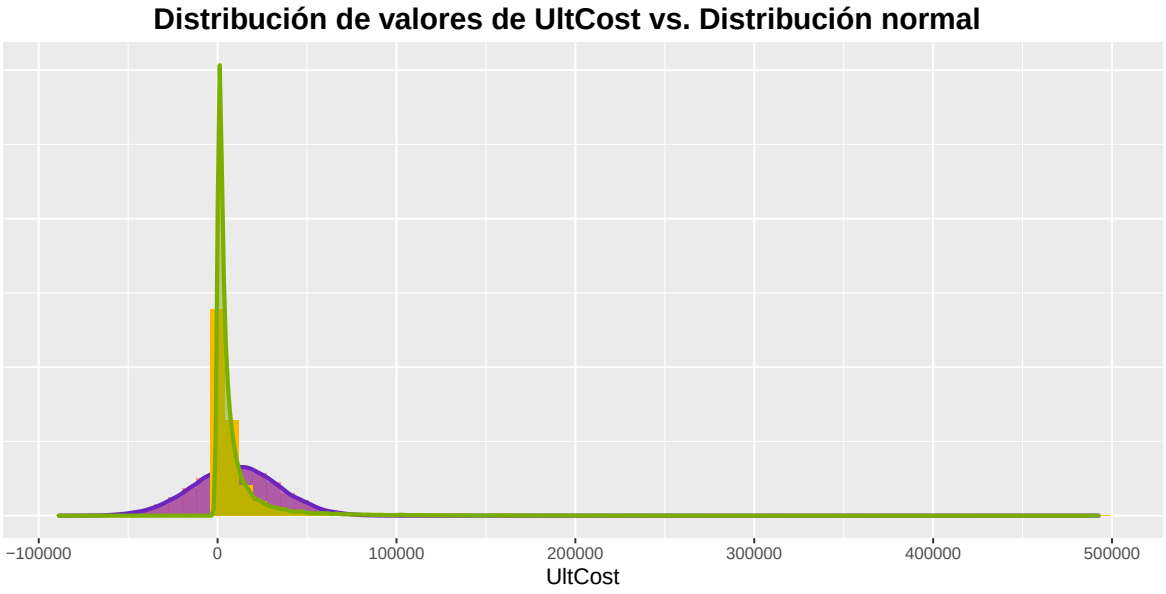
A continuación se plantea comprobar si `UltCost` y $log(UltCost)$ siguen una distribución normal. A priori, observando las gráficas anteriores, se podría asumir que `UltCost`

no sigue una distribución normal; $log(UltCost)$, por el contrario, sí parece seguirla.

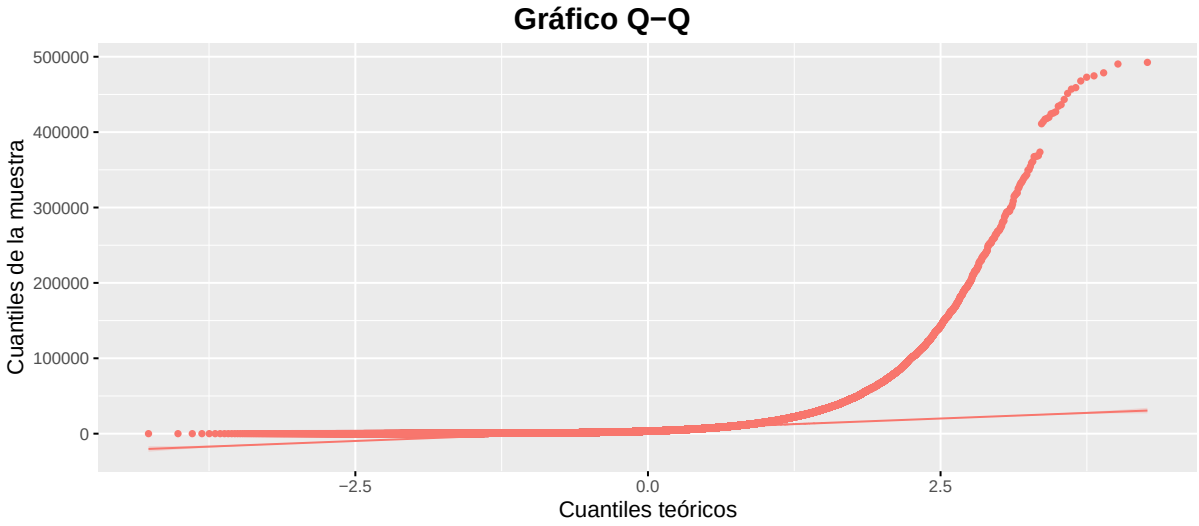
2.2.1 `UltCost`

```
ultcost.n <- length(claim$UltCost)
ultcost.mean <- mean(claim$UltCost)
ultcost.sd <- sd(claim$UltCost)

ultcost.normal.dist <- rnorm(n=ultcost.n, mean=ultcost.mean, sd=ultcost.sd)
```

```
ggplot(data=claim, mapping=aes(x=UltCost)) +
  geom_histogram(fill=default.color.main, bins=75, alpha=0.75,
                 mapping=aes(x=ultcost.normal.dist, y=..density..)) +
  geom_density(alpha=0.5, size=1.05, color=default.color.quat,
               fill=default.color.quat,
               mapping=aes(x=ultcost.normal.dist, y=..density..)) +

  geom_histogram(fill=default.color.cinq, bins=75,
                 mapping=aes(y=..density..)) +
  geom_density(alpha=0.5, size=1.05, color=default.color.terciary,
               fill=default.color.terciary) +
  ylab('') + no.axis.y + title.centered +
  ggtitle('Distribución de valores de UltCost vs. Distribución normal')
```



```
ggqqplot(claim$UltCost, color=default.color.main, ggtheme = theme_gray(),
         xlab='Cuantiles teóricos', ylab='Cuantiles de la muestra',
         title='Gráfico Q-Q', shape=16) +
  title.centered
```



Por análisis visual ya se puede afirmar con seguridad que la variable `UltCost` no sigue una distribución normal. Se contrasta con el test de normalidad de Lilliefors (Kolmogorov-Smirnov).

