Práctica 5. Intervalos de confianza.

Manuela Expósito, Pedro J. Femia, Christian J. Acal González,
Miguel Ángel Montero Alonso, Miguel Ángel Luque y Grupo BioestadísticaR

UNIVERSIDAD
DE GRANADA

Práctica 5. Intervalos de confianza

Manuela Expósito, Pedro J Femia, Christian J. Acal, Miguel Ángel Montero y Miguel Ángel Luque

---

5.1 Intervalo de confianza para la media μ de una variable normal

5.1.1 Intervalo de confianza para la media μ cuando la varianza σ^2 es conocida

Como se ha estudiado en el tema 4, para estimar la media poblacional μ de una variable aleatoria con distribución $N(\mu, \sigma)$ siendo σ conocida, el intervalo de confianza (IC) adopta la siguiente forma:

$$\mu \in \overline{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$$

donde n es el número de observaciones de la variable de interés en la población (muestra), α el nivel de significado fijado de antemano y z_α es el valor correspondiente en la tabla de la distribución *N(0,1)*. R no dispone de ninguna función específica para el cálculo de intervalos de confianza en este tipo de situaciones, aunque puede calcularse a mano. Por ejemplo, tomando a los 10 primeros pacientes del fichero **osteo.sav**, se está interesado en determinar un intervalo al 95% de confianza para la media de la variable *índice de masa corporal* (imc). Se va a suponer que la varianza poblacional es conocida (σ^2=5.80) y se asume que la variable imc tiene una distribución Normal. Las siguientes ordenes en R nos permiten realizar este cálculo:

```
pacientes=osteo[1:10,] #se selecciona a los 10 primeros pacientes
sigma=sqrt(5.80) #desviación típica conocida
media.muestral=mean(pacientes$imc) #estimación puntual de la media muestral
n=length(pacientes$imc) #tamaño de la muestra
z.alfa=1.96 #mirando la tabla de la N(0,1) y tomando alfa=0.05

limite.inferior=media.muestral-z.alfa*sigma/sqrt(n)
limite.superior=media.muestral+z.alfa*sigma/sqrt(n)
intervalo=c(limite.inferior,limite.superior)
intervalo
```

```
## [1] 21.29450 24.27988
```

5.1.2 Intervalo de confianza para la media μ cuando la varianza σ^2 es desconocida

El caso descrito anteriormente es más pedagógico que real, ya que si la media es desconocida, por lo general la varianza también lo será, por lo que en la práctica es muy difícil conocer la varianza de la variable de interés en toda la población. Por tanto, cuando la varianza no es conocida, el IC para μ cambia ligeramente adoptando la siguiente forma:

$$\mu \in \overline{x} \pm t_\alpha \frac{s}{\sqrt{n}}$$

donde s es la varianza muestral (cuasivarianza) y t_α es el valor correspondiente de la tabla de la distribución *t de Student* con *n-1* grados de libertad.

Aunque el intervalo de confianza podría calcularse de forma análoga a la realizada en el caso de varianza conocida, en este caso *R* si dispone de una funcion que permite calcular el intervalo de confianza de forma inmediata, la función **imc()** del paquete **BioestadisticaR2**. Esta función se puede aplicar tanto a vectores, como a datos resumidos. Por ejemplo, el intervalo al 95% de confianza para estimar la media poblacional μ de la variable *imc* del fichero **osteo**, se obtiene escribiendo en la consola de R:

```
icm(osteo$imc)
```

```
##
## Intervalo de confianza bilateral para la media de una VA normal
## ----------------------------------------------------------------
## Información muestral:
##   No hay valores faltantes
##   Tamaño muestral: n =  94
##   Media: m =  23.921
##   Desviación típica: s =  3.748
##   Error estándar de la media: sem =  0.387
##
## Estimación:
##  95%-IC(μ):  (23.153, 24.688)
##   Precisión obtenida: 0.768
```

