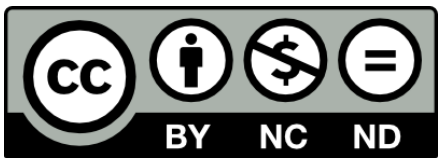
Inferencia Estadística

Los Del DGIIM, losdeldgiim.github.io

Doble Grado en Ingeniería Informática y Matemáticas
Universidad de Granada



Inferencia Estadística

Los Del DGIIM, losdeldgiim.github.io

José Juan Urrutia Milán

Granada, 2025

Índice general

1. Ejercicios de clase

Esta sección tiene el propósito de recoger todos los ejercicios propuestos en clase por parte de la profesora y que fueron resueltos por los alumnos en pizarra.

1.1. Estadísticos muestrales

Ejercicio 1.1.1. Obtener la función masa de probabilidad conjunta de una m.a.s. de $X \rightsquigarrow B(k_0, p)$ y la función de densidad de una m.a.s. de $X \rightsquigarrow U(a, b)$.

Recordamos que si $X \rightsquigarrow B(k_0, p)$, entonces:

$$P[X = x] = \binom{k_0}{x} p^x (1-p)^{n-x} \qquad \forall x \in \{0, \dots, k_0\}$$

Por lo que si tenemos una m.a.s. de n variables independientes e idénticamente distribuidas a X, $(X_1, \dots, X_n)$, su función de densidad vendrá dada por:

$$\begin{aligned} P[X_1 = x_1, \dots, X_n = x_n] &\overset{\text{indep.}}{=} \prod_{i=1}^{n} P[X_i = x_i] \overset{\text{id. d.}}{=} \prod_{i=1}^{n} P[X = x_i] \\ &= \prod_{i=1}^{n} \binom{k_0}{x_i} p^{x_i} (1-p)^{k_0 - x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{nk_0 - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{k_0}{x_i} \\ &\quad \forall x_i \in \{0, \dots, k_0\} \end{aligned}$$

Si ahora $X \rightsquigarrow U(a, b)$ para ciertos $a, b \in \mathbb{R}$ con $a < b$, entonces:

$$f_X(x) = \frac{1}{b-a} \qquad \forall x \in [a, b]$$

de donde:

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \overset{\text{id. d.}}{=} \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n} \qquad \forall x \in [a, b]$$

Ejercicio 1.1.2. Para cada realización muestral, $(x_1, \dots, x_n) \in \mathcal{X}^n$, $F^*_{x_1,\dots,x_n}$ es una función de distribución en $\mathbb{R}$. En particular es una función a saltos, con saltos de amplitud $1/n$ en los sucesivos valores muestrales ordenados de menor a mayor, supuestos que sean distintos, y de saltos múltiplos en el caso de que varios valores muestrales coincidieran.

En las condiciones del enunciado, es decir, suponiendo que $x_1, \ldots, x_n$ están ordenados de menor a mayor y son distintos, entonces es fácil ver que:

$$F^*_{x_1,\ldots,x_n}(x) = \begin{cases} 0 & \text{si } x < x_1 \\ 1/n & \text{si } x_1 \leqslant x < x_2 \\ & \vdots \\ 1 & \text{si } x > x_n \end{cases} \qquad \forall x \in \mathbb{R}$$

Por lo que es claro que $F^*_{x_1,\ldots,x_n}$ es no decreciente, continua por la derecha, con límite 0 en $-\infty$ y con límite 1 en $+\infty$.

Ejercicio 1.1.3. $\forall x \in \mathbb{R}$, $F^*_{X_1,\ldots,X_n}(x)$ es una variable aleatoria tal que $nF^*_{X_1,\ldots,X_n}(x) \rightsquigarrow B(n, F(x))$ y:

$$E[F^*_{X_1,\ldots,X_n}(x)] = F(x), \qquad Var[F^*_{X_1,\ldots,X_n}(x)] = \frac{F(x)(1-F(x))}{n}$$

donde $F(x)$ es la función de distribución de X.

Recordamos que:

$$F^*_{X_1,\ldots,X_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{]-\infty,x]}(X_i) \qquad \forall x \in \mathbb{R}$$

Fijado $x \in \mathbb{R}$, tenemos que $I_{]-\infty,x]}(X) \rightsquigarrow B(1, P[X \leqslant x]) \equiv B(1, F(x))$, por lo que por la propiedad reproductiva de la binomial tenemos que:

$$nF^*_{X_1,\ldots,X_n}(x) \rightsquigarrow B(n, F(x))$$

Por lo que:

$$nE[F^*_{X_1,\ldots,X_n}(x)] = E[nF^*_{X_1,\ldots,X_n}(x)] = nF(x)$$

de donde:

$$E[F^*_{X_1,\ldots,X_n}(x)] = F(x)$$

Para la varianza:

$$n^2 Var[F^*_{X_1,\ldots,X_n}(x)] = Var[nF^*_{X_1,\ldots,X_n}(x)] = nF(x)(1-F(x))$$

de donde:

$$Var[F^*_{X_1,\ldots,X_n}(x)] = \frac{F(x)(1-F(x))}{n}$$

Ejercicio 1.1.4. Para valores grandes de n, en virtual del Teorema Central del Límite:

$$F^*_{X_1,\ldots,X_n}(x) \rightsquigarrow \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Sea $(X_1, \ldots, X_n)$ una m.a.s. de n muestras, sea:

$$S_n = \sum_{i=1}^{n} I_{]-\infty,x]}(X_i) \qquad \forall n \in \mathbb{N}$$

Por el Teorema Central del Límite tenemos que:

$$\frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0,1) \Longrightarrow S_n \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Como $S_n \rightsquigarrow B(n, F(x))$, entonces tenemos que:

$$E[S_n] = nF(x)$$
$$Var[S_n] = nF(x)(1-F(x))$$

Por lo que:

$$F^*_{X_1,\ldots,X_n}(x) = \frac{1}{n} S_n \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Ejercicio 1.1.5. Dada una muestra aleatoria simple formada por las observaciones $(3, 8, 5, 4, 5)$, obtener su función de distribución muestral y realizar la representación gráfica.

Aplicando la definición de la función de distribución muestral obtenemos que:

$$F^*_{(3,8,5,4,5)}(x) = \begin{cases} 0 & \text{si } x < 3 \\ 1/5 & \text{si } 3 \leqslant x < 4 \\ 2/5 & \text{si } 4 \leqslant x < 5 \\ 4/5 & \text{si } 5 \leqslant x < 8 \\ 1 & \text{si } x \geqslant 8 \end{cases}$$

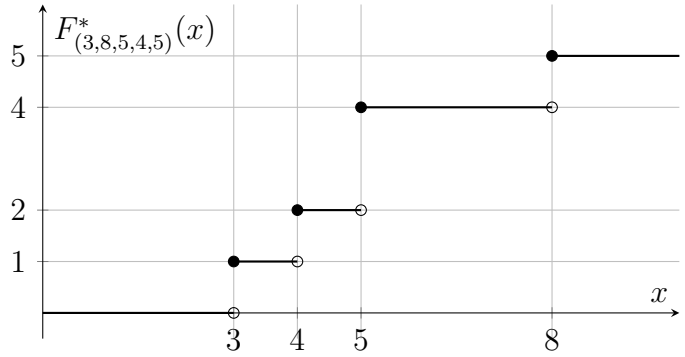


Figura 1.1: Representación gráfica de $F^*_{(3,8,5,4,5)}(x)$.

Ejercicio 1.1.6. Sea X una variable aleatoria con distribución $B(1, p)$ con $p \in (0, 1)$. Se toma una muestra de tamaño 5, $(X_1, X_2, X_3, X_4, X_5)$, y se obtiene la siguiente observación $(0, 1, 1, 0, 0)$. Determinar el valor de los estadísticos estudiados en la observación.

Aplicando las fórmulas vistas en clase obtenemos:

- Media: 0,4.
- Varianza: 0,24.

- Cuasivarianza: 0,3.
- $x_{(1)} = 0$, $x_{(2)} = 0$, $x_{(3)} = 0$, $x_{(4)} = 1$, $x_{(5)} = 1$.

Ejercicio 1.1.7. Sea $(X_1, \ldots, X_n)$ una m.a.s. y $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, entonces:

$$M_{\overline{X}}(t) = (M_X(t/n))^n$$

$$M_{\overline{X}}(t) = E\left[e^{t\overline{X}}\right] = E\left[e^{\frac{t}{n}\sum_{i=1}^{n} X_i}\right] = M_{\sum_{i=1}^{n} X_i}\left(\frac{t}{n}\right) \overset{\text{indep.}}{=} \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) \overset{\text{id. d.}}{=} \left(M_X\left(\frac{t}{n}\right)\right)^n$$

Ejercicio 1.1.8. Obtener la distribución muestral de $\overline{X}$ para $(X_1, \ldots, X_n)$ una m.a.s. de $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$.

$$M_{\overline{X}}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n = \left(e^{\mu t + \frac{\sigma^2 t^2}{2n^2}}\right)^n = e^{\mu t + \frac{\sigma^2 t^2}{2n}}$$

Luego $\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, ya que la función generatriz de momentos caracteriza la distribución.

Proposición 1.1. *Si tenemos una m.a.s. $(X_1, \ldots, X_n)$, entonces:*

$$F_{X_{(n)}}(x) = (F_X(x))^n \qquad \forall x \in \mathbb{R}$$
$$F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n$$

Demostración. Para la distribución del máximo:

$$F_{X_{(n)}}(x) = P[X_{(n)} \leqslant x] = P[X_1 \leqslant x, \ldots, X_n \leqslant x] \overset{\text{indep.}}{=} \prod_{i=1}^{n} P[X_i \leqslant x]$$
$$\overset{\text{id. d.}}{=} \prod_{i=1}^{n} P[X \leqslant x] = (F_X(x))^n$$

Para la del mínimo:

$$F_{X_{(1)}}(x) = P[X_{(1)} \leqslant x] = 1 - P[X_{(1)} > x] = 1 - P[X_1 > x, \ldots, X_n > x]$$
$$\overset{\text{indep.}}{=} 1 - \prod_{i=1}^{n} P[X_i > x] \overset{\text{id. d.}}{=} 1 - (P[X > x])^n = 1 - (1 - F_X(x))^n$$

□

Ejercicio 1.1.9. Obtener las distribuciones muestrales de $X_{(1)}$ y $X_{(n)}$ para $X \rightsquigarrow U(a, b)$.

Si $X \rightsquigarrow U(a, b)$, entonces:

$$F_X(x) = \frac{x - a}{b - a} \qquad \forall x \in [a, b]$$

Por lo que aplicando la Proposición superior:

$$F_{X_{(n)}}(x) = (F_X(x))^n = \left(\frac{x-a}{b-a}\right)^n \qquad \forall x \in [a,b]$$

$$F_{X_{(1)}}(x) = 1-(1-F_X(x))^n = 1-(1-F_X(x))^n = 1-\left(1-\frac{x-a}{b-a}\right)^n$$

$$= 1-\left(\frac{b-x}{b-a}\right)^n \qquad \forall x \in [a,b]$$

1.2. Distribuciones en el muestreo de poblaciones normales

Proposición 1.2. *Sea $X \rightsquigarrow \mathcal{N}(0,1)$, entonces $X^2 \rightsquigarrow \chi^2(1)$.*

Demostración. Sea $Y = X^2 = h(X)$, entonces $X = \pm\sqrt{Y} = h^{-1}(y)$, por lo que:

$$f_Y(y) = f_X(h_1^{-1}(y))\left|\frac{dh_1^{-1}(y)}{dy}\right| + f_X(h_2^{-1}(y)) + \left|\frac{dh_2^{-1}(y)}{dy}\right|$$

Como $X \rightsquigarrow \mathcal{N}(0,1)$, entonces:

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}} \qquad \forall x \in \mathbb{R}$$

De donde:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{\frac{-(\sqrt{y})^2}{2}}\left|\frac{1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}}e^{\frac{-(-\sqrt{y})^2}{2}}\left|\frac{-1}{2\sqrt{y}}\right| = \frac{1}{\sqrt{2\pi y}}e^{\frac{-y}{2}} \qquad \forall y > 0$$

Por lo que $Y \rightsquigarrow \chi^2(1)$. □

Ejercicio 1.2.1. Calcula el valor de k o la probabilidad inducida:

a) $P[\chi^2(10) \geqslant k] = 0{,}005$.
$k = 25{,}1881$.

b) $P[\chi^2(45) \leqslant k] = 0{,}005$.

$$P[\chi^2(45) \geqslant k] = 0{,}995 \Longrightarrow k = 24{,}3110$$

c) $P[\chi^2(14) \geqslant 21{,}06]$
0,1

d) $P[\chi^2(20) \leqslant 12{,}44]$

$$P[\chi^2(20) \leqslant 12{,}44] = 1 - P[\chi^2(20) \geqslant 12{,}44] = 1 - 0{,}9 = 0{,}1$$

Ejercicio 1.2.2. Calcula el valor de k o la probabilidad inducida:

a) $P[t(26) \geqslant k] = 0{,}05$
$k = 1{,}7056$

b) $P[t(20) \leqslant k] = 0{,}25$
$k = -0{,}6870$

c) $P[t(26) \geqslant k] = 0{,}9$
$k = -1{,}3150$

d) $P[t(21) \geqslant 1{,}721]$
$0{,}05$

e) $P[t(11) \leqslant 0{,}697]$
$0{,}75$

f) $P[t(8) \leqslant -2{,}306]$
$0{,}025$

Ejercicio 1.2.3. Calcula el valor de k o la probabilidad inducida:

a) $P[F(7,3) \leqslant k] = 0{,}95$
$k = 8{,}89$

b) $P[F(8,4) \geqslant k] = 0{,}01$

$$0{,}01 = 1 - P[F(8,4) \leqslant k] \Longrightarrow P[F(8,4) \leqslant k] = 0{,}99 \Longrightarrow k = 14{,}8$$

c) $P[F(2,2) \leqslant 19]$
$0{,}95$

d) $P[F(3,5) \geqslant 12{,}1]$

$$P[F(3,5) \geqslant 12{,}1] = 1 - P[F(3,5) \leqslant 12{,}1] = 1 - 0{,}99 = 0{,}01$$

e) $P[F(60,40) \leqslant k] = 0{,}05$
$k = 0{,}627$

1.2.1. Varias demostraciones

Tenemos una $(X_1, \ldots, X_n)$ m.a.s. con $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, si tomamos:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Demostraciones importantes que pueden caer.

Proposición 1.3. *En dichas condiciones, veamos que:*

$$\overline{X}, (X_1 - \overline{X}, \ldots, X_n - \overline{X}) \text{ son independientes}$$

Demostración. Para ellos, usaremos la caracterización por la función generatriz de momentos conjunta:

$$M_{\overline{X},X_1-\overline{X},\ldots,X_n-\overline{X}}(t,t_1,\ldots,t_n) \stackrel{?}{=} M_{\overline{X}}(t)M_{X_1-\overline{X},\ldots,X_n-\overline{X}}(t_1,\ldots,t_n)$$

$$\begin{aligned}
M_{\overline{X},X_1-\overline{X},\ldots,X_n-\overline{X}}(t,t_1,\ldots,t_n) &= E[e^{(t,t_1,\ldots,t_n)\cdot(\overline{X},X_1-\overline{X},\ldots,X_n-\overline{X})}] \\
&= E\left[e^{t\overline{X}+\sum_{i=1}^{n}(X_i-\overline{X})t_i}\right] \\
&= E\left[e^{\frac{t}{n}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}X_it_i-\sum_{i=1}^{n}\overline{X}t_i}\right] \\
&= E\left[e^{\frac{t}{n}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}X_it_i-\overline{X}\sum_{i=1}^{n}t_i}\right] \\
&= E\left[e^{\frac{t}{n}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}X_it_i-\frac{1}{n}\sum_{i=1}^{n}X_i\sum_{i=1}^{n}t_i}\right] \\
&= E\left[e^{\frac{t}{n}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}X_it_i-\sum_{i=1}^{n}X_i\frac{1}{n}\sum_{i=1}^{n}t_i}\right] \\
&= E\left[e^{\frac{t}{n}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}X_it_i-\sum_{i=1}^{n}X_i\bar{t}}\right] \\
&= E\left[e^{\sum_{i=1}^{n}X_i\frac{t}{n}+\sum_{i=1}^{n}X_it_i-\sum_{i=1}^{n}X_i\bar{t}}\right] \\
&= E\left[e^{\sum_{i=1}^{n}X_i\left(\frac{t}{n}+t_i-\bar{t}\right)}\right] \\
&= E\left[\prod_{i=1}^{n}e^{X_i\left(\frac{t}{n}+t_i-\bar{t}\right)}\right] \\
&\stackrel{\text{indep.}}{=} \prod_{i=1}^{n}E\left[e^{X_i\left(\frac{t}{n}+t_i-\bar{t}\right)}\right] = \prod_{i=1}^{n}M_{X_i}\left(\frac{t}{n}+t_i-\bar{t}\right) \\
&\stackrel{(*)}{=} \prod_{i=1}^{n}e^{\left(\frac{t}{n}+t_i-\bar{t}\right)\mu+\left(\frac{t}{n}+t_i-t\right)^2\frac{\sigma^2}{2}} \\
&= e^{\sum_{i=1}^{n}\left[\left(\frac{t}{n}+t_i-\bar{t}\right)\mu+\frac{\sigma^2}{2}\left(\frac{t^2}{n^2}+\left(t_i-\bar{t}\right)^2+2\frac{t}{n}(t_i-\bar{t})\right)\right]} \\
&= e^{\sum_{i=1}^{n}\frac{\mu t}{n}+\sum_{i=1}^{n}\mu(t_i-\bar{t})+\frac{\sigma^2}{2}\left(\sum_{i=1}^{n}\frac{t^2}{n^2}+\sum_{i=1}^{n}(t_i-\bar{t})^2+2\sum_{i=1}^{n}\frac{t}{n}(t_i-\bar{t})\right)} \\
&= e^{\cancel{\mu t}\cancel{\mu\sum t_i-\bar{t}}+\frac{\sigma^2}{n^2}\left(\frac{nt^2}{n^2}+\sum_{i=1}^{n}(t_i-\bar{t})^2+2\frac{t}{n}\sum_{i=1}^{n}(t_i-\bar{t})\right)} \\
&= e^{\mu t+\frac{\sigma^2t^2}{2n}+\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i-\bar{t})^2}
\end{aligned}$$

Sabemos que:

$$\begin{aligned} M_{\overline{X}}(t) &= M_{\left(\overline{X},X_1-\overline{X},\ldots,X_n-\overline{X}\right)}(t,0,...,0) = e^{\mu t+\frac{\sigma^2 t^2}{2n}} \\ M_{(X_1-\overline{X},\ldots,X_n-\overline{X})}(t_1,...,t_n) &= M_{(\overline{X},X_1-\overline{X},\ldots,X_n-\overline{X})}(0,t_1,\ldots,t_n) \\ &= e^{\frac{\sigma^2}{n}\sum\limits_{i=1}^{n}(t_i-\bar{t})^2} \end{aligned}$$

Por lo que es cierto que el producto de las funciones generatrices de momentos es la generatriz de mmoentos conjunta, luego las variables son independientes. □

Corolario 1.3.1. *Como corolario de la Proposición anterior, tenemos que:*

- *Se vio ya, y se saca de la demostración de arriba.*
- *Lema de Fisher:* $\overline{X}$ *y* S^2 *son independientes.*

  Como S^2 *es función del vector de la Proposición anterior, tenemos que es independiente con* $\overline{X}$*, ya que las funciones de variables independientes son independientes.*
- $\dfrac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$

  Para demostrarlo:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \rightsquigarrow \chi^2(n)$$

  Ahora, queremos ver que:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$$

  Para ello:

$$\begin{aligned} \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} &= \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}+\overline{X}-\mu\right)^2}{\sigma^2} \\ &= \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum\limits_{i=1}^{n}\left(\overline{X}-\mu\right)^2+2\sum\limits_{i=1}^{n}(X_i-\overline{X})(\overline{X}-\mu)}{\sigma^2} \\ &= \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2}+n\frac{\left(\overline{X}-\mu\right)^2}{\sigma^2} \end{aligned}$$

  Y como:

$$n\frac{\left(\overline{X}-\mu\right)^2}{\sigma^2} \rightsquigarrow \chi^2(1)$$

  Buscamos ver lo que sigue lo de la derecha $(A = B + C)$*. Para ello, usaremos la función generatriz de momentos. tenemos que* $B = f(S^2)$ *y* $C = f(\overline{X})$*, luego* B *y* C *son independientes, por lo que:*

$$M_{A=B+C}(t) \overset{indep.}{=} M_B(t)M_C(t) = M_B(t)\frac{1}{(1-2t)^{\frac{1}{2}}} \qquad t < \frac{1}{2}$$

Y sabemos que:

$$M_A(t) = \frac{1}{(1-2t)^{\frac{n}{2}}}$$

De donde:

$$M_B(t) = \frac{M_A(t)}{M_C(t)} = \frac{\dfrac{1}{(1-2t)^{\frac{n}{2}}}}{\dfrac{1}{(1-2t)^{\frac{1}{2}}}} = \frac{1}{(1-2t)^{\frac{n-1}{2}}} \qquad t < \frac{1}{2}$$

Por lo que $B \rightsquigarrow \chi^2(n-1)$

- $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \rightsquigarrow t(n-1)$

  Para ello, al igual que la χ^2*, lo más sencillo es ir a la construcción de* t*:*

$$\left.\begin{array}{l} X \rightsquigarrow \mathcal{N}(0,1) \\ Y \rightsquigarrow \chi^2(n) \\ indep \end{array}\right\} \frac{X}{\sqrt{Y/n}} \rightsquigarrow t(n)$$

  Como:

$$\overline{X} \rightsquigarrow \mathcal{N}(\mu, \sigma^2) \Longrightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$$

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$$

  que son independientes por el Lema de Fisher. Si aplicamos la construcción:

$$\frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\dfrac{S}{\sigma}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \rightsquigarrow t(n-1)$$

Este último corolario ayuda a inferir parámetros de las distribuciones. Ver punto 2.3.1. Inferencia sobre x significa que queremos averiguar el valor de x. El corolario de ayer nos sirve para usar otros estadísticos en lugar de otros que deberíamos usar, pero con 1 parámetro desconocido en lugar de 2. Aprenderse fórmulas de 2.3.1.

1.2.2. Dos poblaciones normales

Teorema 1.4 (extensión del Lema de Fisher)**.** *Los vectores* $(\overline{X}, \overline{Y})$*,* (S_1^2, S_2^2) *son independientes.*

Demostración. Usemos la función generatriz de momentos, ya que son independientes si y solo si:

$$M_{\left(\overline{X},\overline{Y},S_1^2,S_2^2\right)}(t_1,t_2,s_1,s_2) \stackrel{?}{=} M_{\left(\overline{X},\overline{Y}\right)}(t_1,t_2) M_{\left(S_1^2,S_2^2\right)}(s_1,s_2)$$

Como las X y las Y son independientes:

$$M_{\left(\overline{X},\overline{Y},S_1^2,S_2^2\right)}(t_1,t_2,s_1,s_2) = M_{\left(\overline{X},S_1^2\right)}(t_1,s_2)M_{\left(\overline{Y},S_2^2\right)}(s_1,s_2) \overset{\text{Lema Fisher}}{=} M_{\overline{X}}(t_1)M_{S_1^2}(s_1)M_{\overline{Y}}(t_2)M_{S_2^2}(s_2)$$

Ahora, como X e Y son independientes:

$$M_{\overline{X}}(t_1)M_{S_1^2}(s_1)M_{\overline{Y}}(t_2)M_{S_2^2}(s_2) = M_{\left(\overline{X},\overline{Y}\right)}(t_1,t_2)M_{\left(S_1^2,S_2^2\right)}(s_1,s_2)$$

□

Corolario 1.4.1. *A partir de entonces (aunque el 3o es el único que necesita el Teorema anterior):*

1. *Tenemos (aunque no sea corolario de Fisher):*

$$\frac{n_2}{n_1}\frac{\sigma_2^2}{\sigma_1^2}\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2} \rightsquigarrow F(n_1,n_2)$$

Que es equivalente a que:

$$\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1\sigma_1^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2/n_2\sigma_2^2} \rightsquigarrow F(n_1,n_2)$$

Por construcción de $F(n_1,n_2)$:

$$\left.\begin{array}{l} X \rightsquigarrow \chi^2(m) \\ \quad independientes \\ Y \rightsquigarrow \chi^2(n) \end{array}\right\} \Longrightarrow \frac{X/m}{Y/n} \rightsquigarrow F(n_1,n_2)$$

Como ayer vimos que:

$$\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sigma_1^2} \rightsquigarrow \chi^2(n_1)$$
$$\frac{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}{\sigma_2^2} \rightsquigarrow \chi^2(n_2)$$

Como X e Y son independientes, tenemos funciones en función de X e Y, luego estas dos variables son independientes. Ahora:

$$\frac{\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{n_1\sigma_1^2}}{\frac{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}{n_2\sigma_2^2}} \rightsquigarrow F(n_1,n_2) \rightsquigarrow F(n_1,n_2)$$

2. *Tenemos:*

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F(n_1-1, n_2-1)$$

Seguimos buscando la construcciónde F, por lo que buscamos aplicar lo que sabemos:

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi^2(n_1-1)$$
$$\frac{(n_2-1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi^2(n_2-1)$$

Que son independientes por ser funciones de X e Y, que son independientes. Dividimos:

$$\frac{\frac{\cancel{(n_1-1)}S_1^2}{\sigma_1^2\cancel{(n_1-1)}}}{\frac{(n_2-1)S_2^2}{\sigma_2^2(n_2-1)}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F(n_1-1, n_2-1)$$

En particular, si $\sigma_1 = \sigma_2$, se tiene:

$$\frac{S_1^2}{S_2^2} \rightsquigarrow F(n_1-1, n_2-1)$$

3. *La construcción de la t-Student era:*

$$\left.\begin{array}{l} X \rightsquigarrow \mathcal{N}(0,1) \\ \quad\quad independientes \\ Y \rightsquigarrow \chi^2(n) \end{array}\right\} \Longrightarrow \frac{X}{\sqrt{Y/n}} \rightsquigarrow t(n)$$

Queremos probar:

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2}}}\sqrt{\frac{n_1+n_2-2}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2}{n_2}}} \rightsquigarrow t(n_1+n_2-2)$$

Ahora:

$$\left.\begin{array}{l} \overline{X} \rightsquigarrow \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \\ \overline{Y} \rightsquigarrow \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \end{array}\right\} \Longrightarrow \overline{X}-\overline{Y} \rightsquigarrow \mathcal{N}\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

Tipificamos:

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0,1)$$

Para el denominador ahora, de la S^2 sabemos:

$$\frac{(n_1-1)S_1^2}{n_1^2} \rightsquigarrow \chi^2(n_1-1)$$
$$\frac{(n_2-1)S_2^2}{n_2^2} \rightsquigarrow \chi^2(n_2-1)$$

Como X e Y son independientes, estas dos son independientes, y al sumar dos χ^2 independientes tenemos la propiedad reproductiva:

$$\frac{(n_1-1)S_1^2}{n_1^2}+\frac{(n_2-1)S_2^2}{n_2^2}\rightsquigarrow\chi^2(n_1+n_2-2)$$

Por la extensión del Lema de Fisher tenemos que las dos variables aleatorias que hemos calculado son independientes. Procedemos ahora a aplicar la construcción de la t:

$$\frac{\dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\dfrac{\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2}}{n_1+n_2-2}}}\rightsquigarrow t(n_1+n_2-2)$$

Las del punto 2.4.1. hay que aprenderlas de memoria también

2. Relaciones de Ejercicios

2.1. Estadísticos muestrales

Ejercicio 2.1.1. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X. Dar el espacio muestral y calcular la función masa de probabilidad de $(X_1, \ldots, X_n)$ en cada uno de los siguientes casos:

a) $X \rightsquigarrow \{B(k_0, p) : p \in (0,1)\}$ Binomial.