Se plantea la siguiente hipótesis nula (se supone normalidad) e hipótesis alternativa (se desea comprobar si la variable no sigue una distribución normal):

$$H_0 : \mu_{UltCost} = \mu_{normal} \wedge \sigma_{UltCost} = \sigma_{normal}$$
$$H_1 : \mu_{UltCost} \neq \mu_{normal} \vee \sigma_{UltCost} \neq \sigma_{normal}$$

```
ultcost.norm.test <- lillie.test(claim$UltCost)
ultcost.norm.test
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  claim$UltCost
## D = 0.33597, p-value < 0.00000000000000022
```

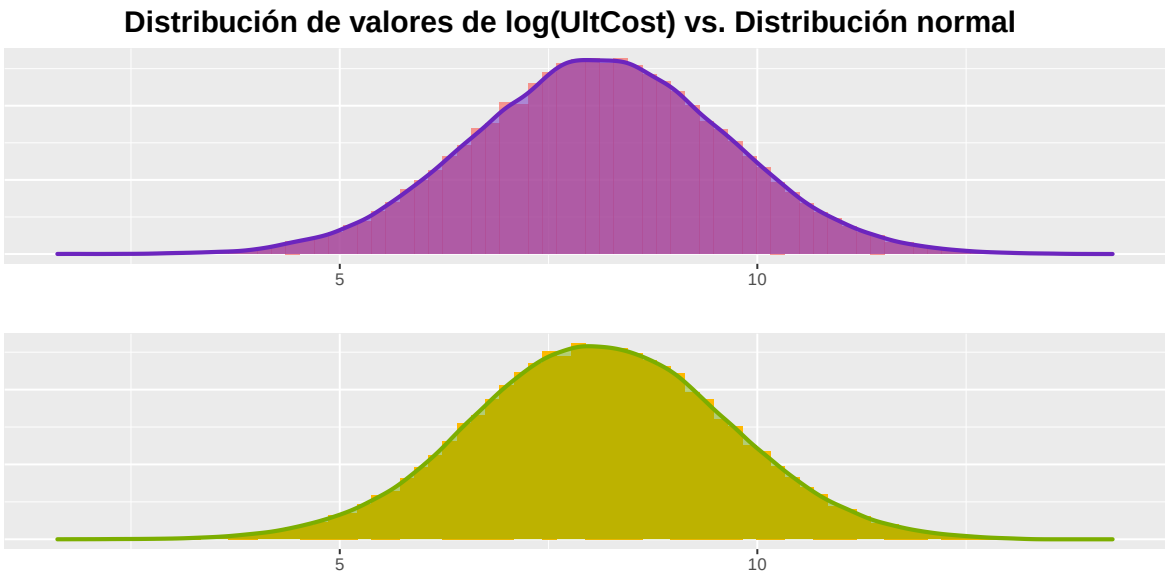
Con un nivel de significación $\alpha = 0.05$ se rechaza la hipótesis nula (H_0), dado que $p - value = 0e + 00 \ll \alpha = 0.05$; por tanto, se puede concluir que `UltCost` **no**

sigue una distribución normal.

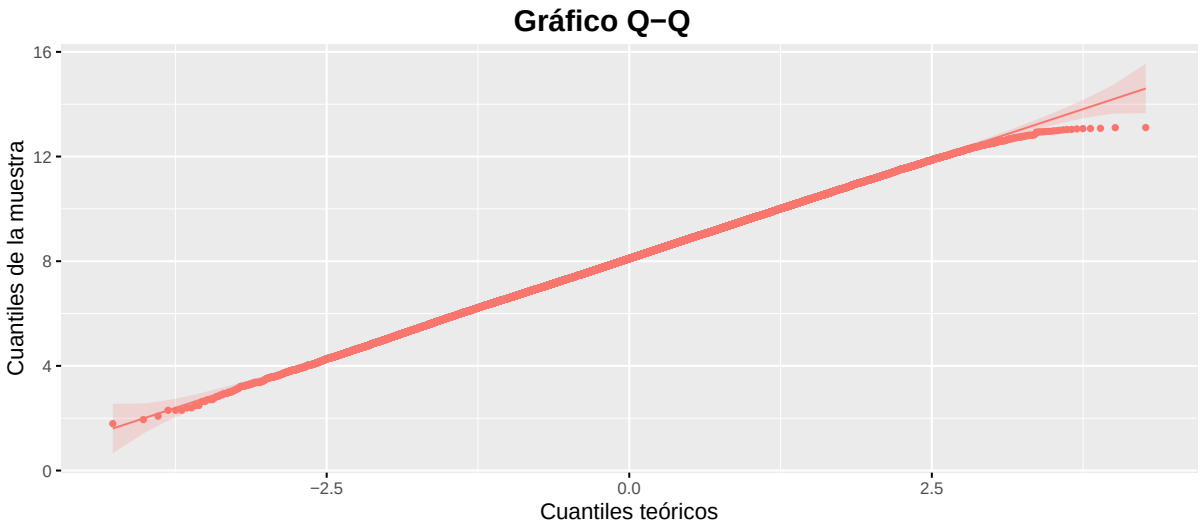
2.2.2 $log(UltCost)$

```
ultcost.log.n <- length(claim$UltCost.log)
ultcost.log.mean <- mean(claim$UltCost.log)
ultcost.log.sd <- sd(claim$UltCost.log)
ultcost.normal.dist <- rnorm(n=ultcost.log.n, mean=ultcost.log.mean,
                             sd=ultcost.log.sd)
```

```
tmp.xmin <- min(ultcost.normal.dist, claim$UltCost.log)
tmp.xmax <- max(ultcost.normal.dist, claim$UltCost.log)
annotate_figure(
  ggarrange(nrow=2, ncol=1, align='hv',
            ggplot(mapping=aes(x=ultcost.normal.dist)) +
              geom_histogram(fill=default.color.main, bins=75, alpha=0.75,
                             mapping=aes(y=..density..)) +
              geom_density(alpha=0.5, size=1.05, color=default.color.quat,
                           fill=default.color.quat,
                           mapping=aes(y=..density..)) +
              xlim(tmp.xmin, tmp.xmax) + xlab('') + ylab('') + no.axis.y,
            ggplot(data=claim, mapping=aes(x=UltCost.log)) +
              geom_histogram(fill=default.color.cinq, bins=75,
                             mapping=aes(y=..density..)) +
              geom_density(alpha=0.5, size=1.05, color=default.color.terciary,
                           fill=default.color.terciary) +
              xlim(tmp.xmin, tmp.xmax) + xlab('') + ylab('') + no.axis.y),
  top=text_grob(size=16, face='bold',
                paste('Distribución de valores de log(UltCost) vs.',
                      'Distribución normal')))
```




```
ggqqplot(claim$UltCost.log, color=default.color.main, ggtheme = theme_gray(),
         xlab='Cuantiles teóricos', ylab='Cuantiles de la muestra',
         title='Gráfico Q-Q', shape=16) +
  title.centered
```



Por análisis visual no se puede afirmar que la transformación logarítmica de la variable `UltCost` no siga una distribución normal (véase la cola del gráfico Q-Q). Se contrasta con el test de normalidad de Lilliefors (Kolmogorov-Smirnov), formulando la hipótesis nula y alternativa como sigue:

$$H_0 : \mu_{log(UltCost)} = \mu_{normal} \wedge \sigma_{log(UltCost)} = \sigma_{normal}$$
$$H_1 : \mu_{log(UltCost)} \neq \mu_{normal} \vee \sigma_{log(UltCost)} \neq \sigma_{normal}$$

```
ultcost.log.norm.test <- lillie.test(claim$UltCost.log)
ultcost.log.norm.test
```

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

Con un nivel de significación $\alpha = 0.05$ se acepta la hipótesis nula (H_0), dado que $p - value = 0.375 \gg \alpha = 0.05$; por tanto, se puede concluir que $log(UltCost)$ **sigue una distribución normal**.