El resultado obtenido indica que la media poblacional del índice de masa corporal, inferida a partir de los datos del fichero **osteo**, es un valor que debe estar comprendido entre 23.2 y 24.7 con un 95% de probabilidad. La precisión de la estimación es de 0.8 unidades (estamos aludiendo a los resultados obtenidos con redondeo a un solo decimal). Por defecto, los cálculos se hacen utilizando una confianza del 95%, es decir, 1-α=0.95. Se puede modificar este valor indicándolo a través del parámetro **conf** o, de forma equivalente, a través del parámetro **alfa**. Por ejemplo, si se desea una confianza del 99%, basta añadir a la función **conf=0.99** o bien **alfa=0.01** (puede omitirse el 0 a la izquierda del decimal).

```
icm(osteo$imc, conf=.99)
```

```
##
## Intervalo de confianza bilateral para la media de una VA normal
## ----------------------------------------------------------------
## Información muestral:
##   No hay valores faltantes
##   Tamaño muestral: n =  94
##   Media: m =  23.921
##   Desviación típica: s =  3.748
##   Error estándar de la media: sem =  0.387
##
## Estimación:
##  99%-IC(μ):  (22.904, 24.937)
##   Precisión obtenida: 1.017
```

Como ejercicio, se deben comparar los resultados obtenidos en estos dos últimos ejemplos. La función **imc()**, y también las otras descritas en este paquete, permiten modificar la precisión decimal con la que se muestran los resultados. Para ello se debe indicar el número deseado de decimales en el parámetro **decs**. Por ejemplo

```
icm(osteo$imc, conf=.99, decs=6)
```

```
##
## Intervalo de confianza bilateral para la media de una VA normal
## -----------------------------------------------------------------
## Información muestral:
##   No hay valores faltantes
##   Tamaño muestral: n =  94
##   Media: m =  23.920649
##   Desviación típica: s =  3.748202
##   Error estándar de la media: sem =  0.386598
##
## Estimación:
##  99%-IC(μ):  (22.904, 24.9373)
##   Precisión obtenida: 1.016648
```

devuelve el intervalo al 99% de confianza con 6 cifras decimales (en lugar de 4, que es el valor por defecto). Del mismo modo que se puede indicar como fuente de datos el nombre de una columna de un data.frame, esta función también puede trabajar sobre un vector de datos introducidos de forma manual con la función concatenación **c()**. Por ejemplo, si disponemos de una pequeña muestra de datos sin estructura de data.frame

```
icm(c(1,4,3,2,5,6,5,4,3,2,6,9))
```

```
##
## Intervalo de confianza bilateral para la media de una VA normal
## -----------------------------------------------------------------
## Información muestral:
##   No hay valores faltantes
##   Tamaño muestral: n =  12
##   Media: m =  4.167
##   Desviación típica: s =  2.209
##   Error estándar de la media: sem =  0.638
##
## Estimación:
##  95%-IC(μ):  (2.763, 5.57)
##   Precisión obtenida: 1.403
```

También se puede hacer el cálculo si la información muestral se presenta de forma resumida, indicando el tamaño de la muestra n, su media $\overline{x}$ y su desviación estándar s=4. Estos valores se deben asignar a los parámetros n, m y s respectivamente. Veamos un ejemplo en el que se dispone de una muestra de tamaño n=250, cuya media es $\overline{x}$=187.0 y su desviación típica es s=25.4

```
icm(n=250, m=187, s=25.4)
```

```
##
## Intervalo de confianza bilateral para la media de una VA normal
## -----------------------------------------------------------------
## Información muestral:
##   Tamaño muestral: n =  250
##   Media: m =  187.000
##   Desviación típica: s =  25.400
```

```
##   Error estándar de la media: sem =  1.606
##
## Estimación:
##  95%-IC(μ):  (183.836, 190.164)
##   Precisión obtenida: 3.164
```

