Estimación

FGK

4/25/2021

La estimación puede ser puntual o por intervalos, hemos trabajado hasta el momento con estimadores puntuales.

Intervalos de confianza

Se muestran funciones que disponibles en R para construir intervalos de confianza para los siguientes estadísticos:

- la media μ,
- la proporción p,
- la varianza σ^2,
- la diferencia de medias $\mu_1 - \mu_2$ puede ser para muestras independientes o pareadas
- la diferencia de proporciones $p_1 - p_2$, y
- un cociente de varianzas σ_1^2/σ_2^2.

Utilizaremos diversos ejemplos para ilustrar el uso de las funciones..

Función `t.test`

La función `t.test` se usa para calcular intervalos de confianza para la media y diferencia de medias, con muestras independientes y dependientes (o pareadas). La función y sus argumentos son los siguientes:

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Intervalo de confianza bilateral para la media μ

Para calcular intervalos de confianza bilaterales para la media a partir de la función `t.test` es necesario definir 2 argumentos:

- `x`: vector numérico con los datos.
- `conf.level`: nivel de confianza a usar (por defecto es 0.95)

Los demás argumentos se usan cuando se desea obtener intervalos de confianza para diferencia de media con muestras independientes y dependientes (o pareadas).

Ejemplo

Suponga que se quiere obtener un intervalo de confianza bilateral del 90% para el salario neto mensual de las trabajadoras de la industria del software.

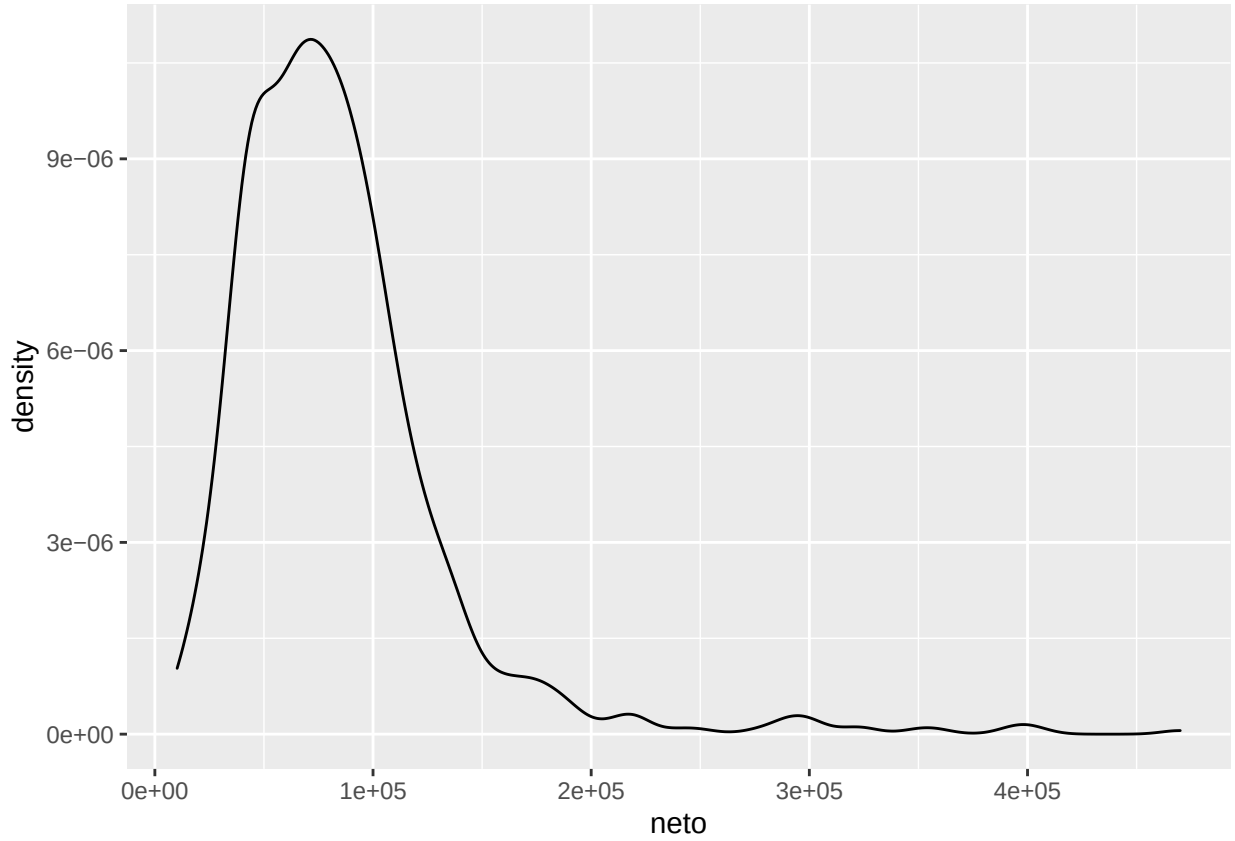
Para calcular el intervalo de confianza, primero llamamos y depuramos la base de datos, luego se crea un subconjunto de `datos` y se aloja en el objeto `mujeres.cis` como sigue a continuación:

```
library(readxl)
library(tidyverse)
sueldossys <- read_excel("datos/sys2021.xlsx")

salario_neto_gen <- sueldossys %>%
  rename(neto =`Salario mensual o retiro NETO (en tu moneda local)`) %>%
  filter(neto< 500000 & neto>10000) %>%
  select(c(neto,`Me identifico`)) %>%
  mutate(genero = fct_lump(`Me identifico`, n = 2,   other_level = "Otre"))


salario_mujerescis <- salario_neto_gen %>%
  filter(genero == "Mujer Cis")

head(salario_mujerescis)
```

