Lección 2

Intervalos de confianza básicos "a mano"

En esta lección explicamos cómo calcular con `R` algunos intervalos de confianza básicos. En la próxima lección estudiaremos algunas funciones de `R` de contraste de hipótesis que, entre otra información, también calculan algunos de estos intervalos de confianza.

Para facilitar el uso de las funciones definidas en esta lección, hemos creado un paquete llamado `IntervalosMatesII` que las contiene; cuando se carga el paquete, las funciones quedan definidas, se puede pedir ayuda sobre las mismas con `help`, etc. Este paquete no se puede instalar desde el repositorio del CRAN, porque todavía es provisional, así que para instalarlo, tenéis que seguir las instrucciones siguientes:

(a) Descargad en vuestro directorio de trabajo de `R` el fichero `IntervalosMatesII.tar.gz` que encontraréis en la zona de recursos.

(b) Abrid *RStudio*, e id a la pestaña «*Packages*» de la ventana inferior derecha.

(c) Pulsad en «*Install*», y en el menú «*Install from:*» escoged la opción «*Package Archive File*».

(d) Pulsando en «*Browse*», seleccionad vuestra copia del fichero `IntervalosMatesII.tar.gz`.

(e) Pulsad en el botón «*Install*».

(f) Una vez instalado, cargadlo marcándolo en la lista de paquetes instalados o usando la función `library`.

En lo que queda de lección, supondremos este paquete instalado y cargado, y que por lo tanto tenemos acceso a sus funciones.

1. Intervalo de confianza para la media de una población normal con varianza poblacional conocida

Supongamos que queremos estimar la media μ de una población que sigue una distribución normal con desviación típica σ conocida y que, para ello, tomamos una muestra aleatoria simple. En esta situación, un intervalo de confianza a un nivel de confianza del $100(1-\alpha)\,\%$ para μ es

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right[,$$

donde:

- $\overline{X}$ es la media muestral de las observaciones.
- n es el tamaño de la muestra.
- $z_{1-\frac{\alpha}{2}}$ es el $\left(1-\frac{\alpha}{2}\right)$-cuantil de una distribución normal estándar Z. En general, z_q es el q-cuantil para la distribución normal estándar Z, es decir, el número real tal que $P(Z \leqslant z_q) = q$.

Supongamos que, a la hora de calcular este intervalo de confianza con `R`, no disponemos de los datos que forman la muestra, sino que sólo conocemos su media muestral $\overline{x}$ y su tamaño n. En este caso, si denotamos esta media muestral $\overline{x}$ por `x`, el tamaño de la muestra por `n`, la desviación típica σ de la población por `sigma`, y el nivel de confianza $1-\alpha$ (expresado en tanto por uno) por `conf.level`, la expresión siguiente calcula un vector formado por los extremos inferior y superior del intervalo de confianza anterior:

```
c(x-qnorm(1-(1-conf.level)/2)*sigma/sqrt(n),
    x+qnorm(1-(1-conf.level)/2)*sigma/sqrt(n))
```

Si, en cambio, disponemos del vector de datos, llamémosle `X`, entonces para calcular el intervalo de confianza anterior basta sustituir en la expresión anterior `x` por `mean(X)` y `n` por `length(X)`:

```
c(mean(X)-qnorm(1-(1-conf.level)/2)*sigma/sqrt(length(X)),
    mean(X)+qnorm(1-(1-conf.level)/2)*sigma/sqrt(length(X)))
```

El paquete `IntervalosMatesII` contiene la función `ICZ.exact` que cubre ambos casos. Su sintaxis es

```
ICZ.exact(x,sigma,n,conf.level=...,na.rm=...)
```

donde:

- `x` puede ser o bien un vector numérico con la muestra aleatoria simple, o bien un número que representa su media muestral;
- `sigma` es la desviación típica σ de la población;

- `n` es el tamaño de la muestra; si como primer parámetro, `x`, hemos entrado un vector, no hace falta especificar este tamaño;
- `conf.level` es el nivel de confianza $1-\alpha$ expresado en tanto por uno; si no se especifica, toma como valor por defecto 0.95.
- `na.rm` tiene el significado usual, y su valor por defecto es `FALSE`. Si como primer parámetro, `x`, hemos entrado un vector que contiene valores `NA`, el intervalo de confianza que obtendremos tendrá extremos `NA` a no ser que especifiquemos `na.rm=TRUE`.

Veamos dos ejemplos de aplicación de esta función.

Ejemplo 2.1. Tenemos una variable aleatoria normal X de media μ desconocida y desviación típica conocida 0.5. Tomamos una m.a.s. de 30 elementos, y obtenemos una media muestral $\overline{x} = 2.36$. Como se cumplen las hipótesis de esta sección, podemos usar la función `ICZ.exact` para calcular un intervalo de confianza al 90 % para μ.

```
> ICZ.exact(2.36,0.5,30,conf.level=0.9)
  mean size    lower    upper conf.level
1 2.36   30 2.209846 2.510154        0.9
```

El resultado es un *data frame*[1] con las columnas siguientes:

- `mean`, la media muestral;
- `size`, el tamaño de la muestra;[2]
- `lower` and `upper`, los extremos inferior y superior del intervalo de confianza;
- `conf.level`, el nivel de confianza.