Finalmente, cuanto la variable bajo estudio es discreta, con salto de una unidad, pero su distribución se está aproximando a la normal (factible si el recorrido de la variable es grande y el *tamaño de muestra es mayor a 60*), lo procedente es aplicar una *corrección por continuidad* (cpc). Se puede hacer que en la estimación intervenga una *cpc* asignando el valor **FALSE** al parámetro **vac**. el nombre del parámetro alude a las siglas de *variable aleatoria continua* y por defecto toma el valor **TRUE** (los valores FALSE o TRUE pueden abreviarse como F o T). Veamos un ejemplo:

```
icm(c(1,4,3,2,5,6,5,4,3,2,6,9), vac=F)
```

```
##
## Intervalo de confianza bilateral para la media de una VA normal
## ---------------------------------------------------------------
## Información muestral:
##   No hay valores faltantes
##   Tamaño muestral: n =  12
##   Media: m =  4.167
##   Desviación típica: s =  2.209
##   Error estándar de la media: sem =  0.638
##
## Estimación:
##   Se aplica cpc = ±1/(2n) = 0.042 para variable discreta
##  95%-IC(μ):  (2.722, 5.612)
##   Precisión obtenida: 1.445
```

También es instructivo comparar este resultado con el obtenido anteriormente sobre estos mismos datos.

5.1.3 Determinación del tamaño de muestra necesario para estimar la media de una variable normal

Para poder decir algo "interesante" acerca del tamaño muestral que es necesario para garantizar cierta precisión deseada al estimar la media de una variable aleatoria normal, es necesario saber "algo" acerca de su variabilidad. En el caso en que la varianza sea conocida, la fórmula para determinar el tamaño muestral para lograr una precisión d sería:

$$n = \left(\frac{z_\alpha \sigma}{d}\right)^2$$

Aunque se puede calcular a mano esta cantidad, es posible utilizar algunas funciones de R, como la del paquete **samplingbook**, que mediante la función *sample.size.mean(e, s, level=0.95)* permite hacer el calculo de tamaño muestral para una precisión e, una desviación típica s y el nivel de confianza (*level*) deseado. Por ejemplo, en el apartado anterior se vio que la precisión de la estimación con la muestra n=250, $\hat{x}$=187,0 y s=25,4, fue de 3,1639 unidades. El tamaño muestral que es necesario para que dicha precisión aumente a 2 unidades (aumentar la precisión es reducir la anchura del intervalo, por lo tanto 2 unidades representa una precisión mayor que 3 unidades) es de 620 casos, tal y como se puede ver en el siguiente resultado:

```
sample.size.mean(2, 25.4, level=0.95)
```

```
##
## sample.size.mean object: Sample size for mean estimate
## Without finite population correction: N=Inf, precision e=2 and standard deviation S=25.4
##
## Sample size needed: 620
```

Como se ha comentado anteriormente, la mayoría de las veces el valor de la varianza poblacional es desconocido. En estos casos, se puede obtener información sobre la variablidad de la variable a través de una *muestra piloto*. Calculando la varianza s^2 en esa muestra, se puede obtener el tamaño muestral necesario mediante la fórmula:

$$n = \left(\frac{t_\alpha s}{d}\right)^2$$

La función **nm()** del paquete **BioestadisticaR2** permite hacer los cálculos necesarios para saber si el tamaño muestral considerado es suficiente o no. La forma de indicar la información muestral a la función es la misma que se ha visto en el apartado anterior con la función **icm()**. Esto es, asignando el vector o la columna del data.frame al parámetro x (no hace falta escribirlo si se pone esta información en primre lugar), o bien indicando las medidas descriptivas de la misma, es decir, el tamaño muestral n, la media $\overline{x}$ y la desviación estándar s a través de los parámetros **n**, **m** y **s** respectivamente. La precisión que se desea obtener en la estiamción se indica a través del parámetro **d**. Siguiendo con el ejemplo anterior, esta vez bajo el supuesto de varianza desconocida, siendo n=250, $\overline{x}$=187,0 y s=25,4, el tamaño muestral necesario para una precisión de 2 unidades se puede calcular como sigue:

```
nm(n=250, m=187, s=25.4, d=2)
```

```
##
##
## # Tamaño de muestra para la estimación de la media de una VA normal o su aproximación
## # ------------------------------------------------------------------------------
##
## # Muestra piloto:
##   Tamaño muestral:  n = 250
##   Media: m = 187.0000
##   Desviación típica: s = 25.4000
##   Error estandar de la media: sem = 1.6064
##   Precisión observada: d = 3.1639
##
## # Estimación del tamaño muestral:
##   Precisión deseada:  = 2.0000
##   Tamaño muestral necesario: n  626
```

Es decir, que se deben considerar los 250 casos de la muestra piloto y muestrear a 626-250=376 casos más. Igual que se vio con la función **imc()**, se puede indicar la muestra piloto en forma de vector o de columna de data.frame, así como modificar el nivel de confianza para la estimación:

```
nm(c(1,4,3,2,5,6,5,4,3,2,6,9), d=0.3, conf=.99)
```

```
##
##
```

```
## # Tamaño de muestra para la estimación de la media de una VA normal o su aproximación
## # ---------------------------------------------------------------------------------
##
## # Muestra piloto:
##   Tamaño muestral:  n = 12
##   Media: m = 4.1667
##   Desviación típica: s = 2.2088
##   Error estandar de la media: sem = 0.6376
##   Precisión observada: d = 1.9803
##
## # Estimación del tamaño muestral:
##   Precisión deseada:  = 0.3000
##   Tamaño muestral necesario: n  523
```

```
nm(osteo$imc, d=0.5, alfa=0.01)
```

```
##
##
## # Tamaño de muestra para la estimación de la media de una VA normal o su aproximación
## # ---------------------------------------------------------------------------------
##
## # Muestra piloto:
##   Tamaño muestral:  n = 94
##   Media: m = 23.9206
##   Desviación típica: s = 3.7482
##   Error estandar de la media: sem = 0.3866
##   Precisión observada: d = 1.0166
##
## # Estimación del tamaño muestral:
##   Precisión deseada:  = 0.5000
##   Tamaño muestral necesario: n  389
```

En los dos ejemplos, el nivel de confianza exigido es del 99%

5.2 Intervalo de confianza para estimar una proporción

5.2.1 Intervalo de confianza para estimar una proporción

Hasta el momento, esta práctica se ha basado en la media de una población normal. Sin embargo, y especialmente en medicina, es bastante habitual trabajar con la determinación de una proporción (proporción de curados con un nuevo tratamiento, proporción de diabéticos, etc.) Siempre que se esté trabajando con una proporción, se tendrá una característica dicotómica y dos grupos: los que presentan y los que no presenta dicha característica. Más formalmente, dada una variable con distribución Binomial $X \sim B(n, p)$, el objetivo es ahora dar un intervalo de confianza para p en base a una observación de X.

Como se ha visto en teoría, existen varias formas de calcular intervalos de confianza para una proporción. Si se considera que el suceso ha ocurrido x veces de un total de n repeticiones del experimento, el estimador de la proporción será $\hat{p} = \frac{x}{n}$. Un intervalo de confianza para esta proporción, según el método aproximado, viene determinado por la fórmula:

$$\hat{p} \in \hat{p} \pm t_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2n}$$

Cuyo valor se puede calcular utilizando R como calculadora con las siguientes instrucciones:

```r
x=56 #número de ocurrencias del suceso
n=458 #número total de repeticiones del experimento
p=x/n #estimación de la proporción muestral
t.alpha=qt(c(.975), df=457) #valor de la distribución t de Student correspondiente al
#nivel de confianza 0.95 con n-1 grados de libertad


limite.inferior<-p-t.alpha*sqrt((p)*(1-p)/n)+(1/(2*n))
limite.superior<-p+t.alpha*sqrt((p)*(1-p)/n)+(1/(2*n))
intervalo=c(limite.inferior,limite.superior)
intervalo
```

```
## [1] 0.09328031 0.15344458
```