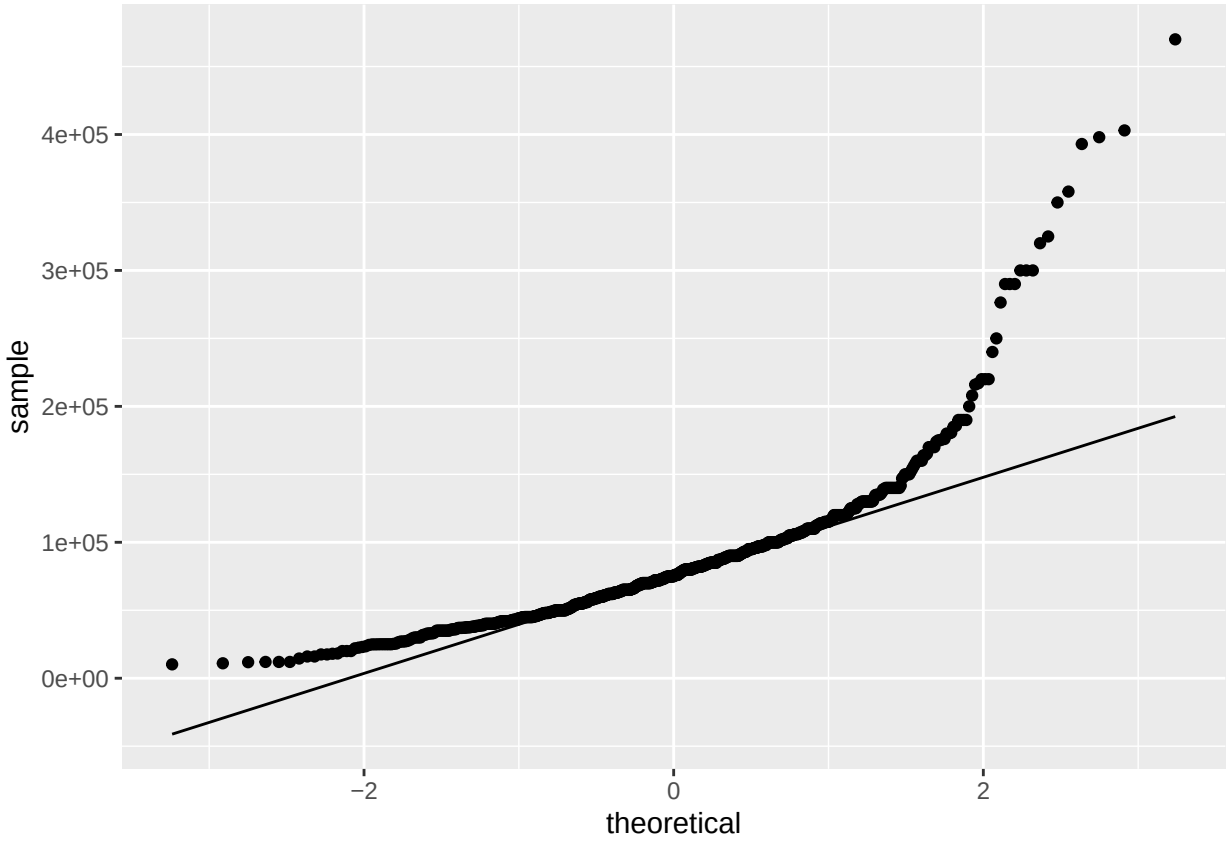
```
## # A tibble: 6 x 3
##     neto `Me identifico` genero
##    <dbl> <chr>           <fct>
## 1 150000 Mujer Cis       Mujer Cis
## 2  91000 Mujer Cis       Mujer Cis
## 3 180632 Mujer Cis       Mujer Cis
## 4  68000 Mujer Cis       Mujer Cis
## 5  42000 Mujer Cis       Mujer Cis
## 6  36000 Mujer Cis       Mujer Cis
```