Por lo tanto, este resultado muestra que el intervalo de confianza pedido es $]2.209846, 2.510154[$.

Si lo que queremos es un vector con los extremos superior e inferior del intervalo de confianza, basta seleccionarlos de este *data frame*:

```
> IC.df=ICZ.exact(2.36,0.5,30,conf.level=0.9)
> IC.df[1,c(3,4)] #Sólo los extremos del intervalo, como
    dataframe
     lower    upper
1 2.209846 2.510154
> c(IC.df[1,3],IC.df[1,4])  #Sólo los extremos del intervalo,
    como vector
[1] 2.209846 2.510154
```

[1] La ventaja de tener este resultado en forma de *data frame* es que si aplicamos esta función a diferentes muestras, luego podemos organizar fácilmente los resultados como filas de un único *data frame*.

[2] Si, como parámetro `x`, hemos entrado un vector que contenía valores `NA` y hemos especificado `na.rm=TRUE`, el tamaño de la muestra será el número de valores definidos en el vector original.

Ejemplo 2.2. Tenemos una muestra aleatoria simple de pesos, en gramos, de 28 recién nacidos con luxación severa de cadera:

$$2466, 3941, 2807, 3118, 3175, 3515, 3317, 3742, 3062, 3033, 2353, 3515, 3260, 2892,$$
$$4423, 3572, 2750, 3459, 3374, 3062, 3205, 2608, 3118, 2637, 3438, 2722, 2863, 3513.$$

Vamos a suponer por ahora que los pesos al nacer de los bebés con esta patología siguen una distribución normal con la misma desviación típica que la población global de recién nacidos, que es de 800 g. A partir de esta muestra, queremos calcular un intervalo de confianza del 95 % para el peso medio de un recién nacido con luxación severa de cadera, y ver si contiene el peso medio de la población global de recién nacidos, que es de unos 3400 g.

Como la variable aleatoria poblacional es normal de desviación típica (supuestamente) conocida, podemos usar la función `ICZ.exact`, aplicándola directamente al vector y a la desviación típica poblacional; no hace falta especificar el nivel de confianza, porque 0.95 es el que toma dicha función por defecto.

```
> pesos=c(2466,3941,2807,3118,3175,3515,3317,3742,3062,3033,
  2353,3515,3260,2892,4423,3572,2750,3459,3374,3062,3205,
  2608,3118,2637,3438,2722,2863,3513)
> ICZ.exact(pesos,800)
      mean size   lower    upper conf.level
1 3176.429   28 2880.11 3472.747       0.95
```

Obtenemos el intervalo de confianza $]2880.11, 3472.747[$, y vemos que sí que contiene el valor 3400.

2. Cálculo del tamaño muestral para la media fijados la amplitud, la desviación típica poblacional y el nivel de confianza

En las condiciones de la sección anterior, supongamos que estamos interesados en un intervalo de confianza para la media poblacional de una amplitud fijada de antemano A,[3] a un nivel de confianza del $100(1-\alpha)\,\%$, y queremos determinar de qué tamaño n hemos de tomar la muestra aleatoria simple para satisfacer este objetivo. Obviamente, por ahorro, lo que queremos es el menor valor de n con esta propiedad. Resulta que el menor tamaño n de la muestra que asegura que el intervalo de confianza para μ al nivel de confianza $1-\alpha$ tenga una amplitud prefijada A (o, equivalentemente, un *error*, o *precisión*, $A/2$) viene dado por la fórmula

$$n = \left\lceil \left(2z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{A} \right)^2 \right\rceil,$$

donde, recordemos, σ denota la desviación típica poblacional.

[3] En realidad, de amplitud *a lo sumo* A, pero usualmente cometeremos el abuso de lenguaje de decir "de amplitud A".

Si denotamos la amplitud máxima deseada A por `A`, σ por `sigma`, y el nivel de confianza $1-\alpha$ expresado en tanto por uno por `conf.level`, esta fórmula se traduce en la expresión siguiente:

```
ceiling((2*qnorm(1-(1-conf.level)/2)*sigma/A)^2)
```

Esta fórmula está implementada en la función

```
NMin.mu(A,sigma,conf.level=...)
```

del paquete `IntervalosMatesII`. Como antes, el valor por defecto de `conf.level` es 0.95.

Ejemplo 2.3. Si en el Ejemplo 2.2, de pesos de recién nacidos, queremos asegurar una amplitud máxima del intervalo de confianza de 200 g (o lo que sería lo mismo, un error máximo de 100 gramos en nuestra estimación) a un nivel de confianza del 95 %, necesitamos una muestra de tamaño como mínimo:

```
> NMin.mu(200,800)
[1] 246
```

Es decir, necesitamos una muestra de al menos 246 bebés para poder estimar el peso medio de un recién nacido con luxación grave de cadera con una precisión de 100 gramos a un nivel de confianza del 95 %.