De forma similar, se puede calcular el intervalo de confianza para una proporción por los métodos de Wald o Wilson, entre otros, si bien es cierto que introducir las formulas en R puede resultar bastante tedioso. Afortunadamente, existen varias funciones que nos permiten calcular intervalos de confianza para una proporción por distintos métodos, como la función *binom.confint* de la librería **binom**. Siguiendo con el ejemplo anterior, para un numero de exitos x=56 de un total de n=458 repeticiones del experimento, considerando el nivel de confianza *conf.level=0.95*, la correspondiente instrucción en R sería:

```r
binom.confint(56,458, conf.level=0.95, methods="all")
```

```
##           method  x   n      mean      lower     upper
## 1  agresti-coull 56 458 0.1222707 0.09520793 0.1556172
## 2     asymptotic 56 458 0.1222707 0.09226828 0.1522732
## 3          bayes 56 458 0.1230937 0.09365364 0.1534592
## 4        cloglog 56 458 0.1222707 0.09425095 0.1541354
## 5          exact 56 458 0.1222707 0.09369823 0.1558268
## 6          logit 56 458 0.1222707 0.09529258 0.1555733
## 7         probit 56 458 0.1222707 0.09480606 0.1548878
## 8        profile 56 458 0.1222707 0.09440081 0.1543536
## 9            lrt 56 458 0.1222707 0.09439903 0.1543533
## 10     prop.test 56 458 0.1222707 0.09440139 0.1566502
## 11        wilson 56 458 0.1222707 0.09537041 0.1554547
```

La salida anterior nos muestra la proporción estimada junto con el limite inferior y superior del intervalo de confianza, según el método de Agresti, asintótico, exacto y Wilson entre otros.

También con el paquete **BioestadisticaR2** es posible calcular el intervalo de confianza para una proporción con la función **icp()**. Los argumentos de la función son el numero de casos que cumplen la condición requerida x del total de casos observados n. A continuación se muestra como se realizaría este cálculo en R:

```r
icp(x=56, n=458)
```

```
##
## Intervalo de confianza para una proporción binomial
## ---------------------------------------------------
##
## Información muestral:
##   Tamaño de muestra: n = 458
##   Estimación puntual clásica: p=x/n = 0.1223, q=(1-p)=0.8777
```

```
##   Casos observados: x = 56
##
## # Método exacto (Clooper-Pearson):
##   Pseudo-estimación puntual: p' = 0.1248, q'=(1-p')=0.8752
##   95%-IC(): (0.0938, 0.1558)
##   Semiamplitud: 0.031
##
## # Método de Wilson (con cpc):
##   Pseudo-estimación puntual: p' = 0.1255, q'=(1-p')=0.8745
##   95%-IC(): (0.0944, 0.1567)
##   Semiamplitud: 0.0311
##
## # Método de Wald (con cpc):
##   Estimación puntual (clásica): p=x/n = 0.1223, q=(1-p)=0.8777
##   95%-IC(): (0.0912, 0.1534)
##   Precisión: 0.0311
##
## # Método de Wald ajustado (Agresti-Coull):
##   Estimación puntual: p=(x+2)/(n+4) = 0.1255, q=(1-p)=0.8745
##   95%-IC(): (0.0953, 0.1558)
##   Precisión: 0.0302
```