El espacio muestral en este caso es $\mathcal{X}^n$, donde:

$$\mathcal{X} = \{0, 1, ..., k_0\}$$

Recordamos que si $X \rightsquigarrow B(k_0, p)$, entonces:

$$P[X = x] = \binom{k_0}{x} p^x (1-p)^{k_0 - x} \qquad \forall x \in \mathcal{X}$$

Por tanto, para nuestra m.a.s. tendremos la función masa de probabilidad:

$$P[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{indep.}}{=} \prod_{i=1}^{n} P[X_i = x_i] \overset{\text{id. d.}}{=} \prod_{i=1}^{n} P[X = x_i]$$

$$= \prod_{i=1}^{n} \binom{k_0}{x_i} p^{x_i} (1-p)^{k_0 - x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{nk_0 - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{k_0}{x_i}$$

$$\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$$

b) $X \rightsquigarrow \{\mathcal{P}(\lambda) : \lambda \in \mathbb{R}^+\}$ Poisson.

El espacio muestral de X es:

$$\mathcal{X} = \mathbb{N} \cup \{0\}$$

Recordamos que si $X \rightsquigarrow \mathcal{P}(\lambda)$, entonces:

$$P[X = x] = e^{-\lambda} \frac{\lambda^x}{x!} \qquad \forall x \in \mathcal{X}$$

Por tanto:

$$P[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{indep.}}{=} \prod_{i=1}^{n} P[X_i = x_i] \overset{\text{id. d.}}{=} \prod_{i=1}^{n} P[X = x_i]$$

$$= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \cdot \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i} \qquad \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$$

c) $X \rightsquigarrow \{BN(k_0, p) : p \in (0,1)\}$ Binomial Negativa.

El espacio muestral de X es:

$$\mathcal{X} = \mathbb{N} \cup \{0\}$$

Recordamos que si $X \rightsquigarrow BN(k_0, p)$, entonces:

$$P[X = x] = \binom{x + k_0 - 1}{x}(1-p)^x p^{k_0} \qquad \forall x \in \mathcal{X}$$

Por tanto:

$$P[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{indep.}}{=} \prod_{i=1}^{n} P[X_i = x_i] \overset{\text{id. d.}}{=} \prod_{i=1}^{n} P[X = x_i]$$
$$= \prod_{i=1}^{n} \binom{x_i + k_0 - 1}{x_i}(1-p)^{x_i} p^{k_0} = p^{nk_0}(1-p)^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{x_i + k_0 - 1}{x_i}$$
$$\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$$

d) $X \rightsquigarrow \{G(p) : p \in (0,1)\}$ Geométrica.

El espacio muestral de X es:

$$\mathcal{X} = \mathbb{N} \cup \{0\}$$

Recordamos que $G(p) \equiv BN(1, p)$, por lo que si sustituimos en la fórmula obtenida en la Binomial Negativa $k_0 = 1$:

$$P[X_1 = x_1, \ldots, X_n = x_n] = p^n (1-p)^{\sum_{i=1}^{n} x_i} \qquad \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$$

e) $X \rightsquigarrow \{P_N : N \in \mathbb{N}\}, \quad P_N(X = x) = \dfrac{1}{N}, \quad x = 1, \ldots, N.$

El espacio muestral ya nos lo dan: $\mathcal{X} = \{1, \ldots, N\}$. Calculemos la masa de probabilidad:

$$P[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{indep.}}{=} \prod_{i=1}^{n} P[X_i = x_i] \overset{\text{id. d.}}{=} \prod_{i=1}^{n} P[X = x_i]$$
$$= \prod_{i=1}^{n} \frac{1}{N} = \left(\frac{1}{N}\right)^n \qquad \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$$

Ejercicio 2.1.2. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X. Dar el espacio muestral y calcular la función de densidad de $(X_1, \ldots, X_n)$ en cada uno de los siguientes casos:

a) $X \rightsquigarrow \{U(a, b) : a, b \in \mathbb{R}, a < b\}$ Uniforme.

El espacio muestral en este caso es $\mathcal{X}^n$, donde:

$$\mathcal{X} = [a, b]$$

Recordamos que si $X \rightsquigarrow U(a,b)$, entonces:

$$f_X(x) = \frac{1}{b-a} \qquad \forall x \in [a,b]$$

Por lo que:

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \overset{\text{id. d.}}{=} \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{b-a}$$
$$= \left(\frac{1}{b-a}\right)^n \qquad \forall (x_1,\dots,x_n) \in \mathcal{X}^n$$

b) $X \rightsquigarrow \{\mathcal{N}(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$ Normal.

El espacio muestral de X es $\mathcal{X} = \mathbb{R}$. Recordamos que si $X \rightsquigarrow \mathcal{N}(\mu,\sigma^2)$, entonces:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \forall x \in \mathbb{R}$$

Por lo que:

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \overset{\text{id. d.}}{=} \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \qquad \forall (x_1,\dots,x_n) \in \mathbb{R}^n$$

c) $X \rightsquigarrow \{\Gamma(p,a) : p,a \in \mathbb{R}^+\}$ Gamma.

El espacio muestral de X es $\mathcal{X} = \mathbb{R}_0^+$. Recordamos que si $X \rightsquigarrow \Gamma(p,a)$, entonces:

$$f_X(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax} \qquad \forall x \in \mathbb{R}_0^+$$

Por lo que:

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \overset{\text{id. d.}}{=} \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{a^p}{\Gamma(p)} x_i^{p-1} e^{-ax_i}$$
$$= \left(\frac{a^p}{\Gamma(p)}\right)^n \cdot e^{-a\sum_{i=1}^{n} x_i} \cdot \prod_{i=1}^{n} x_i^{p-1} \qquad \forall (x_1,\dots,x_n) \in \mathcal{X}^n$$

d) $X \rightsquigarrow \{\beta(p,q) : p,q \in \mathbb{R}^+\}$ Beta.

El espacio muestral de X es $\mathcal{X} = [0,1]$. Recordamos que si $X \rightsquigarrow \beta(p,q)$, entonces:

$$f_X(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1} \qquad \forall x \in [0,1]$$

Donde:

$$\beta(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

Por tanto:

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \overset{\text{id. d.}}{=} \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{\beta(p,q)} x_i^{p-1}(1-x_i)^{q-1}$$
$$= \frac{1}{\beta(p,q)^n} \prod_{i=1}^{n} x_i^{p-1}(1-x_i)^{q-1} \qquad \forall (x_1,\dots,x_n) \in \mathcal{X}^n$$

e) $X \rightsquigarrow \{P_\theta : \theta \in \mathbb{R}^+\}, \quad f_\theta(x) = \dfrac{1}{2\sqrt{x\theta}}, \quad 0 < x < \theta.$

Se nos dice que $\mathcal{X} =]0,\theta[$. Calculamos la función de densidad conjunta:

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \overset{\text{id. d.}}{=} \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{2\sqrt{x_i\theta}}$$
$$= \frac{1}{\left(2\sqrt{\theta}\right)^n} \prod_{i=1}^{n} \frac{1}{\sqrt{x_i}} \qquad \forall (x_1,\dots,x_n) \in \mathcal{X}^n$$

Ejercicio 2.1.3. Se miden los tiempos de sedimentación de una muestra de partículas flotando en un líquido. Los tiempos observados son:

11,5; 1,8; 7,3; 12,1; 1,8; 21,3; 7,3; 15,2; 7,3; 12,1; 15,2;
7,3; 12,1; 1,8; 10,5; 15,2; 21,3; 10,5; 15,2; 11,5

- Construir la función de distribución muestral asociada a a dichas observaciones.

  Si aplicamos la definición de función de distribución muestral obtenemos que esta viene dada por:

$$F_n^*(x) = \begin{cases} 0 & \text{si } x < 1{,}8 \\ 3/20 & \text{si } 1{,}8 \leqslant x < 7{,}3 \\ 7/20 & \text{si } 7{,}3 \leqslant x < 10{,}5 \\ 9/20 & \text{si } 10{,}5 \leqslant x < 11{,}5 \\ 11/20 & \text{si } 11{,}5 \leqslant x < 12{,}1 \\ 14/20 & \text{si } 12{,}1 \leqslant x < 15{,}2 \\ 18/20 & \text{si } 15{,}2 \leqslant x < 21{,}3 \\ 20/20 & \text{si } x \geqslant 21{,}3 \end{cases}$$

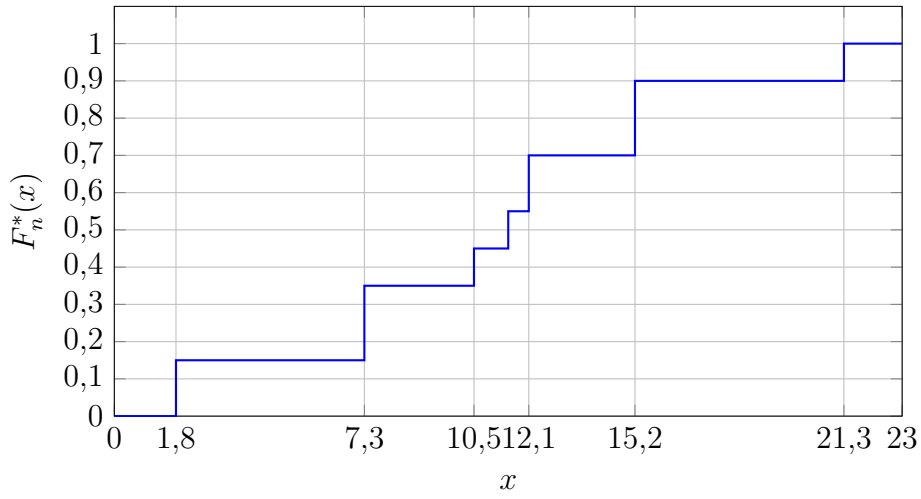


Figura 2.1: Gráfica de la función de distribución muestral.

- Hallar los valores de los tres primeros momentos muestrales respecto al origen y respecto a la media.

  Calculamos primero los tres primeros momentos respecto al origen para luego calcular los centrados respecto a la media a partir de ellos:

$$a_1 = \sum_{i=1}^{n} f_i x_i = 10{,}915 \qquad a_2 = \sum_{i=1}^{n} f_i x_i^2 = 148{,}9325$$

$$a_3 = \sum_{i=1}^{n} f_i x_i^3 = 2280{,}98365$$

$$b_1 = \sum_{i=1}^{n} f_i (x_i - \overline{x}) = 0$$

$$b_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i^2 - 2x_i\overline{x} + \overline{x}^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2\overline{x}}{n}\sum_{i=1}^{n} x_i + \overline{x}^2 = a_2 - 2a_1^2 + a_1^2 = a_2 - a_1^2$$

$$= 148{,}9325 - 10{,}915 = 29{,}795275$$

$$b_3 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^3 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i^3 - 3x_i^2\overline{x} + 3x_i\overline{x}^2 - \overline{x}^3\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^3 - \frac{3\overline{x}}{n}\sum_{i=1}^{n} x_i^2 + \frac{3\overline{x}^2}{n}\sum_{i=1}^{n} x_i - \overline{x}^3 = a_3 - 3a_1a_2 + 3a_1^3 - a_1^3$$

$$= a_3 - 3a_1a_2 + 2a_1^3 = 4{,}95455925$$

- Determinar los valores de los cuartiles muestrales y el percentil 70.

  Para ello, primero ordenamos los datos de menor a mayor y los agrupamos en

grupos de $^{20}/_{4} = 5$ en 5:

$$1{,}8;\ 1{,}8;\ 1{,}8;\ 7{,}3;\ 7{,}3;\ 7{,}3;\ 7{,}3;\ 10{,}5;\ 10{,}5;\ 11{,}5;\ 11{,}5;\ 12{,}1;\ 12{,}1;\ 12{,}1;\ 15{,}2;\ 15{,}2;\ 15{,}2;\ 15{,}2;\ 21{,}3;\ 21{,}3$$

Como en los cambios de agrupaciones de números estos se repiten, hemos obtenido el valor de los cuartiles:

$$q_1 = 7{,}3 \qquad q_2 = 11{,}5 \qquad q_3 = 15{,}2 \qquad q_4 = 21{,}3$$

Para el percentil 70, calculamos:

$$0{,}7 \cdot 20 = 14$$

Como hemos obtenido un número entero, el percentil 70 será:

$$c_{70} = \frac{X_{(14)} + X_{(15)}}{2} = \frac{12{,}1 + 15{,}2}{2} = 13{,}65$$

En el caso de haber obtenido un número no entero (por ejemplo, 14,2), sería $X_{(15)}$.

Ejercicio 2.1.4. Se dispone de una muestra aleatoria simple de tamaño 40 de una distribución exponencial de media 3, ¿cuál es la probabilidad de que los valores de la función de distribución muestral y la teórica, en $x = 1$, difieran menos de 0,01? Aproximadamente, ¿cuál debe ser el tamaño muestral para que dicha probabilidad sea como mínimo 0,98?

Como dice el enunciado, tenemos una m.a.s. $(X_1, \ldots, X_n)$ con $n = 40$, todas ellas idénticamente distribuidas a $X \rightsquigarrow exp(\lambda)$. Sabemos de la asignatura de Probabilidad que:

$$E[X] = \frac{1}{\lambda} = 3 \Longrightarrow \lambda = \frac{1}{3}$$

Por lo que $X \rightsquigarrow exp\left(\frac{1}{3}\right)$. Denotaremos por comodidad:

$$F_n^*(x) = F^*_{(X_1,\ldots,X_n)}(x)$$

Y el enunciado nos pregunta por:

$$P[|F_n^*(1) - F_X(1)| < 0{,}01]$$

Para ello, primero calculamos $F_X(1)$:

$$F_X(1) = 1 - e^{-\lambda \cdot 1} = 1 - e^{-\lambda} = 1 - e^{-1/3} = \alpha$$

Por lo que nos disponemos ya a calcular la probabilidad:

$$\begin{aligned} P[|F_n^*(1) - F_X(1)| < 0{,}01] &= P[|F_n^*(1) - \alpha| < 0{,}01] = P[-0{,}01 < F_n^*(1) - \alpha < 0{,}01] \\ &= P[-0{,}01 + \alpha < F_n^*(1) < 0{,}01 + \alpha] \\ &= P[40(-0{,}01 + \alpha) < 40F_n^*(1) < 40(0{,}01 + \alpha)] \end{aligned}$$

Y como sabemos que $Y = 40F_n^*(1) \rightsquigarrow B(40, F_X(1)) \equiv B(40, \alpha)$:

$$P[|F_n^*(1) - F_X(1)| < 0{,}01] = P[40(-0{,}01 + \alpha) < Y < 40(0{,}01 + \alpha)]$$

Si ahora tomamos:

$$\alpha = 1 - e^{-1/3} \approx 0{,}283469$$

Entonces:

$$40(0{,}01 + \alpha) \approx 40(0{,}01 + 0{,}283469) = 11{,}73876$$
$$40(-0{,}01 + \alpha) \approx 40(-0{,}01 + 0{,}283469) = 10{,}93876$$

Por lo que:

$$P[|F_n^*(1) - F_X(1)| < 0{,}01] \approx P[10{,}93876 < Y < 11{,}73876] = P[Y = 11]$$

De donde usando la masa de probabilidad de la Binomial:

$$P[Y = 11] = \binom{40}{11}(0{,}283469)^{11}(1 - 0{,}283469)^{40-11} \approx 0{,}139$$

Para el segundo apartado, como para $n = 40$ obtenemos una probabilidad de 0,139, podemos intuir que para que dicha probabilidad sea como mínimo 0,98, nos es necesario un valor de n grande, por lo que podemos suponer que:

$$F_n^*(1) \rightsquigarrow \mathcal{N}\left(\alpha, \frac{\alpha(1-\alpha)}{n}\right)$$

De donde:

$$Z = \frac{\sqrt{n}(F_n^*(1) - \alpha)}{\sqrt{\alpha(1-\alpha)}} \rightsquigarrow \mathcal{N}(0, 1)$$

Buscamos el valor de n que verifica:

$$0{,}98 \leqslant P[|F_n^*(1) - F_X(1)| < 0{,}01] = P\left[|Z| < \frac{\sqrt{n}0{,}01}{\sqrt{\alpha(1-\alpha)}}\right]$$

Si aplicamos propiedades conocidas de la Normal, si $a \in \mathbb{R}$, entonces:

$$P[|Z| < a] = P[-a < Z < a] = P[Z < a] - P[Z < -a]$$

Pero:

$$P[Z < -a] = P[Z > a] = 1 - P[Z < a]$$

Por lo que:

$$P[|Z| < a] = P[Z < a] - P[Z < -a] = 2P[Z < a] - 1$$

Volviendo al caso que nos interesa:

$$0{,}98 \leqslant P\left[|Z| < \frac{\sqrt{n}0{,}01}{\sqrt{\alpha(1-\alpha)}}\right] = 2P\left[Z < \frac{\sqrt{n}0{,}01}{\sqrt{\alpha(1-\alpha)}}\right] - 1$$

Luego:

$$0{,}99 = \frac{0{,}98+1}{2} \leqslant P\left[Z < \frac{\sqrt{n}0{,}01}{\sqrt{\alpha(1-\alpha)}}\right]$$

Si consultamos la tabla de la normal $\mathcal{N}(0,1)$, observamos que el primer valor que supera la probbailidad de 0,99 es 2,33, por lo que:

$$2{,}33 = \frac{\sqrt{n}0{,}01}{\sqrt{\alpha(1-\alpha)}} = \frac{\sqrt{n}0{,}01}{\sqrt{0{,}283469(1-0{,}283469)}} \approx 0{,}0221886\sqrt{n}$$

De donde:

$$5{,}4289 = (2{,}33)^2 = (0{,}0221886\sqrt{n})^2 = 0{,}00049233n \Longrightarrow n = \frac{5{,}4289}{0{,}00049233} = 11026{,}95347$$

Por lo que para $n \geqslant 11027$ podemos asegurar que la probabilidad es como mínimo 0,98.

Ejercicio 2.1.5. Se dispone de una muestra aleatoria simple de tamaño 50 de una distribución de Poisson de media 2, ¿cuál es la probabilidad de que los valores de la función de distribución muestral y la teórica, en $x = 2$, difieran menos de 0,02? Aproximadamente, ¿qué tamaño muestral hay que tomar para que dicha probabilidad sea como mínimo 0,99?

Tenemos una m.a.s. $(X_1, \ldots, X_n)$ con $n = 50$ idénticamente distribuidas a $X \rightsquigarrow \mathcal{P}(2)$. Notamos por comodidad:

$$F^*_{(X_1,\ldots,X_n)}(x) = F^*_n(x)$$

Nos preguntan por:

$$P[|F^*_n(2) - F_X(2)| < 0{,}02]$$

Para ello primero calculamos:

$$F_X(2) = \sum_{k=0}^{2} e^{-2}\frac{2^k}{k!} = e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!}\right) = e^{-2}(1+2+2) \approx 0{,}6767$$

Por lo que:

$$P[|F^*_n(2) - 0{,}6767| < 0{,}02] = P[-0{,}02 < F^*_n(2) - 0{,}6767 < 0{,}02] = P[0{,}6567 < F^*_n(2) < 0{,}6967]$$

Como sabemos por lo visto en teoría que:

$$Y = 50F^*_n(2) \rightsquigarrow B(50, F_X(2)) \equiv B(50, \quad 0{,}6767)$$

Multiplicamos por 50 la última expresión:

$$\begin{aligned} P[|F^*_n(2) - 0{,}6767| < 0{,}02] = P[0{,}6567 < F^*_n(2) < 0{,}6967] &= P[32{,}835 < Y < 34{,}835] \\ &= P[Y = 33] + P[Y = 34] \end{aligned}$$

Y calculamos estas dos probabilidades:

$$P[Y = 33] = \binom{50}{33}(0{,}6767)^{33}(1 - 0{,}6767)^{50-33} \approx 0{,}114734$$
$$P[Y = 34] = \binom{50}{34}(0{,}6767)^{34}(1 - 0{,}6767)^{50-34} \approx 0{,}120075$$

Por lo que:

$$P[|F_n^*(2) - 0{,}6767| < 0{,}02] \approx 0{,}114734 + 0{,}120075 = 0{,}234809$$

Para el segundo apartado, como para $n = 50$ obtenemos una probabilidad de 0,234809, podemos intuir que para que dicha probabilidad sea como mínimo 0,99, nos es necesario un valor de n grande, por lo que podemos suponer que:

$$F_n^*(2) \rightsquigarrow \mathcal{N}\left(0{,}6767, \frac{0{,}6767(1 - 0{,}6767)}{n}\right) \equiv \mathcal{N}\left(0{,}6767, \frac{0{,}218777}{n}\right)$$

Por lo que:

$$Z = \frac{\sqrt{n}(F_n^*(2) - 0{,}6767)}{\sqrt{0{,}218777}} \rightsquigarrow \mathcal{N}(0, 1)$$

En dicho caso, buscamos n de forma que:

$$0{,}99 \leqslant P[|F_n^*(2) - F_X(2)| < 0{,}02] = P\left[|Z| < \frac{\sqrt{n}0{,}02}{\sqrt{0{,}218777}}\right]$$

De forma análoga al ejercicio anterior:

$$P\left[|Z| < \frac{\sqrt{n}0{,}02}{\sqrt{0{,}218777}}\right] = 2P\left[Z < \frac{\sqrt{n}0{,}02}{\sqrt{0{,}218777}}\right] - 1$$

Luego:

$$0{,}995 = \frac{0{,}99 + 1}{2} \leqslant P\left[Z < \frac{\sqrt{n}0{,}02}{\sqrt{0{,}218777}}\right]$$

Y si miramos la tabla de la Normal observamos que el primer valor que supera la probabilidad de 0,995 es 2,58, luego:

$$2{,}58 = \frac{\sqrt{n}0{,}02}{\sqrt{0{,}218777}} = 0{,}042759\sqrt{n}$$

Por lo que:

$$6{,}6564 = (2{,}58)^2 = (0{,}042759\sqrt{n})^2 = 0{,}00182833n$$

Luego:

$$n = \frac{6{,}6564}{0{,}00182833} \approx 3640{,}7$$

Por lo que para $n \geqslant 3641$ podemos asegurar que la probabilidad es como mínimo 0,99.

Ejercicio 2.1.6. Sea $X \rightsquigarrow B(1,p)$ y (X_1, X_2, X_3) una muestra aleatoria simple de X. Calcular la función masa de probabilidad de los estadísticos $\overline{X}$, S^2, mín X_i y máx X_i.

Para resolver este ejercicio, como X sigue una distribución discreta, buscamos aplicar el teorema de cambio de variable de discreta a discreta. Para ello, la forma más cómoda será analizar cada uno de los valores que puede tomar la muestra aleatoria simple (X_1, X_2, X_3) y determinar en consecuencia cada uno de los valores que toman $\overline{X}$, S^2, mín X_i y máx X_i. Acompañaremos la tabla junto con la probabilidad de que la muestra tome dicho valor, es decir, en la fila correpondiente a (x_1, x_2, x_3) incluiremos $P[X_1 = x_1, X_2 = x_2, X_3 = x_3]$:

P	(X_1, X_2, X_3)	$\overline{X}$	S^2	mín X_i	máx X_i
$(1-p)^3$	$(0,0,0)$	0	0	0	0
$p(1-p)^2$	$(0,0,1)$	$1/3$	$1/3$	0	1
$p(1-p)^2$	$(0,1,0)$	$1/3$	$1/3$	0	1
$p^2(1-p)$	$(0,1,1)$	$2/3$	$1/3$	0	1
$p(1-p)^2$	$(1,0,0)$	$1/3$	$1/3$	0	1
$p^2(1-p)$	$(1,0,1)$	$2/3$	$1/3$	0	1
$p^2(1-p)$	$(1,1,0)$	$2/3$	$1/3$	0	1
p^3	$(1,1,1)$	1	0	1	1

Podemos ya calcular la función masa de probabilidad de cada uno de los estadísticos, simplemente sumando las probabilidades de la tabla que corresponden a cada valor del espacio muestral de cada estadístico:

- Para $\overline{X}$:

$$P[\overline{X} = 0] = P[X_1 = 0, X_2 = 0, X_3 = 0] = (1-p)^3$$
$$P[\overline{X} = 1/3] = \sum_{i=1}^{3} p(1-p)^2 = 3p(1-p)^2$$
$$P[\overline{X} = 2/3] = \sum_{i=1}^{3} p^2(1-p) = 3p^2(1-p)$$
$$P[\overline{X} = 1] = P[X_1 = 1, X_2 = 1, X_3 = 1] = p^3$$

- Para S^2:

$$P[S^2 = 0] = P[X_1 = 0, X_2 = 0, X_3 = 0] + P[X_1 = 1, X_2 = 1, X_3 = 1] = p^3 + (1-p)^3$$
$$P[S^2 = 1/3] = \sum_{i=1}^{3} p(1-p)^2 + \sum_{i=1}^{3} p^2(1-p) = 3p(1-p)(p+1-p) = 3p(1-p)$$

- Para mín X_i:

$$P[\text{mín } X_i = 1] = P[X_1 = 1, X_2 = 1, X_3 = 1] = p^3$$
$$P[\text{mín } X_i = 0] = 1 - P[\text{mín } X_i = 1] = 1 - p^3$$

- Para máx X_i:

$$P[\text{máx}\, X_i = 0] = P[X_1 = 0, X_2 = 0, X_3 = 0] = (1-p)^3$$
$$P[\text{máx}\, X_i = 1] = 1 - P[\text{máx}\, X_i = 0] = 1 - (1-p)^3$$

Ejercicio 2.1.7. Obtener la función masa de probabilidad o función de densidad de $\overline{X}$ en el muestreo de una variable de Bernoulli, de una Poisson y de una exponencial.