3. Intervalo de confianza para la media de una población normal con varianza poblacional desconocida

Supongamos ahora que la población cuya media μ queremos estimar sigue una distribución normal con desviación típica desconocida. En esta situación, si $\overline{X}$, $\widetilde{S}_X$ y n son, respectivamente, la media muestral, la desviación típica muestral y el número de observaciones de la muestra aleatoria simple, un intervalo de confianza del $(1-\alpha)100\,\%$ para μ es

$$\left]\overline{X} - t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{\widetilde{S}_X}{\sqrt{n}},\ \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{\widetilde{S}_X}{\sqrt{n}}\right[$$

donde $t_{n-1,1-\frac{\alpha}{2}}$ es el $\left(1-\frac{\alpha}{2}\right)$-cuantil de una variable aleatoria con distribución t de Student con $n-1$ grados de libertad.

A la hora de calcular este intervalo de confianza, tenemos las dos mismas posibles situaciones que en la Sección 1. Si de la muestra sólo conocemos su media muestral $\overline{x}$, su desviación típica muestral $\widetilde{s}_x$ y su tamaño n, y los denotamos `x`, `sdm` y `n`, respectivamente, y seguimos denotando el nivel de confianza en tanto por uno $1-\alpha$ por `conf.level`, podemos calcular los extremos de este intervalo de confianza con la expresión siguiente:

```
c(x-qt(1-(1-conf.level)/2,n-1)*sdm/sqrt(n),
    x+qt(1-(1-conf.level)/2,n-1)*sdm/sqrt(n))
```

Si en cambio disponemos de un vector numérico `X` con los valores la muestra, podemos calcular este intervalo de confianza sustituyendo en la expresión anterior `x`, `sdm` y `n` por `mean(X)`, `sd(X)` y `length(X)`, respectivamente:

```
c(mean(X)-qt(1-(1-conf.level)/2,n-1)*sd(X)/sqrt(length(X)),
    mean(X)+qt(1-(1-conf.level)/2,n-1)*sd(X)/sqrt(length(X))
```

El paquete `IntervalosMatesII` contiene la función `ICT.exact` que cubre ambos casos. Su sintaxis es

```
ICT.exact(x,sdm,n,conf.level=...,na.rm=...)
```

donde:

- `x` puede ser o bien un vector numérico con la muestra aleatoria simple, o bien un número que representa su media muestral;
- `sdm` es la desviación típica muestral; si como primer parámetro, `x`, hemos entrado un vector, no hace falta especificar este parámetro;
- `n` es el tamaño de la muestra; de nuevo, si `x` es un vector, no hace falta especificar su tamaño;
- `conf.level` y `na.rm` tienen el mismo significado y uso que en `ICZ.exact`.

Veamos un ejemplo de aplicación de esta función.

Ejemplo 2.4. Supongamos que, en la situación del Ejemplo 2.2, decidimos que, en realidad, no conocemos la desviación típica del peso de los recién nacidos con luxación severa de cadera. En este caso, como seguimos suponiendo que estos pesos siguen una ley normal, para calcular un intervalo de confianza del 95 % para su valor medio podemos usar la fórmula basada en la distribución t de Student, y por lo tanto la función `ICT.exact`:

```
> ICT.exact(pesos)
      mean size    lower    upper conf.level
1 3176.429   28 2997.849 3355.008       0.95
```

El intervalo que obtenemos es $]2997.849, 3355.008[$ y en este caso *no* contiene el peso medio global de 3400 g. Los intervalos son muy diferentes, pero claro, es que si calculáis `sd(pesos)` veréis que es 460.5411, muy lejos de los 800 g que suponíamos en el primer ejemplo. Esto nos lleva a pensar que este último intervalo es más adecuado que el del Ejemplo 2.2.

4. Intervalo de confianza para la media cuando la muestra es grande

Seguimos con el problema de determinar un intervalo de confianza para la media poblacional μ a partir de una muestra aleatoria simple. Supongamos ahora que la distribución de la

población no es necesariamente normal, pero que el tamaño de la muestra es grande: por fijar una cota, de tamaño como mínimo 40.

En esta situación, si $\overline{X}$ denota la media de la muestra, $\widetilde{S}_X$ su desviación típica muestral, y n su tamaño, se puede tomar como intervalo de confianza para la media poblacional μ al nivel $(1-\alpha)100\,\%$ el siguiente:

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\widetilde{S}_X}{\sqrt{n}},\ \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\widetilde{S}_X}{\sqrt{n}}\right[,$$

donde, recordemos, $z_{1-\frac{\alpha}{2}}$ es el $\left(1-\frac{\alpha}{2}\right)$-cuantil de una variable aleatoria normal estándar.

Las instrucciones de `R` para calcular este intervalo son las mismas que las de la sección anterior, cambiando sistemáticamente el cuantil de la t de Student por el correspondiente cuantil de la normal estándar. Así, y con las mismas notaciones que en la sección anterior, si sólo disponemos de los estadísticos de la muestra, el intervalo de confianza se calcula con la expresión

```
c(x-qnorm(1-(1-conf.level)/2)*sdm/sqrt(n),
    x+qnorm(1-(1-conf.level)/2)*sdm/sqrt(n))
```

y si disponemos de un vector `X` con la muestra, se puede usar la expresión

```
c(mean(X)-qnorm(1-(1-conf.level)/2)*sd(X)/sqrt(length(X)),
    mean(X)+qnorm(1-(1-conf.level)/2)*sd(X)/sqrt(length(X))
```

El paquete `IntervalosMatesII` contiene la función `ICZ.approx` que cubre ambos casos. Su sintaxis y uso son los mismos que los de `ICT.exact`, salvo naturalmente por el nombre de la función. Veamos un ejemplo.

