Examen 1 2024

Javier Orín

2025-12-09

Problema 1 (3 puntos). Cierta ceca fabrica monedas de forma que la probabilidad de que salga cara si tiramos al aire la moneda i-esima es p_i. Es decir, distintas monedas tienen distinta probabilidad de cara. Cada p_i puede estimarse con mucha precision, por lo que dada una muestra concreta de n monedas sus p_i se suponen conocidas ($i = 1, ..., n$). Las p_i provienen de una poblacion beta $B(q, q)$, es decir, con ambos parametros iguales. Se dispone de una realización muestral de tamaño $n = 10$, a saber, (p1, . . . , p10) = (0.449, 0.515, 0.432, 0.526, 0.433, 0.539, 0.560, 0.546, 0.476, 0.630).

a) Halla un estadístico suficiente minimal para q (0,6 puntos).

Sabemos que un estadístico es suficiente minimal si y solo si induce la siguiente partición:

$$\vec{x} \sim \vec{y} \iff \frac{f(\vec{x}; q)}{f(\vec{y}; q)} \text{ no depende de } q$$

Veamos cómo es la función de densidad de la variable aleatoria:

$$f_X(x) = \frac{1}{\beta(q, q)}(x - x^2)^{q-1} \cdot I_{(0,1)}(x)$$

Pasemos a acomparar ahora dos m.a.s. de tamaño n de la variable aleatoria:

$$\frac{f(\vec{x}; q)}{f(\vec{y}; q)} = \prod_{i=1}^{n} \left(\frac{x_i - x_i^2}{y_i - y_i^2} \right)^{q-1} = \left(\prod_{i=1}^{n} \frac{x_i - x_i^2}{y_i - y_i^2} \right)^{q-1}$$

Tenemos que el cociente de densidades es una base elevado al exponente $q - 1$. La única forma de que el resultado no dependa de q es hacer que la base sea igual a 1; o, equivalentemente, considerar el siguiente estadístico:

$$T(\vec{x}) = \prod_{i=1}^{n} x_i - x_i^2$$

Otra posible alternativa es expresar la variable aleatoria como parametrización natural de la familia exponencial. Se obtiene una transformación biyectiva de esta solución por medio de logaritmos.

b) Halla una estimacion de q por el metodo de máxima verosimilitud (0,8 puntos).

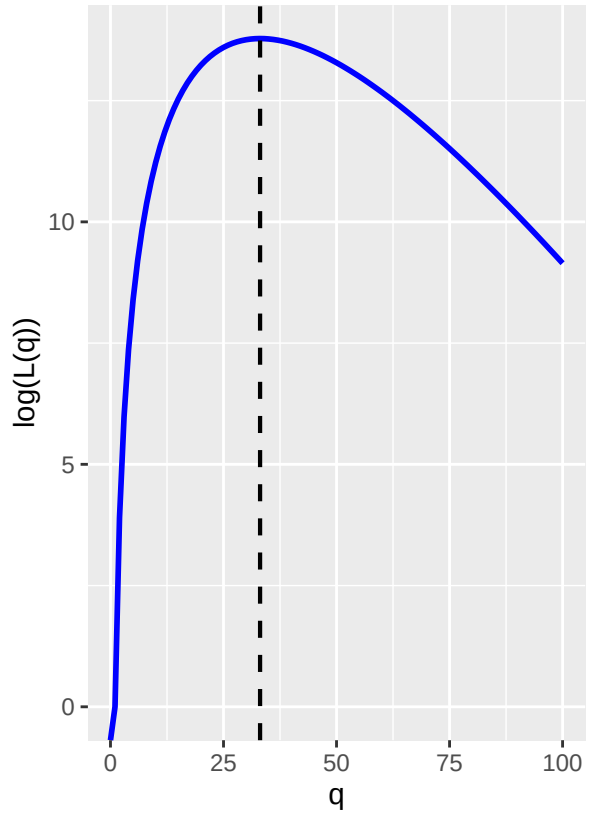
Cargamos la muestra en R:

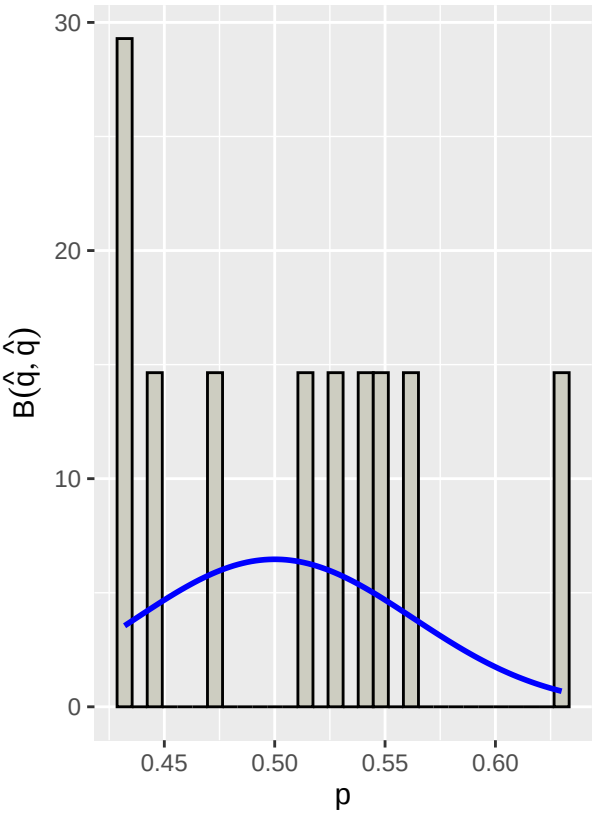
```
prob_caras = c(0.449, 0.515, 0.432, 0.526, 0.433, 0.539, 0.560, 0.546, 0.476, 0.630)
```

Definimos la función `log_-_verosimilitud` para optimizar numéricamente:

```
log_verosimilitud = function(q) {
  n = length(prob_caras)
  s = sum(log(prob_caras) + log(1 - prob_caras))
  (q-1)*s  -n*log(beta(q,q))
}
(q_emv = optimise(f=log_verosimilitud, interval = c(0, 100), maximum=TRUE)$maximum)
```

```
## [1] 33.08617
```





c) Halla una estimación de q por el método de los momentos (0,8 puntos).

Por ser $X \sim B(q, q)$, sabemos

$$E(X) = \frac{q}{q+q} = \frac{1}{2} \qquad\qquad Var(X) = \frac{q^2}{(2q)^2 \cdot (2q+1)} = \frac{1}{8q+4}$$

El método de los momentos aproxima los momentos poblacionales por los muestrales. Por tanto, despejamos q como sigue:

$$s^2 = \frac{1}{8q+4} \Rightarrow 8q+4 = \frac{1}{s^2} \Rightarrow q = \frac{1}{8}\left(\frac{1}{s^2} - 4\right) = \frac{1}{8s^2} - \frac{1}{2}$$

Veamos qué valor toma para la muestra:

```
(q_mom = -0.5 + 1/(8 * var(prob_caras)))
```

```
## [1] 30.57357
```

Es cercano al valor obtenido por el EMV.

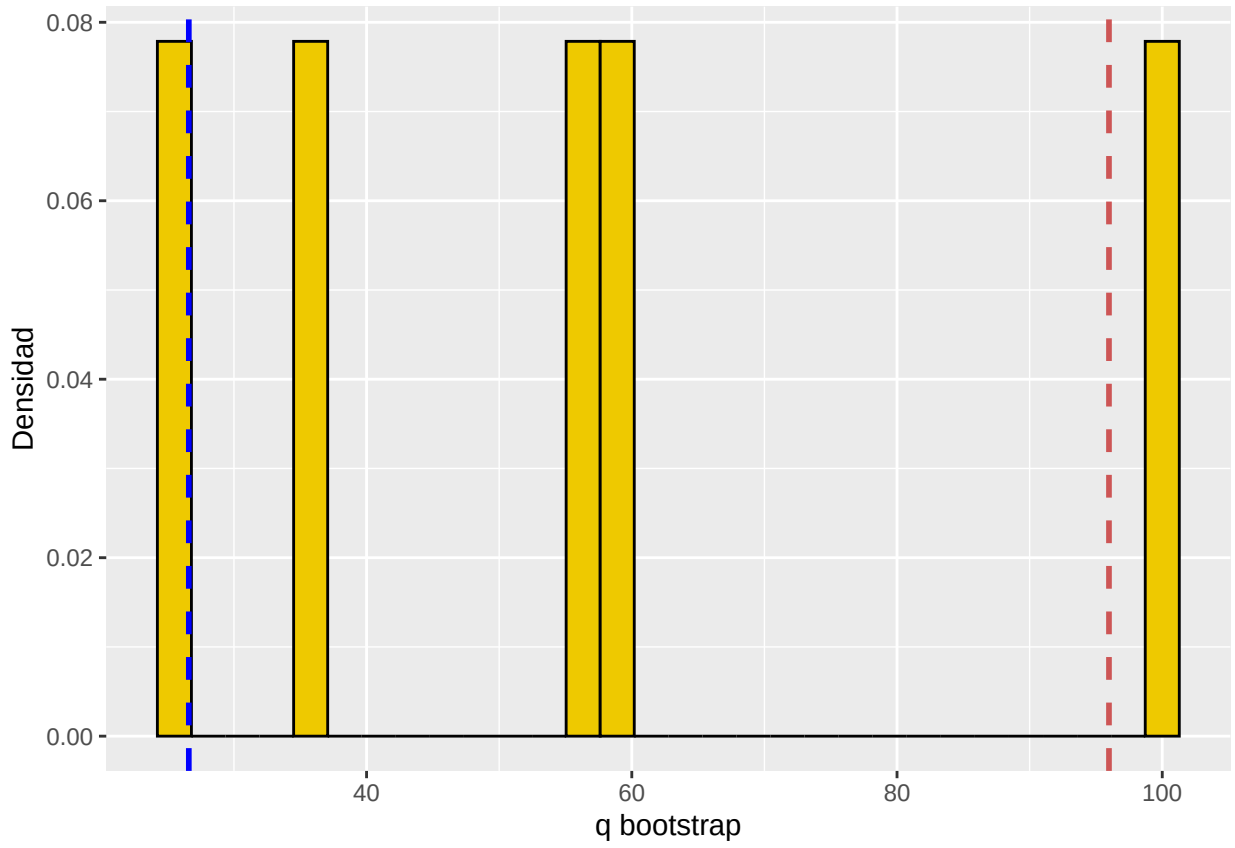
d) Halla un intervalo de confianza del 95% para q.