Calculamos la masa de probabilidad o función de densidad en cada caso, suponiendo que tenemos $(X_1, \ldots, X_n)$ una muestra aleatoria simple con variables aleatorias idénticamente distribuidas a X, que sigue una distribución distinta en cada caso y estaremos interesados en calcular la masa de:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Bernoulli. Supuesto que $X \rightsquigarrow B(1,p)$ para cierto $p \in]0,1[$, si tomamos:

$$Y = \sum_{i=1}^{n} X_i$$

Por la propiedad reproductiva de la Bernoulli, tenemos que $Y \rightsquigarrow B(n,p)$. En dicho caso:

$$P[Y = k] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \forall k \in \{0, \ldots, n\}$$

Por tanto, tendremos que:

$$P\left[\overline{X} = \frac{k}{n}\right] = P[Y = k] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \forall k \in \{0, \ldots, n\}$$

Poisson. Supuesto que $X \rightsquigarrow \mathcal{P}(\lambda)$ para cierto $\lambda \in \mathbb{R}^+$, si tomamos:

$$Y = \sum_{i=1}^{n} X_i$$

Por la propiedad reproductiva de la Poisson, tendremos que:

$$Y \rightsquigarrow \mathcal{P}\left(\sum_{i=1}^{n} \lambda\right) \equiv \mathcal{P}(n\lambda)$$

En dicho caso:

$$P[Y = x] = e^{-n\lambda}\frac{(n\lambda)^x}{x!} \qquad \forall x \in \mathbb{N}$$

Por lo que:

$$P[\overline{X} = {}^{x}/_{n}] = P[Y = x] = e^{-n\lambda}\frac{(n\lambda)^x}{x!} \qquad \forall x \in \mathbb{N}$$

Exponencial. Supuesto ahora que $X \rightsquigarrow exp(\lambda)$ para cierto $\lambda \in \mathbb{R}^+$, tendremos entonces que:

$$M_X(t) = \frac{\lambda}{\lambda - t} \qquad t < \lambda$$

Si aplicamos la igualdad $(*)$ vista en teoría:

$$M_{\overline{X}}(t) \overset{(*)}{=} (M_X(t/n))^n = \left(\frac{\lambda}{\lambda - t/n}\right)^n = \left(\frac{n\lambda}{n\lambda - t}\right)^n$$

Observamos que obtenemos una función generatriz de momentos para $\overline{X}$ igual que para una variable aleatoria de distribución $\Gamma(n, n\lambda)$. Como la función generatriz de momentos de una variable aleatoria caracteriza su distribución, concluimos que $\overline{X} \rightsquigarrow \Gamma(n, n\lambda)$.

Ejercicio 2.1.8. Calcular las funciones de densidad de los estadísticos máx X_i y mín X_i en el muestreo de una variable X con funcion de densidad:

$$f_\theta(x) = e^{\theta - x}, \qquad x > \theta.$$

Calculamos primero la función de distribución, para calcular con mayor comodidad las funciones de distribución de $X_{(n)}$ y $X_{(1)}$:

$$F_\theta(x) = \int_\theta^x f_\theta(t)\ dt = \int_\theta^x e^{\theta - t}\ dt = \left[-e^{\theta - t}\right]_\theta^x = 1 - e^{\theta - x} \qquad \forall x > \theta$$

Supuesto ahora que disponemos de una m.a.s. $(X_1, \ldots, X_n)$ idénticamente distribuidas a X cuya función de densidad es la anteriormente dicha, podemos aplicar las fórmulas obtenidas en teoría para calcular las funciones de distribución del mínimo y del máximo. Para el máximo:

$$F_{X_{(n)}}(x) = (F_X(x))^n = (1 - e^{\theta - x})^n \Longrightarrow f_{X_{(n)}} = n(1 - e^{\theta - x})^{n-1} e^{\theta - x} \qquad \forall x > \theta$$

Para el mínimo:

$$F_{X_{(1)}}(x) = 1 - (1 - F_X(n))^n = 1 - (1 - 1 + e^{\theta - x})^n = 1 - e^{n(\theta - x)} \qquad \forall x > \theta$$

de donde:

$$f_{X_{(1)}}(x) = n e^{n(\theta - x) - 1} \qquad \forall x > \theta$$

Ejercicio 2.1.9. El número de pacientes que visitan diariamente una determinada consulta médica es una variable aleatoria con varianza de 16 personas. Se supone que el número de visitas de cada día es independiente de cualquier otro. Si se observa el número de visitas diarias durante 64 días, calcular aproximadamente la probabilidad de que la media muestral no difiera en más de una persona del valor medio verdadero de visitas diarias.

Sea X una variable aleatoria que indica el número de pacientes que visitan diariamente dicha consulta médica, por cómo nos definen X sabemos que $X \rightsquigarrow \mathcal{P}(\lambda)$. Como además nos dicen que la varianza de dicha variable aleatoria es 16, tenemos

que $Var(X) = \lambda = 16$. Si tenemos ahora una muestra aleatoria simple $(X_1, \ldots, X_n)$ con $n = 64$, nos preguntan por:

$$P[|\overline{X} - E[X]| < 1]$$

Donde $E[X] = \lambda = 16$, ya que $X \rightsquigarrow \mathcal{P}(16)$. Calculamos:

$$P[|\overline{X} - E[X]| < 1] = P[-1 < \overline{X} - 16 < 1] = P[15 < \overline{X} < 17]$$

Aplicamos ahora lo visto en el ejercicio 7, ya que si $X \rightsquigarrow \mathcal{P}(\lambda)$, entonces tendremos que $n\overline{X} \rightsquigarrow \mathcal{P}(n\lambda)$, gracias a la propiedad reproductiva de la Poisson:

$$\begin{aligned} P[|\overline{X} - E[X]| < 1] = P[15 < \overline{X} < 17] &= P[64 \cdot 15 < 64\overline{X} < 64 \cdot 17] \\ &= P[960 < 64\overline{X} < 1088] \end{aligned}$$

Donde $64\overline{X} \rightsquigarrow \mathcal{P}(64 \cdot 16) \equiv \mathcal{P}(1024)$. Para calcular dicha probabilidad, aproximaremos la Poisson a una distribución normal:

$$\mathcal{P}(1024) \approx \mathcal{N}(1024, 1024)$$

Por lo que:

$$\begin{aligned} P[|\overline{X} - E[X]| < 1] = P[960 < 64\overline{X} < 1088] &\approx P\left[\frac{960 - 1024}{\sqrt{1024}} < Z < \frac{1088 - 1024}{\sqrt{1024}}\right] \\ &= P[-2 < Z < 2] = 2P[Z < 2] - 1 \\ &= 2 \cdot 0{,}97725 - 1 = 0{,}9545 \end{aligned}$$

Ejercicio 2.1.10. Una máquina de refrescos está arreglada para que la cantidad de bebida que sirve sea una variable aleatoria con media 200 ml. y desviación típica 15 ml. Calcular de forma aproximada la probabilidad de que la cantidad media servida en una muestra aleatoria de tamaño 36 sea al menos 204 ml.

Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple idénticamente distribuida a la variable aleatoria X de tamaño $n = 36$, aplicando el Teorema Central del Límite obtenemos que $\overline{X} \rightsquigarrow \mathcal{N}\left(200, \frac{15^2}{36}\right)$. Calculamos la probabilidad:

$$\begin{aligned} P[\overline{X} \geqslant 204] \overset{\text{tipificamos}}{=} P\left[Z \geqslant \frac{\sqrt{36}(204 - 200)}{15}\right] &= P[Z \geqslant 1{,}6] = 1 - P[Z < 1{,}6] \\ &= 1 - 0{,}9452 = 0{,}0548 \end{aligned}$$

2.2. Distribuciones en el muestreo de poblaciones normales

Ejercicio 2.2.1. Se toma una muestra aleatoria simple de tamaño 5 de una variable aleatoria con distribución $\mathcal{N}(2{,}5,\ 36)$. Calcular:

a) Probabilidad de que la cuasivarianza muestral esté comprendida entre 1,863 y 2,674.

Tenemos una m.a.s. $(X_1, X_2, X_3, X_4, X_5)$, todas ellas idénticamente distribuidas a $X \rightsquigarrow \mathcal{N}(2{,}5,\ 36)$. Tomamos la cuasivarianza de dichos datos:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{4}\sum_{i=1}^{5}(X_i - 2{,}5)^2$$

Y queremos calcular la probabilidad:

$$P\left[1{,}863 < S^2 < 2{,}674\right]$$

Para ello, en teoría hemos visto que como tenemos una población normal, se cumple que:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{4S^2}{36} \rightsquigarrow \chi^2(4) \equiv \chi^2(n-1)$$

Por tanto:

$$\begin{aligned} P[1{,}863 < S^2 < 2{,}674] &= P\left[\frac{4\cdot 1{,}863}{36} < \frac{4S^2}{36} < \frac{4\cdot 2{,}674}{36}\right] \\ &= P\left[0{,}207 < \frac{4S^2}{36} < 0{,}2971\right] \\ &= P\left[\frac{4S^2}{36} > 0{,}207\right] - P\left[\frac{4S^2}{36} > 0{,}2971\right] \end{aligned}$$

Si consultamos la tabla de la $\chi^2(4)$:

$$P\left[\frac{4S^2}{36} > 0{,}207\right] = 0{,}995, \qquad P\left[\frac{4S^2}{36} > 0{,}2971\right] = 0{,}99$$

Por lo que:

$$\begin{aligned} P[1{,}863 < S^2 < 2{,}674] &= P\left[\frac{4S^2}{36} > 0{,}207\right] - P\left[\frac{4S^2}{36} > 0{,}2971\right] \approx 0{,}995 - 0{,}99 \\ &= 0{,}005 \end{aligned}$$

b) Probabilidad de que la media muestral esté comprendida entre 1,3 y 3,5, supuesto que la cuasivarianza muestral está entre 30 y 40.

En la situación del apartado anterior, ahora tenemos también en consideración la media muestral:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Bajo estas condiciones, el Lema de Fisher nos dice que S^2 y $\overline{X}$ son independientes, por lo que la probabilidad de que la media muestral esté comprendida entre 1,3 y 3,5 suponiendo que la cuasivarianza muestral está comprendida entre 30 y 40 es igual a la probabilidad de que la media muestral esté comprendida entre 1,3 y 3,5:

$$P[1{,}3 < \overline{X} < 3{,}5 \mid 30 < S^2 < 40] = P[1{,}3 < \overline{X} < 3{,}5]$$

Sabemos que $\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, nos disponemos a calcular dicha probabilidad:

$$\begin{aligned} P[1{,}3 < \overline{X} < 3{,}5] &\overset{\text{tipificamos}}{=} P\left[\frac{\sqrt{5}(1{,}3-2{,}5)}{\sqrt{36}} < Z < \frac{\sqrt{5}(3{,}5-2{,}5)}{\sqrt{36}}\right] \\ &= P[-0{,}447214 < Z < 0{,}372678] \\ &= P[Z < 0{,}372678] - P[Z > 0{,}447214] \\ &= P[Z < 0{,}372678] - 1 + P[Z < 0{,}447214] \\ &= 0{,}64431 - 1 + 0{,}67003 = 0{,}31434 \end{aligned}$$

Ejercicio 2.2.2. La longitud craneal en una determinada población humana es una variable aleatoria que sigue una distribución normal con media 185,6 mm. y desviación típica 12,78 mm. ¿Cuál es la probabilidad de que una muestra aleatoria simple de tamaño 20 de esa población tenga media mayor que 190 mm.?

Tenemos una población que sigue una variable aleatoria $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ con parámetros $\mu = 185{,}6$ y $\sigma = 12{,}78$. Tomamos una m.a.s. $(X_1, \ldots, X_n)$ con $n = 20$ y tomamos la media muestral:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Queremos calcular:

$$P\left[\overline{X} > 190\right]$$

Para ello, sabemos por lo visto en teoría que $\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, por lo que:

$$\begin{aligned} P\left[\overline{X} > 190\right] &\overset{\text{tipificamos}}{=} P\left[Z > \frac{\sqrt{20}(190-185{,}6)}{12{,}78}\right] = P[Z > 1{,}5397] \\ &= 1 - P[Z < 1{,}5397] = 1 - 0{,}93822 = 0{,}06178 \end{aligned}$$

Ejercicio 2.2.3. ¿De qué tamaño mínimo habría que seleccionar una muestra de una variable con distribución normal $\mathcal{N}(\mu, 4)$ para poder afirmar, con probabilidad mayor que 0,9, que la media muestral diferirá de la poblacional menos de 0,1?

Supuesto que tenemos una m.a.s. $(X_1, \ldots, X_n)$ de tamaño $n \in \mathbb{N}$ de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow \mathcal{N}(\mu, 4)$, queremos buscar el menor n de forma que:

$$0{,}9 \leqslant P\left[|\overline{X} - \mu| < 0{,}1\right]$$

Para ello, usaremos que hemos visto en teoría que $\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ con $\sigma = 2$:

$$0{,}9 \leqslant P[|\overline{X} - \mu| < 0{,}1] = P[-0{,}1 < |\overline{X} - \mu| < 0{,}1] = P\left[\frac{-\sqrt{n}0{,}1}{2} < Z < \frac{\sqrt{n}0{,}1}{2}\right]$$
$$= P\left[-\sqrt{n}0{,}05 < Z < \sqrt{n}0{,}05\right] = 2P\left[Z < \sqrt{n}0{,}05\right] - 1$$

De donde:

$$0{,}95 = \frac{0{,}9 + 1}{2} \leqslant P[Z < \sqrt{n}0{,}05]$$

De donde mirando la tabla de la normal $\mathcal{N}(0, 1)$:

$$\sqrt{n}0{,}05 \geqslant 1{,}65 \Longrightarrow n \geqslant \left(\frac{1{,}65}{0{,}05}\right)^2 = 1089$$

Por lo que habría que seleccionar como mínimo una muestra de tamaño 1089 para poder afirmar con probabilidad mayor que 0,9 que la media muestral difiere de la poblacional menos de 0,1.

Ejercicio 2.2.4. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable con distribución normal. Calcular la probabilidad de que la cuasivarianza muestral sea menor que un 50 % de la varianza poblacional para $n = 16$ y para $n = 1000$.

Dicha muestra aleatoria simple es de variables idénticamente distribuidas a $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$. Bajo estas condiciones, sabemos por lo visto en teoría que:

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$$

Y queremos calcular:

$$P[S^2 < 0{,}5\sigma^2] = P\left[\frac{(n-1)S^2}{\sigma^2} < \frac{(n-1)0{,}5\cancel{\sigma^2}}{\cancel{\sigma^2}}\right] = 1 - P\left[\frac{(n-1)S^2}{\sigma^2} > (n-1)0{,}5\right]$$

- Para $n = 16$, mirando en la tabla de $\chi^2(15)$ tenemos que:

$$P[S^2 < 0{,}5\sigma^2] = 1 - P\left[\frac{(n-1)S^2}{\sigma^2} > 7{,}5\right] = 1 - 0{,}95 = 0{,}05$$

- Para $n = 1000$, no disponemos de tabla para $\chi^2(999)$, por lo que aproximaremos $\chi^2(999)$ por $\mathcal{N}(999, 2 \cdot 999) \equiv \mathcal{N}(999, 1998)$, gracias al Teorema de Lévy. De esta forma, si tomamos:

$$Y = \frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \mathcal{N}(999, 1998)$$

Tenemos entonces que:

$$P[999 \cdot 0{,}5] = P[Y > 499{,}5] = P\left[Z > \frac{499{,}5 - 999}{\sqrt{1998}}\right] = 1 - P[Z < -11{,}17]$$

Por lo que:

$$P[S^2 < 0{,}5\sigma^2] \approx P[Z < -11{,}17] \approx 0$$

Ejercicio 2.2.5. Sean S_1^2 y S_2^2 las cuasivarianzas muestrales de dos muestras independientes de tamaños $n_1 = 5$ y $n_2 = 4$ de dos poblaciones normales con la misma varianza. Calcular la probabilidad de que S_1^2/S_2^2 sea menor que 5,34 o mayor que 9,12.

Bajo estas hipótesis, sabemos por lo visto en teoría que (como $\sigma_1 = \sigma_2$):

$$\frac{S_1^2}{S_2^2} \rightsquigarrow F(n_1 - 1, n_2 - 1) \equiv F(4, 3)$$

Por tanto, notando $Y = \frac{S_1^2}{S_2^2}$ por comodidad, calculamos (son sucesos disjuntos):

$$P[Y < 5{,}34 \text{ o } Y > 9{,}12] = P[Y < 5{,}34] + P[Y > 9{,}12]$$

Observamos la tabla de $F(4, 3)$:

$$\begin{aligned} &P[Y < 5{,}34] = 0{,}9 \\ &P[Y > 9{,}12] = 1 - P[Y < 9{,}12] = 1 - 0{,}95 = 0{,}05 \end{aligned}$$

En definitiva:

$$P[Y < 5{,}34 \text{ o } Y > 9{,}12] = P[Y < 5{,}34] + P[Y > 9{,}12] = 0{,}95$$

Ejercicio 2.2.6. Se consideran dos poblaciones de bombillas cuyas longitudes de vida siguen una ley normal con la misma media y desviaciones típicas 425 y 375 horas, respectivamente. Con objeto de realizar un estudio comparativo de ambas poblaciones, se considera una muestra aleatoria simple de 10 bombillas en la primera población y una de tamaño 6 en la segunda. ¿Cuál es la probabilidad de que la media muestral del primer grupo menos la del segundo sea menor que la observada en dos realizaciones muestrales que dieron 1325 horas y 1215 horas, respectivamente?

Como describe el enunciado, tenemos una m.a.s. $(X_1, \ldots, X_n)$ de tamaño $n = 10$ de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow \mathcal{N}\left(\mu, 425^2\right)$; y otra m.a.s. $(Y_1, \ldots, Y_m)$ de tamaño $m = 6$ de variables aleatorias idénticamente distribuidas a $Y \rightsquigarrow \mathcal{N}\left(\mu, 375^2\right)$. Notaremos $\mu_1 = 1325$, $\mu_2 = 1215$. Bajo estas condiciones, sabemos por lo visto en teoría que:

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X} - \overline{Y} - (\mu - \mu)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

Luego si notamos a esta última variable como Z:

$$\begin{aligned} P[\overline{X} - \overline{Y} < 1325 - 1215] &= P\left[Z < \frac{1325 - 1215}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right] = P\left[Z < \frac{1325 - 1215}{\sqrt{\frac{425^2}{10} + \frac{375^2}{6}}}\right] \\ &= P[Z < 0{,}5399] = 0{,}7054 \end{aligned}$$

Ejercicio 2.2.7. Sean $X_1, \ldots, X_n, X_{n+1}$ variables aleatorias independientes e idénticamente distribuidas según una $\mathcal{N}(\mu, \sigma^2)$, y sean $\overline{X}$ y S^2 la media y la cuasivarianza muestral de $(X_1, \ldots, X_n)$. Calcular la distribución de

$$\frac{X_{n+1} - \overline{X}}{S}\sqrt{\frac{n}{n+1}}$$

Bajo dichas hipótesis, tenemos:

$$\left.\begin{array}{r} X_{n+1} \rightsquigarrow \mathcal{N}(\mu, \sigma^2) \\ \overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \end{array}\right\} \Longrightarrow X_{n+1} - \overline{X} \rightsquigarrow \mathcal{N}\left(\mu - \mu, \sigma^2 + \frac{\sigma^2}{n}\right) \equiv \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$$

de donde si tipificamos la variable:

$$\frac{X_{n+1} - \overline{X}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} = \frac{X_{n+1} - \overline{X}}{\sigma\sqrt{\frac{n+1}{n}}} = \frac{X_{n+1} - \overline{X}}{\sigma}\sqrt{\frac{n}{n+1}} \rightsquigarrow \mathcal{N}(0,1)$$

Además, se ha visto en teoría que:

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$$

Como tanto X_{n+1} como $\overline{X}$ son independientes de S^2, podemos aplicar la construcciones de la distribución t de Student:

$$\left.\begin{array}{l} U \rightsquigarrow \mathcal{N}(0,1) \\ \qquad \text{independientes} \\ V \rightsquigarrow \chi^2(n) \end{array}\right\} \Longrightarrow \frac{U}{\sqrt{V/n}} \rightsquigarrow t(n)$$

Si lo aplicamos a nuestras variables:

$$\frac{\frac{X_{n+1} - \overline{X}}{\sigma}\sqrt{\frac{n}{n+1}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\frac{X_{n+1} - \overline{X}}{\sigma}\sqrt{\frac{n}{n+1}}}{\frac{S}{\sigma}} = \frac{X_{n+1} - \overline{X}}{S}\sqrt{\frac{n}{n+1}} \rightsquigarrow t(n-1)$$

Ejercicio 2.2.8. Sean $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_m)$ muestras aleatorias simples independientes de poblaciones $\mathcal{N}(\mu_1, \sigma^2)$ y $\mathcal{N}(\mu_2, \sigma^2)$, respectivamente. Sean $\alpha, \beta \in \mathbb{R}$ y $\overline{X}, \overline{Y}, S_1^2, S_2^2$ las medias y cuasivarianzas de las dos muestras. Calcular la distribución de

$$\frac{\alpha\left(\overline{X} - \mu_1\right) + \beta\left(\overline{Y} - \mu_2\right)}{\sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}\sqrt{\frac{\alpha^2}{n} + \frac{\beta^2}{m}}}$$

Sabemos que:

$$\left.\begin{array}{r} \overline{X} \rightsquigarrow \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right) \\ \overline{Y} \rightsquigarrow \mathcal{N}\left(\mu_2, \frac{\sigma^2}{m}\right) \end{array}\right\} \Longrightarrow \left.\begin{array}{r} \overline{X} - \mu_1 \rightsquigarrow \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \\ \overline{Y} - \mu_2 \rightsquigarrow \mathcal{N}\left(0, \frac{\sigma^2}{m}\right) \end{array}\right\} \Longrightarrow \left.\begin{array}{r} \alpha(\overline{X} - \mu_1) \rightsquigarrow \mathcal{N}\left(0, \frac{\alpha^2\sigma^2}{n}\right) \\ \beta(\overline{Y} - \mu_2) \rightsquigarrow \mathcal{N}\left(0, \frac{\beta^2\sigma^2}{m}\right) \end{array}\right\}$$

Como ambas son independientes, podemos aplicar la propeidad reproductiva de la normal:

$$\alpha(\overline{X}-\mu_1)+\beta(\overline{Y}-\mu_2) \rightsquigarrow \mathcal{N}\left(0, \frac{\alpha^2\sigma^2}{n}+\frac{\beta^2\sigma^2}{m}\right) \equiv \mathcal{N}\left(0, \sigma^2\left(\frac{\alpha^2}{n}+\frac{\beta^2}{m}\right)\right)$$

Si tipificamos la variable:

$$\frac{\alpha(\overline{X}-\mu_1)+\beta(\overline{Y}-\mu_2)}{\sigma\sqrt{\frac{\alpha^2}{n}+\frac{\beta^2}{m}}} \rightsquigarrow \mathcal{N}(0,1)$$

Por otra parte, tenemos que:

$$\frac{(n-1)S_1^2}{\sigma^2} \rightsquigarrow \chi^2(n-1), \qquad \frac{(m-1)S_2^2}{\sigma^2} \rightsquigarrow \chi^2(m-1)$$

Como ambas son independientes por ser S_1^2 y S_2^2 independientes (ya que las muestras aleatorias simples eran independientes), podemos aplicar la propiedad reproductiva de χ^2, obteniendo que:

$$\frac{(n-1)S_1^2}{\sigma^2}+\frac{(m-1)S_2^2}{\sigma^2} = \frac{(n-1)S_1^2+(m-1)S_2^2}{\sigma^2} \rightsquigarrow \chi^2(n+m-2)$$

Finalmente, la extensión del Lema de Fisher nos dice que $(\overline{X}, \overline{Y})$ es independiente de (S_1^2, S_2^2), por lo que las dos variables aleatorias con las que trabajamos son independientes, lo que nos permite aplicar la construcción de la distribución t de Student:

$$\frac{\dfrac{\alpha(\overline{X}-\mu_1)+\beta(\overline{Y}-\mu_2)}{\cancel{\sigma}\sqrt{\frac{\alpha^2}{n}+\frac{\beta^2}{m}}}}{\sqrt{\dfrac{\dfrac{(n-1)S_1^2+(m-1)S_2^2}{\cancel{\sigma^2}}}{n+m-2}}} = \frac{\alpha\left(\overline{X}-\mu_1\right)+\beta\left(\overline{Y}-\mu_2\right)}{\sqrt{\dfrac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}\sqrt{\dfrac{\alpha^2}{n}+\dfrac{\beta^2}{m}}} \rightsquigarrow t(n+m-2)$$

2.3. Suficiencia y completitud

Ejercicio 2.3.1. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable $X \rightsquigarrow \{B(k,p) : p \in]0,1[\}$ y sea $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$. Probar

a) usando la definición

b) aplicando el teorema de factorización

que T es suficiente para p.

a) Si notamos para abreviar $T = T(X_1, \ldots, X_n)$, tenemos que probar que la distribución de la muestra condicionada a cualquier valor del estadístico no depende del parámetro p para probar que T es suficiente para p. Para ello:

$$P_p[X_1 = x_1, \ldots, X_n = x_n \mid T = t] = \frac{P_p[X_1 = x_1, \ldots, X_n = x_n, T = t]}{P_p[T = t]}$$
$$= \begin{cases} 0 & \text{si } T(x_1, \ldots, x_n) \neq t \\ \dfrac{P_p[X_1 = x_1, \ldots, X_n = x_n]}{P_p[T = t]} & \text{si } T(x_1, \ldots, x_n) = t \end{cases}$$

Como 0 obviamente no depende de p, el caso $T(x_1, \ldots, x_2) \neq t$ se encuentra ya estudiado, por lo que nos centramos en el caso $T(x_1, \ldots, x_2) = t$:

$$\frac{P_p[X_1 = x_1, \ldots, X_n = x_n]}{P_p[T = t]} \overset{\text{iid.}}{=} \frac{\prod_{i=1}^{n} P_p[X = x_i]}{P_p[T = t]} \overset{(*)}{=} \frac{\prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}}{\binom{nk}{t} p^t (1-p)^{nk-t}}$$
$$= \frac{p^{\sum_{i=1}^{n} x_i} (1-p)^{nk - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{k}{x_i}}{\binom{nk}{t} p^t (1-p)^{nk-t}} \overset{(**)}{=} \frac{\prod_{i=1}^{n} \binom{k}{x_i}}{\binom{nk}{t}}$$

Donde en $(*)$ hemos usado que $(X_1, \ldots, X_n)$ es una m.a.s. (variables independientes) y la reproductividad de la Binomial, por lo que $T \rightsquigarrow B(nk, p)$; y en $(**)$ hemos usado que $t = T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$. En definitiva, hemos obtenido que la distribución de la muestra condicionada a cualquier valor del estadístico no dependende del parámetro p, por lo que T es suficiente para p.

b) Si podemos usar el Teorema de factorización, escribimos la función masa de probabilidad de la distribución conjunta de la muestra aleatoria simple:

$$P[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{iid.}}{=} \prod_{i=1}^{n} P[X = x_i] = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}$$
$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{nk - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{k}{x_i}$$

Si tomamos:

$$h(x_1, \ldots, x_n) = \prod_{i=1}^{n} \binom{k}{x_i}, \qquad T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i, \qquad g_p(t) = p^t (1-p)^{nk-t}$$

Podemos aplicar el Teorema de Factorización de Neymann-Fisher, obteniendo que T es un estadístico suficiente para p.