Ejemplo 2.5. Se ha realizado un estudio del efecto del calor en la movilidad de los caracoles comunes de jardín de la especie *Helix aspersa* (llamados "caragols bovers" en Mallorca). Se han medido las distancias en cm recorridas por 75 caracoles a una temperatura de 24º, y han sido las siguientes:

> 4.9, 5.6, 5.8, 4.5, 5.1, 4.8, 5.6, 5.7, 5.3, 4.6, 5.3, 4.9, 5.8, 6.8, 4.6, 3.8, 5.2, 4.5, 5.3, 4.8, 4.4, 4.8, 4.5, 4.7, 4.7, 5.3, 3.9, 4.1, 6.0, 5.8, 4.6, 4.7, 5.7, 5.6, 5.9, 6.1, 5.4, 4.3, 5.1, 5.1, 3.2, 3.7, 4.3, 4.6, 4.7, 4.9, 4.9, 4.4, 6.2, 4.5, 4.9, 5.4, 4.3, 4.4, 5.5, 5.1, 4.4, 5.1, 5.8, 5.2, 5.2, 5.4, 4.8, 4.6, 4.9, 4.6, 5.4, 4.4, 4.9, 4.0, 5.0, 3.8, 3.9, 6.0, 5.7.

A partir de estos datos, y sabiendo que a 18º se estima que recorren una media de 2.8 cm, queremos estimar si el aumento de temperatura influye en la distancia recorrida. Para ello, calcularemos un intervalo de confianza del 95 % para la distancia media recorrida por un caracol de esta especie a 24º y comprobaremos si contiene la distancia media recorrida a 18º.

Como la muestra es grande, podemos usar la fórmula anterior para calcular este intervalo de confianza.

```
> dists=c(4.9,5.6,5.8,4.5,5.1,4.8,5.6,5.7,5.3,4.6,5.3,4.9,5.8,
6.8,4.6,3.8,5.2,4.5,5.3,4.8,4.4,4.8,4.5,4.7,4.7,5.3,3.9,4.1,
6.0,5.8,4.6,4.7,5.7,5.6,5.9,6.1,5.4,4.3,5.1,5.1,3.2,3.7,4.3,4.6,
4.7,4.9,4.9,4.4,6.2,4.5,4.9,5.4,4.3,4.4,5.5,5.1,4.4,5.1,5.8,5.2,
5.2,5.4,4.8,4.6,4.9,4.6,5.4,4.4,4.9,4.0,5.0,3.8,3.9,6.0,5.7)
> #Calculamos el IC, redondeado a 2 cifras decimales
> round(ICZ.approx(dists),2)
  mean size lower upper conf.level
1 4.96   75   4.8  5.11       0.95
```

Así pues, un intervalo de confianza del 95 % para la distancia media recorrida por un caracol de esta especie a 24º es $]4.8, 5.11[$. Esto nos dice que, con una probabilidad del 95 %, esta distancia media cae dentro de este intervalo, y por lo tanto es muy superior a los 2.8 cm recorridos de media a 18º.

5. Intervalos de confianza para la proporción poblacional

En esta sección consideramos el caso en que la población objeto de estudio sigue una distribución Bernoulli y que queremos estimar su probabilidad de éxito (o *proporción poblacional*) p. Para ello, tomamos una muestra aleatoria simple de tamaño n y número de éxitos x, y, por lo tanto, de *proporción muestral* de éxitos $\widehat{p}_X = x/n$.

El *método "exacto" de Clopper-Pearson* para calcular un intervalo de confianza del $(1-\alpha)100\,\%$ para p, que se puede usar siempre, sin ninguna restricción sobre la muestra, consiste básicamente en resolver las ecuaciones

$$\sum_{k=x}^{n} p_0^k (1-p_0)^{n-k} = \frac{\alpha}{2}, \qquad \sum_{k=0}^{x} p_1^k (1-p_1)^{n-k} = \frac{\alpha}{2}$$

y dar el intervalo $]p_0, p_1[$. Como es muy difícil de explicar a este nivel cómo se resuelven estas ecuaciones, lo más práctico es usar la función `binom.exact` del paquete `epitools`, que calcula este intervalo. Su sintaxis es

```
binom.exact(x,n,conf.level)
```

donde `x` y `n` representan, respectivamente, el número de éxitos y el tamaño de la muestra, y `conf.level` $= 1-\alpha$. El valor por defecto de `conf.level` es 0.95.

Ejemplo 2.6. Supongamos que, de una muestra de 15 enfermos tratados con un cierto medicamento, 1 ha desarrollado taquicardia. Querríamos conocer un intervalo de confianza del 95 % para la proporción de enfermos tratados con este medicamento que presentan este efecto adverso.