Una vez leídos los datos, se analiza la normalidad de la variable a podemos usar un QQplot y un histograma

```
ggplot(salario_mujerescis, aes(x=neto)) +
  geom_density()
```




```
ggplot(salario_mujerescis, aes(sample=neto)) +
  geom_qq()+
  geom_qq_line()
```



El QQplot y el histograma para la variable no muestra un claro patrón normal, pero vamos a asumir que si y vamos utilizar la función `t.test` sobre la variable de interés para construir el intervalo de confianza. El resultado de usar `t.test` es una lista, uno de los elementos de esa lista es justamente el intevalo de confianza y para extraerlo es que se usa `$conf.int` al final de la instrucción. A continuación se muestra el código utilizado.

```
t.test(x=salario_mujerescis$neto, conf.level=0.90)$conf.int
```

```
## [1] 81502.10 87329.25
## attr(,"conf.level")
## [1] 0.9
```

A partir del resultado obtenido se puede concluir, con un nivel de confianza del 90%, que el salario neto promedio de las mujeres cis que trabajan en la industria del software en Argentina se encuentra entre 81502 y 87329 pesos.

- Qué sucede con los varones?
- Porqué utilizamos una distribución t

Intervalo de confianza bilateral para la diferencia de medias ($\mu_1 - \mu_2$) de muestras independientes

Para construir intervalos de confianza bilaterales para la diferencia de medias ($\mu_1 - \mu_2$) de muestras independientes se usa la función `t.test` y es necesario definir 5 argumentos:

- `x`: vector numérico con la información de la muestra 1,
- `y`: vector numérico con la información de la muestra 2,

- `paired=FALSE`: indica que el intervalo de confianza se hará para muestras independientes, en el caso de que sean dependientes (o pareadas) este argumento será `paired=TRUE`,
- `var.equal=FALSE`: indica que las varianzas son desconocidas y diferentes, si la varianzas se pueden considerar iguales se coloca `var.equal=TRUE`.
- `conf.level`: nivel de confianza.

Ejemplo

Se quiere saber si existe diferencia estadísticamente significativa entre el salario neto entre mujeres y hombres. Para responder esto se va a construir un intervalo de confianza del 95% para la `diferencia` de los salarios neto promedio de los hombres y de las mujeres ($\mu_{hombres} - \mu_{mujeres}$).