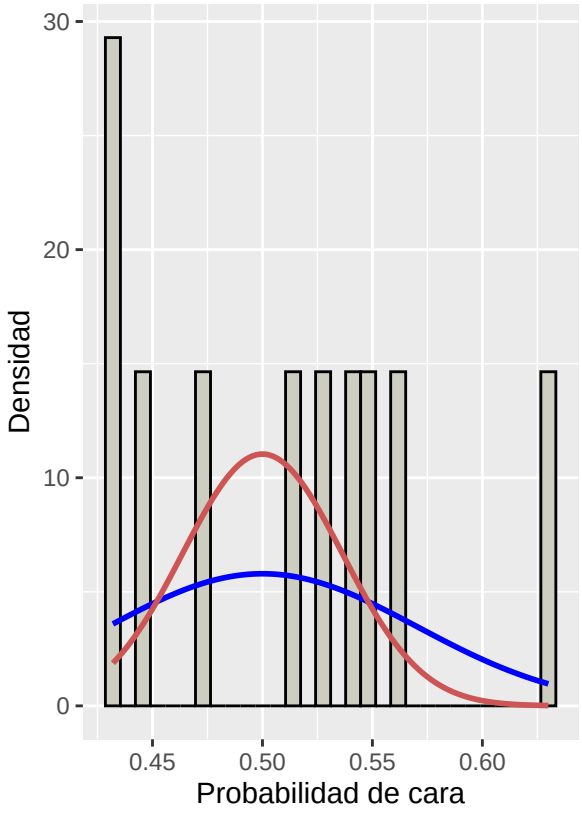
Debido a la alta simetría de la función de densidad, podemos usar el TCL aun teniendo un tamaño muestral bajo. En este caso, tenemos que despejar q de la varianza en lugar de la muestra o trabajar con una $N(0,1)^2 \equiv \chi^2_1$.

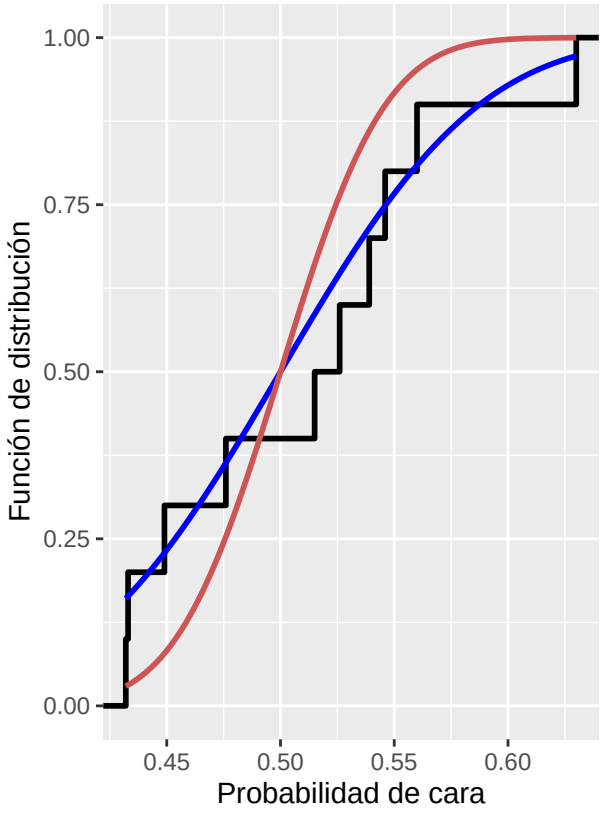
Alternativamente, podemos utilizar un prodecidimento de remuestreo como bootstrap. Al saber la distribución, optaremos por un enfoque paramétrico.

```
log_verosimilitud = function(q, muestra) {
  n = length(muestra)
  s = sum(log(muestra) + log(1 - muestra))
  (q-1)*s  -n*log(beta(q,q))
}
B = 5
n = length(prob_caras)
emv_bootstrap = replicate(B, {
  muestra = rbeta(n=n, shape1=q_emv, shape2=q_emv)
  optimise(f=function(q) log_verosimilitud(q, muestra), interval = c(0, 100), maximum=TRUE)$maximum
})

# 1 - \alpha = 0.95
alpha = 0.05
(intervalo_bootstrap = quantile(emv_bootstrap, c(alpha/2, 1-alpha/2)))
```

```
##     2.5%    97.5%
## 26.59624 95.98161
```