Tenemos una población Bernoulli, formada por los enfermos tratados con el medicamento en cuestión, donde los éxitos son los enfermos que desarrollan taquicardia. La fracción de éstos es la fracción poblacional p para la que queremos calcular el intervalo de confianza del 95 %. Para ello cargamos el paquete `epitools` y usamos `binom.exact`:

```
> #Instalamos y cargamos el paquete epitools
...
> round(binom.exact(1,15),3)
  x  n proportion lower upper conf.level
1 1 15      0.067 0.002 0.319       0.95
```

Como podéis observar, el resultado de la función `binom.exact` es un *data frame* similar a los que producen las funciones de `IntervalosMatesII`; el intervalo de confianza deseado está formado por los números en las columnas `lower` (extremo inferior) y `upper` (extremo superior). Hemos redondeado el resultado a 3 cifras decimales (algo muy razonable en general al manejar estimaciones de proporciones: corresponde a redondear a décimas de punto porcentual), obteniendo el intervalo de confianza $]0.002, 0.319[$: podemos afirmar con un nivel de confianza del 95 % que la proporción de enfermos tratados con este medicamento que presentan este efecto adverso está entre el 0.2 % y el 31.9 %.

Supongamos ahora que el tamaño n de la muestra aleatoria simple es grande; de nuevo, pongamos, $n \geqslant 40$. En esta situación, podemos usar el *Método de Wilson* para aproximar, a partir del Teorema Central del Límite, un intervalo de confianza del parámetro p al nivel de confianza $(1-\alpha)100\,\%$, mediante la fórmula

$$\left]\frac{\widehat{p}_X + \frac{z^2_{1-\alpha/2}}{2n} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_X\widehat{q}_X}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{z^2_{1-\alpha/2}}{n}}, \frac{\widehat{p}_X + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_X\widehat{q}_X}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{z^2_{1-\alpha/2}}{n}}\right[$$

donde $\widehat{q}_X = 1 - \widehat{p}_X$ y, como antes, $z_{1-\frac{\alpha}{2}}$ es el $\left(1-\frac{\alpha}{2}\right)$-cuantil de una variable aleatoria normal estándar.

Sería fácil traducir esta fórmula en una función de `R`, pero la función resultante sería muy larga y complicada, y ya está hecho en el paquete `epitools`. Se trata de la función `binom.wilson`, cuya sintaxis es

```
binom.wilson(x,n,conf.level)
```

con los mismos parámetros que `binom.exact`.

Ejemplo 2.7. Supongamos que tratamos 45 ratones con un agente químico, y 10 de ellos desarrollan un determinado cáncer de piel. Queremos estimar si, a un nivel de confianza del 90 %, podemos concluir que la proporción p de ratones que desarrollan este cáncer de piel al ser tratados con este agente químico es superior al 15 %.

Como 45 es relativamente grande, usaremos el método de Wilson. Para comparar los resultados, usaremos también el método exacto. Por comodidad, redondearemos a 3 cifras decimales.

```
> round(binom.wilson(10,45,0.9),3)
   x  n proportion lower upper conf.level
1 10 45      0.222 0.138 0.338        0.9
```

```
> round(binom.exact(10,45,0.9),3)
   x  n proportion lower upper conf.level
1 10 45      0.222 0.126 0.348        0.9
```

Con el método de Wilson obtenemos el intervalo $]0.138, 0.338[$ y con el método exacto, el intervalo $]0.126, 0.348[$. Hay una diferencia en los extremos de alrededor de un punto porcentual. En todo caso, ambos intervalos contienen valores inferiores a 0.15, por lo que no podemos concluir a este nivel de confianza que $p > 0.15$.

Supongamos finalmente que la muestra aleatoria simple es considerablemente más grande que la usada en el método de Wilson y que, además, la proporción muestral de éxitos $\widehat{p}_X$ está alejada de 0 y de 1. Una posible manera de formalizar estas condiciones es requerir que $n \geqslant 100$ y que $n\widehat{p}_X \geqslant 10$ y $n(1-\widehat{p}_X) \geqslant 10$; observaréis que estas dos últimas condiciones son equivalentes a que tanto el número de éxitos como el número de fracasos en la muestra sean como mínimo 10. En este caso, se puede usar la *fórmula de Laplace*, que simplifica la de Wilson (aunque, en realidad, la precede en más de 100 años): un intervalo de confianza del parámetro p al nivel de confianza $(1-\alpha)100\,\%$ viene dado aproximadamente por la fórmula

$$\left]\widehat{p}_X - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}, \widehat{p}_X + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}\right[.$$

Esta fórmula sí que es fácil de traducir a `R`: si la muestra aleatoria tiene tamaño `n` y contiene `x` éxitos y denotamos $1-\alpha$ por `conf.level`, los extremos inferior y superior del intervalo de confianza son los elementos del siguiente vector:

```
c((x/n)-qnorm(1-(1-conf.level)/2)*sqrt((x/n)*(1-(x/n))/n),
    (x/n)+qnorm(1-(1-conf.level)/2)*sqrt((x/n)*(1-(x/n))/n))
```

Este intervalo de confianza se obtiene con la función

```
binom.approx(x,n,conf.level)
```

del paquete `epitools`, de uso similar al de las dos funciones anteriores. También se puede obtener con la función

```
ICZ.p(x,n,conf.level=...)
```

de `IntervalosMatesII`, con la diferencia respecto de `binom.approx` que permite entrar como `x` un vector de ceros y unos, en cuyo caso no hay que especificar su tamaño `n`, y que genera un mensaje de advertencia si las condiciones explicadas más arriba para poder aplicar la fórmula de Laplace no se satisfacen.