Ejercicio 2.3.2. Sea $(X_1, \dots, X_n)$ una muestra aleatoria simple de una variable $X \rightsquigarrow \{\mathcal{P}(\lambda) : \lambda \in \mathbb{R}^+\}$ y sea $T(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i$. Probar

a) usando la definición

b) aplicando el teorema de factorización

que T es suficiente para λ.

a) Notando $T = T(X_1, \dots, X_n)$, seguimos los mismos pasos que en el ejercicio anterior:

$$P_\lambda[X_1 = x_1, \dots, X_n = x_n \mid T = t] = \frac{P_\lambda[X_1 = x_1, \dots, X_n = x_n, T = t]}{P_\lambda[T = t]}$$
$$= \begin{cases} 0 & \text{si } T(x_1, \dots, x_n) \neq t \\ \dfrac{P_\lambda[X_1 = x_1, \dots, X_n = x_n]}{P_\lambda[T = t]} & \text{si } T(x_1, \dots, x_n) = t \end{cases}$$

Y ahora nos interesamos por el segundo término, que es el que puede depender de λ:

$$\frac{P_\lambda[X_1 = x_1, \dots, X_n = x_n]}{P_\lambda[T = t]} \overset{\text{iid.}}{=} \frac{\prod_{i=1}^{n} P_\lambda[X = x_i]}{P_\lambda[T = t]} \overset{(*)}{=} \frac{\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}}{e^{-n\lambda}\frac{(n\lambda)^t}{t!}} = \frac{e^{-n\lambda} \cdot \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}}{e^{-n\lambda} \cdot \frac{n^t\lambda^t}{t!}}$$
$$\overset{(**)}{=} \frac{t!}{n^t \prod_{i=1}^{n} x_i!}$$

Donde en $(*)$ usamos que $T \rightsquigarrow P(n\lambda)$, por la reproductividad de la Poisson y en $(**)$ usamos que $t = T(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$. Obtenemos una cantidad que no depende de λ, por lo que T es suficiente para λ.

b) Si podemos aplicar el Teorema de factorización, escribimos:

$$P_\lambda[X_1 = x_1, \dots, X_n = x_n] \overset{\text{iid.}}{=} \prod_{i=1}^{n} P_\lambda[X = x_i] = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \cdot \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Si tomamos:

$$h(x_1, \dots, x_n) = \frac{1}{\prod_{i=1}^{n} x_i!}, \qquad T(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i, \qquad g_\lambda(t) = e^{-n\lambda} \cdot \lambda^t$$

Por el Teorema de Factorización de Neymann-Fisher obtenemos que T es suficiente para λ.

Ejercicio 2.3.3. Sea (X_1, X_2, X_3) una muestra aleatoria simple de una variable $X \rightsquigarrow \{B(1,p) : p \in]0,1[\}$. Probar que el estadístico $X_1+2X_2+3X_3$ no es suficiente.

Consideramos el estadístico $T(X_1, X_2, X_3) = X_1+2X_2+3X_3$. El espacio muestral de X es $\mathcal{X} = \{0,1\}$, por lo que el espacio muestral de T es $\mathcal{T} = \{0,1,2,3,4,5,6\}$. Sabemos por un ejemplo visto en teoría que el "truco" para demostrar que $T(X_1, X_2, X_3)$ no es suficiente para p es buscar un valor del espacio muestral $\mathcal{T}$ que provenga de varias combinaciones de estados del espacio muestral $\mathcal{X}^3$. Como $0, 1, 2$ solo provienen de una combinación del espacio muestral $\mathcal{X}^3$ (son $(0,0,0)$, $(1,0,0)$ y $(0,1,0)$ respectivamente), probamos buscar el contraejemplo con $t = 3$, que proviene de considerar las observaciones de la muestra $(1,1,0)$ y $(0,0,1)$.

Una vez explicado el procedimiento para buscar cuál es el valor de t que funciona, procedemos a probar que la distribución de la muestra condicionada a dicho valor de t depende del parámetro p. Para ello:

$$P_p[X_1 = x_1, X_2 = x_2, X_3 = x_3 \mid T = 3] = \frac{P_p[X_1 = x_1, X_2 = x_2, X_3 = x_3, T = 3]}{P_p[T = 3]}$$
$$= \begin{cases} 0 & \text{si } T(x_1, x_2, x_2) \neq 3 \\ \dfrac{P_p[X_1 = x_1, X_2 = x_2, X_3 = x_3]}{P_p[T = 3]} & \text{si } T(x_1, x_2, x_3) = 3 \end{cases}$$

Vemos que el primer caso no puede depender nunca de p, por lo que buscamos probar que el segundo caso sí que depende de p. Para ello:

$$\frac{P_p[X_1 = x_1, X_2 = x_2, X_3 = x_3]}{P_p[T = 3]} \overset{\text{iid.}}{=} \frac{P_p[X = x_1]P_p[X = x_2]P_p[X = x_3]}{P_p[T = 3]}$$
$$\overset{(*)}{=} \frac{p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}p^{x_3}(1-p)^{1-x_3}}{P_p[X_1 = 1, X_2 = 1, X_3 = 0] + P_p[X_1 = 0, X_2 = 0, X_3 = 1]}$$
$$= \frac{p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}p^{x_3}(1-p)^{1-x_3}}{p \cdot p \cdot (1-p) + (1-p)(1-p)p}$$

Donde en $(*)$ he usado la propiedad reproductiva de la Binomial, por lo que $T \rightsquigarrow B(3,p)$, así como que la condición $t = 3$ provenía de los valores $(1,1,0)$ y $(0,0,1)$. Ahora, si tomamos $(x_1, x_2, x_3) = (1,1,0)$, tenemos que:

$$\frac{P_p[X_1 = 1, X_2 = 1, X_3 = 0]}{P_p[T = 3]} = \frac{p^1(1-p)^0p^1(1-p)^0p^0(1-p)^1}{p \cdot p \cdot (1-p) + (1-p)(1-p)p}$$
$$= \frac{pp(1-p)}{pp(1-p) + (1-p)(1-p)p}$$
$$= \frac{p^2(1-p)}{p(1-p)\cancel{(p + 1 - p)}^{1}} = p$$

Que claramente depende de p, por lo que T no es suficiente para p.

Ejercicio 2.3.4. Aplicando el teorema de factorización, y basándose en una muestra de tamaño arbitrario, encontrar un estadístico suficiente para cada una de las siguientes familias de distribuciones (en las familias biparamétricas, suponer los casos de sólo un parámetro desconocido y de los dos desconocidos).

a) $X \rightsquigarrow \{U(-\theta/2, \theta/2) : \theta > 0\}$

b) $X \rightsquigarrow \{\Gamma(p, a) : p, a > 0\}$

c) $X \rightsquigarrow \{\beta(p, q) : p, q > 0\}$

d) $X \rightsquigarrow \{P_{N_1,N_2} : N_1, N_2 \in \mathbb{N}, N_1 \leqslant N_2\}$ y la masa de probabilidad viene dada por:

$$P_{N_1,N_2}[X = x] = \frac{1}{N_2 - N_1 + 1} \qquad x \in \{N_1, \dots, N_2\}$$

En lo que sigue, supondremos que tenemos una m.a.s. $(X_1, \dots, X_n)$ de variables idénticamente distribuidas a la respectiva variable X:

a) Si $X \rightsquigarrow U(-\theta/2, \theta/2)$ con $\theta > 0$:

$$f(x_1, \dots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n} \qquad x_i \in \,]-\theta/2, \theta/2[\,, \quad \forall i \in \{1, \dots, n\}$$

Por lo que tendremos $-\theta/2 < X_{(1)} \leqslant X_{(n)} < \theta/2$:

$$f(x_1, \dots, x_n) = \frac{1}{\theta^n} I_{]0,+\infty[}(X_{(1)} + \theta/2) I_{]-\infty,0[}(X_{(n)} - \theta/n)$$

Si tomamos:

$$h(x_1, \dots, x_n) = 1, \qquad T(X_1, \dots, X_n) = (X_{(1)}, X_{(n)})$$
$$g_\theta(t_1, t_2) = \frac{1}{\theta^n} I_{]0,+\infty[}(t_1 + \theta/2) I_{]-\infty,0[}(t_2 - \theta/2)$$

Por el Teorema de factorización, tenemos que $T(X_1, \dots, X_n)$ es un estadístico suficiente para θ.

b) Si $X \rightsquigarrow \Gamma(p, a)$ con $p, a > 0$:

$$f(x_1, \dots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{a^p}{\Gamma(p)} x_i^{p-1} e^{-a x_i} = \left(\frac{a^p}{\Gamma(p)}\right)^n e^{-a \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i^{p-1}$$
$$= \left(\frac{a^p}{\Gamma(p)}\right)^n e^{-a \sum_{i=1}^{n} x_i} \left(\prod_{i=1}^{n} x_i\right)^{p-1} \qquad x_i \geqslant 0 \qquad \forall i \in \{1, \dots, n\}$$

- Suponiendo que p es conocida, podemos tomar:

$$h(x_1, \dots, x_n) = \left(\prod_{i=1}^{n} x_i\right)^{p-1}, \qquad T(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i$$
$$g_a(t) = \left(\frac{a^p}{\Gamma(p)}\right)^n e^{-at}$$

- Suponiendo ahora que a es conocida:

$$h(x_1, \dots, x_n) = e^{-a \sum_{i=1}^{n} x_i}, \qquad T(X_1, \dots, X_n) = \prod_{i=1}^{n} x_i$$
$$g_p(t) = \left(\frac{a^p}{\Gamma(p)}\right)^n t^{p-1}$$

- Si ahora tanto p como a son desconocidas, podemos tomar:

$$h(x_1,\dots,x_n)=1,\qquad T(X_1,\dots,X_n)=\left(\sum_{i=1}^n X_i,\prod_{i=1}^n X_i\right)$$
$$g_{(a,p)}(t_1,t_2)=\left(\frac{a^p}{\Gamma(p)}\right)^n e^{-at_1}t_2^{p-1}$$

c) Si $X\rightsquigarrow\beta(p,q)$ con $p,q>0$:

$$f(x_1,\dots,x_n)\overset{\text{iid.}}{=}\prod_{i=1}^n f(x_i)=\prod_{i=1}^n\frac{1}{\beta(p,q)}x_i^{p-1}(1-x_i)^{q-1}$$
$$=\frac{1}{\beta(p,q)^n}\left(\prod_{i=1}^n x_i\right)^{p-1}\left(\prod_{i=1}^n(1-x_i)\right)^{q-1}\qquad x_i\in[0,1],\forall i\in\{1,\dots,n\}$$

- Si p es conocida, tomamos:

$$h(x_1,\dots,x_n)=\left(\prod_{i=1}^n x_i\right)^{p-1},\qquad T(X_1,\dots,X_n)=\prod_{i=1}^n(1-X_i)$$
$$g_q(t)=\frac{1}{\beta(p,q)^n}t^{q-1}$$

- Si q es conocida:

$$h(x_1,\dots,x_n)=\left(\prod_{i=1}^n(1-x_i)\right)^{q-1},\qquad T(X_1,\dots,X_n)=\prod_{i=1}^n X_i$$
$$g_p(t)=\frac{1}{\beta(p,q)^n}t^{p-1}$$

- Si tanto p como q son parámetros:

$$h(x_1,\dots,x_n)=1,\qquad T(X_1,\dots,X_n)=\left(\prod_{i=1}^n X_i,\prod_{i=1}^n(1-X_i)\right)$$
$$g_{(p,q)}(t_1,t_2)=\frac{1}{\beta(p,q)^n}t_1^{p-1}t_2^{q-1}$$

d) Si $X\rightsquigarrow P_{N_1,N_2}$ con $N_1,N_2\in\mathbb{N}$, $N_1\leqslant N_2$, entonces:

$$P[X_1=x_1,\dots,X_n=x_n]\overset{\text{iid.}}{=}\prod_{i=1}^n P[X=x_i]=\prod_{i=1}^n\frac{1}{N_2-N_1+1}$$
$$=\frac{1}{(N_2-N_1+1)^n},\quad x_i\in\{N_1,\dots N_2\}$$

Como se tiene $N_1\leqslant X_{(1)}\leqslant X_{(n)}\leqslant N_2$, entonces:

$$P[X_1=x_1,\dots,X_n=x_n]=\frac{I_{-\mathbb{N}_0}(X_{(1)}-N_1)I_{\mathbb{N}_0}(X_{(n)}-N_2)}{(N_2-N_1+1)^n}$$

- Si N_1 es conocido:

$$h(x_1,\dots,x_n) = I_{-\mathbb{N}_0}(X_{(1)} - N_1), \qquad T(X_1,\dots,X_n) = X_{(n)}$$
$$g_{N_2}(t) = \frac{I_{\mathbb{N}_0}(t - N_2)}{(N_2 - N_1 + 1)^n}$$

- Si N_2 es conocida:

$$h(x_1,\dots,x_n) = I_{\mathbb{N}_0}(X_{(n)} - N_2), \qquad T(X_1,\dots,X_n) = X_{(1)}$$
$$g_{N_1}(t) = \frac{I_{-\mathbb{N}_0}(t - N_1)}{(N_2 - N_1 + 1)^n}$$

- Si tanto N_1 como N_2 son parámetros:

$$h(x_1,\dots,x_n) = 1, \qquad T(X_1,\dots,X_n) = \big(X_{(1)}, X_{(n)}\big)$$
$$g_{(N_1,N_2)}(t_1,t_2) = \frac{I_{-\mathbb{N}_0}(t_1 - N_1) I_{\mathbb{N}_0}(t_2 - N_2)}{(N_2 - N_1 + 1)^n}$$

Por lo que en cualquier caso obtenemos un estadístico suficiente, por el Teorema de Factorización.

Ejercicio 2.3.5. Sea $X \rightsquigarrow \{P_N : N \in \mathbb{N}\}$, siendo P_N la distribución uniforme en los puntos $\{1,\dots,N\}$, y sea $(X_1,\dots,X_n)$ una muestra aleatoria simple de X. Probar que máx$(X_1,\dots,X_n)$ es un estadístico suficiente y completo.

Buscamos aplicar el Teorema de factorización de Neymann-Fisher:

$$P[X_1 = x_1,\dots,X_n = x_n] \overset{\text{iid.}}{=} \prod_{i=1}^{n} P[X = x_i] = \prod_{i=1}^{n} \frac{1}{N}, \quad x_i \in \{1,\dots,N\}, \forall i \in \{1,\dots,n\}$$

Por lo que $1 \leqslant X_{(1)} \leqslant X_{(n)} \leqslant N$:

$$P[X_1 = x_1,\dots,X_n = x_n] = \frac{I_{\mathbb{N}_0}(X_{(1)} - 1) I_{-\mathbb{N}_0}(X_{(n)} - N)}{N^n}$$

De donde podemos tomar:

$$h(x_1,\dots,x_n) = I_{\mathbb{N}_0}(X_{(1)} - 1), \qquad T(X_1,\dots,X_n) = X_{(n)}$$
$$g_N(t) = \frac{I_{-\mathbb{N}_0}(t - N)}{N^n}$$

Por el Teorema de factorización de Neymann-Fisher, tenemos que el estadístico $T(X_1,\dots,X_n) = X_{(n)}$ es suficiente. Comprobamos que también es completo: sea g cualquier función medible, supongamos que (abreviaremos $T(X_1,\dots,X_n) = T$):

$$0 = E[g(T)] = \sum_{t=1}^{N} g(t) P[T = t] \qquad \forall N \in \mathbb{N}$$

Calculamos la función masa de probabilidad de T:

$$F_T(t) = (F_X(t))^n \Longrightarrow P[T = t] = P[T \leqslant t] - P[T \leqslant t-1] = (F_X(t))^n - (F_X(t-1))^n$$

Como $F_X(t) = \frac{t}{N}$, tenemos entonces que:

$$P[T = t] = (F_X(t))^n - (F_X(t-1))^n = \frac{t^n}{N^n} - \frac{(t-1)^n}{N^n} = \frac{t^n - (t-1)^n}{N^n}$$

Por lo que:

$$0 = E[g(T)] = \sum_{t=1}^{N} g(t)\frac{t^n - (t-1)^n}{N^n} = \frac{1}{N^n}\sum_{t=1}^{N} g(t)(t^n - (t-1)^n) \qquad \forall N \in \mathbb{N}$$

de donde:

$$\sum_{t=1}^{N} g(t)(t^n - (t-1)^n) = 0 \qquad \forall N \in \mathbb{N}$$

Probemos por inducción sobre N que $g(t) = 0 \quad \forall t \in \mathbb{N}$:

- Para $N = 1$: tenemos que:

$$g(1) = g(1)(1^n - (1-1)^n) = \sum_{t=1}^{1} g(t)(t^n - (t-1)^n) = 0$$

- Supuesto que $g(t) = 0$ para $t < N$:

$$g(N)(N^n - (N-1)^n) = \sum_{t=1}^{N} g(t)(t^n - (t-1)^n) = 0$$

  Por lo que:

  - $g(N) = 0$.
  - $N^n - (N-1)^n = 0$, que es imposible, puesto que la potencia $n-$ésima es una función estrictamente creciente en el intervalo $[0, +\infty[$.

  En definitiva, tenemos que:

$$\mathbb{N} \subseteq \{t : g(t) = 0\}$$

  Por lo que:

$$1 \geqslant P[g(T) = 0] \geqslant P[T \in \mathbb{N}] = 1 \Longrightarrow P[g(T) = 0] = 1$$

  Lo que demuestra que $X_{(n)}$ es completo.

Ejercicio 2.3.6. Basándose en una muestra de tamaño arbitrario, obtener un estadístico suficiente y completo para la familia de distribuciones definidas por todas las densidades de la forma

$$f_\theta(x) = e^{\theta - x}, \qquad x > \theta$$

Sea $(X_1, \ldots, X_n)$ una m.a.s. de tamaño $n \in \mathbb{N}$:

$$f_\theta(x_1, \ldots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} e^{\theta - x_i} = e^{n\theta - \sum_{i=1}^{n} x_i} = \frac{e^{n\theta}}{e^{\sum_{i=1}^{n} x_i}}, \quad x_i > \theta \ \forall i \in \{1, \ldots, n\}$$

Por lo que ha de ser $\theta < X_{(1)}$:

$$f_\theta(x_1, \ldots, x_n) = \frac{e^{n\theta} \cdot I_{]-\infty,0[}(X_{(1)} - \theta)}{e^{\sum_{i=1}^{n} x_i}}$$

Si tomamos:

$$h(x_1, \ldots, x_n) = e^{-\sum_{i=1}^{n} x_i}, \qquad T(X_1, \ldots, X_n) = X_{(1)}$$
$$g_\theta(t) = e^{n\theta} \cdot I_{]-\infty,0[}(t - \theta)$$

Por el Teorema de factorización de Neymann-Fisher tenemos que el estadístico $X_{(1)}$ es suficiente para θ. Comprobemos si es completo: sea g cualquier función medible, suponemos que (y escribimos $T = T(X_1, \ldots, X_n)$ para abreviar):

$$0 = E[g(T)] = \int_\theta^{+\infty} g(t) f_T(t)\ dt \qquad \forall \theta \in \mathbb{R}$$

Como $T = X_{(1)}$, tenemos que:

$$F_T(t) = 1 - (1 - F_\theta(t))^n \Longrightarrow f_T(t) = n(1 - F_\theta(t))^{n-1} f_\theta(t)$$

donde:

$$F_\theta(t) = \int_\theta^t e^{\theta - x}\ dx = \left[-e^{\theta - x}\right]_\theta^t = 1 - e^{\theta - t}, \qquad t > \theta$$

por lo que:

$$f_T(t) = n\left(e^{\theta - t}\right)^{n-1} e^{\theta - t} = n\left(e^{\theta - t}\right)^n, \qquad t > \theta$$

volviendo al caso que nos interesa:

$$0 = E[g(T)] = \int_\theta^{+\infty} g(t) n\left(e^{\theta - t}\right)^n\ dt = n e^{n\theta} \int_\theta^{+\infty} g(t) e^{-nt}\ dt \qquad \forall \theta \in \mathbb{R}$$

por lo que:

$$\int_\theta^{+\infty} g(t) e^{-nt}\ dt = 0 \qquad \forall \theta \in \mathbb{R}$$

Por simplicidad de cálculos, supondremos que g es continua, con lo que si $G(t)$ es una primitiva de $g(t)e^{-nt}$, entonces:

$$\int_\theta^{+\infty} g(t) e^{-nt}\ dt = \lim_{n\to\infty} \int_\theta^n g(t) e^{-nt}\ dt = \lim_{n\to\infty} G(n) - G(\theta) = 0 \qquad \forall \theta \in \mathbb{R}$$

Si derivamos ahora respecto a θ, tenemos que:

$$-g(\theta) e^{-n\theta} = 0 \qquad \forall \theta \in \mathbb{R}$$

con lo que $g(\theta) = 0 \quad \forall \theta \in \mathbb{R}$. Es decir:

$$\mathbb{R} \subseteq \{t : g(t) = 0\}$$

luego:

$$1 \geqslant P[g(T) = 0] \geqslant P[T \in \mathbb{R}] = 1 \Longrightarrow P[g(T) = 0] = 1$$

por lo que T es completo.

Ejercicio 2.3.7. Comprobar que las siguientes familias de distribuciones son exponenciales uniparamétricas y, considerando una muestra aleatoria simple de una variable con distribución en dicha familia, obtener, si existe, un estadístico suficiente y completo.

a) $\{B(k_0, p) : 0 < p < 1\}$

Comprobamos todas las condiciones:

1. El espacio paramétrico es: $]0, 1[\subseteq \mathbb{R}$.
2. El espacio muestral es $\mathcal{X} = \{0, \ldots, k_0\}$, que no depende de p.
3. Para la tercera condición:

$$\begin{aligned} P[X = x]_p &= \binom{k_0}{p} p^x (1-p)^{k_0 - x} = exp\left[\ln\left(\binom{k_0}{p} p^x (1-p)^{k_0-x}\right)\right] \\ &= exp\left[\ln\binom{k_0}{p} + x\ln p + (k_0 - x)\ln(1-p)\right] \\ &= exp\left[\ln\binom{k_0}{p} + x\ln p + k_0\ln(1-p) - x\ln(1-p)\right] \\ &= exp\left[\ln\binom{k_0}{p} + k_0\ln(1-p) + x\ln\left(\frac{p}{1-p}\right)\right] \end{aligned}$$

Tomando:

$$\begin{aligned} T(x) = x, \qquad & S(x) = 0 \\ Q(p) = \ln\left(\frac{p}{1-p}\right), \qquad & D(p) = \ln\binom{k_0}{p} + k_0\ln(1-p) \end{aligned}$$

obtenemos la tercera condición.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow B(k_0, p)$ con $p \in]0, 1[$, el Teorema visto en teoría para las familias de distribuciones exponenciales nos dice que el estadístico:

$$T = T(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$$

es suficiente para p. Para ver que T es también completo, hemos de ver que ImQ contiene un abierto de $\mathbb{R}$:

Opción 1. Como $Q :]0, 1[\to \mathbb{R}$ es continua, no constante y definida sobre un intervalo, por el Teorema del Valor Intermedio su imagen ha de ser un intervalo, que es un abierto de $\mathbb{R}$, por lo que T es completo.

Opción 2.

$$ImQ = \left\{\ln\left(\frac{p}{1-p}\right) : p \in]0, 1[\right\}$$

Si definimos la función $f :]0, 1[\to \mathbb{R}^+$, tenemos que f es sobreyectiva (de hecho es biyectiva):

- Está bien definida, puesto que si $x \in]0,1[$, entonces $1-x>0$, con lo que $f(x) \in \mathbb{R}^+$.
- Sea $y \in \mathbb{R}^+$, tenemos que:

$$\frac{t}{1-t} = y \iff t = y(1-t) \iff t = y - yt \iff t(1+y) = y \iff t = \frac{y}{y+1} < 1$$

Por lo que $f(t) = y$ con $t \in]0,1[$, con lo que f es sobreyectiva.

Por tanto, $Imf = \mathbb{R}^+$, de donde deducimos que:

$$ImQ = \left\{ \ln\left(\frac{p}{1-p}\right) : p \in]0,1[\right\} = \{\ln(f(p)) : p \in]0,1[\} = \ln(Imf) = \ln(\mathbb{R}^+) = \mathbb{R}$$

Como obviamente $\mathbb{R}$ contiene algún abierto de $\mathbb{R}$, deducimos que T era un estadístico completo.

b) $\{\mathcal{P}(\lambda) : \lambda > 0\}$

Comprobamos las condiciones:

1. El espacio paramétrico es $\mathbb{R}^+ \subseteq \mathbb{R}$.
2. El espacio muestral es $\mathcal{X} = \mathbb{N} \cup \{0\}$, que no depende de λ.
3. Para la tercera condición:

$$P_\lambda[X = x] = e^{-\lambda}\frac{\lambda^x}{x!} = exp\left[\ln\left(e^{-\lambda}\frac{\lambda^x}{x!}\right)\right] = exp[-\lambda + x\ln\lambda - \ln(x!)]$$

Tomando:

$$\begin{aligned} T(x) &= x, & S(x) &= -\ln(x!) \\ Q(\lambda) &= \ln\lambda, & D(\lambda) &= -\lambda \end{aligned}$$

obtenemos la tercera condición.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow \mathcal{P}(\lambda)$ con $\lambda \in \mathbb{R}^+$, el Teorema visto en teoría para las familias de distribuciones exponenciales nos dice que el estadístico:

$$T = T(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$$

es suficiente para λ. Para ver que T es también completo, hemos de ver que ImQ contiene un abierto de $\mathbb{R}$. Como:

$$ImQ = \{\ln(\lambda) : \lambda \in \mathbb{R}^+\} = \ln(\mathbb{R}^+) = \mathbb{R}$$

Tenemos que $ImQ = \mathbb{R}$ claramente contiene un abierto de $\mathbb{R}$, por lo que T es completo.

c) $\{BN(k_0, p) : 0 < p < 1\}$

Comprobamos las condiciones:

1. El espacio paramétrico es $]0,1[\subseteq \mathbb{R}$.
2. El espacio muestral es $\mathcal{X} = \mathbb{N} \cup \{0\}$, que no depende de p.
3. Para la tercera condición:

$$P_p[X = x] = \binom{x + k_0 - 1}{x}(1-p)^x p^{k_0} = exp\left[\ln\left(\binom{x + k_0 - 1}{x}(1-p)^x p^{k_0}\right)\right]$$
$$= exp\left[\ln\binom{x + k_0 - 1}{x} + x\ln(1-p) + k_0 \ln p\right]$$

Tomando:

$$T(x) = x, \qquad S(x) = \ln\binom{x + k_0 - 1}{x}$$
$$Q(p) = \ln(1-p), \qquad D(p) = k_0 \ln p$$

Obtenemos la tercera condición.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow BN(k_0, p)$ con $p \in]0,1[$, el Teorema visto en teoría para las familias de distribuciones exponenciales nos dice que el estadístico:

$$T = T(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$$

es suficiente para p. Para ver que T es también completo, hemos de ver que ImQ contiene un abierto de $\mathbb{R}$. Como $Q :]0,1[\to \mathbb{R}$ es una función continua, no constante y definida en un intervalo, tenemos que su imagen es un intervalo, por lo que contiene abiertos de $\mathbb{R}$, de donde T es completo.

d) $\{exp(\lambda) : \lambda > 0\}$

Comprobamos las condiciones:

1. El espacio paramétrico es $\mathbb{R}^+ \subseteq \mathbb{R}$.
2. El espacio muestral es $\mathcal{X} = \mathbb{R}^+$, que no depende de λ.
3. Para la tercera condición:

$$f_\lambda(x) = \lambda e^{-\lambda x} = exp\left[\ln\left(\lambda e^{-\lambda x}\right)\right] = exp\left[\ln(\lambda) - \lambda x\right]$$

Tomando:

$$T(x) = x, \qquad S(x) = 0$$
$$Q(\lambda) = -\lambda, \qquad D(\lambda) = \ln(\lambda)$$

tenemos la tercera condición.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow exp(\lambda)$ con $\lambda \in \mathbb{R}^+$, el Teorema visto en teoría para las familias de distribuciones exponenciales nos dice que el estadístico:

$$T = T(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$$

es suficiente para λ. Para ver que T es también completo, hemos de ver que ImQ contiene un abierto de $\mathbb{R}$. Como $Q : \mathbb{R}^+ \to \mathbb{R}$ es una función continua, no constante y definida en un intervalo, tenemos que su imagen es un intervalo, por lo que contiene abiertos de $\mathbb{R}$, de donde T es completo.