Para construir el intervalo de confianza, primero generamos el subconjuntos de datos como sigue a continuación:

```
summary(salario_mujerescis)
```

```
##       neto        Me identifico            genero
##  Min.   : 10200   Length:835         Mujer Cis:835
##  1st Qu.: 51353   Class :character   Varón Cis:  0
##  Median : 75000   Mode  :character   Otre     :  0
##  Mean   : 84416
##  3rd Qu.:100000
##  Max.   :470000
```

```
salario_hombrescis <- salario_neto_gen %>%
  filter(genero == "Varón Cis")
```

Utilizamos la función `t.test` para construir el intervalo de confianza requerido. A continuación se muestra el código

```
t.test(x=salario_hombrescis$neto, y=salario_mujerescis$neto,
       paired=FALSE, var.equal=FALSE,
       conf.level = 0.95)$conf.int
```

```
## [1] 15358.22 23284.75
## attr(,"conf.level")
## [1] 0.95
```

A partir del intervalo de confianza anterior se puede concluir, con un nivel de confianza del 95%, que el salario neto promedio de los hombres es superior al de las mujeres, ya que el intervalo de confianza para la diferencia de medias **NO** incluye el cero y por ser positivos sus limites se puede afirmar con un nivel de confianza del 95% que $\mu_{hombres} > \mu_{mujeres}$.

Intervalo de confianza bilateral para la diferencia de medias ($\mu_1 - \mu_2$) de muestras dependientes o pareadas

Para construir intervalos de confianza bilaterales para la diferencia de medias de muestras dependientes a partir de la función `t.test` es necesario definir 4 argumentos:

- `x`: vector numérico con la información de la muestra 1,
- `y`: vector numérico con la información de la muestra 2, `paired=TRUE` indica que el intervalo de confianza se hará para muestras dependientes o pareadas.
- `conf.level`: nivel de confianza.

Ejemplo

Los desórdenes musculoesqueléticos del cuello y hombro son comunes entre empleados de oficina que realizan tareas repetitivas mediante pantallas de visualización. Se reportaron los datos de un estudio para determinar si condiciones de trabajo más variadas habrían tenido algún impacto en el movimiento del brazo. Los datos

que siguen se obtuvieron de una muestra de $n = 16$ sujetos. Cada observación es la cantidad de tiempo, expresada como una proporción de tiempo total observado, durante el cual la elevación del brazo fue de menos de 30 grados. Las dos mediciones de cada sujeto se obtuvieron con una separación de 18 meses. Durante este período, las condiciones de trabajo cambiaron y se permitió que los sujetos realizaran una variedad más amplia de tareas. ¿Sugieren los datos que el tiempo promedio verdadero durante el cual la elevación es de menos de 30 grados luego del cambio difiere de lo que era antes? Calcular un intervalo de confianza del 95% para responder la pregunta.

Sujeto	1	2	3	4	5	6	7	8
Antes	81	87	86	82	90	86	96	73
Después	78	91	78	78	84	67	92	70
Diferencia	3	-4	8	4	6	19	4	3