Ejemplo 2.8. En una muestra aleatoria de 500 familias con niños en edad escolar de una determinada ciudad se ha observado que 340 introducen fruta de forma diaria en la dieta de sus hijos. Se pide encontrar un intervalo de confianza del 95 % para la proporción real de familias de esta ciudad con niños en edad escolar que incorporan fruta fresca de forma diaria en la dieta de sus hijos.

Tenemos una población Bernoulli donde los éxitos son las familias que aportan fruta de forma diaria a la dieta de sus hijos, y la fracción de estas familias en el total de la población es la fracción poblacional p para la que queremos calcular el intervalo de confianza. Como n es muy grande y los números de éxitos y fracasos también lo son, podemos emplear el método de Laplace.

```
> round(ICZ.p(340,500),3)
  proportion size lower upper conf.level
1       0.68  500 0.639 0.721       0.95
> round(binom.approx(340,500),3) #Con binom.approx de epitools
    x   n proportion lower upper conf.level
1 340 500       0.68 0.639 0.721       0.95
```

Por lo tanto, según la fórmula de Laplace, un intervalo de confianza al 95 % para la proporción poblacional es $]0.639, 0.721[$. ¿Qué hubiéramos obtenido con los otros dos métodos?

```
> round(binom.wilson(340,500,0.95),3)
    x   n proportion lower upper conf.level
1 340 500       0.68 0.638 0.719       0.95
> round(binom.exact(340,500,0.95),3)
    x   n proportion lower upper conf.level
1 340 500       0.68 0.637 0.721       0.95
```

Como veis, los resultados son muy parecidos, con diferencias de unas pocas milésimas.

6. Cálculo del tamaño muestral para una proporción fijados la amplitud y el nivel de confianza

En las condiciones de la sección anterior, supongamos que estamos interesados en calcular un intervalo de confianza para la proporción poblacional de amplitud (a lo sumo) A con un nivel de confianza $1-\alpha$, y queremos determinar el menor tamaño de la muestra n para satisfacer este objetivo. Como esperamos que la n sea grande, tiene sentido emplear la fórmula de Laplace. La amplitud del intervalo de confianza al nivel de confianza $1-\alpha$ en este caso es

$$A = 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}.$$

Ahora bien, no conocemos $\widehat{p}_X$, porque de hecho aún no hemos tomado ninguna muestra. Por lo tanto, nos tenemos que poner en el peor de los casos, que es cuando $\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}$ toma su valor máximo. Este máximo se alcanza en $\widehat{p}_X = 0.5$ y vale $0.5/\sqrt{n}$. Por consiguiente, en el peor de los casos ($\widehat{p}_X = 1-\widehat{p}_X = 0.5$), el menor tamaño de la muestra que nos garantiza la amplitud y nivel de confianza fijados es

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{A^2} \right\rceil.$$

Podemos calcular este valor con la expresión siguiente (donde `A` representa la amplitud máxima deseada A y `conf.level` el nivel de confianza $1-\alpha$):

```
ceiling((qnorm(1-(1-conf.level)/2)/A)^2)
```

Esta fórmula está implementada en la instrucción

```
NMin.p(A,conf.level=...)
```

del paquete `IntervalosMatesII`.

Ejemplo 2.9. Supongamos que estamos en una situación similar a la del último ejemplo. Queremos calcular el tamaño mínimo de la muestra n que nos asegure un intervalo de confianza con una precisión de una centésima (es decir con una amplitud de $A = 0.02$) a un nivel de confianza del 90 %.

```
> NMin.p(0.02,0.9)
[1] 6764
```

Por lo tanto, en el peor de los casos, el tamaño muestral que nos asegura una precisión de 0.01 con nivel de confianza del 90 % es $n = 6764$, lo que significa muchísimas familias (para que os hagáis una idea, en Palma hay unas 35 000 familias con niños en edad escolar). En general, la aproximación del tamaño muestral de esta manera tiende a dar muestras muy grandes.

7. Intervalo de confianza para la varianza de una población normal

Supongamos ahora que queremos estimar la varianza σ^2, o la desviación típica σ, de una población que sigue una distribución normal. Tomamos una muestra aleatoria simple de tamaño n, y sea $\widetilde{S}_X$ su desviación típica muestral. En esta situación, un intervalo de confianza del $(1-\alpha)100\,\%$ para σ^2 es

$$\left]\frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}},\ \frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right[,$$

donde $\chi^2_{n-1,\frac{\alpha}{2}}$ y $\chi^2_{n-1,1-\frac{\alpha}{2}}$ son, respectivamente, los cuantiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$ de una variable aleatoria que sigue una distribución χ^2 con $n-1$ grados de libertad.