Ejercicio 2.3.8. Estudiar si las siguientes familias de distribuciones son exponenciales biparamétricas. En caso afirmativo, considerando una muestra aleatoria simple de una variable con distribución en dicha familia, obtener, si existe, un estadístico suficiente y completo.

a) $\{\Gamma(p,a) : p,a > 0\}$

Comprobamos las condiciones:

1. El espacio paramétrico es $\mathbb{R}^+ \times \mathbb{R}^+ \subseteq \mathbb{R}^2$.
2. El espacio muestral es $\mathbb{R}^+$, que no depende de p ni de a.
3. Para la tercera condición:

$$\begin{aligned} f_{(p,a)}(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax} &= exp\left[\ln\left(\frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax}\right)\right] \\ &= exp\left[\ln\left(\frac{a^p}{\Gamma(p)}\right) + (p-1)\ln x - ax\right] \end{aligned}$$

Tomando:

$$\begin{aligned} T_1(x) = \ln x, \qquad & T_2(x) = x, \qquad & S(x) = 0 \\ Q_1(p,a) = (p-1), \qquad & Q_2(p,a) = -a, \qquad & D(p,a) = \ln\left(\frac{a^p}{\Gamma(p)}\right) \end{aligned}$$

Tenemos la tercera condición.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow \Gamma(p,a)$ con $p, a \in \mathbb{R}^+$, el Teorema visto en teoría para las familias de distribuciones exponenciales multiparamétricas nos dice que el estadístico:

$$T = T(X_1, \ldots, X_n) = \left(\sum_{i=1}^n T_1(X_i), \sum_{i=1}^n T_2(X_i)\right) = \left(\sum_{i=1}^n \ln(X_i), \sum_{i=1}^n X_i\right)$$

es suficiente para (p,a). Para ver que T es también completo, hemos de ver que ImQ contiene un abierto de $\mathbb{R}^2$, donde $Q : (\mathbb{R}^+)^2 \to \mathbb{R}^2$, con $Q = (Q_1, Q_2)$. Para ello:

$$\begin{aligned} ImQ = \left\{(p-1, -a) : (p,a) \in (\mathbb{R}^+)^2\right\} &= \{(x-1, y) : x \in \mathbb{R}^+, y \in \mathbb{R}^-\} \\ &=]-1, +\infty[\times \mathbb{R}^- \end{aligned}$$

Como claramente $]-1, +\infty[\times \mathbb{R}^-$ contiene un abierto de $\mathbb{R}^2$, tenemos que T es completo.

Observemos que también podríamos haber tomado:

$$T(X_1, \ldots, X_n) = \left(\prod_{i=1}^n X_i, \sum_{i=1}^n X_i\right)$$

b) $\{\beta(p,q) : p,q > 0\}$

Comprobamos las condiciones:

1. El espacio paramétrico es $\mathbb{R}^+ \times \mathbb{R}^+ \subseteq \mathbb{R}^2$.
2. El espacio muestral es $[0,1]$, que no depende de p ni de q.
3. Para la tercera condición:

$$f_{(p,q)}(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1} = exp\left[\ln\left(\frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1}\right)\right]$$
$$= exp\left[\ln\left(\frac{1}{\beta(p,q)}\right) + (p-1)\ln x + (q-1)\ln(1-x)\right]$$

Tomando:

$$T_1(x) = \ln x, \qquad T_2(x) = \ln(1-x), \qquad S(x) = 0$$
$$Q_1(p,q) = p-1, \qquad Q_2(p,q) = q-1, \qquad D(p,q) = \ln\left(\frac{1}{\beta(p,q)}\right)$$

Tenemos la tercera condición.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de variables aleatorias idénticamente distribuidas a $X \rightsquigarrow \beta(p,q)$ con $p,q \in \mathbb{R}^+$, el Teorema visto en teoría para las familias de distribuciones exponenciales multiparamétricas nos dice que el estadístico:

$$T = T(X_1, \ldots, X_n) = \left(\sum_{i=1}^n T_1(X_i), \sum_{i=1}^n T_2(X_i)\right) = \left(\sum_{i=1}^n \ln(X_i), \sum_{i=1}^n \ln(1-X_i)\right)$$

es suficiente para (p,q). Para ver que T es también completo, hemos de ver que ImQ contiene un abierto de $\mathbb{R}^2$, donde $Q : (\mathbb{R}^+)^2 \to \mathbb{R}^2$, con $Q = (Q_1, Q_2)$. Para ello:

$$ImQ = \{(p-1, q-1) : p,q \in \mathbb{R}^+\} =]-1, +\infty[\times]-1, +\infty[$$

Como claramente este conjunto contiene un abierto de $\mathbb{R}^2$, tenemos que T es completo.

2.4. Estimación puntual. Insesgadez y mínima varianza

Ejercicio 2.4.1. Sea $(X_1, \ldots, X_n)$ una muestra de una variable $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ con $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$. Probar que

$$T(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \overline{X} \leqslant 0 \\ 0 & \text{si } \overline{X} > 0 \end{cases}$$

es un estimador insesgado de la función paramétrica $\Phi\left(\frac{-\mu\sqrt{n}}{\sigma}\right)$, siendo Φ la función de distribución de la $\mathcal{N}(0,1)$.

Tenemos $T(X_1, \ldots, X_n) = I_{]-\infty,0]}(\overline{X})$. Como $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, sabemos por lo visto en el Tema 1 que entonces:

$$\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

de donde (escribiendo $T = T(X_1, \ldots, X_n)$):

$$T = I_{]-\infty,0]}(\overline{X}) \rightsquigarrow B(1, P[\overline{X} \leqslant 0])$$

estamos ya en condiciones de ver que T es insesgado para dicha función:

$$E[T] \overset{(*)}{=} P[\overline{X} \leqslant 0] \overset{\text{tipif.}}{=} P\left[Z \leqslant \frac{-\mu\sqrt{n}}{\sigma}\right] = \Phi\left(\frac{-\mu\sqrt{n}}{\sigma}\right)$$

donde en $(*)$ usamos que conocemos bien la esperanza de una distribución Bernoulli.

Ejercicio 2.4.2. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de $X \rightsquigarrow B(1,p)$ con $p \in]0,1[$ y sea $T = \sum_{i=1}^{n} X_i$.

a) Probar que si $k \in \mathbb{N}$ y $k \leqslant n$, el estadístico

$$\frac{T(T-1)\cdot \ldots \cdot (T-k+1)}{n(n-1)\cdot \ldots \cdot (n-k+1)}$$

es un estimador insesgado de p^k. ¿Es este estimador el UMVUE?.

b) Probar que si $k > n$, no existe ningún estimador insesgado para p^k.

c) ¿Puede afirmarse que $\frac{T}{n}\left(1 - \frac{T}{n}\right)^2$ es insesgado para $p(1-p)^2$?

Veamos cada apartado:

a) Sea $k \in \mathbb{N}$ con $k \leqslant n$, definimos:

$$h(T) = \frac{T(T-1)\cdot \ldots \cdot (T-k+1)}{n(n-1)\cdot \ldots \cdot (n-k+1)}$$

Veamos que $h(T)$ es insesgado para p^k. En primer lugar, observemos que por la reproductividad de la binomial $T \rightsquigarrow B(n,p)$:

$$E[h(T)] = \frac{E[T(T-1)\cdot\ldots\cdot(T-k+1)]}{n(n-1)\cdot\ldots\cdot(n-k+1)} = \frac{(n-k)!}{n!}E[T(T-1)\cdot\ldots\cdot(T-k+1)]$$

Calculamos ahora la esperanza:

$$\begin{aligned}
E[T(T-1)\cdot\ldots\cdot(T-k+1)] &= \sum_{t=0}^{n} t(t-1)\cdot\ldots\cdot(t-k+1)P[T=t] \\
&= \sum_{t=0}^{n} t(t-1)\cdot\ldots\cdot(t-k+1)\binom{n}{t}p^t(1-p)^{n-t} \\
&= \sum_{t=k}^{n} t(t-1)\cdot\ldots\cdot(t-k+1)\binom{n}{t}p^t(1-p)^{n-t} \\
&= \sum_{t=k}^{n} \frac{t!}{(t-k)!}\frac{n!}{t!(n-t)!}p^t(1-p)^{n-t} \\
&= \sum_{t=k}^{n} \frac{n!}{(t-k)!(n-t)!}p^t(1-p)^{n-t}
\end{aligned}$$

si desarrollamos ahora los primeros términos, observamos que:

$$\begin{aligned}
E[T(T-1)\cdot\ldots\cdot(T-k+1)] &= \sum_{t=k}^{n} \frac{n!}{(t-k)!(n-t)!}p^t(1-p)^{n-t} \\
&= \frac{n!}{(n-k)!}p^k(1-p)^{n-k} + \frac{n!}{(n-k-1)!}p^{k+1}(1-p)^{n-k-1} + \ldots
\end{aligned}$$

donde podemos ver que podemos sacar factor común de la sumatoria ciertos términos:

$$\begin{aligned}
E[T(T-1)\cdot\ldots\cdot(T-k+1)] &= \sum_{t=k}^{n} \frac{n!}{(t-k)!(n-t)!}p^t(1-p)^{n-t} \\
&= \frac{n!\cdot p^k}{(n-k)!}\sum_{t=k}^{n} p^{t-k}(1-p)^{n-t}\frac{(n-k)!}{(t-k)!(n-t)!} \\
&= \frac{n!\cdot p^k}{(n-k)!}\sum_{t=0}^{n-k} p^{t}(1-p)^{n-k-t}\frac{(n-k)!}{t!(n-k-t)!} \\
&= \frac{n!\cdot p^k}{(n-k)!}\sum_{t=0}^{n-k} \binom{n-k}{t}p^t(1-p)^{n-k-t} \\
&= \frac{n!\cdot p^k}{(n-k)!}\sum_{t=0}^{n-k} P[S=t] = \frac{n!\cdot p^k}{(n-k)!}P[0\leqslant S\leqslant n-k] \overset{(*)}{=} \frac{n!\cdot p^k}{(n-k)!}
\end{aligned}$$

para cierta variable aleatoria $S \rightsquigarrow B(n-k,p)$, donde en $(*)$ usamos que $P[0\leqslant S\leqslant n-k]=1$. Ahora, vemos que:

$$E[h(T)] = \frac{(n-k)!}{n!}\frac{n!}{(n-k)!}p^k = p^k$$

por lo que $h(T)$ es insesgado para p^k. Veamos ahora que $h(T)$ es un estimador. Para ello, observamos primero que:

$$h(T) = \frac{T}{n} \cdot \frac{T-1}{n-1} \cdot \ldots \cdot \frac{T-k+1}{n-k+1} = \prod_{j=0}^{k-1} \frac{T-j}{n-j}$$

Para cada $j \in \{0, \ldots, k-1\}$, observemos que $T \in \{0, \ldots, n\}$, por lo que $T - j \in \{-j, \ldots, n-j\}$ con $j < k \leqslant n$, de donde deducimos que:

$$\frac{T-j}{n-j} \in [-1, 1] \quad \Longrightarrow \quad \prod_{j=0}^{k-1} \frac{T-j}{n-j} \in [-1, 1]$$

Tenemos que ver finalmente que dicho producto es positivo, con lo que habremos probado que $h(T)$ es un estimador. Para ello, si el prodcuto no fuera positivo es por la existencia de $j \in \{0, \ldots, k-1\}$ de forma que $T - j < 0$, es decir, tenemos entonces que $T \in \{0, \ldots, j-1\}$ supuesto que $T = l \in \{0, \ldots, j-1\}$, tendremos entonces que:

$$h(T) = \prod_{j=0}^{k-1} \frac{T-j}{n-j} = \left(\prod_{j=0}^{l-1} \frac{T-j}{n-j} \right) \frac{T-l}{n-l} \left(\prod_{j=l+1}^{k-1} \frac{T-j}{n-j} \right) = 0$$

es decir, siempre que un término del producto sea negativo el producto entero se anula, por lo que siempre el producto es positivo, de donde $h(T) \in [0, 1]$, por lo que $h(T)$ es un estimador.

Finalmente, como:

$$E\left[(h(T))^2\right] = \sum_{t=0}^{n} (h(t))^2 P[T = t] < \infty$$

y teníamos que T era un estimador suficiente y completo (cuando vimos que $\{B(1, p) : p \in]0, 1[\}$ era una familia exponencial), tenemos entonces que $E[h(T)/T] = h(T)$ es el UMVUE de p^k.

b) Sea ahora $k > n$, veamos que no puede existir ningun estimador insesgado para p^k. Para ello, por reducción al absurdo, supongamos que $h(T)$ es un estimador insesgado para p^k, con lo que:

$$p^k = E[h(T)] = \sum_{t=0}^{n} h(t) P[T = t] = \sum_{t=0}^{n} h(t) \binom{n}{t} p^t (1-p)^{n-t}$$

de donde:

$$1 = \sum_{t=0}^{n} h(t) \binom{n}{t} p^{t-k} (1-p)^{n-t} \qquad \forall p \in]0, 1[$$

en particular, tomando $p \to 0$, como $t - k < 0$ para todo $t \in \{0, \ldots, k\}$, tenemos que:

$$1 = \lim_{p \to 0} \sum_{t=0}^{n} h(t) \binom{n}{t} p^{t-k} (1-p)^{n-t} = \infty$$

contradicción, con lo que para $k > n$ no puede existir un estimador insesgado para p^k. También podríamos haber justificado que son dos polinomios de distinto grado, para llegar a contradicción.

c) Buscamos ahora comprobar si $h(T) = \frac{T}{n}\left(1 - \frac{T}{n}\right)^2$ es insesgado para $p(1-p)^2$. Resulta que esto no puede afirmarse para todos los valores de n y p. Como un contraejemplo, para $n = 3$ resulta que tenemos:

$$\begin{aligned} E[h(T)] &= \sum_{t=0}^{3} \frac{t}{3}\left(1 - \frac{t}{3}\right)^2 \binom{3}{t} p^t (1-p)^{3-t} \overset{(*)}{=} \sum_{t=1}^{2} \frac{t}{3}\left(1 - \frac{t}{3}\right)^2 \binom{3}{t} p^t (1-p)^{3-t} \\ &= \frac{1}{3}\left(1 - \frac{1}{3}\right)^2 \binom{3}{1} p(1-p)^2 + \frac{2}{3}\left(1 - \frac{2}{3}\right)^2 \binom{3}{2} p^2 (1-p) \\ &= p(1-p)\left(\frac{1}{9}(1-p) + \frac{2}{9}p\right) = p(1-p)\frac{(1+p)}{9} \qquad \forall p \in]0,1[\end{aligned}$$

donde en $(*)$ hemos usado que la función $t \mapsto \frac{t}{n}\left(1 - \frac{t}{n}\right)^2$ evaluada en 0 y 3 es igual a cero. Observamos que:

$$E[h(T)] = p(1-p)\frac{(1+p)}{9} = p(1-p)^2 \iff \frac{1+p}{9} = 1 - p$$

resultado que no es cierto para todo $p \in]0,1[$ (basta considerar $p = 1/2$), por lo que dicho estimardor no es insesgado para $p(1-p)^2$.

Ejercicio 2.4.3. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable $X \rightsquigarrow \mathcal{P}(\lambda)$ con $\lambda \in \mathbb{R}^+$. Encontrar, si existe, el UMVUE para λ^s, siendo $s \in \mathbb{N}$ arbitrario.

Veamos que $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ es un estadístico suficiente y completo. Para ello, recordemos que $\{\mathcal{P}(\lambda) : \lambda > 0\}$ es una familia exponencial:

1. El espacio paramétrico es $\mathbb{R}^+ \subseteq \mathbb{R}$.
2. El espacio muestral es $\mathcal{X} = \mathbb{N} \cup \{0\}$, que no depende de λ.
3. Observamos que:

$$P_\lambda[X = x] = e^{-\lambda}\frac{\lambda^x}{x!} = exp\left[\ln\left(e^{-\lambda}\frac{\lambda^x}{x!}\right)\right] = exp\left(-\lambda + x\ln\lambda - \ln(x!)\right)$$

por lo que basta tomar:

$$Q(\lambda) = \ln\lambda, \qquad T(x) = x \qquad D(\lambda) = -\lambda, \qquad S(x) = -\ln(x!)$$

En consecuencia, por un Teorema visto en teoría, tenemos que el estadístico:

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$$

es suficiente y completo para λ. Observemos que por la reproductividad de la Poisson tenemos que (notando $T = T(X_1, \ldots, X_n)$):

$$T \rightsquigarrow \mathcal{P}\left(\sum_{i=1}^{n} \lambda\right) \equiv \mathcal{P}(n\lambda)$$

Ahora, para buscar el UMVUE, buscamos una función h medible de forma que:

$$\lambda^s = E[h(T)] = \sum_{t \in \mathbb{N} \cup \{0\}} h(t) P[T = t] = \sum_{t \in \mathbb{N} \cup \{0\}} h(t) e^{-n\lambda} \frac{(n\lambda)^t}{t!}$$

por lo que:

$$\lambda^s e^{n\lambda} = \sum_{t \in \mathbb{N} \cup \{0\}} h(t) \frac{(n\lambda)^t}{t!}$$

y si aplicamos el desarrollo en serie de la exponencial, obtenemos:

$$\lambda^s \sum_{t \in \mathbb{N} \cup \{0\}} \frac{(n\lambda)^t}{t!} = \lambda^s e^{n\lambda} = \sum_{t \in \mathbb{N} \cup \{0\}} h(t) \frac{(n\lambda)^t}{t!}$$

si desarrollamos cada uno de los términos:

$$\lambda^s + \lambda^{s+1} n + \frac{\lambda^{s+2} n^2}{2!} + \ldots = h(0) + h(1)(n\lambda) + \ldots + h(s) \frac{(n\lambda)^s}{s!} + \ldots$$

observamos que tomando:

$$h(0) = \ldots = h(s-1) = 0, \quad h(s) = \frac{s!}{n^s}$$
$$h(s+1) = \frac{(s+1)!}{n^s}, \quad \ldots \quad h(s+k) = \frac{(s+k)!}{n^s k!}$$

es decir:

$$h(T) = \begin{cases} 0 & \text{si } T < s \\ \dfrac{T!}{n^s (T-s)!} & \text{si } T \geqslant s \end{cases}$$

tenemos que $h(T)$ es insesgado para λ^s. Es claro además que $h(t) \in \mathbb{R}^+$ para cualquier valor de t, con lo que $h(T)$ es un estimador de λ^s. Finalmente, observemos que:

$$E\left[(h(T))^2\right] = \sum_{t \in \mathbb{N} \cup \{0\}} (h(t))^2 P[T = t] = \sum_{t \geqslant s} \left(\frac{t!}{n^s (t-s)!}\right)^2 e^{-n\lambda} \frac{(n\lambda)^t}{t!}$$
$$= \frac{1}{n^s e^{n\lambda}} \sum_{t \geqslant s} \frac{(n\lambda)^t t!}{((t-s)!)^2}$$

como:

$$\frac{\dfrac{(n\lambda)^{t+1}(t+1)!}{((t+1-s)!)^2}}{\dfrac{(n\lambda)^t t!}{((t-s)!)^2}} = \frac{(n\lambda)^{t+1}(t+1)!((t-s)!)^2}{(n\lambda)^t t!((t+1-s)!)^2} = \frac{n\lambda(t+1)}{(t+1-s)^2} \to 0 < 1$$

por el Criterio del cociente, tenemos que:

$$\sum_{t\geqslant s}\frac{(n\lambda)^t t!}{((t-s)!)^2}<\infty \Longrightarrow E\left[(h(T))^2\right]=\frac{1}{n^s e^{n\lambda}}\sum_{t\geqslant s}\frac{(n\lambda)^t t!}{((t-s)!)^2}<\infty$$

en consecuencia, tenemos que $h(T)$ es un estimador insesgado para λ^s y de momento de segundo orden finito y es función de un estadístico suficiente y completo, con lo que el Teorema de Lehmann-Scheffé nos dice que:

$$E[h(T)/T]=h(T)$$

es un UMVUE para λ^s.

Ejercicio 2.4.4. Sea $(X_1,\ldots,X_n)$ una muestra aleatoria simple de una variable con distribución uniforme discreta en los puntos $\{1,\ldots,N\}$, siendo N un número natural arbitrario. Encontrar el UMVUE para N.

En el Ejercicio 2.3.5 vimos que $T(X_1,\ldots,X_n)=X_{(n)}$ era un estadístico suficiente y completo. Si notamos $T=T(X_1,\ldots,X_n)$, tenemos que:

$F_T(t)=(F_X(t))^n \Longrightarrow P[T=t]=P[T\leqslant t]-P[T\leqslant t-1]=(F_X(t))^n-(F_X(t-1))^n$

como $F_X(t)=\frac{t}{N}$ para $t\in\{1,\ldots,N\}$, tenemos:

$$P[T=t]=(F_X(t))^n-(F_X(t-1))^n=\frac{t^n-(t-1)^n}{N^n}$$

Buscamos ahora una función h medible de forma que:

$$N=E[h(T)]=\sum_{t=1}^{N}h(t)P[T=t]=\sum_{t=1}^{N}h(t)\frac{t^n-(t-1)^n}{N^n}$$

Si desarrollamos la suma para $h(t)=1$, observamos un comportamiento telescópico:

$$\begin{aligned}\frac{1}{N^n}\sum_{t=1}^{N}(t^n-(t-1)^n)&=\frac{1}{N^n}(N^n-(N-1)^n+(N-1)^n-(N-2)^n+\ldots+2^n-1+1-0)\\&=\frac{N^n}{N^n}=1\end{aligned}$$

No hemos obteniedo lo que queríamos, puesto que queríamos que el resultado de la suma fuera N^{n+1}. Si tomamos sin embargo:

$$h(t)=\frac{t^{n+1}-(t-1)^{n+1}}{t^n-(t-1)^n}$$

tenemos entonces que:

$$\begin{aligned}E[h(T)]&=\frac{1}{N^n}\sum_{t=1}^{N}h(t)(t^n-(t-1)^n)=\frac{1}{N^n}\sum_{t=1}^{N}t^{n+1}-(t-1)^{n+1}\\&=\frac{1}{N^n}(N^{n+1}-(N-1)^{n+1}+(N-1)^{n+1}+\ldots+2^{n+1}-1+1-0)\\&=\frac{N^{n+1}}{N^n}=N\end{aligned}$$

Por lo que el estadístico $h(T)$ es insesgado para N. Nos falta comprobar si es un estimador y si tiene momento de segundo orden finito:

- No es estimador, puesto que $h(\mathbb{N}) \nsubseteq \mathbb{N}$. Sin embargo, podemos realizar aproximaciones y quedarnos con el natural más próximo. Es decir, no va a haber un UMVUE pero podemos quedarnos con este estadístico, que es aquello que podemos conseguir más próximo a un UMVUE.
- Para el momento de segundo orden, observamos que:

$$E[(h(T))^2] = \frac{1}{N^n}\sum_{t=1}^{N}(h(t))^2(t^n-(t-1)^n) < \infty$$

al ser una suma finita.

Ejercicio 2.4.5. Sea $(X_1,\dots,X_n)$ una muestra aleatoria simple de una variable aleatoria X cuya función de densidad es de la forma

$$f_\theta(x) = \frac{1}{2\sqrt{x\theta}}, \quad 0 < x < \theta$$

Calcular, si existe, el UMVUE para θ.