0.08
0.06
0.04
0.02
0.00
Densidad
40
60
80
100
q bootstrap


30
20
10
0
Densidad
0.45
0.50
0.55
0.60
Probabilidad de cara


1.00
0.75
0.50
0.25
0.00
Función de distribución
0.45
0.50
0.55
0.60
Probabilidad de cara

Problema 2 (3 puntos). Se quiere estudiar el tiempo de vida (en años) de los dispositivos eléctricos de la marca Chispazo. Para ello se estudia una muestra de 100 de estos dispositivos electricos, recogiendo los tiempos de vida resultantes en el conjunto de datos `Chispazo.RData`.

a) Basandote en los datos recogidos, estudia desde un punto de vista descriptivo si la distribución del tiempo de vida sigue aproximadamente una distribucion de Weibull $W(\alpha, \lambda)$ con parametro de forma $\alpha = 2$. Para ello, estima puntualmente el parametro de escala λ de la distribucion haciendo uso de su estimador maximo-verosímil.

La densidad de una variable aleatoria Weibull $X \sim W(\alpha, \lambda)$ viene dada por:

$$f_X(x) = \frac{\alpha}{\lambda} \cdot \left(\frac{x}{\lambda}\right)^{\alpha-1} \cdot \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right) \cdot I_{(0,+\infty)}(x) \overset{\alpha=2}{=} \frac{2x}{\lambda^2} \cdot \exp\left(\frac{-x^2}{\lambda^2}\right) \cdot I_{(0,+\infty)}(x)$$

De este modo, sea $\vec{x}_n = (x_1, \ldots, x_n)$ una m.a.s. extraída de X, su log-verosimilitud viene dada por:

$$\log L(\lambda; \vec{x}_n) = \sum_{i=1}^{n} \left[\log(2x_i) - 2\log(\lambda) - \frac{x_i^2}{\lambda^2}\right] = -2n\log(\lambda) + \sum_{i=1}^{n} \left[\log(2x_i) - \frac{x_i^2}{\lambda^2}\right]$$

Derivamos respecto del parámetro desconocido para hallar el EMV:

$$\frac{\partial}{\partial \lambda} \log L(\lambda; \vec{x}_n) = -\frac{2n}{\lambda} + \frac{2\sum_{i=1}^{n} x_i^2}{\lambda^3} = 0 \iff \frac{n}{\lambda} = \frac{\sum_{i=1}^{n} x_i^2}{\lambda^3} \iff \lambda\left(n\lambda^2 - \sum_{i=1}^{n} x_i^2\right) = 0$$

Como el parámetro λ ha de estar en $(0, +\infty)$, esto equivale a su vez a exigir:

$$n\lambda^2 = \sum_{i=1}^{n} x_i^2 \Rightarrow \lambda_{MV} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

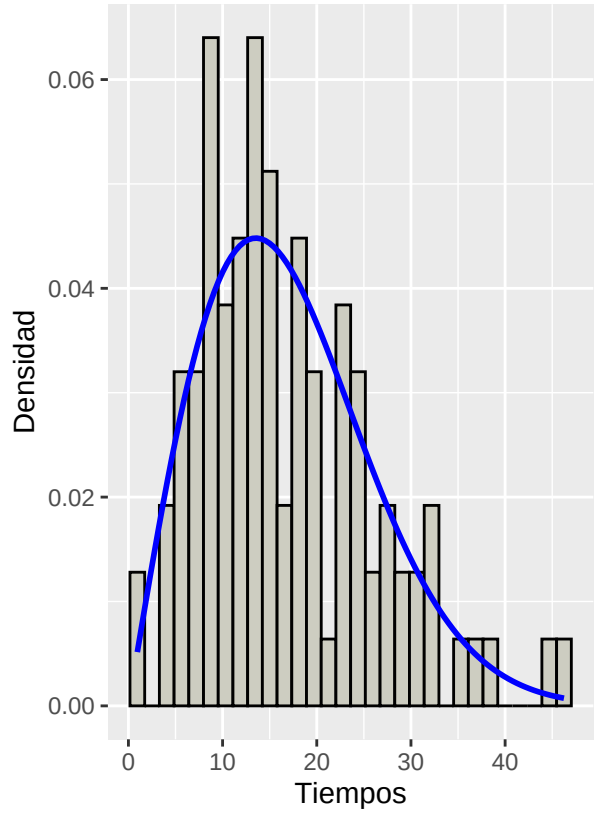
Para ver que es un máximo, necesitamos analizar el signo de la segunda derivada:

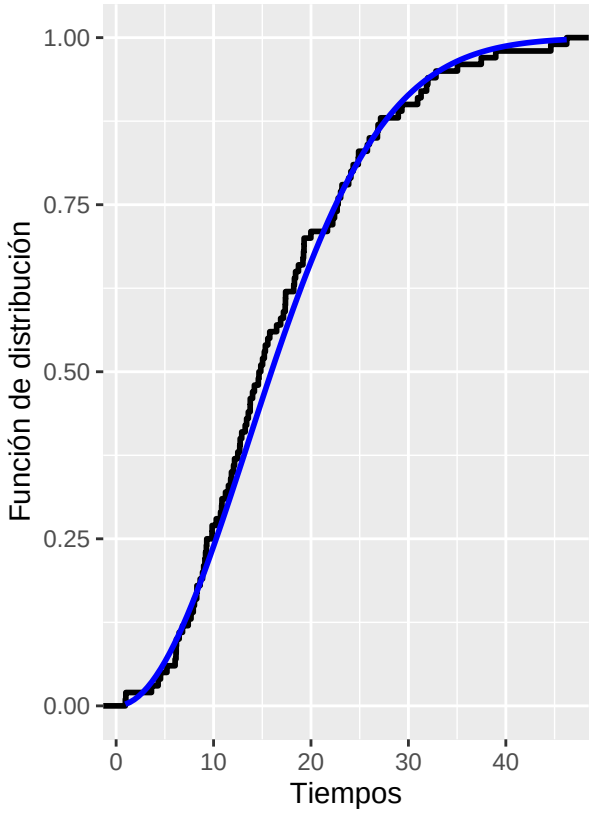
$$\frac{\partial^2}{\partial \lambda^2} \log L(\lambda_{MV}; \vec{x}_n) = \frac{2n}{\lambda_{MV}^2} - \frac{6\sum_{i=1}^{n} x_i^2}{\lambda_{MV}^4} = \frac{2n^2}{\sum_{i=1}^{n} x_i^2} - \frac{6n^2}{\sum_{i=1}^{n} x_i^2} < 0$$

Es en efecto un máximo. Veamos qué valor toma para la muestra:

```
load("Chispazo.RData")
est_lambda_mv = function(muestra) {
  sqrt(mean(muestra^2))
}
(lambda_mv = est_lambda_mv(Chispazo))
```

```
## [1] 19.14119
```





Gráficamente, se ve una clara semejanza. Veamos si se cumple también para un análisis cuantitativo.

Simulemos muchas muestras del mismo tamaño y recopilemos medidas de error para ellas:

```
nr = 1e3
n = length(Chispazo)
results = data.frame()
for (rep in 1:nr) {
  muestra = rweibull(n=n, shape=2, scale=lambda_mv)
  error = sort(Chispazo) - sort(muestra)
  results[rep, "sesgo"] = mean(error)
  results[rep, "error_absoluto"] = mean(sapply(error, abs))
  results[rep, "ECM"] = mean(sapply(error, function(x) x^2))
}
summary(results)
```

```
##      sesgo          error_absoluto       ECM
##  Min.   :-3.3906   Min.   :0.573   Min.   : 0.7107
##  1st Qu.:-0.8261   1st Qu.:1.029   1st Qu.: 2.0142
##  Median :-0.2240   Median :1.262   Median : 2.9456
##  Mean   :-0.2253   Mean   :1.317   Mean   : 3.3249
##  3rd Qu.: 0.3942   3rd Qu.:1.543   3rd Qu.: 4.2234
##  Max.   : 2.1065   Max.   :3.391   Max.   :13.4809
```

Se observan errores muy bajos.

b) Suponiendo que el tiempo de vida sigue una distribucion de Weibull con parámetro de forma $\alpha = 2$, calcula un intervalo de confianza al nivel de confianza 0,99 para el parametro de escala λ basándote en la distribución asintótica de su estimador máximo-verosímil.