Por lo tanto, si denotamos la varianza muestral por `S2`, el tamaño n de la muestra por `n`, y $1-\alpha$ por `conf.level`, los extremos inferior y superior de este intervalo de confianza se obtienen como los elementos del vector siguiente:

```
c((n-1)*S2/qchisq(1-(1-conf.level)/2,n-1),
    (n-1)*S2/qchisq((1-conf.level)/2,n-1))
```

Si conocemos el vector de datos `X`, la fórmula para calcular este intervalo de confianza con `R` se obtiene sustituyendo en la expresión anterior `S2` por `var(X)` y `n` por `length(X)`. Ambos casos quedan cubiertos por la función

```
IC.var(x,n,conf.level=...,na.rm=...)
```

del paquete `IntervalosMatesII`, donde:

- `x` puede ser o bien un vector numérico con la muestra aleatoria simple, o bien un número que representa su varianza muestral;
- `n` es el tamaño de la muestra; si como primer parámetro, `x`, hemos entrado un vector, no hace falta especificar este tamaño;
- `conf.level` y `na.rm` tienen el mismo significado y uso que en las funciones anteriores.

Ejemplo 2.10. Un índice de calidad de un reactivo químico es el tiempo que tarda en actuar. Se supone que la distribución de este tiempo de actuación del reactivo es aproximadamente normal.

Se realizan 30 pruebas, que forman una muestra aleatoria simple, en las que se mide el tiempo de actuación del reactivo. Los tiempos obtenidos son

> 12, 13, 13, 14, 14, 14, 15, 15, 16, 17, 17, 18, 18, 19, 19, 25, 25, 26, 27, 30, 33, 34, 35, 40, 40, 51, 51, 58, 59, 83.

Se pide calcular un intervalo de confianza para la varianza de este tiempo de actuación al nivel 95 %.

El siguiente código carga los tiempos obtenidos en la variable `reactivo` y calcula este intervalo, redondeado a 3 cifras decimales:

```
> reactivo = c(12,13,13,14,14,14,15,15,16,17,17,18,18,19,19,
  25,25,26,27,30,33,34,35,40,40,51,51,58,59,83)
> round(IC.var(reactivo),3)
  variance size   lower   upper conf.level
1  301.551   30 191.263 544.957       0.95
```

Por lo tanto un intervalo de confianza para la varianza poblacional al nivel de confianza del 95 % es $]191.263, 544.957[$.

8. Ejercicios resueltos

En esta sección presentamos varios ejercicios sencillos de intervalos de confianza resueltos con `R`.

Ejemplo 2.11. En una muestra aleatoria simple de 45 personas, se ha estudiado el porcentaje de aumento del contenido de alcohol en la sangre después de ingerir cuatro cervezas, obteniéndose una media de $\overline{x} = 41.2$ con una desviación típica muestral de $\widetilde{s} = 2.1$.

Calcula un intervalo de confianza del 90 % para el porcentaje medio de aumento del contenido de alcohol en la sangre de una persona, después de tomar cuatro cervezas. ¿Creerías la afirmación de que el incremento medio es menor del 35 %? ¿Por qué?

Como la muestra es de tamaño grande, 45, podemos utilizar la función `ICZ.approx`:

```
> round(ICZ.approx(41.2,2.1,45,0.9),2)
  mean size lower upper conf.level
1 41.2   45 40.69 41.71        0.9
```

Obtenemos el intervalo de confianza $]40.69, 41.71[$, muy por encima de 35. Por lo tanto, no nos creeríamos que el incremento medio es menor del 35 %.

Ejemplo 2.12. La agencia de Protección del Medio Ambiente identificó recientemente en Estados Unidos 30000 vertederos de basura considerados al menos potencialmente peligrosos. ¿Qué tamaño muestral se necesita para estimar el porcentaje de estos lugares que suponen una amenaza para la salud, con un error de a lo sumo 2 puntos porcentuales y con una confianza del 90 %?

Queremos saber cuántos vertederos tenemos que muestrear para estimar el porcentaje de peligrosos con un error de, como máximo, 2 % (dos puntos porcentuales), es decir, de 0.02. Por lo tanto nos piden el tamaño muestral para obtener un intervalo de confianza de amplitud a lo sumo $A = 0.04$ a un nivel de confianza del 90 %. Estamos en la situación de la Sección 6, por lo que tomaremos

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{A^2} \right\rceil$$

```
> NMin.p(0.04,0.9)
[1] 1691
```

Por lo tanto el tamaño mínimo de la muestra, supuesto el peor caso, es $n = 1691$.

Ejemplo 2.13. La empresa *RX-print* ofrece una impresora de altísima calidad para la impresión de radiografías. En su publicidad afirma (incluyendo la nota a pie de página), que "sus cartuchos imprimen un promedio de 500 radiografías*."

Una organización de radiólogos desea comprobar esta afirmación, y toma una muestra aleatoria de $n = 25$ cartuchos, obteniendo una media de $\overline{X} = 518$ radiografías por cartucho y una desviación estándar muestral $\widetilde{S}_X = 40$. Con esta muestra, ¿cae la media poblacional que afirma el fabricante dentro del intervalo de confianza del 90 %?

Este problema se reduce a calcular el intervalo de confianza para el número medio μ de radiografías por cartucho a un nivel de confianza $1 - \alpha = 0.9$ y comprobar si contiene el valor 500 anunciado por el fabricante. Como el fabricante supone que el número de

*Datos técnicos: Muestra mensual de tamaño $n = 25$, población supuesta normal, nivel de confianza del 90 %.

radiografías por cartucho sigue una ley normal, lo usaremos para calcular el intervalo de confianza, por lo que podemos emplear la función `ICT.exact`:

```
> round(ICT.exact(518,40,25,0.9),2)
  mean size  lower  upper conf.level
1  518   25 504.31 531.69        0.9
```

El intervalo que obtenemos es $]504.31, 531.69[$. En este caso la afirmación del fabricante queda contradicha por la muestra, puesto que el valor 500 cae fuera de este intervalo de confianza. De todas formas, el error favorece al consumidor: parece que la media de radiografías por cartucho es superior a 500.