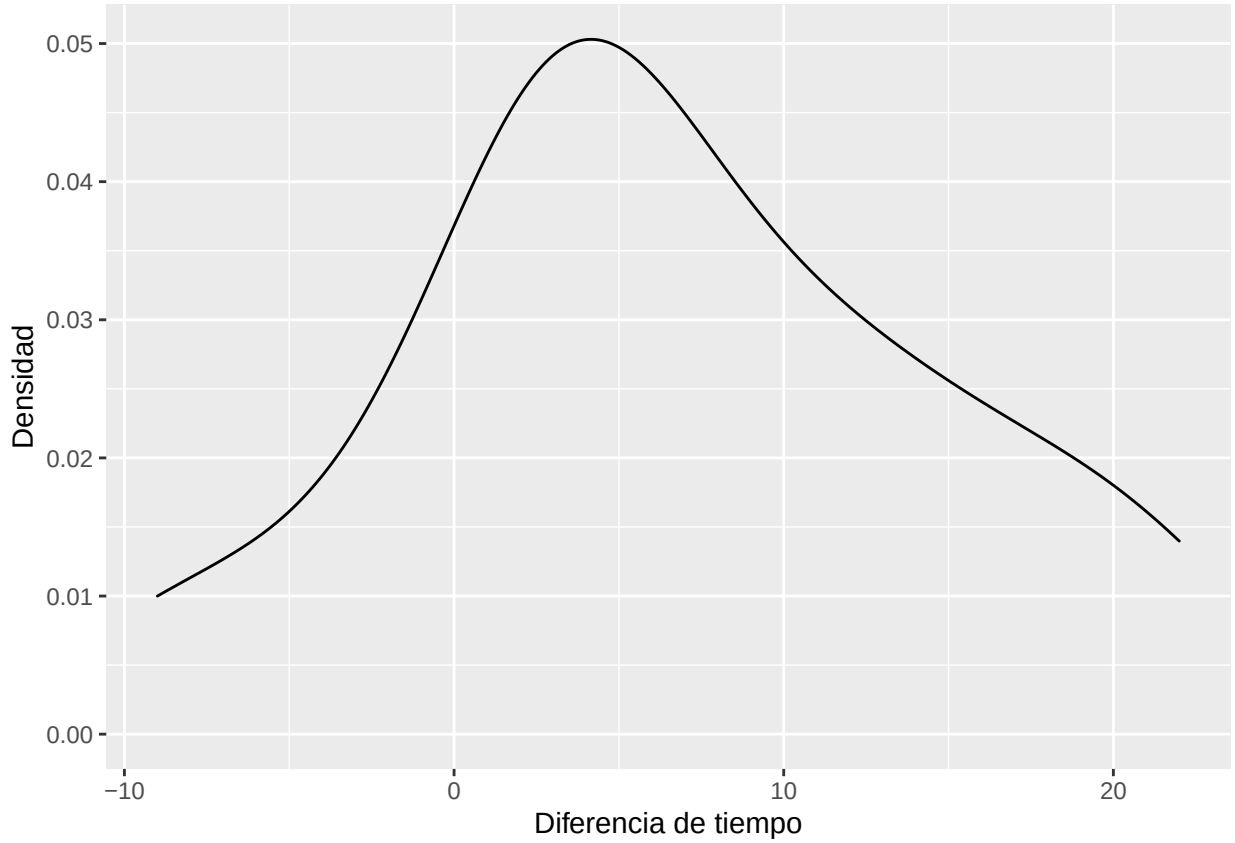
Sujeto	9	10	11	12	13	14	15	16
Antes	74	75	72	80	66	72	56	82
Después	58	62	70	58	66	60	65	73
Diferencia	16	13	2	22	0	12	-9	9

Para construir el intervalo de confianza primero se crean dos vectores con los datos y se nombran `Antes` y `Despues`, luego se calcula la diferencia y se aloja en el vector `Diferencia`, como sigue a continuación:

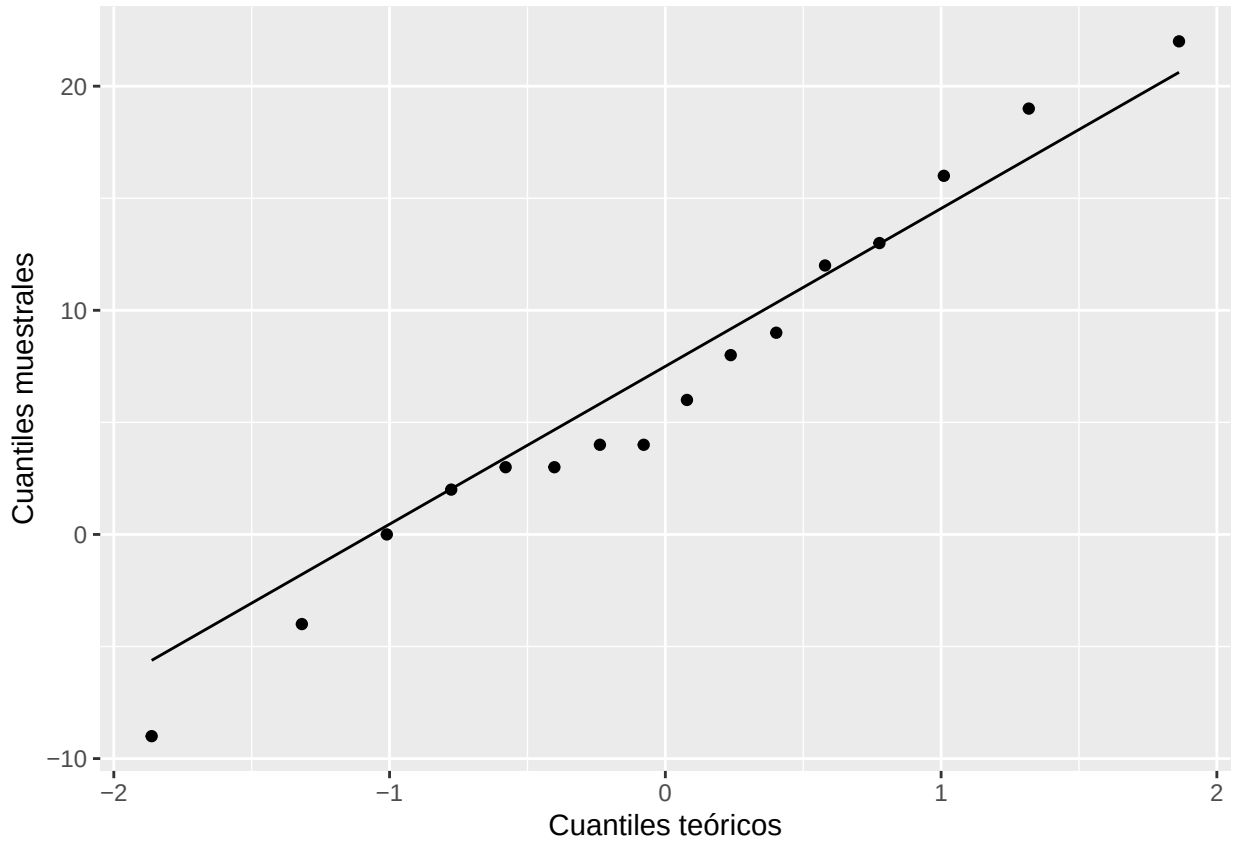
```
Antes   <- c(81, 87, 86, 82, 90, 86, 96, 73,
             74, 75, 72, 80, 66, 72, 56, 82)
Despues <- c(78, 91, 78, 78, 84, 67, 92, 70,
             58, 62, 70, 58, 66, 60, 65, 73)
Diferencia <- Antes - Despues

df <- as.data.frame(cbind("Antes"=Antes,
                          "Despues"=Despues,
                          "Diferencia"=Diferencia))
```

Se analiza la normalidad de la variable `Diferencia` de los cambios en las condiciones de trabajo, a partir de un qqplot y una densidad.

```
ggplot(df, aes(x=Diferencia)) +
  geom_density()+
  xlab('Diferencia de tiempo')+
  ylab('Densidad')
```




```
ggplot(df, aes(sample=Diferencia)) +
  geom_qq()+
  geom_qq_line()+
  xlab('Cuantiles teóricos')+
  ylab('Cuantiles muestrales')
```



Ahora sí, se observa que la diferencia de los tiempos sigue una distribución normal, debido a que en el QQplot se observa un patron lineal y la densidad muestra una forma cercana a la simétrica.

Luego de chequear la normalidad de la variable `Diferencia` se usa la función `t.test` para construir el intervalo. A continuación se muestra el código utilizado.

```
t.test(x=df$Antes, y=df$Despues, paired=TRUE, conf.level=0.95)$conf.int
```

```
## [1]  2.362371 11.137629
## attr(,"conf.level")
## [1] 0.95
```

A partir del resultado obtenido se puede concluir con un nivel de confianza del 95%, que el tiempo promedio verdadero durante el cual la elevación es de menos de 30 grados luego del cambio difiere de lo que era antes del mismo. Como el intervalo de confianza es $2.362 < \mu_D < 11.138$, esto indica que $\mu_{antes} - \mu_{despues} > 0$ y por lo tanto $\mu_{antes} > \mu_{despues}$.

Intervalo de confianza unilateral para la media μ

Tambien se pueden construir intervalos de confianza unilaterales para eso se usa el argumento `alternative = 'less'` o `alternative='greater'` veremos implementaciones la clase próxima, ahora utilizamos un ejemplo de simulación.