La función **icp()** devuelve la estimación para la proporción según tres métodos de estimación. El de Wilson se suele considerar el mejor de ellos, pero no siempre es aplicable. Debe verificarse que tanto x como $n - x$ sean mayores a 5. Cuando no ocurre esto, este método no se aplica y la función informa de ello:

```
icp(x=4, n=110)
```

```
##
## Intervalo de confianza para una proporción binomial
## ---------------------------------------------------
##
## Información muestral:
##   Tamaño de muestra: n = 110
##   Estimación puntual clásica: p=x/n = 0.0364, q=(1-p)=0.9636
##   Casos observados: x = 4
##
## # Método exacto (Clooper-Pearson):
##   Pseudo-estimación puntual: p' = 0.0502, q'=(1-p')=0.9498
##   95%-IC(): (0.01, 0.0905)
##   Semiamplitud: 0.0402
##
## # Método de Wilson (con cpc):
##   No aplicable: x=4<5, n-x=106
##
## # Método de Wald (con cpc):
##    No aplicable: x= 4 <20 , n-x= 106
##
## # Método de Wald ajustado (Agresti-Coull):
##   Estimación puntual: p=(x+2)/(n+4) = 0.0526, q=(1-p)=0.9474
##   95%-IC(): (0.0116, 0.0936)
##   Precisión: 0.041
```

Nótese que existe cierta diferencia entre el resultado que arroja esta función, y el obtenido anteriormente con la función **binom.confint**. Esta se debe a que la fórmula de Wilson de **BioestadisticaR2** lleva implementada la *corrección por continuidad*, como se ha visto en teoría, a diferencia de los otros paquetes, en las que esta no se ha tenido en cuenta.

Por otro lado, el método de Wald es más clásico y, en general, el peor. Su validez descansa en que sean x y $n - x$ valores mayores a 20. Observemos que en el ejemplo anterior, este método tampoco se puede aplicar. Finalmente, el método de Wald ajustado siempre da buenos resultados y no requiere de ninguna condición de validez, asi que siempre es una buena opción a utilizar cuando el método de Wilson no es procedente. Cuando el tamaño de muestra es grande y la estimación puntual de la proporción no es un valor extremo (cercano al 0% o al 100%), los tres métodos dan resultados similares. En los tres casos, la función **icp()** informa de la precisión obtenida (expresada en tanto por uno, se puede multiplicar por 100 y expresarla en tanto por ciento). Nótese que el intervalo de Wilson no es simétrico respecto al estimador puntual, de ahí que la función devuelva la amplitud dividida por dos. Esto es una aproximación a la precisión de un intervalo simétrico. Igual que ocurría con las funciones anteriores, el nivel de confianza, o de error, se puede modificar con los parámetros **conf** o **alfa** respectivamente.

```
icp(x=56, n=458, conf=.90)
```

```
##
## Intervalo de confianza para una proporción binomial
## ---------------------------------------------------
##
## Información muestral:
##   Tamaño de muestra: n = 458
##   Estimación puntual clásica: p=x/n = 0.1223, q=(1-p)=0.8777
##   Casos observados: x = 56
##
## # Método exacto (Clooper-Pearson):
##   Pseudo-estimación puntual: p' = 0.1242, q'=(1-p')=0.8758
##   90%-IC(): (0.0979, 0.1504)
##   Semiamplitud: 0.0262
##
## # Método de Wilson (con cpc):
##   Pseudo-estimación puntual: p' = 0.1246, q'=(1-p')=0.8754
##   90%-IC(): (0.0983, 0.1509)
##   Semiamplitud: 0.0263
##
## # Método de Wald (con cpc):
##   Estimación puntual (clásica): p=x/n = 0.1223, q=(1-p)=0.8777
##   90%-IC(): (0.096, 0.1485)
##   Precisión: 0.0263
##
## # Método de Wald ajustado (Agresti-Coull):
##   Estimación puntual: p=(x+2)/(n+4) = 0.1255, q=(1-p)=0.8745
##   90%-IC(): (0.1002, 0.1509)
##   Precisión: 0.0254
```