Se cumplen las condiciones de regularidad necesarias para afirmar que el EMV es asintóticamente normal con varianza igual a la cota FCR. Veamos cómo calcular la información de X sobre el parámetro. Comparando con la segunda derivada respecto del parámetro calculada anteriormete (pues hay regularidad suficiente),

$$I_\lambda(X) = -E\left[\frac{\partial^2}{\partial\lambda^2}\log f(X;\lambda)\right] = -E\left[\frac{2}{\lambda^2} - \frac{6X^2}{\lambda^4}\right]$$

Calculemos la esperanza de X^2 según la definición:

$$E[X^2] = \int_0^\infty \frac{2x^3}{\lambda^2}\exp\left(\frac{-x^2}{\lambda^2}\right)dx \overset{t=\frac{x}{\lambda}}{=} 2\lambda^2\int_0^\infty t^3\exp(-t^2)\,dt \overset{u=t^2}{=} \lambda^2\int_0^\infty u^{2-1}\exp(-u)\,du = \lambda^2\cdot\Gamma(2)$$

Aplicando las propiedades de la función gamma, $\Gamma(2) = (2-1)! = 1 \Rightarrow E[X^2] = \lambda^2$. Sustituyendo para hallar la información de Fisher del parámetro:

$$I_\lambda(X) = E\left[\frac{-2}{\lambda^2} + \frac{6X^2}{\lambda^4}\right] = \frac{-2}{\lambda^2} + \frac{6\lambda^2}{\lambda^4} = \frac{4}{\lambda^2} \Rightarrow I_\lambda(X)^{\frac{-1}{2}} = \frac{\lambda}{2}$$

Construimos por tanto la siguiente función pivote:

$$\sqrt{n}\frac{\frac{1}{n}\sqrt{\sum_{i=1}^n x_i^2} - \lambda}{\frac{\lambda}{2}} = \frac{2\sqrt{\sum_{i=1}^n x_i^2} - 2\lambda\sqrt{n}}{\lambda} \approx N(0,1)$$

Hallamos los cuantiles y deshacemos el pivote:

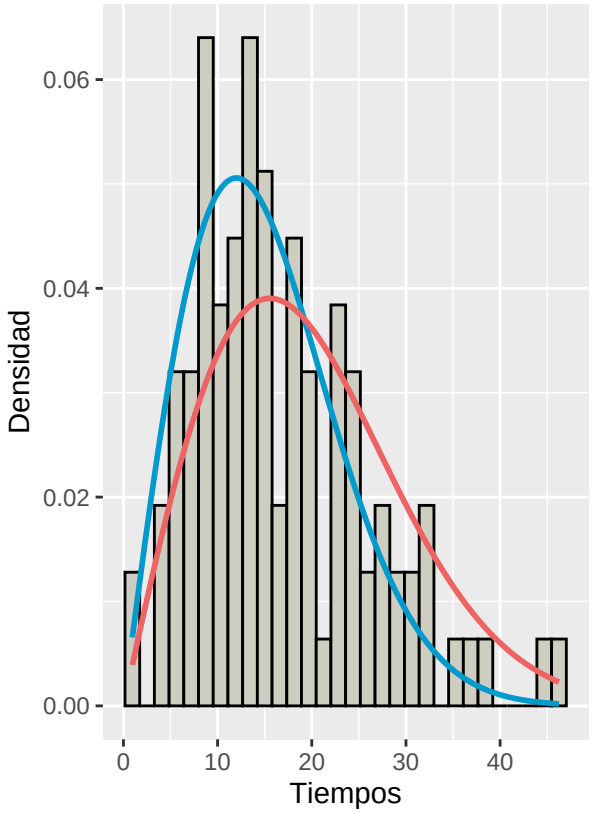
```
alpha = 0.01
c1 = qnorm(p=alpha/2, mean=0, sd=1)
c2 = qnorm(p=1-alpha/2, mean=0, sd=1)
des_pivote = function(q) {
  n = length(Chispazo)
  factor = 2 * sqrt(sum(sapply(Chispazo, function(x) x^2)))
  factor * 1/(q + 2*sqrt(n))
}

(intervalo = c(i1=des_pivote(c2), i2=des_pivote(c1)))
```