Ejemplo

Simule una muestra aleatoria de una $N(18, 3)$ y calcule un intervalo de confianza unilateral superior del 90% para la media

```r
x <- rnorm(50, mean = 18, sd =3)
t.test(x, alternative = "greater", conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 44.726, df = 49, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 90 percent confidence interval:
##  17.4474     Inf
## sample estimates:
## mean of x
##  17.96932
```

En el resultado anterior se muestra el intervalo de confianza unilateral.

Función `var.test`

Para construir intervalos de confianza para la varianza se usa la función `var.test`

Intervalo de confianza bilateral para el cociente de varianzas σ_1^2/σ_2^2

Para calcular intervalos de confianza bilaterales para la razón de varianzas a partir de la función `var.test` es necesario definir 3 argumentos:

- `x`: vector numérico con la información de la muestra 1,
- `y`: vector numérico con la información de la muestra 2,
- `conf.level`: nivel de confianza.

Ejemplo

Usando la información del ejemplo de diferencia de medias para muestras independientes (salario) se quiere obtener un intervalo de confianza del 95% para la razón de las varianzas del salario neto de mujeres y hombres.

```r
var.test(x=salario_hombrescis$neto,
         y=salario_mujerescis$neto,
         conf.level=0.95)$conf.int
```

```
## [1] 1.411430 1.740911
## attr(,"conf.level")
## [1] 0.95
```

El intervalo de confianza del 95% indica que la razón de varianzas se encuentra entre 1.41 y 1.74. Puesto que el intervalo de confianza no incluye el 1 se concluye que las varianzas de los salarios netos son estadísticamente diferentes.

Función `prop.test`

La función `prop.test` se usa para calcular intervalos de confianza para la porporción y diferencia de proporciones. La función y sus argumentos son los siguientes:

```r
prop.test(x, n, p=NULL,
          alternative=c("two.sided", "less", "greater"),
          conf.level=0.95, correct=TRUE)
```

Intervalo de confianza bilateral para la proporción

p

Para calcular intervalos de confianza bilaterales para la proporción a partir de la función `prop.test` es necesario definir 3 argumentos: `x` considera el conteo de éxitos, `n` indica el número de eventos o de forma equivalente corresponde a la longitud de la variable que se quiere analizar, y `conf.level` corresponde al nivel de confianza.

Ejemplo

Covid-Arg hoy. Queremos calcular un IC para la tasa de mortalidad de los ultimos 7 días en Arg. Buscamos en la página de la OMS y observamos la cantidad de positivos de los ultimos 7 días y de la cantidad de fallecidos de los últimos 7 días ¿Cuál es el intervalo de confianza del 90% para estimar la tasa de mortalidad?

muertes = 298 casos = 168125

```
prop.test(x=2310, n=168125, conf.level=0.90)$conf.int
```

```
## [1] 0.01327764 0.01421767
## attr(,"conf.level")
## [1] 0.9
```

A partir del resultado obtenido se puede concluir, con un nivel de confianza del 90%, que la proporción p de muertos se encuentra entre 0.013 y 0.014.

Intervalo de confianza bilateral para la diferencia de proporciones

$p_1 - p_2$

Para construir intervalos de confianza bilaterales para la proporción a partir de la función `prop.test` es necesario definir 3 argumentos:

- `x`: vector con el conteo de éxitos de las dos muestras,
- `n`: vector con el número de ensayos,
- `conf.level`: nivel de confianza.

Ejemplo

Se quiere determinar si existe diferencia en la tasa de mortalidad entre Argentina y Brasil. De la pagina se obtienen los siguientes valores.

n	Arg	Br
Muertes	2310	17814
Contagios	168125	408124

Construir un intervalo de confianza del 90% para decidir si son diferentes estas proporciones o no.

```
prop.test(x=c(2310, 17814), n=c(168125, 408124), conf.level=0.95)$conf.int
```

```
## [1] -0.03075109 -0.02906635
## attr(,"conf.level")
## [1] 0.95
```

A partir del resultado obtenido se puede concluir, con un nivel de confianza del 95%, que la diferencia de proporción de muertos ($p_1 - p_2$) se encuentra entre -0.031 y -0.029. Como el cero no está dentro del intervalo se concluye que existen diferencia.