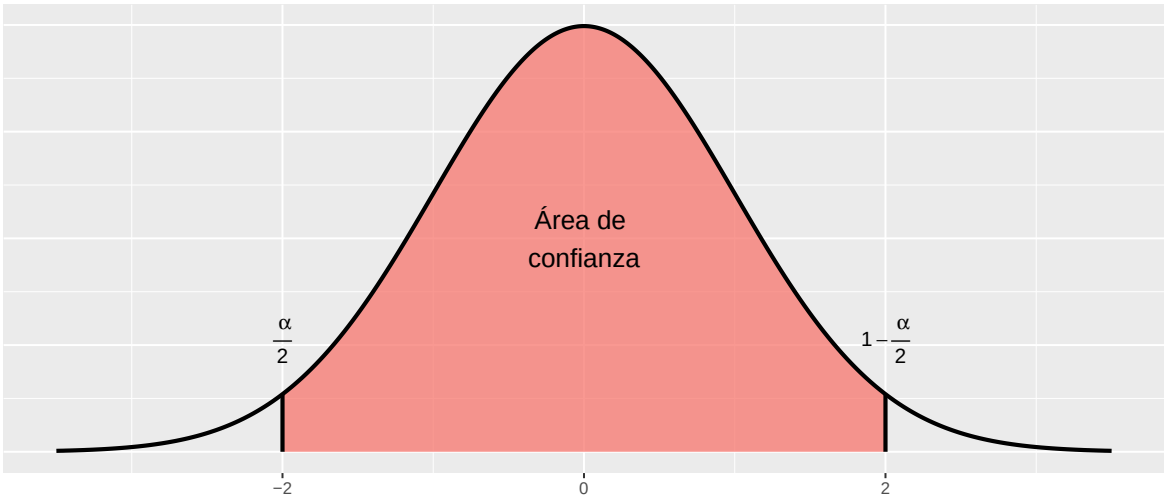
2.3 Intervalo de confianza de la media poblacional de la variable `UltCost`

Tras realizar el contraste de normalidad de la variable `UltCost` (véase *Comprobación de normalidad de* `UltCost`), se asume que la variable no sigue una distribución normal. Además, se desconoce la varianza dado que se trata de una muestra y no una población.

La variable $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ se comporta como una distribución t de Student con $n-1$ grados de libertad; y por tanto, se puede calcular el intervalo de confianza como:

$$\left[\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \quad \bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

También se sabe que si el tamaño de la muestra es suficientemente grande ($n = 50526 \gg 50$), por el **Teorema del Límite Central** se puede asumir la normalidad de la media muestral, dado que $t_{n>50} \sim z$. En este caso, se procede a calcular el intervalo de confianza utilizando la fórmula anterior (t de *Student*).




```
ultcost.interval <- confidence.interval.mean(nc=0.95, dist_mean=ultcost.mean,
                          dist_sd=ultcost.sd, dist_n=ultcost.n, 'two.sided')
ultcost.interval
```

```
## [1]  9938.855 10356.479
```

El intervalo de confianza al 95% de la media poblacional de `UltCost` es [9938.86, 10356.48]; esto significa que, con un nivel de confianza del 95%, la media de la población se encuentra entre 9938.86 y 10356.48.

Patricia Lázaro Tello
plazarotello@uoc.edu

3 Coste inicial y final de los siniestros

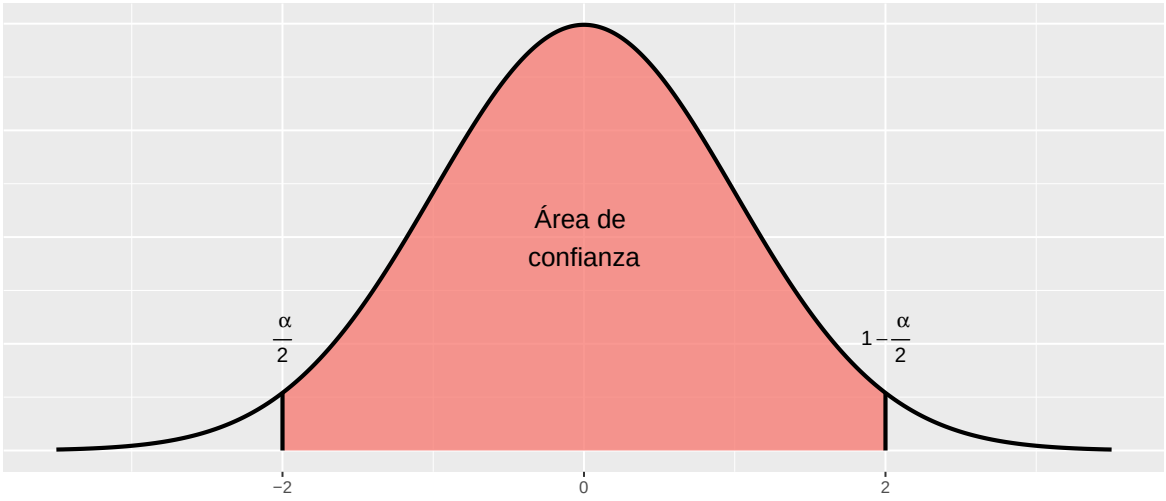
La pregunta planteada es: *¿Podemos aceptar que no hay diferencias entre* `IniCost` *y* `UltCost`*?*. Se exponen algunas observaciones para poder elegir correctamente el contraste de hipótesis:

- `IniCost` y `UltCost` son muestras distintas; se concluye que se trata de un contraste de dos muestras.
- Se sabe que `IniCost` hace referencia a la estimación inicial del siniestro, y `UltCost` es el coste total pagado. Por tanto, sabiendo la semántica de ambas variables, se puede asumir que están emparejadas.
- Se habla también de investigar si no hay diferencia entre las variables; en consecuencia, se trata de un contraste bilateral de medias.
- No se conoce la varianza de la población.