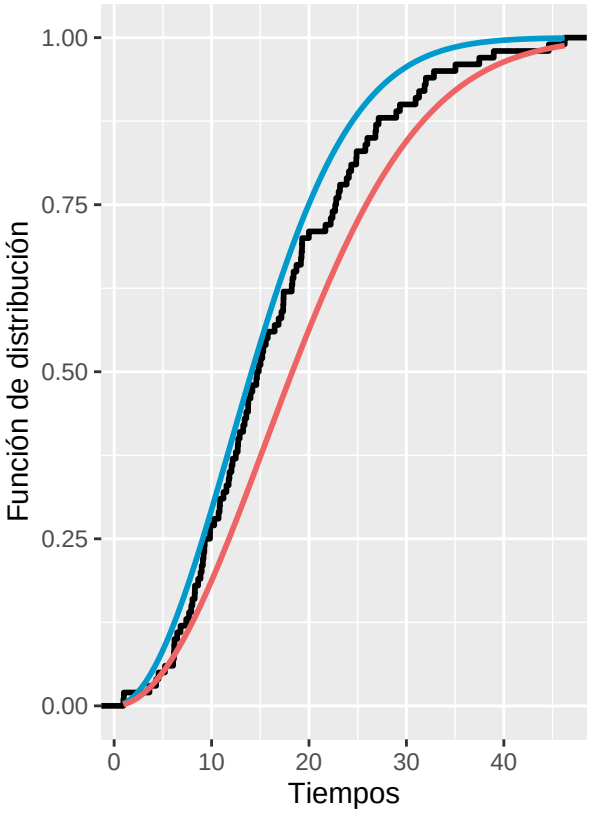
```
##       i1       i2
## 16.95724 21.97084
```

```
lambda_mv > intervalo[1] && lambda_mv < intervalo[2]
```

```
## [1] TRUE
```

Veamos cómo son las distribuciones Weibull de parámetro $\alpha = 2$ y λ valores extremos del intervalo:





c) Se quiere realizar un nuevo estudio mas preciso que consiga estimar λ por medio de un intervalo de confianza asintotico para λ al nivel de confianza 0,99 que tenga una longitud esperada inferior a 0,1 unidades. Suponiendo que el tiempo de vida sigue una distribucion de Weibull con parámetro de forma $\alpha = 2$ y haciendo uso de la estimacion de λ en el estudio preliminar, determina de manera aproximada el tamano de muestra que se necesita considerar para lograr el objetivo.

Veamos qué expresión tiene el intervalo de confianza calculado en el apartado anterior. Como hemos usado la propiedad de normalidad asintótica, tenemos que $-c_1 = c_2 = c$ según la notación anterior, pues hemos tomado cuantiles simétricos. El intervalo calculado es el siguiente:

$$\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \left(\frac{2}{c + 2\sqrt{n}}\right) < \lambda < \sqrt{\sum_{i=1}^{n} x_i^2} \cdot \left(\frac{2}{-c + 2\sqrt{n}}\right)$$

Por tanto, la longitud del intervalo viene dada por:

$$\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \left(\frac{2}{-c + 2\sqrt{n}} - \frac{2}{c + 2\sqrt{n}}\right) = \sqrt{\sum_{i=1}^{n} x_i^2} \cdot \left(\frac{2c + 4\sqrt{n} + 2c - 4\sqrt{n}}{(2\sqrt{n} + c)(2\sqrt{n} - c)}\right) = \sqrt{\sum_{i=1}^{n} x_i^2} \cdot \frac{4c}{4n2 - c^2}$$

Para tomar esperanzas, reescribimos para que la expresión dependa de λ_{MV}:

$$\sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2} = \sqrt{n} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} = \sqrt{n} \cdot \lambda_{MV}$$

Basándonos en el estudio anterior, nos pondremos en el peor caso. Es decir, λ será el extremo derecho del intervalo de confianza calculado anteriormente, pues cuanto mayor sea el valor del EMV mayor será la longitud del intervalo. Despejamos el tamaño muestral a partir de la longitud del intervalo.

$$\frac{4c}{4n-c^2}\lambda_p\sqrt{n} = L \Rightarrow \left(\frac{4c\cdot\lambda_p}{L}\right)^2 n = 16n^2 - 8c^2n + c^4 \Rightarrow 16n^2 - n\left(8c^2 + \left(\frac{4c\cdot\lambda_p}{L}\right)^2\right) + c^4 = 0$$

```
c = qnorm(1-alpha/2)
lambda_p = max(intervalo)
longitud = 0.1
b = 8*c^2 +(4*c*lambda_p/longitud)^2
(tam_muestra = (b+sqrt(b^2-64*c^4))/32)
```