En primer lugar, buscamos un estadístico suficiente y completo:

$$f_\theta(x_1,\dots,x_n) \overset{\text{iid.}}{=} \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n}\frac{1}{2\sqrt{x_i\theta}} = \frac{1}{\left(2\sqrt{\theta}\right)^n}\prod_{i=1}^{n}\frac{1}{\sqrt{x_i}} \qquad x_i \in]0,\theta[$$

Si consideramos $0 < X_{(1)} \leqslant X_{(n)} < \theta$, podemos escribir:

$$f_\theta(x_1,\dots,x_n) = \frac{I_{]0,+\infty[}(X_{(1)}-0)\cdot I_{]-\infty,0[}(X_{(n)}-\theta)}{\left(2\sqrt{\theta}\right)^n}\prod_{i=1}^{n}\frac{1}{\sqrt{x_i}}$$

Tomando:

$$h(x_1,\dots,x_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{x_i}}, \qquad T(X_1,\dots,X_n) = X_{(n)}$$
$$g_\theta(t) = \frac{I_{]0,+\infty[}(X_{(1)}-0)\cdot I_{]-\infty,0[}(t-\theta)}{\left(2\sqrt{\theta}\right)^n}$$

por el Teorema de factorización de Neymann-Fisher tenemos que el estadístico $X_{(n)}$ es suficiente para θ. Notando $T = X_{(n)}$ para abreviar, calculamos la distribución de T:

$$F_T(t) = (F_X(t))^n \quad \Longrightarrow \quad f_T(t) = n(F_X(t))^{n-1}f_\theta(t)$$

Calculamos la distribución de X:

$$F_X(t) = \int_0^t \frac{1}{2\sqrt{x\theta}}\,dx = \frac{1}{2\sqrt{\theta}}\int_0^t \frac{1}{\sqrt{x}}\,dx = \sqrt{\frac{t}{\theta}} \qquad t \in]0,\theta[$$

Por lo que:

$$f_T(t) = n(F_X(t))^{n-1}f_\theta(t) = n\left(\sqrt{\frac{t}{\theta}}\right)^{n-1}\frac{1}{2\sqrt{t\theta}} = \frac{n\left(\sqrt{t}\right)^{n-2}}{2\left(\sqrt{\theta}\right)^n} \qquad t \in]0,\theta[$$

Sea ahora h una función medible de forma que:

$$0 = E[h(T)] = \int_0^\theta h(t) f_T(t)\ dt = \frac{n}{2\Big(\sqrt{\theta}\Big)^n} \int_0^\theta h(t)\Big(\sqrt{t}\Big)^{n-2}\ dt \qquad \forall \theta \in \mathbb{R}^+$$

tenemos entonces que:

$$\int_0^\theta h(t)\Big(\sqrt{t}\Big)^{n-2}\ dt = 0 \qquad \forall \theta \in \mathbb{R}^+$$

Sea ahora $H(t)$ una primitiva de $h(t)\big(\sqrt{t}\big)^{n-2}$, tenemos entonces que:

$$H(\theta) - H(0) = 0 \qquad \forall \theta \in \mathbb{R}^+$$

por lo que derivando respecto a θ:

$$h(\theta)\Big(\sqrt{\theta}\Big)^{n-2} = 0 \qquad \forall \theta \in \mathbb{R}^+ \quad \Longrightarrow \quad h(\theta) = 0 \qquad \forall \theta \in \mathbb{R}^+$$

En conclusión, tenemos que:

$$\mathbb{R}^+ \subseteq \{t : h(t) = 0\}$$

por lo que:

$$1 \geqslant P[h(T) = 0] \geqslant P[T \in \mathbb{R}^+] = 1 \quad \Longrightarrow \quad P[h(T) = 0] = 1$$

lo que demustra que T es un estadístico completo. Ahora, tratamos de buscar un estimador insesgado para θ, que será nuestro candidato a UMVUE. Buscamos una función medible h de forma que:

$$\theta = E[h(T)] = \frac{n}{2\Big(\sqrt{\theta}\Big)^n} \int_0^\theta h(t)\Big(\sqrt{t}\Big)^{n-2}\ dt$$

Observemos que:

$$\int_0^\theta \Big(\sqrt{t}\Big)^n\ dt = \frac{2}{n+2}\left[\Big(\sqrt{t}\Big)^{n+2}\right]_0^\theta = \frac{2}{n+2}\Big(\sqrt{\theta}\Big)^{n+2} = \frac{2}{n+2}\Big(\sqrt{\theta}\Big)^n \theta$$

Por lo que si tomamos:

$$h(t) = \frac{n+2}{n} \cdot t$$

tenemos que:

$$\begin{aligned} E[h(T)] &= \int_0^\theta \frac{n+2}{n} t \frac{n}{2} \frac{\big(\sqrt{t}\big)^{n-2}}{\Big(\sqrt{\theta}\Big)^n}\ dt = \frac{n+2}{2\Big(\sqrt{\theta}\Big)^n} \int_0^\theta t\Big(\sqrt{t}\Big)^{n-2}\ dt \\ &= \frac{n+2}{2\Big(\sqrt{\theta}\Big)^n} \int_0^\theta \Big(\sqrt{t}\Big)^n\ dt = \frac{n+2}{2\Big(\sqrt{\theta}\Big)^n} \frac{2}{n+2}\Big(\sqrt{\theta}\Big)^n \theta = \theta \qquad \forall \theta \in \mathbb{R}^+ \end{aligned}$$

Por lo que $h(T)$ es insesgado para θ. Ahora, si $t \in \mathbb{R}^+$, tendremos entonces que:

$$h(t) = \frac{n+2}{n} \cdot t \geqslant t \in \mathbb{R}^+ \quad \Longrightarrow \quad h(t) \in \mathbb{R}^+$$

por lo que $h(T)$ es además un estimador. Si calculamos ahora:

$$E\left[(h(T))^2\right] = \frac{n}{2\left(\sqrt{\theta}\right)^n} \int_0^\theta (h(t))^2 (\sqrt{t})^{n-2}\, dt$$

Tenemos la integral en un compacto de una función continua, por lo que sabemos que $E[(h(T))^2] < \infty$. En conclusión, tenemos por el Teorema de Lehmann-Scheffé que: $E[h(T)/T] = h(T)$ es el UMVUE.

Ejercicio 2.4.6. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X con función de densidad

$$f_\theta(x) = \frac{\theta}{x^2}, \quad x > \theta > 0$$

Calcular, si existen, los UMVUE para θ y para $1/\theta$.

Calculamos en primer lugar un estadístico suficiente y completo para θ:

$$f_\theta(x_1, \ldots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{\theta \cdot I_{]0,+\infty[}(X_{(1)} - \theta)}{x_i^2} = \theta^n \cdot I_{]0,+\infty[}(X_{(1)} - \theta) \prod_{i=1}^n \frac{1}{x_i^2}$$

Tomando:

$$h(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{x_i^2}, \qquad T(X_1, \ldots, X_n) = X_{(1)}$$
$$g_\theta(t) = \theta^n \cdot I_{]0,+\infty[}(t - \theta)$$

Podemos aplicar el Teorema de factorización de Neymann-Fisher, obteniendo que T es suficiente para θ. Notando $T = X_{(1)}$ para abreviar, calculamos la distribución de T:

$$F_T(t) = 1 - (1 - F_X(t))^n \quad \Longrightarrow \quad f_T(t) = n(1 - F_X(t))^{n-1} f_\theta(t)$$

Si calculamos:

$$F_X(t) = \int_\theta^t \frac{\theta}{x^2}\, dx = \theta \int_\theta^t \frac{1}{x^2}\, dx = \theta \left[\frac{-1}{x}\right]_\theta^t = \theta \left(\frac{-1}{t} + \frac{1}{\theta}\right) = \frac{t-\theta}{t} \qquad t > \theta$$

Tenemos que:

$$f_T(t) = n\left(1 - \frac{t-\theta}{t}\right)^{n-1} \frac{\theta}{t^2} = n \cdot \frac{\theta^n}{t^{n+1}}$$

Sea h una función medible de forma que:

$$0 = E[h(T)] = \int_\theta^{+\infty} h(t) f_T(t)\, dt = \lim_{n\to\infty} \int_\theta^n \underbrace{h(t) \cdot n \cdot \frac{\theta^n}{t^{n+1}}}_{(*)}\, dt \qquad \forall \theta \in \mathbb{R}^+$$

Sea $H(t)$ una primitiva de $(*)$, tenemos entonces que:

$$0 = \lim_{n\to\infty} H(n) - H(\theta) \qquad \forall\theta \in \mathbb{R}^+$$

con lo que derivando respecto θ:

$$0 = h(\theta)\cdot n \cdot \frac{\theta^n}{\theta^{n+1}} = h(\theta)\cdot\frac{n}{\theta} \quad \Longrightarrow \quad h(\theta) = 0 \qquad \forall\theta\in\mathbb{R}^+$$

Por lo que:

$$\mathbb{R}^+ \subseteq \{t : h(t) = 0\}$$

de donde:

$$1 \geqslant P[h(T)=0] \geqslant P[T\in\mathbb{R}^+] = 1 \quad \Longrightarrow \quad P[h(T)=0] = 1$$

Luego tenemos que T es completo. Si calculamos la esperanza de T:

$$\begin{aligned} E[T] &= \int_\theta^\infty t\cdot n\cdot\frac{\theta^n}{t^{n+1}}\,dt = \int_\theta^{+\infty}\frac{n\theta^n}{t^n}\,dt = n\theta^n\int_\theta^{+\infty}\frac{1}{t^n}\,dt = \frac{-n\theta^n}{n-1}\left[\frac{1}{t^{n-1}}\right]_\theta^\infty \\ &= \frac{n\theta^n}{(n-1)\theta^{n-1}} = \frac{n\theta}{n-1} \end{aligned}$$

Por lo que tomando:

$$h(t) = \frac{n-1}{n}\cdot t$$

tenemos que:

$$E[h(T)] = \frac{n-1}{n}E[T] = \theta$$

Es decir, $h(T)$ es un estadístico insesgado para θ. Veamos que es estimador y que tiene momento de segundo orden:

- Si $t\in\mathbb{R}^+$, tenemos entonces que $h(t)\in\mathbb{R}_0^+$ para $n>1$, por lo que $h(T)$ es estimador.
- Calculamos para $n>2$:

$$\begin{aligned} E[(h(T))^2] &= \int_\theta^{+\infty}(h(t))^2\frac{n\theta^n}{t^{n+1}}\,dt = \left(\frac{n-1}{n}\right)^2\int_\theta^{+\infty}\frac{n\theta^n t^2}{t^{n+1}}\,dt \\ &= n\theta^n\left(\frac{n-1}{n}\right)^2\int_\theta^{+\infty}\frac{1}{t^{n-1}}\,dt = n\theta^n\left(\frac{n-1}{n}\right)^2\left[\frac{1}{(-n+2)t^{n-2}}\right]_\theta^{+\infty} \\ &= n\theta^n\left(\frac{n-1}{n}\right)^2\frac{1}{(n-2)\theta^{n-2}} = \left(\frac{n-1}{n}\right)^2\frac{n}{(n-2)}\theta^2 < \infty \end{aligned}$$

Por lo que $h(T) = \frac{n-1}{n}\cdot t$ es UMVUE para θ. Si buscamos ahora un UMVUE para $1/\theta$, observamos que tomando $h(t) = \frac{n+1}{n}\cdot\frac{1}{t}$, tenemos que:

$$\begin{aligned} E[h(T)] &= \int_\theta^{+\infty}h(t)\frac{n\theta^n}{t^{n+1}}\,dt = (n+1)\theta^n\int_\theta^{+\infty}\frac{1}{t^{n+2}}\,dt = (n+1)\theta^n\left[\frac{1}{-(n+1)t^{n+1}}\right]_\theta^{+\infty} \\ &= \frac{\theta^n}{\theta^{n+1}} = \frac{1}{\theta} \end{aligned}$$

por lo que $h(T)$ es insesgado para T. Observamos además que:

- Si $t \in \mathbb{R}^+$, entonces $h(t) \in \mathbb{R}^+$, por lo que $h(T)$ es un estimador.
- Calculamos si $n+3>1$:

$$E[(h(T))^2] = \int_\theta^{+\infty} (h(t))^2 \frac{n\theta^n}{t^{n+1}}\, dt = \frac{\theta^n (n+1)^2}{n} \int_\theta^{+\infty} \frac{1}{t^{n+3}}\, dt$$
$$= \frac{\theta^n (n+1)^2}{n} \left[\frac{1}{-(n+2)t^{n+2}}\right]_\theta^{+\infty} = \frac{(n+1)^2}{n(n+2)\theta^2} < \infty$$

Aplicando el Teorema de Lehmann-Scheffé, tenemos que $E[h(T)/T] = h(T)$ es UMVUE para $1/\theta$.

Ejercicio 2.4.7. Sea $X \rightsquigarrow P_\theta$ siendo P_θ una distribución con función de densidad

$$f_\theta(x) = e^{\theta - x}, \quad x \geqslant \theta$$

Dada una muestra aleatoria simple de tamaño arbitrario, encontrar los UMVUE de θ y de e^θ.

Calculamos en primer lugar un estadístico suficiente y completo para θ:

$$f_\theta(x_1, \ldots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n e^{\theta - x_i} \cdot I_{[0,+\infty[}(X_{(1)} - \theta) = I_{[0,+\infty[}(X_{(1)} - \theta) e^{n\theta} \prod_{i=1}^n e^{-x_i}$$

Tomando:

$$h(x_1, \ldots, x_n) = \prod_{i=1}^n e^{-x_i}, \qquad T(X_1, \ldots, X_n) = X_{(1)}$$
$$g_\theta(t) = I_{[0,+\infty[}(t - \theta) \cdot e^{n\theta}$$

Tenemos por el Teorema de factorización de Neymann-Fisher que $T = X_{(1)}$ es suficiente para θ. Calculamos la distribución de T:

$$F_T(t) = 1 - (1 - F_X(t))^n \quad \Longrightarrow \quad f_T(t) = n(1 - F_X(t))^{n-1} f_\theta(t)$$

Si calculamos:

$$F_X(t) = \int_\theta^t e^{\theta - x}\, dx = e^\theta \int_\theta^t e^{-x}\, dx = e^\theta [-e^{-x}]_\theta^t = e^\theta \left(e^{-\theta} - e^{-t}\right) = 1 - e^{\theta - t}, \quad t \geqslant \theta$$

Tenemos entonces que:

$$f_T(t) = n\left(e^{\theta - t}\right)^{n-1} e^{\theta - t} = n\left(e^{\theta - t}\right)^n$$

Sea h una función medible de forma que:

$$0 = E[h(T)] = \int_\theta^{+\infty} h(t) n \left(e^{\theta - t}\right)^n\, dt = n e^{n\theta} \int_\theta^{+\infty} h(t) e^{-nt}\, dt \qquad \forall \theta \in \mathbb{R}$$

Tenemos entonces que:

$$0 = \lim_{n \to \infty} \int_\theta^n h(t) e^{-nt}\, dt \qquad \forall \theta \in \mathbb{R}$$

Sea $H(t)$ una primiriva de $h(t)e^{-nt}$, tenemos que:

$$0 = \lim_{n\to\infty} H(n) - H(\theta) \qquad \forall\theta \in \mathbb{R}$$

derivando respecto θ obtenemos:

$$0 = -h(\theta)e^{-n\theta} \quad \Longrightarrow \quad h(\theta) = 0 \qquad \forall\theta \in \mathbb{R}$$

Por lo que:

$$\mathbb{R} \subseteq \{t : h(t) = 0\}$$

de donde deducimos que T es completo:

$$1 \geqslant P[h(t) = 0] \geqslant P[T \in \mathbb{R}] = 1 \quad \Longrightarrow \quad P[h(t) = 0] = 1$$

Buscamos ahora un estadístico insesgado para θ, es decir, buscamos una función h medible de forma que:

$$\theta = E[h(T)] = \int_\theta^{+\infty} h(t)n(e^{\theta-t})^n \, dt = ne^{n\theta} \int_\theta^{+\infty} h(t)e^{-nt} \, dt$$

Sea $H(t)$ una primitiva de $h(t)e^{-nt}$, tenemos entonces que:

$$\theta = ne^{n\theta}\left(\lim_{n\to\infty} H(n) - H(\theta)\right) \Longrightarrow \frac{\theta}{ne^{n\theta}} = \left(\lim_{n\to\infty} H(n) - H(\theta)\right)$$

y si derivamos a ambos lados:

$$\frac{e^{-n\theta}}{n} - \theta e^{-n\theta} = -h(\theta)e^{-n\theta} \quad \Longrightarrow \quad h(\theta) = \theta - \frac{1}{n}$$

Por lo que si tomamos:

$$h(T) = T - \frac{1}{n}$$

Tenemos que $E[T] = \theta$, con lo que $h(T)$ es insesgado para θ. Además, se trata de un estimador (puesto que el espacio paramétrico es $\mathbb{R}$). Finalmente, vemos que:

$$E[(h(T))^2] = \int_\theta^{+\infty} (h(t))^2 n\big(e^{\theta-t}\big)^n \, dt = ne^{n\theta} \int_\theta^{+\infty} \left(t^2 + \frac{1}{n^2} - \frac{2t}{n}\right) e^{-nt} \, dt$$

calculamos la integral a parte:

$$\begin{aligned}\int_\theta^{+\infty} \left(t^2 + \frac{1}{n^2} - \frac{2t}{n}\right) e^{-nt} \, dt &= \left\{ \begin{array}{ll} u = (t - 1/n)^2 & du = 2t - 2/n \\ dv = e^{-nt} & v = -e^{-nt}/n \end{array} \right\} \\ &= \left[\frac{-(t - 1/n)^2 e^{-nt}}{n}\right]_\theta^{+\infty} + \frac{1}{n}\int_\theta^{+\infty} (2t - 2/n)e^{-nt} \, dt\end{aligned}$$

tenemos que:

$$\left[\frac{-(t - 1/n)^2 e^{-nt}}{n}\right]_\theta^{+\infty} < \infty$$

y volvemos a calcular a parte:

$$\int_\theta^{+\infty} (2t - 2/n)e^{-nt}\, dt = \left\{ \begin{array}{ll} u = 2t - 2/n & du = 2 \\ dv = e^{-nt} & v = -e^{-nt}/n \end{array} \right\}$$
$$= \left[\frac{-(2t - 2/n)e^{-nt}}{n} \right]_\theta^{+\infty} + \frac{2}{n} \int_\theta^{+\infty} e^{-nt}\, dt$$

tenemos que:

$$\left[\frac{-(2t - 2/n)e^{-nt}}{n} \right]_\theta^{+\infty} < \infty$$

y también que:

$$\frac{2}{n} \int_\theta^{+\infty} e^{-nt}\, dt = \frac{2}{n} \left[\frac{-e^{-nt}}{n} \right]_\theta^{+\infty} < \infty$$

En definitiva, $E[(h(T))^2] < \infty$, por lo que aplicando el Teorema de Lehmann-Scheffé, obtenemos que $E[h(T)/T] = h(T)$ es UMVUE.

Si buscamos ahora un estadístico insesgado para e^θ, sea h una función medible de forma que:

$$e^\theta = E[h(T)] = \int_\theta^{+\infty} h(t) n (e^{\theta - t})^n\, dt = n e^{n\theta} \int_\theta^{+\infty} h(t) e^{-nt}\, dt$$

Sea $H(t)$ una primitiva de $h(t)e^{-nt}$ tenemos entonces que:

$$e^\theta = n e^{n\theta} \left(\lim_{n\to\infty} H(n) - H(\theta) \right) \Longrightarrow \frac{e^{\theta(1-n)}}{n} = \lim_{n\to\infty} H(n) - H(\theta)$$

y si derivamos a ambos lados:

$$\frac{(1-n)e^{\theta(1-n)}}{n} = -h(\theta)e^{-n\theta} \Longrightarrow h(t) = \frac{n-1}{n} e^t$$

Obtenemos que claramente $h(T)$ es estimador, así como que:

$$E[(h(T))^2] = \int_\theta^{+\infty} \left(\frac{n-1}{n} \right)^2 e^{2t} n e^{n\theta - nt}\, dt = \left(\frac{n-1}{n} \right)^2 n e^{n\theta} \int_\theta^{+\infty} e^{(2-n)t}\, dt$$
$$= \frac{1}{2-n} \left[e^{(2-n)t} \right]_\theta^\infty = \frac{e^{(2-n)\theta}}{n-2} < \infty \qquad \text{si } n > 2$$

Por lo que aplicando el Teorema de Lehmann-Scheffé, obtenemos que $E[h(T)/T] = h(T)$ es UMVUE para e^θ.

Ejercicios de estimadores eficientes

Ejercicio 2.4.8. Sea X la variable que describe el número de fracasos antes del primer éxito en una sucesión de pruebas de Bernoulli con probabilidad de éxito $\theta \in]0, 1[$, y sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de X.

a) Probar que la familia de distribuciones de X es regular y calcular la función de información asociada a la muestra.

Según el enunciado, "X modela el número de fracasos antes del primer éxito en una sucesión de pruebas de Bernoulli con probabilidad de éxito $\theta \in]0,1[$", por lo que X sigue una distribución geométrica con probabilidad de éxito θ, $X \rightsquigarrow G(\theta)$, $\theta \in]0,1[$. Para ver que la familia de distribuciones de X es regular:

i) El espacio paramétrico es $\Theta =]0,1[$, un intervalo abierto de $\mathbb{R}$.

ii) El espacio muestral es $\mathcal{X} = \mathbb{N} \cup \{0\}$, que no depende de θ.

iii) Para comprobar la tercera condición, recordamos que la función masa de probabilidad viene dada por:

$$P_\theta[X = x] = (1-\theta)^x \theta$$

que es derivable respecto θ. Calculamos la derivada del logaritmo de la función masa de probabilidad:

$$\ln P_\theta[X = x] = \ln\left((1-\theta)^x\theta\right) = x\ln(1-\theta) + \ln\theta$$
$$\frac{\partial \ln P_\theta[X = x]}{\partial\theta} = \frac{-x}{1-\theta} + \frac{1}{\theta} = \frac{1-\theta-x\theta}{\theta(1-\theta)}$$

Y si calculamos su esperanza:

$$E\left[\frac{1-\theta-X\theta}{\theta(1-\theta)}\right] = \frac{1-\theta-\theta E[X]}{\theta(1-\theta)} \overset{(*)}{=} \frac{1-\theta-(1-\theta)}{\theta(1-\theta)} = 0$$

donde en $(*)$ usamos que $E[X] = \frac{1-\theta}{\theta}$. En definitiva, se cumple también la tercera condición de las familias exponenciales.

Calculamos ahora la función de información asociada a la muestra:

$$I_X(\theta) = Var\left(\frac{1-\theta-X\theta}{\theta(1-\theta)}\right) = \frac{\theta^2}{\theta^2(1-\theta)^2}Var(X)$$
$$\overset{(*)}{=} \frac{1-\theta}{(1-\theta)^2\theta^2} = \frac{1}{\theta^2(1-\theta)}$$

donde en $(*)$ hemos usado que $Var(X) = \frac{1-\theta}{\theta^2}$.

b) Especificar la clase de funciones paramétricas que admiten estimadores eficientes y los correspondientes estimadores.

Una vez que sabemos que la familia es exponencial y que $0 < I_X(\theta) < \infty$, lo que hacemos es buscar las funciones $a(\theta)$ y $g(\theta)$ que se usan en el enunciado del Teorema de caracterización de los estimadores eficientes. Para ello:

$$\frac{\partial \ln P_\theta[X_1 = x_1, \ldots, X_n = x_n]}{\partial\theta} = \sum_{i=1}^n \frac{\partial \ln P_\theta[X = x_i]}{\partial\theta} = \sum_{i=1}^n \left(\frac{1-\theta-x_i\theta}{\theta(1-\theta)}\right)$$
$$= \frac{n(1-\theta) - \theta\sum_{i=1}^n x_i}{\theta(1-\theta)} = \frac{\sum_{i=1}^n x_i - \frac{n(1-\theta)}{\theta}}{\theta - 1}$$

Por lo que tomando:

$$T(X_1,\dots,X_n)=\sum_{i=1}^{n}X_i,\qquad g(\theta)=\frac{n(1-\theta)}{\theta},\qquad a(\theta)=\frac{1}{\theta-1}$$

Tenemos que $a(\theta)\neq 0\quad \forall\theta\in\,]0,1[$, que g es derivable, con:

$$g'(\theta)=\frac{-n}{\theta^2}\neq 0\qquad \forall\theta\in\,]0,1[$$

Además, observamos que:

$$a(\theta)g'(\theta)=\frac{-n}{\theta^2(\theta-1)}=\frac{n}{\theta^2(1-\theta)}=nI_X(\theta)=I_{(X_1,\dots,X_n)}(\theta)\qquad \forall\theta\in\,]0,1[$$

Así como que T es un estimador, pues:

$$T(\mathcal{X}^n)=[0,n]\cap\mathbb{N}\subset g(]0,1[)$$

Finalmente, por un Corolario visto en teoría sabemos que las únicas funciones paramétricas que admiten estimadores eficientes son de la forma:

$$a\cdot\frac{n(1-\theta)}{\theta}+b,\qquad a,b\in\mathbb{R},\quad a\neq 0$$

y que sus estimadores eficientes son de la forma:

$$a\cdot\sum_{i=1}^{n}X_i+b$$

c) Calcular la varianza de cada estimador eficiente y comprobar que coincide con las correspondiente cota de Fréchet-Cramér-Rao.

Si calculamos ahora la varianza de $T=a\cdot\sum_{i=1}^{n}X_i$ y la cota de Fréchet-Cramér-Rao:

$$\begin{aligned}Var(T)&=Var\left(a\cdot\sum_{i=1}^{n}X_i+b\right)\overset{\text{indep.}}{=}a^2\sum_{i=1}^{n}Var(X_i)=a^2\sum_{i=1}^{n}\left(\frac{1-\theta}{\theta^2}\right)\\&=\frac{a^2n(1-\theta)}{\theta^2}\end{aligned}$$

$$\frac{(ag'(\theta))^2}{I_{(X_1,\dots,X_n)}(\theta)}=\frac{\left(\frac{-an}{\theta^2}\right)^2}{\frac{n}{\theta^2(1-\theta)}}=\frac{a^2n(1-\theta)}{\theta^2}$$

d) Calcular, si existen, los UMVUE para $P_\theta[X=0]$ y para $E_\theta[X]$ y decir si son eficientes.

- Para $P_\theta[X=0]$:

$$P_\theta[X=0]=(1-\theta)^0\theta=\theta$$

Como ya hemos demostrado que $T = \sum_{i=1}^{n} X_i$ es un estimador eficiente para $g(\theta) = \frac{n(1-\theta)}{\theta}$, tenemos automáticamente que T es un estadístico suficiente. Tomando ahora $Q(\theta)$ como una primitiva de $a(\theta)$, por ejemplo $Q(\theta) = \ln(|\theta - 1|)$, como:

$$Q(]0,1[) =]-\infty, 0[$$

claramente contiene un abierto de $\mathbb{R}$ concluimos que T es completo. Buscamos ahora una función h medible de forma que $h(T)$ sea insesgado para θ, sabiendo que $T \rightsquigarrow BN(n,\theta)$, con función masa de probabilidad:

$$P_\theta[T = t] = \binom{t+n-1}{t}(1-\theta)^t\theta^n$$

Buscamos pues una función h medible de forma que:

$$E[h(T)] = \sum_{t=0}^{\infty} h(t)\frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^t\theta^n = \theta$$

Es decir, que cumpla (dividiendo todo entre θ):

$$E[h(T)] = \sum_{t=0}^{\infty} h(t)\frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^t\theta^{n-1} = 1$$

Notemos que si $Y \rightsquigarrow B(n-1,\theta)$, tendremos entonces que:

$$P[Y = t] = \binom{t+n-2}{t}(1-\theta)^t\theta^{n-1} = \frac{(t+n-2)!}{t!(n-2)!}(1-\theta)^t\theta^{n-1}$$

Por lo que ha de ser (por ser una función masa de probabilidad):

$$1 = \sum_{t=0}^{\infty} \frac{(t+n-2)!}{t!(n-2)!}(1-\theta)^t\theta^{n-1}$$

Por tanto, si tomamos:

$$h(t) = \frac{n-1}{t+n-1}$$

obtenemos lo buscado. Observamos ahora que:

- $h(T)$ es un estimador, puesto que $h(T) \in]0,1[$.
- Calculamos su momento de segundo orden:

$$\begin{aligned} E[(h(t))^2] &= \sum_{t=0}^{\infty} (h(t))^2 \frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^t\theta^n \\ &= \sum_{t=0}^{\infty} \left(\frac{n-1}{t+n-1}\right)^2 \frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^t\theta^n \end{aligned}$$

  Como tenemos que $h(t) \in]0,1[$ para todo $t \in \mathbb{N}$, tenemos que $(h(t))^2 \in]0,1[$, por lo que cada sumando de esta nueva serie es menor o igual que cada sumando de la serie $E[h(t)]$, que era convergente, por lo que por el criterio de comparación esta nueva serie también será convergente, $E[(h(t))^2] < \infty$.

En definitiva, el UMVUE para $P_\theta[X=0]=\theta$ es $h(T)$, por el Teorema de Lehmann-Scheffé.