A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de dos muestras emparejas sobre la media**, siendo $d = UltCost - IniCost$:

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

Por tanto, el intervalo de confianza queda:

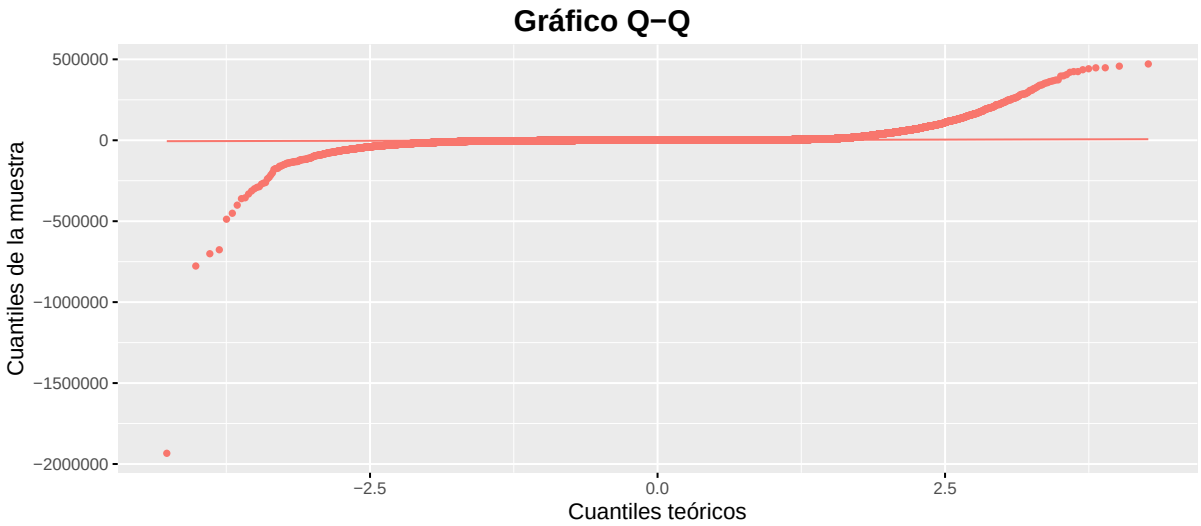


En primer lugar, se procede a comprobar la normalidad de la diferencia de variables mediante el test de normalidad de *Lilliefors* y un gráfico Q-Q.

```
ult.ini.diff <- claim$UltCost - claim$IniCost

lillie.test(ult.ini.diff)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  ult.ini.diff
## D = 0.35186, p-value < 0.00000000000000022
```

```
ggqqplot(ult.ini.diff, color=default.color.main, ggtheme = theme_gray(),
         xlab='Cuantiles teóricos', ylab='Cuantiles de la muestra',
         title='Gráfico Q-Q', shape=16) + title.centered
```



Con un nivel de significancia $\alpha = 0.05$ se rechaza la hipótesis nula (H_0), dado que $p - value = 0e + 00 \ll \alpha = 0.05$.

Por tanto, se puede concluir que la variable $UltCost - IniCost$ **no sigue una distribución normal**. La siguiente fórmula sigue una distribución *t* de Student con $n - 1$ grados de libertad:

$$t_{obs} = \frac{\bar{X}}{S/\sqrt{n}} \sim t_{n-1}$$

Por tanto, el intervalo de confianza se formula como:

$$\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$$

```
diff.nc <- 0.95
diff.alpha <- 1.0 - diff.nc

diff.media <- mean(ult.ini.diff)
diff.sd <- sd(ult.ini.diff)

diff.n <- length(ult.ini.diff)
diff.df <- diff.n - 1

diff.se <- diff.sd/sqrt(diff.n)

diff.tobs <- diff.media / diff.se
diff.pvalue <- pt(diff.tobs, df=diff.df, lower.tail=FALSE)
diff.crit <- qt(diff.alpha/2, df=diff.df, lower.tail=FALSE)

diff.ic <- confidence.interval.mean(diff.nc, diff.media, diff.sd,
                                    diff.n, 'two.sided')
```

Se comprueban los resultados obtenidos con los resultados de la función `t.test`:

```
t.test(claim$UltCost, claim$IniCost, paired=TRUE, alternative='two.sided',
       conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  claim$UltCost and claim$IniCost
## t = 20.801, df = 50525, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1955.884 2362.820
## sample estimates:
## mean of the differences
##                2159.352
```

De los cálculos anteriores se obtiene que $t_{obs} = 20.801$, el valor crítico es $t_{\alpha,n-1} = 1.96$ y el valor p es $p - value = 5.34e - 96$.

Por tanto, se rechaza la hipótesis nula (H_0) con un nivel de significancia del 95%, dado que $p - value = 0 \ll \alpha = 0.05$; **hay diferencias entre `IniCost` y `UltCost`**.

Patricia Lázaro Tello
plazarotello@uoc.edu

Con un nivel de significancia del 95%, la media de la diferencia $UltCost - IniCost$ se encuentra en el intervalo $[1955.8844, 2362.8204]$; es decir, de media la estimación inicial del coste del siniestro no cubre el coste final del mismo.

	Valor
Estadístico	$t_{obs} = 20.801$
Valor crítico	$t_{\alpha,n-1} = \pm 1.96$
Valor p	$p - value = 5.34e - 96$
Intervalo de confianza	$[1955.8844, 2362.8204]$

Patricia Lázaro Tello
plazarotello@uoc.edu

4 Diferencia de salario según género

El salario semanal se encuentra almacenado en la variable `WeeklyWages`. Se desea comprobar si las mujeres cobran menos que los hombres; para ello se utiliza la variable `Gender` con el propósito de obtener los salarios semanales por género.

```
wages.male <- claim$WeeklyWages[claim$Gender == 'M']
wages.female <- claim$WeeklyWages[claim$Gender == 'F']
```

Se obtienen 38904 muestras del salario de los hombres y 11620 muestras del salario de las mujeres. A continuación se observa la distribución de los salarios según el género y una análisis de la misma.