También es posible estimar una proporción de una variable en un data.frame o un vector, indicando el nivel del factor cuya proporción queremos estimar. Por ejemplo, en el archivo **osteo**, la variable **sexo** toma valores 1 (hombres) y 2 (mujeres), asi que para estimar la proporción de "unos" (hombres) podemos escribir:

```r
icp(osteo$sexo, level=1)
```

```
##
## Intervalo de confianza para una proporción binomial
## ---------------------------------------------------
##
## Información muestral:
##   Tamaño de muestra: n = 94
##   Estimación puntual clásica: p=x/n = 0.4787, q=(1-p)=0.5213
##   Casos observados: (nivel =1)x = 45
##
## # Método exacto (Clooper-Pearson):
##   Pseudo-estimación puntual: p' = 0.4823, q'=(1-p')=0.5177
##   95%-IC(): (0.3804, 0.5843)
##   Semiamplitud: 0.102
##
## # Método de Wilson (con cpc):
##   Pseudo-estimación puntual: p' = 0.4796, q'=(1-p')=0.5204
##   95%-IC(): (0.3755, 0.5837)
##   Semiamplitud: 0.1041
##
## # Método de Wald (con cpc):
##   Estimación puntual (clásica): p=x/n = 0.4787, q=(1-p)=0.5213
##   95%-IC(): (0.3724, 0.585)
##   Precisión: 0.1063
##
## # Método de Wald ajustado (Agresti-Coull):
##   Estimación puntual: p=(x+2)/(n+4) = 0.4796, q=(1-p)=0.5204
##   95%-IC(): (0.3807, 0.5785)
##   Precisión: 0.0989
```

Si dicha variable estuviera presente como un factor con niveles **"Hombre"** y **"Mujer"** (en lugar de 1 y 2), podríamos estimar la proporción de hombres asignando este valor al parámetro **level** mediante la expresión **icp(osteo$sexo, level="Hombre")**

5.2.2 Determinación del tamaño de muestra necesario para estimar una proporción

A diferencia de lo que ocurría con la media de una variable normal, el tamaño muestral necesario para estimar una proporción binomial con una precisión establecida de antemano, se puede determinar sin la necesidad de realizar una muestra piloto. Se habla entonces del *método sin información*. Este método consiste en considerar que la proporción poblacional es del 50%, lo que representa el "peor de los casos" en el sentido de que dicha proporción tiene implícita la mayor icertidumbre y es la que mayor tamaño de muestra requiere para garantizar cierta precisión. Basta entonces con aplicar la fórmula:

$$n = (\frac{z_\alpha}{2d})^2$$

donde $z_\alpha = 1.96$ si consideramos el nivel de confianza del 95%, y d es la precisión deseada. Es posible salir de este "peor de los casos" si se dispone de información muestral y esta pone de manifiesto que la proporción poblacional no debe estar próxima al 50%. A esta conclusión se llega si el intervalo de confianza obtenido de la muestra piloto no contiene al valor 0.5. En ese caso, el tamaño de muestra necesario no es tan grande, pero la estimación, por basarse en la información de una muestra piloto, no es exacta, sino aproximada, y vendrá dada por la expresión:

$$n = \frac{z_\alpha^2 pq}{d^2}$$

En ambos casos es posible realizar los cálculos con la función *np* del paquete **BioestadisticaR2**, de forma similar a como se hizo en el caso de la determinación del tamaño muestral para la media. La información de la muestra piloto se puede indicar de la misma forma que se ha expuesto en la función *icp*: mediante los parámetros **x** y **n**, mediante una columna de un data.frame o un vector junto a la indicación del valor **x**=valor o del nivel del factor **level**="nivel". La forma de indicar la precisión deseada es asignando su valor (en tanto por uno ¡cuidado!) al parámetro **d**. Veamos algunos ejemplos. En el primero, se indica que la precisión deseada es del 3%, no se aporta información de ninguna muestra piloto y la confianza es del 95% (valor por defecto):

```
np(d=.03)
```

```
##
## Tamaño de muestra para estimar una proporción binomial
## ------------------------------------------------------
##
## Tamaño muestral requerido para   = 0.03(3.00%), conf.= 95%
##   sin información previa: n = 1068
```