- Para $E_\theta[X]$:

$$E[X]=\frac{1-\theta}{\theta}$$

Por el apartado anterior tenemos ya que $T=\sum_{i=1}^{n}X_i$ es un estadístico suficiente y completo. Buscamos ahora una función h medible tal que $h(T)$ sea insesgado para $\frac{1-\theta}{\theta}$, es decir:

$$E[h(T)]=\sum_{t=0}^{\infty}h(t)\frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^t\theta^n=\frac{1-\theta}{\theta}$$

Escribiendo los términos en el mismo lado para igualar a 1:

$$\sum_{t=0}^{\infty}h(t)\frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^{t-1}\theta^{n+1}=1 \tag{2.1}$$

Si $Y\rightsquigarrow BN(n+1,\theta)$, entonces:

$$P[Y=t]=\binom{n+t}{t}(1-\theta)^t\theta^{n+1}=\frac{(t+n)!}{t!n!}(1-\theta)^t\theta^{n+1}$$

De donde sabemos que:

$$\sum_{t=0}^{\infty}\frac{(t+n)!}{t!n!}(1-\theta)^t\theta^{n+1}=1$$

Si sacamos fuera de la suma de (2.1) el primer sumando, tenemos:

$$h(0)\frac{\theta}{1-\theta}+\sum_{t=1}^{\infty}h(t)\frac{(t+n-1)!}{t!(n-1)!}(1-\theta)^{t-1}\theta^{n+1}$$

y podemos comenzar a numerar la serie desde 0:

$$h(0)\frac{\theta}{1-\theta}+\sum_{t=0}^{\infty}h(t+1)\frac{(t+n)!}{(t+1)!(n-1)!}(1-\theta)^t\theta^{n+1}$$

De esta forma, si tomamos:

$$h(t)=\frac{t}{n}$$

tendremos en particular que $h(0)=0$, con lo que el primer sumando se anula y el segundo coincide con la suma en $\mathbb{N}$ de la función masa de probablidad de una variable aleatoria con distribución binomial negativa, que ha de sumar 1, como queríamos en la igualdad (2.1). Observemos ahora que:

- $h(t)\in[0,1]$ para todo $t\in\{0,\dots,n\}$, por lo que $h(T)$ es un estimador.

- Para el momento de segundo orden:

$$E[(h(t))^2] = \sum_{t=0}^{\infty} \frac{t^2}{n^2} \frac{(t+n-1)!}{t!(n-1)!} (1-\theta)^t \theta^n$$

Aplicamos el criterio del cociente:

$$\frac{\frac{(t+1)^2(t+n)!}{\cancel{n^2}(t+1)!\cancel{(n-1)!}}(1-\theta)^{t+1}\cancel{\theta^n}}{\frac{t^2(t+n-1)!}{\cancel{n^2}t!\cancel{(n-1)!}}(1-\theta)^t\cancel{\theta^n}} = \frac{(t+1)^2(t+n)!t!(1-\theta)}{(t+1)!t^2(t+n-1)!}$$

$$= \frac{(t+1)^2(t+n)(1-\theta)}{t^2(t+1)} = \frac{(t+1)(t+n)(1-\theta)}{t^2} \to 1-\theta$$

Con $\theta \in]0,1[$, por lo que $1-\theta \in]0,1[$, de donde la serie es convergente.

En definitiva, el UMVUE para $E_\theta[X] = \frac{1-\theta}{\theta}$ es $h(T)$, por el Teorema de Lehmann-Scheffé.

En los apartados anteriores observamos que la familia de funciones paramétricas que admitían estimadores eficientes eran las de la forma:

$$a \cdot \frac{n(1-\theta)}{\theta} + b \qquad a, b \in \mathbb{R}, \quad a \neq 0$$

Y sus estimadores eficientes eran:

$$a \cdot \sum_{i=1}^{n} X_i + b$$

Por lo que:

- Como θ no está en dicha familia de funciones paramétricas, el UMVUE obtenido para θ no puede ser eficiente.
- Si tomamos $a = \frac{1}{n}$ y $b = 0$, tenemos que l UMVUE obtenido para $\frac{1-\theta}{\theta}$ coincide con el estimador eficiente en dicho caso.

Ejercicio 2.4.9. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X con distribución exponencial.

a) Probar que la familia de distribuciones de X es regular.

Supuesto que $X \rightsquigarrow exp(\lambda)$ con $\lambda \in \mathbb{R}^+$, la función de densidad de X viene dada por:

$$f_\lambda(x) = \lambda e^{-\lambda x} \qquad x \geqslant 0$$

Y se cumple:

$$E[X] = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2}$$

Comprobamos cada una de las propiedades:

i) El espacio paramétrico es $\Theta = \mathbb{R}^+$, intervalo abierto de $\mathbb{R}$.

ii) El espacio muestral es $\mathbb{R}_0^+$, que no depende de λ.

iii) Para la tercera, observamos que f_λ es derivable respecto λ, con:

$$\ln f_\lambda(X) = \ln\left(\lambda e^{-\lambda X}\right) = \ln\lambda - \lambda X \quad \Longrightarrow \quad \frac{\partial \ln f_\lambda(X)}{\partial \lambda} = \frac{1}{\lambda} - X$$

Y si calculamos:

$$E\left[\frac{\partial \ln f_\lambda(X)}{\partial \lambda}\right] = E\left[\frac{1}{\lambda} - X\right] = \frac{1}{\lambda} - E[X] = \frac{1}{\lambda} - \frac{1}{\lambda} = 0$$

b) Encontrar la clase de funciones paramétricas que admiten estimador eficiente y el estimador correspondiente. Calcular la varianza de estos estimadores.

Calculamos primero la función de información de Fisher:

$$I_X(\lambda) = Var\left(\frac{\partial \ln f_\lambda(X)}{\partial \lambda}\right) = Var\left(\frac{1}{\lambda} - X\right) = Var(X) = \frac{1}{\lambda^2}$$

que como vemos verifica $0 < I_X(\lambda) < \infty$. Buscamos ahora las funciones a y g:

$$\frac{\partial \ln f_\lambda^n(x_1,\ldots,x_n)}{\partial \lambda} = \sum_{i=1}^n \frac{\partial \ln f_\lambda(x_i)}{\partial \lambda} = \sum_{i=1}^n \left(\frac{1}{\lambda} - x_i\right) = \frac{n}{\lambda} - \sum_{i=1}^n x_i = -\left(\sum_{i=1}^n x_i - \frac{n}{\lambda}\right)$$

Por lo que tomando:

$$T(X_1,\ldots,X_n) = \sum_{i=1}^n X_i, \qquad g(\lambda) = \frac{n}{\lambda}, \qquad a(\lambda) = -1$$

Tenemos que $a(\lambda) \neq 0 \quad \forall \lambda \in \mathbb{R}^+$, que g es derivable con:

$$g'(\lambda) = \frac{-n}{\lambda^2} \neq 0 \qquad \forall \lambda \in \mathbb{R}^+$$

y que T es un estimador, pues $T(\mathbb{R}^+) = \mathbb{R}^+ = g(\mathbb{R}^+)$. Finalmente, falta comprobar que:

$$a(\lambda)g'(\lambda) = \frac{n}{\lambda^2} = nI_X(\lambda) = I_{(X_1,\ldots,X_n)}(\lambda)$$

Por lo que podemos deducir ya por el Teorema de caracterización de los estimadores eficientes que T es un estimador eficiente. Más aún, por un Corolario del mismo, tenemos que las únicas funciones paramétricas de dicha familia que admiten estimadores eficientes son de la forma:

$$a \cdot \frac{n}{\lambda} + b, \qquad a, b \in \mathbb{R} \quad a \neq 0$$

y que dichos estimadores son $a \cdot T + b$. Nos disponemos a calcular la varianza de estos estimadores:

$$Var\left(a \cdot T + b\right) = a^2 Var\left(\sum_{i=1}^n X_i\right) = a^2 \sum_{i=1}^n Var(X) = a^2 \sum_{i=1}^n \left(\frac{1}{\lambda^2}\right) = \frac{na^2}{\lambda^2}$$

Veamos que coincide con la cota de Fréchet-Cramér-Rao, para comprobar nuestro trabajo calculando la familia de estimadores eficientes:

$$\frac{(ag'(\lambda))^2}{I_{(X_1,\dots,X_n)}(\lambda)} = \frac{\left(\frac{-an}{\lambda^2}\right)^2}{\frac{n}{\lambda^2}} = \frac{na^2}{\lambda^2} = Var(a\cdot T+b)$$

c) Basándose en el apartado anterior, encontrar el UMVUE para la media de X.

Como ya sabemos que $T=\sum_{i=1}^{n} X_i$ es un estimador eficiente para $g(\lambda)$, tenemos automáticamente que T es suficiente. Para ver que T es completo, sea Q una primitiva de a, podemos tomar $Q(\lambda)=-\lambda$, tenemos que claramente $Q(\mathbb{R}^+)$ contiene un abierto de $\mathbb{R}$, por lo que T es completo.

Como en el apartado anterior vimos que T es eficiente para $g(\lambda)=\frac{n}{\lambda}$, tenemos entonces que $\frac{T}{n}$ es eficiente para $\tilde{g}(\lambda)=\frac{g(\lambda)}{n}=\frac{1}{\lambda}$. Como la correspondencia $t\mapsto\frac{t}{n}$ es biunívoca, $\frac{T}{n}$ también será suficiente y completo (para ver que es completo no hace falta que la corresponencia sea unívoca). Por un Corolario del Teorema de caracterización de estimadores eficientes tenemos que T es UMVUE para $\frac{1}{\lambda}=E[X]$.

d) Dar la cota de Fréchet-Cramér-Rao para la varianza de estimadores insesgados y regulares de λ^3. ¿Es alcanzable dicha cota?

Si tomamos $g(\lambda)=\lambda^3$, la cota de Fréchet-Cramér-Rao tiene la forma:

$$\frac{(g'(\lambda))^2}{I_{(X_1,\dots,X_n)}(\lambda)} = \frac{(3\lambda^2)^2}{\frac{n}{\lambda^2}} = \frac{9\lambda^6}{n}$$

y como $g(\lambda)$ no es de la forma $a\cdot\frac{n}{\lambda}+b$, para ciertos $a,b\in\mathbb{R}$, sabemos que esta cota no es alcanzable.

Ejercicio 2.4.10. Sea X una variable aleatoria con función de densidad de la forma

$$f_\theta(x)=\theta x^{\theta-1},\quad 0<x<1$$

a) Sabiendo que $E_\theta[\ln X]=-\frac{1}{\theta}$ y $Var_\theta[\ln X]=\frac{1}{\theta^2}$, comprobar que esta familia de distribuciones es regular.

Veamos las propiedades, suponiendo que $\theta>0$.

i) El espacio paramétrico es $\Theta=\mathbb{R}^+$, intervalo abierto de $\mathbb{R}$.

ii) El espacio muestral es $\mathcal{X}=]0,1[$, que no depende de θ.

iii) La función de densidad es derivable respecto θ, calculamos la derivada del logaritmo de la función de densidad:

$$\ln f_\theta(X)=\ln(\theta X^{\theta-1})=\ln\theta+(\theta-1)\ln X \quad\Longrightarrow\quad \frac{\partial \ln f_\theta(X)}{\partial\theta}=\frac{1}{\theta}+\ln X$$

Con lo que:

$$E\left[\frac{\partial \ln f_\theta(X)}{\partial\theta}\right]=E\left[\frac{1}{\theta}+\ln X\right]=\frac{1}{\theta}+E[\ln X]=\frac{1}{\theta}-\frac{1}{\theta}=0$$

b) Basándose en una muestra aleatoria simple de X, dar la clase de funciones paramétricas con estimador eficiente, los estimadores y su varianza.

Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de X, en primer lugar calculamos la función de información de Fisher:

$$I_X(\theta) = Var\left(\frac{\partial \ln f_\theta(X)}{\partial \theta}\right) = Var\left(\frac{1}{\theta} + \ln X\right) = Var\left(\ln X\right) = \frac{1}{\theta^2}$$

y observamos que $0 < I_X(\theta) < \infty$. Ahora, buscamos las funciones g y a:

$$\frac{\partial \ln f_\theta^n(x_1, \ldots, x_n)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ln f_\theta(x_i)}{\partial \theta} = \sum_{i=1}^n \left(\frac{1}{\theta} + \ln x_i\right) = \frac{n}{\theta} + \sum_{i=1}^n \ln x_i$$

Tomando:

$$T(X_1, \ldots, X_n) = \sum_{i=1}^n \ln X_i, \qquad g(\theta) = \frac{-n}{\theta}, \qquad a(\theta) = 1$$

tenemos que $a(\theta) \neq 0$ para todo $\theta \in \mathbb{R}^+$, que g es derivable con:

$$g'(\theta) = \frac{n}{\theta^2} \neq 0 \qquad \forall \theta \in \mathbb{R}^+$$

además de que T es un estimador, puesto que:

$$T((]0,1[)^n) \subseteq]-\infty, 0[= g(\mathbb{R}^+)$$

Comprobemos que:

$$a(\theta)g'(\theta) = \frac{n}{\theta^2} = nI_X(\theta) = I_{(X_1,\ldots,X_n)}(\theta)$$

Por lo que T es un estimador eficiente para $g(\theta) = \frac{-n}{\theta}$, luego las únicas funciones paramétricas que admiten estimadores eficientes son las de la forma:

$$a \cdot \frac{-n}{\theta} + b, \qquad a, b \in \mathbb{R}, \quad a \neq 0$$

cuyos estimadores eficientes son:

$$a \cdot T + b$$

Calculemos sus varianzas y veamos que coinciden con la cota de Fréchet-Cramér-Rao:

$$Var(a \cdot T + b) = a^2 Var\left(\sum_{i=1}^n \ln X_i\right) \overset{\text{indep.}}{=} a^2 \sum_{i=1}^n Var(\ln X_i) = a^2 \sum_{i=1}^n \frac{1}{\theta^2} = \frac{a^2 n}{\theta^2}$$

$$\frac{\left(ag'(\theta)\right)^2}{I_{(X_1,\ldots,X_n)}(\theta)} = \frac{\left(\frac{an}{\theta^2}\right)^2}{\frac{n}{\theta^2}} = \frac{a^2 n}{\theta^2}$$

2.5. Estimación de máxima verosimilitud y otros métodos

Ejercicio 2.5.1. Sea $X \rightsquigarrow P_\theta$ con $\theta \in \mathbb{R}$ siendo P_θ una distribución con función de densidad

$$f_\theta(x) = e^{\theta - x}, \qquad x \geqslant \theta$$

Dada una muestra aleatoria simple de tamaño n, encontrar los estimadores máximo verosímiles de θ y de e^θ. Basándose en los resultados del Ejercicio 2.4.7, decir si estos estimadores son insesgados.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de $X \rightsquigarrow P_\theta$, tenemos que:

$$f_\theta^n(x_1, \ldots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n e^{\theta - x} = e^{n\theta - \sum_{i=1}^n x_i} \qquad \forall x_i \geqslant \theta$$

Por lo que si consideramos $x_{(1)} = \min\limits_{i \in \{1,\ldots,n\}} \{x_i\}$:

$$L_{x_1,\ldots,x_n}(\theta) = f_\theta^n(x_1, \ldots, x_n) = \begin{cases} e^{n\theta - \sum_{i=1}^n x_i} & \text{si } \theta \leqslant x_{(1)} \\ 0 & \text{en otro caso} \end{cases}$$

Observamos que $L_{x_1,\ldots,x_n}(\theta)$ es una función creciente en $\left]-\infty, x_{(1)}\right]$, por lo que alcanza su máximo en $x_{(1)}$. Como el espacio paramétrico es $\Theta = \mathbb{R}$, tenemos que $T(X_1, \ldots, X_n) = X_{(1)}$ es un estimador de θ, con lo que es el EMV, al maximizar la función de verosimilitud. Además, el Teorema de Zehna nos dice que:

$$\widehat{e^\theta} = e^{\hat{\theta}} = e^{X_{(1)}}$$

Por lo que $e^{X_{(1)}}$ es EMV para e^θ.

Buscamos ahora razonar si los estimadores obtenidos son o no insesgados, a partir de los resultados del Ejercicio 2.4.7. Si suponemos que los EMVs que hemos obtenido son insesgados, vimos anteriormente en dicho ejercicio que el mínimo era suficiente y completo, por lo que tendríamos una función del suficiente y completo que es insesgada. Si comprobamos que también tiene momento de segundo orden finito y que es estimador (hágase), tendríamos que es un UMVUE distinto al anterior, pero la unicidad del UMVUE nos dice que estos han de ser iguales, contradicción que viene de suponer que los EMVs son insesgados.

Ejercicio 2.5.2. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria X con distribución exponencial. Basándose en los resultados del Ejercicio 2.4.9, encontrar los estimadores máximo verosímiles de la media y de la varianza de X.

Sea $X \rightsquigarrow exp(\lambda)$ con $\lambda \in \mathbb{R}^+$, nos piden encontrar EMVs para:

$$E[X] = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2}$$

En el Ejercicio 2.4.9 obtuvimos que las únicas funciones paramétricas de dicha familia de distribuciones que admiten estimadores eficientes son las de la forma:

$$a \cdot \frac{n}{\lambda} + b \qquad a, b \in \mathbb{R}, \quad a \neq 0$$

y sus estimadores eficientes son:

$$a \cdot \sum_{i=1}^{n} X_i + b$$

Tomando $a = \frac{1}{n}, b = 0$, obtenemos que el estimador $T = \frac{\sum_{i=1}^{n} X_i}{n}$ es eficiente para $E[X] = 1/\lambda$, por lo que un Teorema visto en teoría nos asegura que T es el único EMV que podemos considerar para $1/\lambda$. Para $1/\lambda^2$ podemos aplicar el Teorema de Zehna, obteniendo:

$$\widehat{Var(X)} = \widehat{\left(\frac{1}{\lambda^2}\right)} = \frac{1}{\left(\hat{\lambda}\right)^2} = \frac{n}{\left(\sum_{i=1}^{n} X_i\right)^2}$$

que es estimador, luego es EMV.

Ejercicio 2.5.3. Sea X una variable aleatoria con función de densidad de la forma

$$f_\theta(x) = \theta x^{\theta-1}, \qquad 0 < x < 1$$

a) Calcular un estimador máximo verosímil para θ.

Maximizamos la función de verosimilitud de θ:

$$L_{x_1,\ldots,x_n}(\theta) = f_\theta^n(x_1, \ldots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

Si aplicamos logaritmo no cambia la abscisa a maximizar:

$$\ln L_{x_1,\ldots,x_n}(\theta) = n \ln \theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i$$

hayamos los puntos críticos:

$$\frac{\partial \ln L_{x_1,\ldots,x_n}(\theta)}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i = \frac{n + \theta \sum_{k=1}^{n} \ln x_i}{\theta} = 0 \iff \theta = \frac{-n}{\sum_{k=1}^{n} \ln x_i}$$

Por lo que tomando:

$$\hat{\theta}(X_1, \ldots, X_n) = \frac{-n}{\sum_{i=1}^{n} \ln X_i}$$

tenemos que $\hat{\theta}$ es un estimador de θ, luego es EMV para θ.

b) Deducir dicho estimador a partir de los resultados del Ejercicio 2.4.10.

En el Ejercicio 2.4.10 vimos que las únicas funciones paramétricas que admiten un estimador eficiente son las de la forma:

$$a \cdot \frac{-n}{\theta} + b \qquad a, b \in \mathbb{R}, \quad a \neq 0$$

y sus estimadores eficientes son:

$$a \cdot \sum_{i=1}^{n} \ln X_i + b$$

Tomando $a = \frac{-1}{n}, b = 0$, tenemos que el estimador:

$$T = \frac{\sum_{i=1}^{n} X_i}{-n}$$

es eficiente para $\frac{1}{\theta}$, por lo que T es el EMV de θ. Si aplicamos el Teorema de Zehna, tenemos que:

$$\widehat{\left(\frac{1}{\theta}\right)} = \frac{1}{\hat{\theta}} \Longrightarrow \hat{\theta} = \frac{1}{\widehat{\left(\frac{1}{\theta}\right)}} = \frac{-n}{\sum_{i=1}^{n} \ln X_i}$$

Observamos que obtenemos el mismo resultado que en el primer apartado.

Ejercicio 2.5.4. Sea $(X_1, \ldots, X_n)$ una muestra de una variable $X \rightsquigarrow B(k_0, p)$ para cierto $k_0 \in \mathbb{N}$ y $p \in]0, 1[$. Estimar, por máxima verosimilitud y por el método de los momentos, el parámetro p y la varianza de X.

Aplicación: Se lanza 10 veces un dado cargado y se cuenta el número de veces que sale un 4. Este experimento se realiza 100 veces de forma independiente, obteniéndose los siguientes resultados:

nº de 4	0	1	2	3
frecuencia	84	15	1	0

Estimar, a partir de estos datos, la probabilidad de salir un cuatro.

Por máxima verosimilitud. Tenemos:

$$P[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{iid.}}{=} \prod_{i=1}^{n} P[X = x_i] = \prod_{i=1}^{n} \binom{k_0}{x_i} p^{x_i} (1-p)^{k_0 - x_i}$$
$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{nk_0 - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{k_0}{n}$$

Por lo que:

$$\ln L_{x_1,\dots,x_n}(p) = \ln p \sum_{i=1}^{n} x_i + \ln(1-p)\left(nk_0 - \sum_{i=1}^{n} x_i\right) + \sum_{k=1}^{n} \ln \binom{k_0}{n}$$

Luego:

$$\frac{\partial \ln L_{x_1,\dots,x_n}(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(nk_0 - \sum_{i=1}^{n} x_i\right) = \frac{(1-p)\sum\limits_{i=1}^{n} x_i - pnk_0 + p\sum\limits_{i=1}^{n} x_i}{p(1-p)}$$
$$= \frac{\sum\limits_{i=1}^{n} x_i - pnk_0}{p(1-p)} = 0 \iff p = \frac{\sum\limits_{i=1}^{n} x_i}{nk_0}$$

Como $\sum\limits_{i=1}^{n} x_i \leqslant nk_0$, tenemos entonces que $p \in [0,1]$, por lo que esta fórmula nos dará una estimador, con lo que el EMV es:

$$\hat{p} = \frac{1}{nk_0}\sum_{i=1}^{n} X_i = \frac{1}{k_0}\overline{X}$$

Para $Var(X) = k_0 p(1-p)$, si aplicamos el Teorema de Zehna tenemos que:

$$\widehat{Var(X)} = k_0\widehat{p(1-p)} = k_0\hat{p}(1-\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} X_i\left(1 - \frac{1}{nk_0}\sum_{i=1}^{n} X_i\right) = \overline{X}\left(1 - \frac{1}{k_0}\overline{X}\right)$$

Por el método de los momentos. Por el método de los momentos:

$$\begin{cases} k_0 p = E[X] = \overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i \\ k_0 p(1-p) = Var(X) = \frac{1}{n}\left(\sum\limits_{k=1}^{n} X_i^2 - \overline{X}^2\right) \end{cases}$$

Del primero de ducimos que:

$$\hat{p} = \frac{1}{nk_0}\sum_{i=1}^{n} X_i = \frac{1}{k_0}\overline{X}$$

Y del segundo que:

$$\widehat{Var(X)} = \frac{1}{n}\left(\sum_{k=1}^{n} X_i^2 - \overline{X}^2\right)$$

Para estimar a partir de los datos mencionados la probabilidad de salir un cuatro, lo que hacemos primero es considerar:

$Y \equiv$ "Número de veces que sale un 4 en 10 tiradas del dado" $\rightsquigarrow B(10,p)$

donde estimaremos el valor de p mediante la estimación obtenida:

$$p = \frac{1}{nk_0}\sum_{i=1}^{n} x_i = \frac{1}{nk_0}\sum_{i=1}^{n} x_i = \frac{1}{1000}(0\cdot 84 + 1\cdot 15 + 2\cdot 1 + 3\cdot 0) = \frac{15+2}{1000} = \frac{17}{1000}$$

Por lo que la probabilidad de obtener un 4 es $\frac{17}{1000} = 0{,}017$.

Ejercicio 2.5.5. Se lanza un dado hasta que salga un 4 y se anota el número de lanzamientos necesarios; este experimento se efectúa veinte veces de forma independiente. A partir de los resultados obtenidos, estimar la probabilidad de sacar un 4 por máxima verosimilitud.

Sea:

$$X \equiv \text{“número de lanzamientos antes del lanzamiento en el que sale 4”}$$

tenemos que si p es la probabilidad de que salga un 4, entonces X sigue una distribución geométrica de parámetro p: $X \rightsquigarrow G(p)$. Como se realiza el experimento 20 veces de forma independientes, disponemos de una muestra $x_1, \ldots, x_n$ donde $n = 20$. Estimamos p por máxima verosimilitud:

$$L_{x_1,\ldots,x_n}(p) = P_p[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{indep.}}{=} \prod_{i=1}^{n} P_p[X = x_i] = \prod_{i=1}^{n} (1-p)^{x_i} p = p^n (1-p)^{\sum_{i=1}^{n} x_i}$$

de donde al aplicar logaritmos:

$$\ln L_{x_1,\ldots,x_n}(p) = n \cdot \ln p + \ln(1-p) \sum_{i=1}^{n} x_i$$

si derivamos:

$$\frac{\partial \ln L_{x_1,\ldots,x_n}(p)}{\partial p} = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i}{1-p} = \frac{n - np - p\left(\sum_{i=1}^{n} x_i\right)}{p(1-p)} = \frac{n - p\left(\sum_{i=1}^{n} x_i + n\right)}{p(1-p)}$$

Obtenemos que:

$$\frac{\partial \ln L_{x_1,\ldots,x_n}(p)}{\partial p} = 0 \iff p = \frac{n}{\sum_{i=1}^{n} x_i + n}$$

Como siempre tenemos que:

$$\frac{n}{\sum_{i=1}^{n} x_i + n} \in [0, 1]$$

Tenemos entonces que el EMV es:

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i + n} = \frac{20}{\sum_{i=1}^{20} X_i + 20}$$

o bien:

$$\hat{p} = \frac{1}{1 + \overline{X}}$$

Ejercicio 2.5.6. En 20 días muy fríos, una granjera pudo arrancar su tractor en el primer, tercer, quinto, primer, segundo, primer, tercer, séptimo, segundo, cuarto, cuarto, octavo, primer, tercer, sexto, quinto, segundo, primer, sexto y segundo intento. Suponiendo que la probabilidad de arrancar en cada intento es constante,

y que las observaciones se han obtenido de forma independiente, dar la estimación más verosímil de la probabilidad de que el tractor arranque en el segundo intento.

Si reunimos en una tabla la información del enunciado:

Primer	Segundo	Tercero	Cuarto	Quinto	Sexto	Séptimo	Octavo
5	4	3	2	2	2	1	1

Tabla 2.1: Veces que arrancó en cada intento.