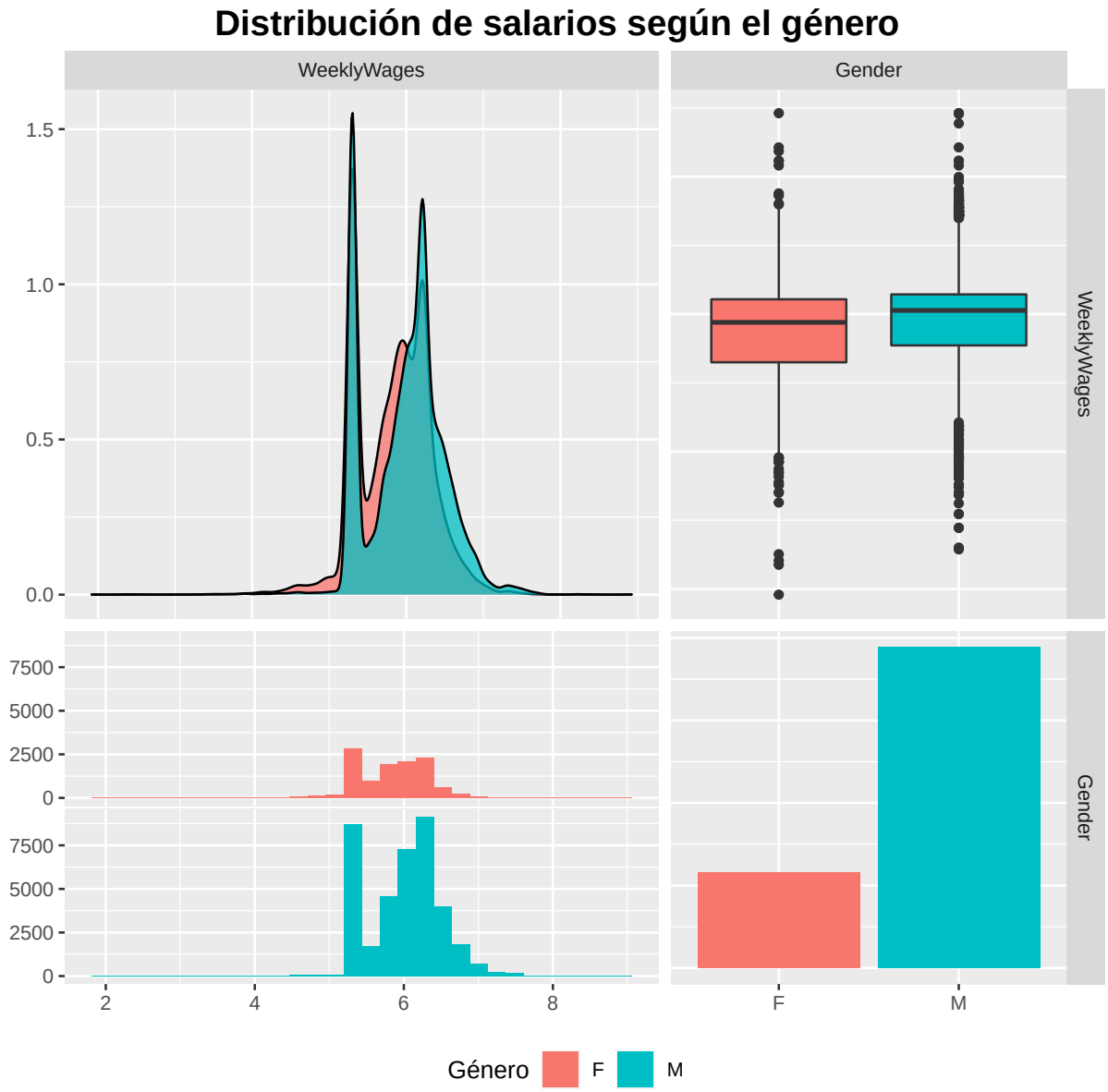
En el histograma de la figura se observa que los valores máximos que alcanzan los salarios de los hombres son significativamente más altos que los de los salarios de las mujeres.

En el gráfico de densidad, la cola derecha de los salarios de los hombres es más pronunciada que la de los salarios de las mujeres.

Además, en el *boxplot* se observa que el primer cuartil (Q1), el tercer cuartil (Q3) y la media son ligeramente superiores en los salarios de hombres que en los de mujeres.

```
wages.log <- claim %>%
  select(Gender, WeeklyWages) %>%
  mutate(WeeklyWages=log(WeeklyWages))
GGally::ggpairs(wages.log[wages.log$Gender %in% c('M', 'F'),],
                mapping=aes(fill=factor(Gender)),
                columns=c('WeeklyWages', 'Gender'),
                title='Distribución de salarios según el género',
                legend=4, proportions=c(1.5,1),
                diag=list(continuous = wrap("densityDiag", alpha = 0.75))) +
  scale_colour_manual(name='Género', values=palette) +
  scale_fill_manual(name='Género', values=palette) +
  theme(legend.position='bottom') + title.centered
```

Patricia Lázaro Tello
plazarotello@uoc.edu



La pregunta planteada es *¿Podemos aceptar que los hombres cobran más que las mujeres en promedio a la semana?*. Algunas consideraciones a la hora de elegir el tipo de contraste a aplicar son:

- Aunque se hace referencia únicamente a la variable `WeeklyWages`, al filtrarla respecto al género (`Gender`), se obtienen dos variables distintas. Se trata, por tanto, de un contraste de dos muestras.
- El salario del hombre no depende del salario de la mujer: son muestras independientes.

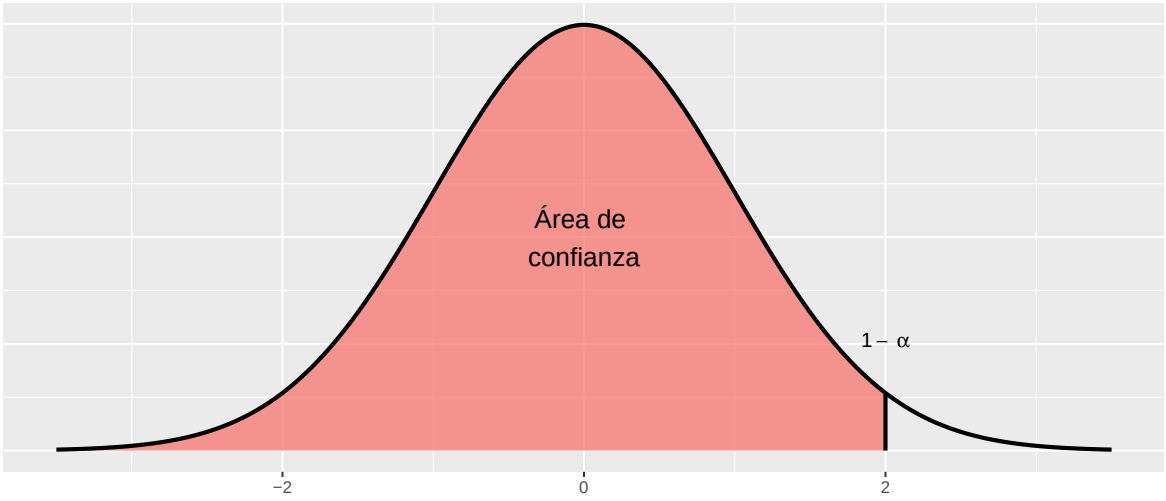
Patricia Lázaro Tello
plazarotello@uoc.edu

- Se trata de un contraste unilateral por la derecha de medias, ya que se busca saber si el salario medio de los hombres es superior al de las mujeres.

A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de dos muestras independientes sobre la media con varianzas desconocidas**:

$$H_0 : \mu_{hombre} = \mu_{mujer}$$
$$H_1 : \mu_{hombre} > \mu_{mujer}$$

Por tanto el intervalo de confianza quedará:



4.1 Comprobación de la igualdad de varianzas de dos muestras

En primer lugar se ha de realizar un test de **homocedasticidad** o igualdad de varianzas de dos muestras. Para ello, se plantea el siguiente contraste de hipótesis:

$$H_0 : s^2_{hombre} = s^2_{mujer}$$
$$H_1 : s^2_{hombre} \neq s^2_{mujer}$$

El test estadístico es:

$$f_{obs} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$$

```
wages.var <- variance.equals.2.samples(0.95, wages.male, wages.female)
```

No se puede asumir que las varianzas de los salarios de hombres y mujeres sean iguales ($p-value = 4.94e-109 \ll \alpha = 0.05$) con un nivel de confianza del 95%. Analizando $F_{obs} = 1.4084413$, se observa que la varianza muestral de los hombres es superior a la de las mujeres.