Ejemplo 2.14. En los inviernos rigurosos, se utiliza sal para quitar el hielo de las carreteras. Para hallar la cantidad aproximada de sal que se está introduciendo en el medio ambiente por esta causa, se realizó un estudio en Nueva Inglaterra. Se obtuvieron las siguientes observaciones sobre la variable aleatoria X, número de toneladas de sal utilizadas sobre las carreteras en una semana, en distritos aleatoriamente seleccionados a lo largo de la región:

3900, 3875, 3820, 3860, 3840, 3852, 3800, 3825, 3790

a) Calculad una estimación puntual de la media μ de X.

b) Calculad una estimación puntual de la varianza σ^2 y la desviación típica σ de X.

c) Suponed a partir de ahora que X está normalmente distribuida. Calculad un intervalo de confianza para μ del $90\,\%$.

d) Calculad intervalos de confianza del $90\,\%$ para σ^2 y para σ.

a) Una estimación puntual de μ sería la media de la muestra.

```
> datos=c(3900,3875,3820,3860,3840,3852,3800,3825,3790)
> mean(datos)
 [1] 3840.222
```

b) Una estimación puntual de σ^2 sería la varianza muestral, y de σ, la desviación típica muestral.

```
> var(datos)
[1] 1261.694
> sd(datos)
[1] 35.52034
```

c) Si X está normalmente distribuida, pero no conocemos su varianza, para encontrar un intervalo de confianza para la media poblacional hemos de proceder como en la Sección 3, usando la función `ICT.exact`:

```
> round(ICT.exact(datos,conf.level=0.9),3)
      mean size    lower    upper conf.level
1 3840.222    9 3818.205 3862.239        0.9
```

d) Como la población es normal, para obtener un intervalo de confianza para la varianza poblacional podemos proceder como usar la función `IC.var` de la Sección 7:

```
> round(IC.var(datos,conf.level=0.9),3)
  variance size  lower    upper conf.level
1 1261.694    9 650.89 3693.706        0.9
```

El intervalo de confianza para σ tendrá como extremos las raíces cuadradas de los extremos del intervalo de confianza para σ^2:

```
> round(sqrt(IC.var(datos,conf.level=0.9)[1,3:4]),3)
   lower  upper
1 25.513 60.776
```

Ejemplo 2.15. Se ha efectuado un estudio sobre la obesidad en niños menores de 12 años. Se ha obtenido una muestra aleatoria de 100 niños obesos y de cada uno se ha averiguado la edad en la que comenzó a sufrir la obesidad. Se ha determinado que la media muestral es de 4 años, con una desviación típica muestral de 1.5 años.

a) Encontrad un intervalo de confianza del $95\,\%$ para la edad media del inicio de la obesidad de los niños.

b) Suponiendo que la edad en la que los niños obesos empiezan a sufrir obesidad sigue una distribución normal, determinad un intervalo de confianza del $95\,\%$ para su desviación típica.

(Basado en la información publicada en R. Unger, L. Kreeger, C. Christoffel, "Childhood Obesity", *Clinical Pediatrics,* 1990, págs. 368-373.)

La siguiente sesión de `R` resuelve este ejercicio.

```
> #Apartado a)
> round(ICZ.approx(4,1.5,100),2)
  mean size lower upper conf.level
1    4  100  3.71  4.29       0.95
> #Apartado b)
> round(IC.var(1.5^2,100),4)
  variance size  lower  upper conf.level
1     2.25  100 1.7345 3.0364       0.95
> round(sqrt(IC.var(1.5^2,100)[1,3:4]),2)
  lower upper
1  1.32  1.74
```

Modelo de test

Son cuatro preguntas de uso de las funciones explicadas en esta lección. Damos algunos ejemplos del estilo de las preguntas propuestas, pero os pueden salir otras.

(1) Tenemos una población normal de media μ desconocida y desviación típica 1.2. Tomamos una muestra aleatoria simple de 20 individuos y obtenemos una media muestral de 6.2. Calculad el extremo inferior de un intervalo de confianza para la μ a un nivel de confianza del 92 %. Dadlo redondeado a dos cifras decimales.

(2) Tenemos una población Bernoulli de proporción poblacional p desconocida. Tomamos una muestra aleatoria simple de 80 individuos y obtenemos una proporción muestral de 35 % de éxitos. Calculad el extremo inferior de un intervalo de confianza para p a un nivel de confianza del 92 % usando el método de Wilson. Dadlo redondeado a cuatro cifras decimales.

(3) Tenemos una población Bernoulli de proporción poblacional p desconocida. Calculad el tamaño mínimo de una muestra que garantice un intervalo de confianza del 92 % para p de amplitud $A = 0.3$.

Respuestas

(1) `5.73`

(2) `0.2637`

(3) `35`