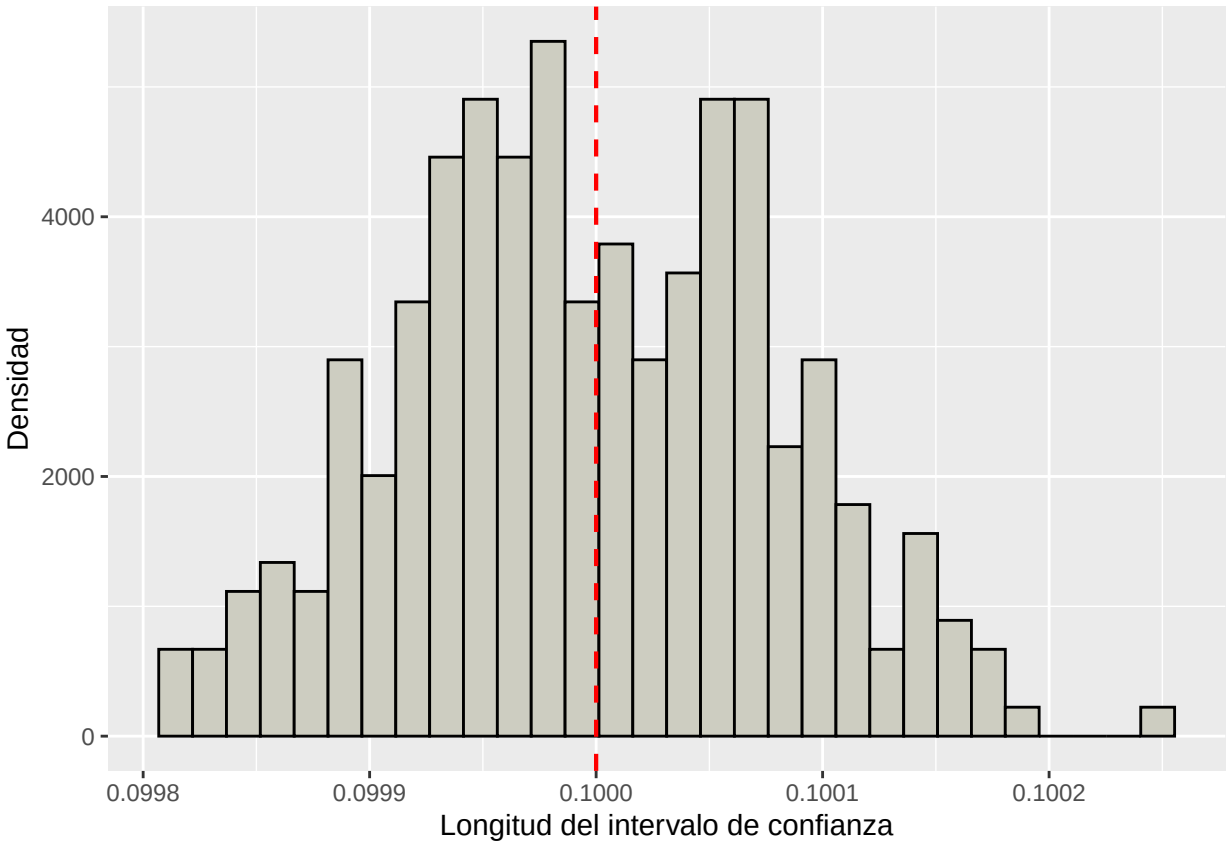
```
## [1] 320281.7
```

Para comprobar este resultado vamos a simular muestras de este tamaño de $X \sim W(\alpha = 2, \lambda)$ con λ desconocido diferente en cada iteración. Para cada una de ellas, calcularemos el intervalo de confianza del parámetro según el EMV, registrando su longitud y la cobertura del parámetro.

```
nr = 300
n = ceiling(tam_muestra)
alpha = 0.01
c1 = qnorm(alpha/2)
c2 = qnorm(1-alpha/2)
calc_intervalo = function(q, muestra) {
  n = length(muestra)
  factor = 2 * sqrt(sum(sapply(muestra, function(x) x^2)))
  factor * 1/(q + 2*sqrt(n))
}

results = data.frame()
for(rep in 1:nr) {
  muestra = rweibull(n = n, shape = 2, scale=lambda_p)
  i1 = calc_intervalo(c2, muestra)
  i2 = calc_intervalo(c1, muestra)
  results[rep, "i1"] = i1
  results[rep, "i2"] = i2
  results[rep, "longitud"] = i2-i1
  results[rep, "cobertura"] = i1 < lambda_p && lambda_p < i2
}
summary(results)
```

```
##        i1              i2          longitud       cobertura
##  Min.   :21.88   Min.   :21.98   Min.   :0.09981   Mode :logical
##  1st Qu.:21.91   1st Qu.:22.01   1st Qu.:0.09994   FALSE:1
##  Median :21.92   Median :22.02   Median :0.09999   TRUE :299
##  Mean   :21.92   Mean   :22.02   Mean   :0.10000
##  3rd Qu.:21.93   3rd Qu.:22.03   3rd Qu.:0.10006
##  Max.   :21.98   Max.   :22.08   Max.   :0.10025
```

Debido al elevado tamaño muestral, no podemos generar muchas muestras. Los resultados de la simulación parecen corresponderse con las especificaciones del enunciado.


Densidad
4000
2000
0
0.0998
0.0999
0.1000
0.1001
0.1002
Longitud del intervalo de confianza