Se comprueba con la función del test estadístico de R:

```
var.test(wages.male, wages.female, conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  wages.male and wages.female
## F = 1.4084, num df = 38903, denom df = 11619, p-value <
## 0.00000000000000022
## 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
```

Por tanto, para el **contraste de dos muestras independientes sobre la media con varianzas desconocidas *diferentes*** se utilizará el siguiente estadístico:

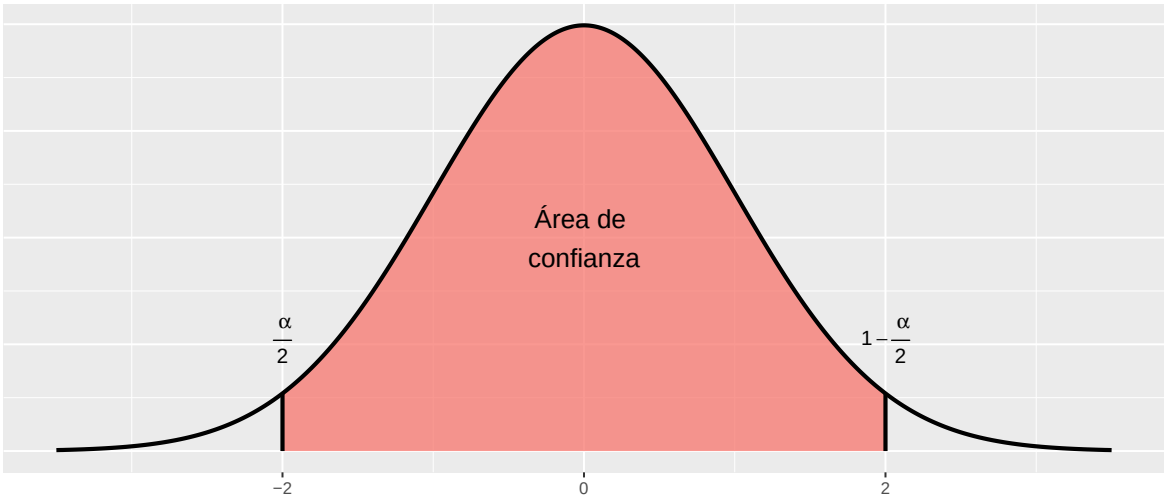
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_v$$

siendo t_v una distribución *t* de Student con v grados de libertad, calculados según:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}}$$

Y el intervalo de confianza queda:

$$\left[(\overline{X}_1 - \overline{X}_2) - t_{1-\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + t_{1-\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$




```
wages.medias.iguales <- means.contrast.2.samples.vars.unknown.dif(
  wages.male, wages.female, 0, 0.95, 'right')
```

De los cálculos anteriores se obtiene que $t_v = 28.813$ y el valor p es $p - value = 1.44e - 179 \ll 0.05$; por tanto la hipótesis nula (H_0) se rechaza.

La media del salario de los hombres es mayor que la media del salario de las mujeres con un nivel de confianza del 95%. Analizando el intervalo de confianza, se puede afirmar que, con ese nivel de significancia, los hombres cobran al menos 64.1803€ más que las mujeres.

	Valor
Estadístico	$t_{obs} = 28.813$
Grados de libertad v	$v = 22274$
Valor crítico	$t_v = 1.64$
Valor p	$p - value = 1.44e - 179$
Intervalo de confianza	$[64.1803, \infty)$

Se comprueban los resultados con los que ofrece la función `t.test` de R:

Patricia Lázaro Tello
plazarotello@uoc.edu

```
t.test(wages.male, wages.female, alternative='greater', var.equal=FALSE,
       conf.level=0.95)
```

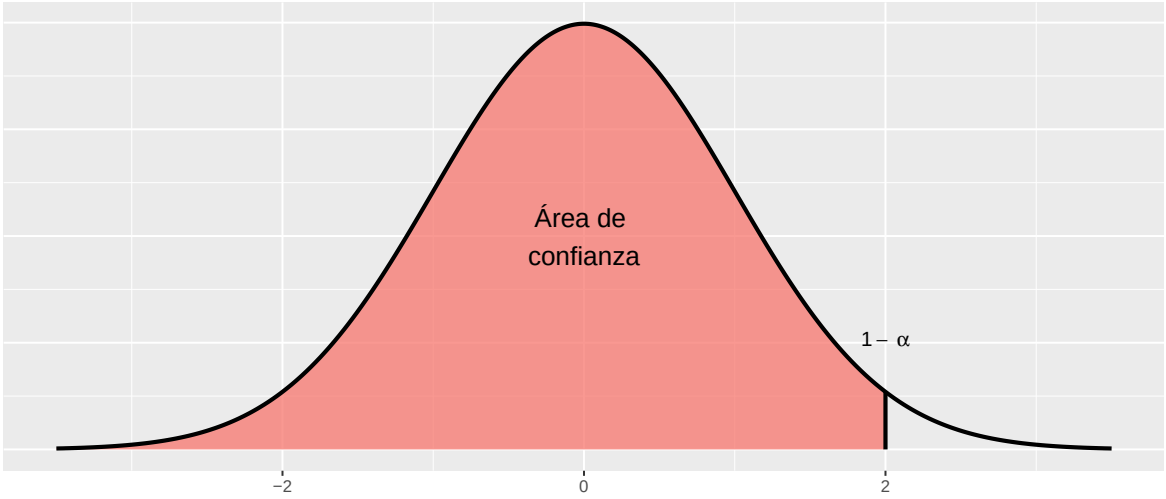
```
##
##  Welch Two Sample t-test
##
## data:  wages.male and wages.female
## t = 28.813, df = 22274, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  64.18029      Inf
## sample estimates:
## mean of x mean of y
##  449.4311  381.3649
```

Patricia Lázaro Tello
plazarotello@uoc.edu

5 Salario semanal (II)

La pregunta planteada es *¿Podemos aceptar que los hombres cobran al menos 50 euros más que las mujeres en promedio a la semana?*. Las consideraciones para elegir el contraste correcto son:

- Como en la *sección anterior*, se trata de un contraste de dos muestras independientes.
- La diferencia entre las medias de las muestras ha de ser más de 50.0. Se trata de un contraste unilateral por la derecha.
- Como en la *sección anterior*, la varianza es desconocida y diferente.



A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de dos muestras independientes sobre la media con varianza desconocida diferente**:

$$H_0 : \mu_{hombre} - \mu_{mujer} = 50$$
$$H_1 : \mu_{hombre} - \mu_{mujer} > 50$$

Por tanto, se utilizará el mismo estadístico utilizado en la *sección anterior* así como el intervalo de confianza.

Patricia Lázaro Tello
plazarotello@uoc.edu

```
wages.medias.50 <- means.contrast.2.samples.vars.unknown.dif(
  wages.male, wages.female, 50, 0.95, type='right')
```

Con un nivel de confianza del 95% se rechaza la hipótesis nula H_0 ($p - value = 1.07e-14 \ll \alpha = 0.05$). **Los hombres cobran de media al menos 50€ más que las mujeres a la semana**. Como se observaba en la sección anterior, los hombres cobran al menos 64.1803€ más que las mujeres.