En este ejemplo se ha considerado "el peor de los casos" en el que la proporción muestral es del 50%. Si ahora contamos con una muestra piloto de tamaño n=90 en la que se observa que el número de casos con la característica de interés es x=15, entonces se puede indicar:

```
np(x=15, n=90, d=.03)
```

```
##
## Tamaño de muestra para estimar una proporción binomial
## ------------------------------------------------------
##
## Información muestral
##   Tamaño de la muestra: n = 90
##   Casos: x = 15
##   Inferencia para la proporción basada en el método de Wald ajustado:
##   95%-IC(): (0.1030, 0.2587)
##   precisión observada: d = 0.0778 (7.78%)
##
## Tamaño muestral requerido para  = 0.03 (3.00%), conf.= 95%
##   - Basado en la muestra actual (po = 0.2587):   n  819
##   - Sin considerar la información previa: n  1068
```

Con la información muestral, se ha elaborado el intervalo de confianza por el método -siempre válido- de Wald ajustado y se observa que el valor 0.5 (50%) no está contenido en dicho intervalo, por lo tanto, la información permite reducir el tamaño muestral necesario para obtener una precisión del 3% de 1068 casos a 819. En el siguiente ejemplo se determina el tamaño muestral necesario para estimar la proporción de mujeres (sexo=2) a partir de los datos del archivo **osteo**

```
np(osteo$sexo, x=2, n=94, d=.03)
```

```
##
## Tamaño de muestra para estimar una proporción binomial
## -------------------------------------------------------
##
## Información muestral
##   Tamaño de la muestra: n = 94
##   Casos: x = 49
##   Inferencia para la proporción basada en el método de Wald ajustado:
##   95%-IC(): (0.4215, 0.6193)
##   precisión observada: d = 0.0989 (9.89%)
##
## Tamaño muestral requerido para  = 0.03 (3.00%), conf.= 95%
##   No se distinguen casos con y sin información (la muestra actual es compatible con p =0.5): n= 1068
```

Como el intervalo de confianza contiene al valor 0.5, la información muestral no aporta indicios de que se pueda "salir del peor de los casos", asi que el tamaño obtenido es el que se considera, precisamente que $\hat{p}$=0.5. Finalmente, si la información muestral se presenta en forma de factor, por ejemplo, la variable sexo tomase los niveles **"Hombre"** y **"Mujer"**, entonces la función tomaría la expresión **np(osteo$sexo, level="Mujer", d=.03)**

Apéndice: el paquete BioestadisticaR2

El paquete **BioestadisticaR2** no está disponible en la red CRAN de R, es necesario descargarlo de la plataforma docente. Para instalarlo en *R Studio* es necesario hacerlo a través de la opción del menú **Tools**→Install Packages como se vio en la práctica 0. Una vez habilitado el paquete, estarán disponibles cuatro funciones dedicadas a la obtención de intervalos de confianza y a la estimación del tamaño muestral necesario para garantizar una determinada precisión.

- **icm()** obtiene el intervalo de confianza para estimar una media μ de una variable alaeatoria con distribución normal
- **nm()** permite determinar el tamaño muestral necesario para que el intervalo anterior tenga una precisión fijada de antemano
- **icp()** obtiene los intervalos de confianza para estimar una proporción binomial π según los métodos de Wilson, Wald y Wald ajustado
- **np()** permite determinar el tamaño muestral necesario para que el intervalo para la proporción binomial tenga la precisión deseada

Es posible obtener ayuda sobre estas funciones escribiendo el nombre de la función tras el signo "?", por ejemplo la orden

```
?icm()
```

provoca que se abra el navegador informando sobre los parámetros y el cometido de la función **icm()** (esto suele ser algo común a todas las funciones de cualquier paquete de R, siempre que su autor se haya preocupado de escribir la información de ayuda).