	Valor
Estadístico	$t_{obs} = 7.647$
Grados de libertad υ	$\upsilon = 22274$
Valor crítico	$t_{\upsilon} = 1.64$
Valor p	$p - value = 1.07e-14$
Intervalo de confianza	$[64.1803, \infty]$

Se comprueban los resultados utilizando la función `t.test` de R:

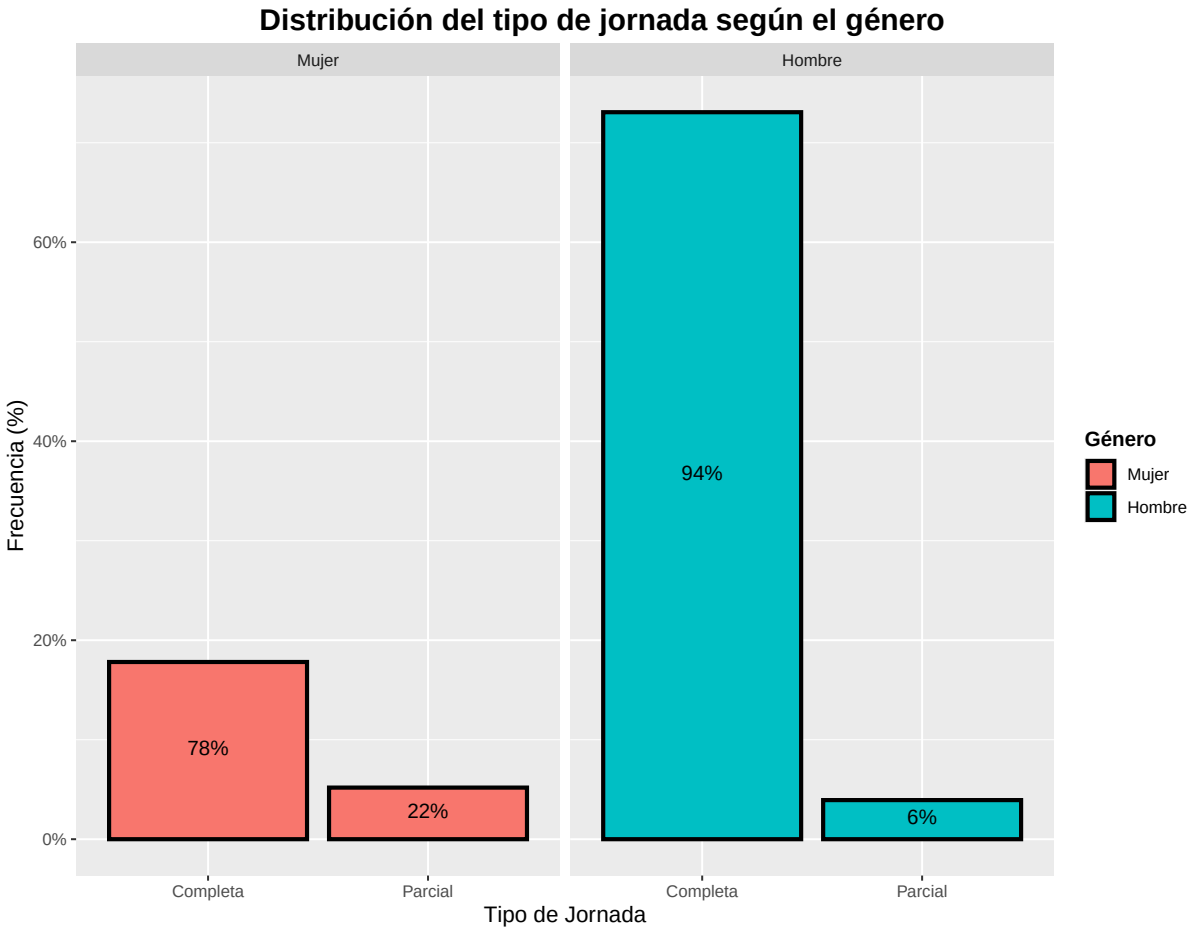
```
t.test(wages.male, wages.female, mu=50, alternative='greater', var.equal=FALSE,
       conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wages.male and wages.female
## t = 7.6475, df = 22274, p-value = 0.00000000000001066
## alternative hypothesis: true difference in means is greater than 50
## 95 percent confidence interval:
##  64.18029      Inf
## sample estimates:
## mean of x mean of y
##  449.4311  381.3649
```

Patricia Lázaro Tello
plazarotello@uoc.edu

6 Diferencia de jornada según género

Se desea averiguar si las mujeres realizan más frecuentemente una jornada a tiempo parcial que los hombres. Para ello, se observarán los valores de la variable `PartTimeTullTime` respecto al género.



Se aprecia que, aunque no existen tantas muestras de género femenino como de género masculino, la proporción de mujeres que se adhieren a la jornada parcial es superior a la de hombres (22% del total de mujeres contra el 6% del total de hombres).

A su vez, existe una proporción significativamente mayor de hombres que trabajan a jornada completa (94% de del total de hombres contra el 78% del total de mujeres).

Se puede asumir que las mujeres realizan más frecuentemente una jornada a tiempo parcial que los hombres.

Patricia Lázaro Tello
plazarotello@uoc.edu

La pregunta planteada para el contraste de hipótesis es *¿La proporción de personas que trabajan a tiempo completo es diferente para hombres que para mujeres?*. Las consideraciones para elegir el tipo de contraste son:

- Se trata de un contraste entre dos muestras independientes (la proporción de hombres que trabaja a tiempo completo no tiene relación con la proporción de mujeres que trabaja a tiempo completo) sobre la proporción.
- Se busca averiguar si la proporción de trabajadores a tiempo completo, que se cree igual, es en realidad distinta para hombres y mujeres. Se trata de un contraste bilateral.

A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de dos muestras sobre la proporción**:

$$H_0 : p_{hombre} = p_{mujer}$$
$$H_1 : p_{hombre} \neq p_{mujer}$$

Se utilizará el siguiente estadístico:

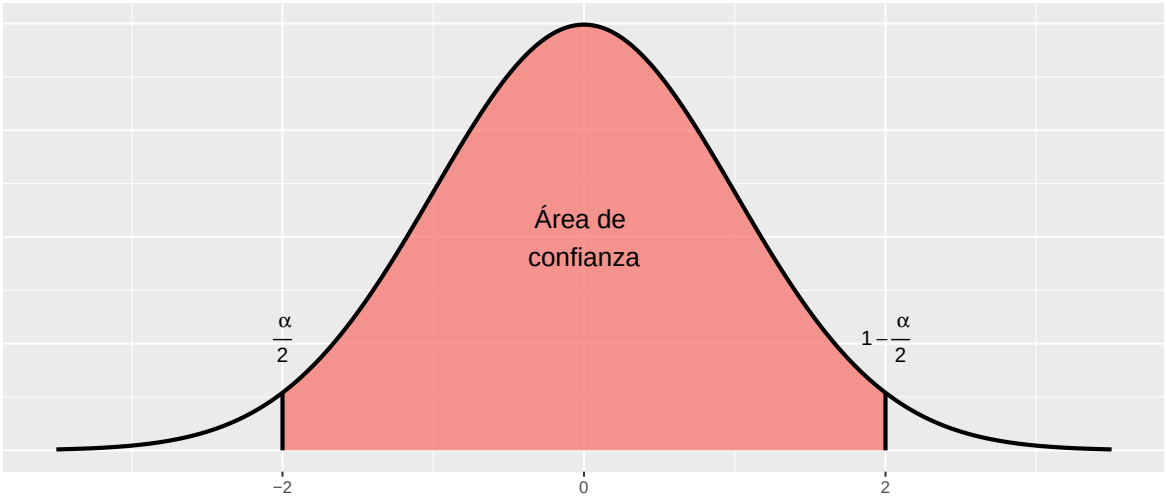
$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Siendo $\hat{p}$:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Por tanto, el intervalo de confianza queda como $[IC_1, IC_2]$, siendo:

$$IC_1 = (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$
$$IC_2 = (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$




```
ptft.males <- claim$PartTimeFullTime[claim$Gender == 'M']
ptft.females <- claim$PartTimeFullTime[claim$Gender == 'F']

p.males <- sum(ptft.males == 'F')/length(ptft.males)
n.males <- length(ptft.males)
p.females <- sum(ptft.females == 'F')/length(ptft.females)
n.females <- length(ptft.females)

p <- (n.males*p.males + n.females*p.females) / (n.males+n.females)

zobs <- (p.males-p.females) / sqrt(p*(1-p)*(1/n.males + 1/n.females))
zcrit <- qnorm(0.025, lower.tail=FALSE)
pvalue <- pnorm(zobs, lower.tail=FALSE)
ic <- confidence.interval.2.props(0.95, p.males, p.females,
                                  n.males, n.females, 'two.sided')
```

Con un nivel de confianza del 95% se rechaza la hipótesis nula H_0 ($p - value = 0 \ll \alpha = 0.05$). **La proporción de hombres que trabajan a tiempo completo es diferente que la proporción de mujeres que trabajan a tiempo completo**.

	Valor
Estadístico	$z_{obs} = 57.342$
Valor crítico	$z_{\alpha/2} = \pm 1.96$
Valor p	$p - value = 0$
Intervalo de confianza	$[0.1666, 0.1824]$

Patricia Lázaro Tello
plazarotello@uoc.edu

Se comprueban los resultados utilizando la función `prop.test` de R:

```
prop.test(c(n.males*p.males, n.females*p.females),
          c(n.males, n.females), alternative='two.sided', correct=FALSE,
          conf.level=0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(n.males * p.males, n.females * p.females) out of c(n.males, n.females)
## X-squared = 3288.1, df = 1, p-value < 0.00000000000000022
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1666030 0.1824183
## sample estimates:
##    prop 1    prop 2
## 0.9489513 0.7744406
```

Patricia Lázaro Tello
plazarotello@uoc.edu

7 Salario por hora

La pregunta planteada es *¿Podemos afirmar que los hombres cobran más que las mujeres por hora trabajada?*. Se exponen las consideraciones a tomar en cuenta para elegir el tipo de contraste:

- El salario por hora de los hombres no tiene relación con el salario por hora de las mujeres. Se trata de dos muestras independientes.
- Se busca averiguar si los hombres cobran más de media por hora trabajada que las mujeres; por tanto se habla de un contraste sobre la media unilateral.
- Las varianzas de las muestras son desconocidas y diferentes (test de homocedasticidad (igualdad de varianzas) de las dos muestras):

```
wages.default <- claim %>%
  mutate(HourlyWages = WeeklyWages/HoursWeek) %>%
  dplyr::select(Gender, HourlyWages)

wages.male <- wages.default$HourlyWages[wages.default$Gender=='M']
wages.female <- wages.default$HourlyWages[wages.default$Gender=='F']

vars <- variance.equals.2.samples(0.95, wages.male, wages.female)
```

```
## F observada: 0.624
## Valores críticos: [0.9712, 1.0299]
## Valor p: 2.83e-236
```

```
var.test(wages.male, wages.female, conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  wages.male and wages.female
## F = 0.624, num df = 38903, denom df = 11619, p-value <
## 0.00000000000000022
## 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
```

A continuación se plantea la hipótesis nula y la hipótesis alternativa del **contraste de**

dos muestras sobre la media con varianzas desconocidas diferentes:

$$H_0 : \mu_{hombre} = \mu_{mujer}$$
$$H_1 : \mu_{hombre} > \mu_{mujer}$$

El estadístico e intervalo de confianza quedará igual que el del contraste de hipótesis del *Salario semanal (II)* y *Salario según género*.

```
wages.hourly <- means.contrast.2.samples.vars.unknown.dif(
  wages.male, wages.female, 0, 0.95, 'right')
```

Con un nivel de confianza del 95% se acepta la hipótesis nula H_0 ($p - value = 0.239 \gg \alpha = 0.05$). **Los hombres cobran de media igual que las mujeres por hora**.

	Valor
Estadístico	$t_{obs} = 0.71$
Grados de libertad υ	$\upsilon = 16186$
Valor crítico	$t_{\upsilon} = 1.64$
Valor p	$p - value = 0.239$
Intervalo de confianza	$[-0.0879, \infty]$

Se comprueban los resultados utilizando la función `t.test` de R:

```
t.test(wages.male, wages.female, alternative='greater', conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wages.male and wages.female
## 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
```

Patricia Lázaro Tello
plazarotello@uoc.edu

8 Resumen ejecutivo

A continuación se exponen las principales conclusiones del análisis estadístico inferencial sobre las variables referentes al coste de los siniestros, el tipo de jornada y los salarios semanales y por hora.

- Se ha cargado el fichero `train_clean2.csv`, que contiene 50526 observaciones con 17 atributos cada una.
- El coste final de los siniestros no sigue una distribución normal.
- La estimación inicial del coste de los siniestros no coincide con el coste final de los mismos. Es decir, las estimaciones iniciales no son suficientemente precisas.
- De media, el salario semanal de los hombres es mayor que el de las mujeres, y es al menos 50€ superior al de ellas.
- Hombres y mujeres cobran de media el mismo salario por hora.
- La proporción de hombres que trabaja a tiempo completo es diferente a la proporción de mujeres que trabaja a tiempo completo; por análisis visual se tiene la noción de que las mujeres se adhieren a la jornada parcial más que los hombres.
- De los puntos anteriores se puede inferir que el salario semanal de los hombres es superior al de las mujeres debido a que ellos trabajan más a tiempo completo que ellas; por tanto, aunque a la hora cobran lo mismo, ellos trabajan más horas que ellas.