Si consideramos:

$$X \equiv \text{“Número de intentos fallidos de arrancar el tractor”}$$

tenemos que $X \rightsquigarrow G(p)$, donde p es la probabilidad de arrancar el tractor en un intento. tenemos a nuestra disposición de una muestra de X $x_1, \ldots, x_n$ de tamaño $n = 20$. Si observamos la similitud con el Ejercicio 2.5.5, estamos bajo las mismas hipótesis, por lo que al calcular el EMV para p obtendremos:

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i + n} = \frac{20}{\sum_{i=1}^{20} X_i + 20}$$

Sin embargo, queremos calcularlo para la probabilidad de que el tractor arranque en el segundo intento, es decir, para tener exactamente un intento fallido de arranzar el tractor:

$$P_p[X = 1] = (1 - p)p$$

Si amplicamos el Teorema de Zehna, podemos obtener el EMV de $P_p[X = 1]$ a partir de $\hat{p}$:

$$\widehat{P_p[X = 1]} = \widehat{(1 - p)}p = (1 - \hat{p})\hat{p} = \left(1 - \frac{20}{\sum_{i=1}^{20} X_i + 20}\right) \frac{20}{\sum_{i=1}^{20} X_i + 20}$$

Si sustituimos ahora en los valores de la muestra, observamos que:

$$\sum_{i=1}^{n} x_i = 0 \cdot 5 + 1 \cdot 4 + 2 \cdot 3 + 3 \cdot 2 + 4 \cdot 2 + 5 \cdot 2 + 6 \cdot 1 + 7 \cdot 1 = 47$$

Por lo que la estimación máximo verosímil de la probabilida de que el tractor arranque en el segundo intento es igual a:

$$\left(1 - \frac{20}{\sum_{i=1}^{20} X_i + 20}\right) \frac{20}{\sum_{i=1}^{20} X_i + 20} = \left(1 - \frac{20}{47 + 20}\right) \frac{20}{47 + 20} = \frac{940}{4489} \approx 0{,}2094$$

Ejercicio 2.5.7. Una variable aleatoria discreta toma los valores 0, 1 y 2 con las siguientes probabilidades

$$P_p[X=0]=p^2, \qquad P_p[X=1]=2p(1-p), \qquad P_p[X=2]=(1-p)^2$$

siendo p un parámetro desconocido. En una muestra aleatoria simple de tamaño 100, se ha presentado 22 veces el 0, 53 veces el 1 y 25 veces el 2. Calcular la función de verosimilitud asociada a dicha muestra y dar la estimación más verosímil de p.

Los datos obtenidos son:

Dato	0	1	2
Veces	22	53	25

Sea X una variable aleatoria cuya función masa de probabilidad nos viene dada, tenemos una muestra $x_1,\dots,x_n$ de tamaño $n=100$. La función de verosimilitud asociada a la muestra es:

$$\begin{aligned} L_{x_1,\dots,x_n}(p) &= P_p[X_1=x_1,\dots,X_n=x_n] \overset{\text{iid.}}{=} \prod_{i=1}^{n} P_p[X=x_i] \\ &= \prod_{i=1}^{22} P_p[X=0] \prod_{i=1}^{53} P_p[X=1] \prod_{i=1}^{25} P_p[X=2] \\ &= \prod_{i=1}^{22} p^2 \prod_{i=1}^{53} 2p(1-p) \prod_{i=1}^{25} (1-p)^2 \\ &= p^{44}\cdot 2^{53}p^{53}(1-p)^{53}\cdot(1-p)^{50} = 2^{53}p^{97}(1-p)^{103} \end{aligned}$$

Si aplicamos logaritos:

$$\ln L_{x_1,\dots,x_n}(p) = 53\ln 2 + 97\ln p + 103\ln(1-p)$$

y derivamos:

$$\frac{\partial \ln L_{x_1,\dots,x_n}(p)}{\partial p} = \frac{97}{p} - \frac{103}{1-p} = \frac{97(1-p)-103p}{p(1-p)} = \frac{97-200p}{p(1-p)} = 0 \iff p = \frac{97}{200}$$

Por lo que la estimación por máxima verosimilitud es $\frac{97}{200} = 0{,}485$.

Este ejercicio se podría haber resuelto también usando la multinomial.

Ejercicio 2.5.8. En el muestreo de una variable aleatoria con distribución $\mathcal{N}(\mu,1)$, $\mu\in\mathbb{R}$, se observa que no se obtiene un valor menor que -1 hasta la quinta observación. Dar una estimación máximo verosímil de μ.

Sea $Y \rightsquigarrow \mathcal{N}(\mu,1)$, si tomamos:

$X \equiv$ "número de observaciones hasta obtener una menor que -1 (incluida)"

tenemos que:

$$P_p[X=x] = (1-p)^{x-1}p$$

donde $p = P[Y < -1]$. Tenemos una muestra de tamaño 1 ($x = 5$) de X, con lo que la función de verosimilitud es:

$$L_5(p) = P_p[X = 5] = (1-p)^{5-1}p = p(1-p)^4$$

aplicando logaritmos:

$$\ln L_5(p) = 4\ln(1-p) + \ln p$$

de donde:

$$\frac{\partial \ln L_5(p)}{\partial p} = \frac{-4}{1-p} + \frac{1}{p} = \frac{-4p+1-p}{p(1-p)} = \frac{1-5p}{p(1-p)} = 0 \Longleftrightarrow p = \frac{1}{5}$$

Si notamos por Φ la función de distribución de una $\mathcal{N}(0,1)$, tenemos que:

$$p = P[Y < -1] = P[Z < -1-\mu] = \Phi(-1-\mu) = 1 - \Phi(1+\mu)$$

de donde (usando que Φ es biyectiva):

$$\mu = \Phi^{-1}(1-p) - 1$$

por lo que si aplicamos el Teorema de Invarianza de Zehna:

$$\hat{\mu} = \Phi^{-1}(1-\hat{p}) - 1 = \Phi^{-1}\left(1 - \frac{1}{5}\right) - 1 = \Phi^{-1}(0{,}8) - 1$$

Y si miramos en la tabla de la normal $\mathcal{N}(0,1)$ la abscisa en la que se da la probabilidad 0,8, obtenemos que se alcanza en la abscisa 0,85, por lo que:

$$\hat{\mu} = \Phi^{-1}(0{,}8) - 1 = 0{,}85 - 1 = -0{,}15$$

es la estimación máximo verosímil de μ.

Ejercicio 2.5.9. En la producción de filamentos eléctricos la medida de interés, X, es el tiempo de vida de cada filamento, que tiene una distribución exponencial de parámetro θ. Se eligen n de tales filamentos de forma aleatoria e independiente, pero, por razones de economía, no conviene esperar a que todos se quemen y la observación acaba en el tiempo T. Dar el estimador máximo verosímil para la media de X a partir del número de filamentos quemados durante el tiempo de observación.

Sea $X \rightsquigarrow exp(\theta)$, si consideramos:

$$Y \equiv \text{“número de filamentos con tiempo de vida menor que } T \text{ en una muestra de tamaño } n\text{”}$$

tenemos que $Y \rightsquigarrow B(n,p)$ donde p es la probabilidad de que el tiempo de vida de un filamento sea menor que T, $p = P[X < T]$. Disponemos de una muestra de tamaño 1 de Y, y. Tratamos de calcular el EMV de p:

$$L_y(p) = P_p[Y = y] = \binom{n}{y} p^y (1-p)^{n-y}$$

si aplicamos logaritmo:

$$\ln L_y(p) = \ln \binom{n}{y} + y \ln p + (n-y)\ln(1-p)$$

y ahora derivamos:

$$\frac{\partial \ln L_y(p)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} = \frac{y - yp - np + yp}{p(1-p)} = \frac{y-np}{p(1-p)} = 0 \iff p = \frac{y}{n}$$

tenemos que el EMV de p es:

$$\hat{p} = \frac{Y}{n}$$

Si observamos la definición de p:

$$p = P[X < T] = F_X(T) = 1 - e^{-\theta T} \Longrightarrow e^{-\theta T} = 1 - p \Longrightarrow -\theta T = \ln(1-p)$$

de donde:

$$\theta = \frac{-\ln(1-p)}{T}$$

Sin embargo, lo que nos interesa es calcuar el EMV de $E[X] = \frac{1}{\theta}$, por lo que:

$$E[X] = \frac{1}{\theta} = \frac{-T}{\ln(1-p)}$$

Si aplicamos ahora el Teorema de Zehna obtenemos finalmente el EMV de $E[X]$:

$$\widehat{E[X]} = \frac{\widehat{1}}{\theta} = \frac{\widehat{-T}}{\ln(1-p)} = \frac{-T}{\ln(1-\hat{p})} = \frac{-T}{\ln\left(1-\frac{Y}{n}\right)}$$

Ejercicio 2.5.10. Sean $X_1, \ldots, X_n$ observaciones independientes de una variable $X \rightsquigarrow \{\Gamma(p,a) : p, a > 0\}$. Estimar ambos parámetros mediante el método de los momentos.

Aplicación: Ciertos neumáticos radiales tuvieron vidas útiles de 35200, 41000, 44700, 38600 y 41500 kilómetros. Suponiendo que estos datos son observaciones independientes de una variable con distribución exponencial de parámetro θ, dar una estimación de dicho parámetro por el método de los momentos.

Planteamos el sistema a resolver, sabiendo que si $X \rightsquigarrow \Gamma(p,a)$, entonces:

$$E[X] = \frac{p}{a}, \qquad Var(X) = \frac{p}{a^2}$$

por lo que:

$$\begin{cases} p/a = E[X] = \overline{X} \\ p/a^2 = Var(X) = A_2 - \left(\overline{X}\right)^2 \end{cases}$$

de la primera ecuación tenemos que:

$$p = a\overline{X}$$

y si sustituimos en la segunda:

$$\frac{\overline{X}}{a} = A_2 - \left(\overline{X}\right)^2 \Longrightarrow a = \frac{\overline{X}}{A_2 - \left(\overline{X}\right)^2}$$

Sustituyendo ahora a en la primera:

$$p = \frac{\left(\overline{X}\right)^2}{A_2 - \left(\overline{X}\right)^2}$$

Por lo que las estimaciones a considerar son:

$$\hat{a} = \frac{\overline{X}}{A_2 - \left(\overline{X}\right)^2}, \qquad \hat{p} = \frac{\left(\overline{X}\right)^2}{A_2 - \left(\overline{X}\right)^2}$$

Sea $X \rightsquigarrow exp(\theta)$, por el método de los momentos tenemos que:

$$\frac{1}{\theta} = E[X] = \overline{X} \Longrightarrow \theta = \frac{1}{\overline{X}}$$

Por tanto, para estimar θ lo primero que hacemos es calcular la media de los valores:

$$\frac{35200 + 41000 + 44700 + 38600 + 41500}{5} = 40200$$

y la estimación será:

$$\hat{\theta} = \frac{1}{\overline{x}} = \frac{1}{40200} \approx 2{,}488 \cdot 10^{-5}$$

2.6. Estimación por intervalos de confianza

Ejercicio 2.6.1. Sea $\overline{X}$ la media de una muestra aleatoria de tamaño n de una población $\mathcal{N}(\mu, 16)$ Encontrar el menor valor de n para que $\left]\overline{X}-1, \overline{X}+1\right[$ sea un intervalo de confianza para μ al nivel de confianza 0,9.

Si $\overline{X}$ es la media de una m.a.s. de tamaño n de una población $\mathcal{N}(\mu, 16)$, tenemos entonces que $\overline{X} \rightsquigarrow \mathcal{N}(\mu, \frac{16}{n})$. Calculamos n para que el intervalo $\left]\overline{X}-1, \overline{X}+1\right[$ sea un intervalo de confianza para μ a nivel 0,9, buscando:

$$\begin{aligned} P_\mu\left[\overline{X}-1 \leqslant \mu \leqslant \overline{X}+1\right] \geqslant 0{,}9 &\iff P_\mu[\mu-1 \leqslant \overline{X} \leqslant \mu+1] \geqslant 0{,}9 \\ &\overset{\text{tipif.}}{\iff} P_\mu\left[\frac{-\sqrt{n}}{4} \leqslant Z \leqslant \frac{\sqrt{n}}{4}\right] \geqslant 0{,}9 \\ &\iff 2P_\mu\left[Z \leqslant \frac{\sqrt{n}}{4}\right] - 1 \geqslant 0{,}9 \\ &\iff 2P_\mu\left[Z \leqslant \frac{\sqrt{n}}{4}\right] \geqslant 1{,}9 \\ &\iff P_\mu\left[Z \leqslant \frac{\sqrt{n}}{4}\right] \geqslant 0{,}95 \end{aligned}$$

Consultando la tabla de la normal $\mathcal{N}(0,1)$, tenemos que la primera abscisa en la que se alcanza una probabilidad superior a 0,95 es 1,65, por lo que:

$$\frac{\sqrt{n}}{4} = 1{,}65 \iff \sqrt{n} = 6{,}6 \iff n = 43{,}56$$

Por tanto, el menor valor de n para el cual el intervalo $\left]\overline{X}-1, \overline{X}+1\right[$ es un intervalo de confianza para μ a nivel de confianza 0,9 es 44.

Ejercicio 2.6.2. La altura en cm. de los individuos varones de una población sigue una distribución $\mathcal{N}(\mu,\ 56{,}25)$. Si en una muestra aleatoria simple de tamaño 12 de dicha población se obtiene una altura media de 175 cm., determinar un intervalo de confianza para μ al nivel de confianza 0,95. ¿Qué tamaño de muestra es necesario para que el intervalo de confianza a dicho nivel tenga longitud menor que 1 cm?

Buscamos un intervalo de confianza para μ en una población normal con varianza σ_0 conocida a nivel de confianza $1-\alpha = 0{,}95$ (por lo que $\alpha = 0{,}05$). Tenemos pues una muestra aleatoria simple $(X_1, \ldots, X_n)$ de $X \rightsquigarrow \mathcal{N}(\mu, 56{,}25)$. Usaremos el método de la cantidad pivotal, usando para ello la función pivote (donde $n = 12$):

$$T(X_1, \ldots, X_n; \mu) = \frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$$

Que:

- Es estrictamente decreciente respecto μ.
- Si tomamos λ de forma que:

$$\lambda = \frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} \Longrightarrow \mu = \overline{X} - \lambda\frac{\sigma_0}{\sqrt{n}}$$

Por lo que el intervalo de confianza que consideraremos viene dado por:

$$\left]\overline{X} - \lambda_2\frac{\sigma_0}{\sqrt{n}}, \overline{X} - \lambda_1\frac{\sigma_0}{\sqrt{n}}\right[$$

donde λ_1 y λ_2 están sujetos a la restricción $P_\mu[\lambda_1 < T < \lambda_2] = 1 - \alpha$. Obtenemos un intervalo de longitud esperada:

$$E[L] = E\left[(\lambda_2 - \lambda_1)\frac{\sigma_0}{\sqrt{n}}\right] = (\lambda_2 - \lambda_1)\frac{\sigma_0}{\sqrt{n}}$$

Como $\sigma_0/\sqrt{n}$ es una constante positiva, bastará minimizar la cantidad $\lambda_2 - \lambda_1$ (sujeta a la restricción $P_\mu[\lambda_1 < T < \lambda_2] = 1 - \alpha$) para minimizar la longitud del intervalo. La restricción mencionada puede reescribirse como:

$$1 - \alpha = P_\mu[\lambda_1 < T < \lambda_2] = F_Z(\lambda_2) - F_Z(\lambda_1)$$

Consideramos por el método de los multiplicadores de Lagrange:

$$F(\lambda_1, \lambda_2) = (\lambda_2 - \lambda_1) - \lambda(F_Z(\lambda_2) - F_Z(\lambda_1) - (1 - \alpha))$$

y buscamos los valores de λ_1 y λ_2 que minimicen la expresión, por lo que calculamos sus derivadas parciales:

$$\frac{\partial F}{\partial \lambda_1} = -1 + \lambda f_Z(\lambda_1) \qquad \frac{\partial F}{\partial \lambda_2} = 1 - \lambda f_Z(\lambda_2)$$

Si igualamos ambas a cero y tratamos de despejar λ:

$$\left.\begin{array}{l} 0 = \dfrac{\partial F}{\partial \lambda_1} = -1 + \lambda f_Z(\lambda_1) \\ 0 = \dfrac{\partial F}{\partial \lambda_2} = 1 - \lambda f_Z(\lambda_2) \end{array}\right\} \Longrightarrow \left.\begin{array}{l} \lambda = \frac{1}{f_Z(\lambda_1)} \\ \lambda = \frac{1}{f_Z(\lambda_2)} \end{array}\right\} \Longrightarrow f_Z(\lambda_1) = f_Z(\lambda_2)$$

Como f_Z es una función simétrica respecto al origen, las únicas posibilidades son bien $\lambda_1 = \lambda_2$ bien $\lambda_1 = -\lambda_2$. Como estos valores han de cumplir la restricción $P_\mu[\lambda_1 < T < \lambda_2] = 1 - \alpha > 0$, la primera opción es imposible, por lo que tiene que ser $\lambda_1 = -\lambda_2$. Como han de dejar entre ellos un valor de $1 - \alpha$, han de ser:

$$\lambda_2 = Z_{\alpha/2}, \qquad \lambda_1 = -\lambda_2 = -Z_{\alpha/2}$$

Por lo que el intervalo a considerar es:

$$\left]\overline{X} - Z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right[$$

A partir de los datos proporcionados, tenemos que:

- $\alpha = 0{,}05$.
- $\sigma_0 = \sqrt{56{,}25} = 7{,}5$.
- $\sqrt{n} = \sqrt{12} = 3{,}4641$.

- $\overline{x} = 175$.
- $Z_{\alpha/2} = Z_{0,025} = 1{,}96$.

Por lo que el intervalo a considerar es:

$$\left]175 - 1{,}96 \cdot \frac{7{,}5}{3{,}4641}, 175 + 1{,}96 \cdot \frac{7{,}5}{3{,}4641}\right[=]170{,}7564755, 179{,}2435245[$$

Buscamos ahora el tamaño de la muestra para que la longitud del intervalo sea menor que 1cm. Para ello, vemos que nuestro intervalo tiene longitud:

$$2Z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} = 2 \cdot 1{,}96 \cdot \frac{7{,}5}{\sqrt{n}} = \frac{29{,}4}{\sqrt{n}} \leqslant 1 \iff 29{,}4 \leqslant \sqrt{n} \iff 864{,}36 \leqslant n$$

Por lo que nicesitaremos al menos una nuestra de tamaño 865 para que el intervalo de confianza a dicho nivel tenga longitud menor que 1 cm.

Ejercicio 2.6.3. Una fábrica produce tornillos cuyo diámetro medio es 3 mm. Se seleccionan aleatoria e independientemente 12 de estos tornillos y se miden sus diámetros, que resultan ser 3,01, 3,05, 2,99, 2,99, 3,00, 3,02, 2,98, 2,99, 2,97, 2,97, 3,02 y 3,01. Suponiendo que el diámetro es una variable aleatoria con distribución normal, determinar un intervalo de confianza para la varianza al nivel de confianza 0,99, y una cota superior de confianza al mismo nivel. Interpretar los resultados en términos de la desviación típica del diámetro de los tornillos.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de $X \rightsquigarrow \mathcal{N}(3, \sigma^2)$, vamos a calcular un intervalo de confianza para σ^2 en una población normal con media μ_0 conocida a nivel de confianza $1 - \alpha$ mediante el método de la cantidad pivotal, usando la función pivote:

$$T(X_1, \ldots, X_n; \sigma^2) = \frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{\sigma^2} \rightsquigarrow \chi^2(n)$$

Que:

- Es estrictamente decreciente en σ^2.
- Si tenemos λ con:

$$\lambda = \frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{\sigma^2} \rightsquigarrow \chi^2(n) \Longrightarrow \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{\lambda}$$

Por lo que el intervalo a considerar será:

$$\left]\frac{1}{\lambda_2}\sum_{i=1}^{n} (X_i - \mu_0)^2, \frac{1}{\lambda_1}\sum_{i=1}^{n} (X_i - \mu_0)^2\right[$$

donde λ_1 y λ_2 están sujetos a la restricción:

$$1 - \alpha \leqslant P_{\sigma^2}[\lambda_1 < T < \lambda_2] = F_T(\lambda_2) - F_T(\lambda_1) \qquad (T \rightsquigarrow \chi^2(n))$$

Tenemos que la longitud esperada del intervalo es:

$$E[L] = E\left[\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)\sum_{i=1}^{n}(X_i - \mu_0)^2\right] = \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)E\left[\sum_{i=1}^{n}(X_i - \mu_0)^2\right]$$

Donde la última esperanza es una cantidad constante y positiva, por lo que trataremos de minimizar simplemente $1/\lambda_1 - 1/\lambda_2$. Mediante el método de los multiplicadores de Lagrange:

$$F(\lambda_1, \lambda_2) = \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) - \lambda(F_T(\lambda_2) - F_T(\lambda_1) - (1-\alpha))$$

Calculando las derivadas parciales e igualando a cero llegamos a:

$$\frac{f_T(\lambda_1)}{f_T(\lambda_2)} = \frac{\lambda_2^2}{\lambda_1^2}$$

Se trata de un despeje difícil, por lo que en la práctica se toman los valores:

$$\lambda_1 = \chi^2_{n,1-\alpha/2}, \qquad \lambda_2 = \chi^2_{n,\alpha/2}$$

En definitiva, consideraremos el intervalo:

$$\left]\frac{1}{\chi^2_{n,\alpha/2}}\sum_{i=1}^{n}(X_i - \mu_0)^2, \frac{1}{\chi^2_{n,1-\alpha/2}}\sum_{i=1}^{n}(X_i - \mu_0)^2\right[$$

Para determinar una cota superior, lo que hacemos es tomar $\lambda_2 = \infty$, por lo que:

$$1 - \alpha = F_T(\lambda_2) - F_T(\lambda_1) = 1 - F_T(\lambda_1) \Longrightarrow F_T(\lambda_1) = \alpha$$

Luego podemos tomar $\lambda_1 = \chi^2_{n,1-\alpha}$, obteniendo la cota superior:

$$\frac{1}{\chi^2_{n,1-\alpha}}\sum_{i=1}^{n}(X_i - \mu_0)^2$$

Si situimos ahora en los parámetros que tenemos:

- $\mu_0 = 3$.
- $n = 12$.
- $\alpha = 0{,}01$.
- $\chi^2_{n,\alpha/2} = \chi^2_{12,\ 0{,}005} = 28{,}2997$.
- $\chi^2_{n,1-\alpha/2} = \chi^2_{12,\ 0{,}995} = 3{,}0738$.
- $\chi^2_{n,1-\alpha} = \chi^2_{12,0{,}99} = 3{,}5706$.
- Finalmente:

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = 0{,}0059$$

El intervalo de confianza a nivel de confianza 0,99 será:

$$\left]\frac{1}{28{,}2997}0{,}0059, \frac{1}{3{,}0738}0{,}0059\right[=]0{,}0002084, 0{,}00191945[$$

Y la cota superior será:

$$\frac{1}{3{,}5706}0{,}0059 = 0{,}00165238$$

La desviación típica de la muestra dada es:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2 = \frac{1}{12}\cdot 0{,}0059 = 0{,}000491667$$

Que como veos pertence al intervalo y es menor que la cota dada.

Ejercicio 2.6.4. Las notas en cierta asignatura de 7 alumnos de una clase, elegidos de forma aleatoria e independiente son: 4,5, 3, 6, 7, 1,5, 5,2 y 3,6. Suponiendo que las notas tienen distribución normal, dar un intervalo de confianza para la varianza de las mismas al nivel de confianza 0,95.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de $X \rightsquigarrow \mathcal{N}(\mu_0, \sigma^2)$. Tenemos que el intervalo de confianza para σ^2 a nivel de confianza $1-\alpha$ es:

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

Pero ahora tenemos los valores:

- $n = 7$.
- $\alpha = 0{,}05$.
- $\chi^2_{n,\alpha/2} = \chi^2_{7,0{,}025} = 16{,}0128$.
- $\chi^2_{n,1-\alpha/2} = \chi^2_{7,0{,}975} = 1{,}6899$.

Por lo que el intervalo será:

$$]1{,}451936, 16{,}959525[$$

Ejercicio 2.6.5. Dos muestras independientes, cada una de tamaño 7, de poblaciones normales con igual varianza, producen medias 4,8 y 5,4 y cuasivarianzas muestrales 8,38 y 7,62, respectivamente. Encontrar un intervalo de confianza para la diferencia de medias al nivel de confianza 0,95.

Si tenemos $(X_1, \ldots, X_{n_1})$ m.a.s. de $X \rightsquigarrow \mathcal{N}(\mu_1, \sigma_1^2)$ y $(Y_1, \ldots, Y_{n_2})$ m.a.s. de $Y \rightsquigarrow \mathcal{N}(\mu_2, \sigma_2^2)$. Si queremos calcular el intervalo de confianza de $\mu_1 - \mu_2$ a nivel de confianza $1-\alpha$, hemos visto en teoría que se obtiene:

$$\left]\overline{X} - \overline{Y} - t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \overline{X} - \overline{Y} + t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right[$$

donde:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Calculamos el intervalo, usando para ello que:

- $n_1 = 7 = n_2$.
- $\overline{x} = 4{,}8$, $\overline{y} = 5{,}4$.
- $s_1^2 = 8{,}38$, $s_2^2 = 7{,}62$.
- $\alpha = 0{,}05$.

Por lo que:

$$S_p = \sqrt{\frac{6 \cdot 8{,}38 + 6 \cdot 7{,}62}{7+7-2}} = 2{,}828427$$

De donde el intervalo será, usando que:

$$t_{n_1+n_2-2,\alpha/2} = t_{12,\ 0{,}025} = 2{,}1788$$

$$\left]4{,}8 - 5{,}4 - 2{,}1788 \cdot 2{,}828427 \cdot \sqrt{\frac{1}{7}+\frac{1}{7}}, 4{,}8 - 5{,}4 + 2{,}1788 \cdot 2{,}828427 \cdot \sqrt{\frac{1}{7}+\frac{1}{7}}\right[$$
$$=]-3{,}894036, 2{,}694036[$$

Ejercicio 2.6.6. La siguiente tabla presenta los salarios anuales (en miles de euros) de dos grupos de recién graduados de dos carreras diferentes. Suponiendo normalidad en los salarios de ambos grupos, determinar un intervalo de confianza para el cociente de las varianzas al nivel de confianza 0,90.

GRUPO 1	16.3	18.2	17.5	16.1	15.9	15.4	15.8	17.3	14.9	15.1
GRUPO 2	13.2	15.1	13.9	14.7	15.6	15.8	14.9	18.1	15.6	15.3
	16.2	15.2	15.4	16.6						

Si tenemos $(X_1, \ldots, X_{n_1})$ m.a.s. de $X \rightsquigarrow \mathcal{N}(\mu_1, \sigma_1^2)$ (que se corresponderá con el grupo 1) y $(Y_1, \ldots, Y_{n_2})$ m.a.s. de $Y \rightsquigarrow \mathcal{N}(\mu_2, \sigma_2^2)$ (que se corresponderá con el grupo 2). Si queremos calcular el intervalo de confianza de $\frac{\sigma_1^2}{\sigma_2^2}$ a nivel de confianza $1-\alpha$, hemos visto en teoría que este será:

$$\left]F_{n_1-1,n_2-1,1-\alpha/2}\frac{S_1^2}{S_2^2}, F_{n_1-1,n_2-1,\alpha/2}\frac{S_1^2}{S_2^2}\right[$$

Con los datos:

- $\alpha = 0{,}1$.
- $n_1 = 10$.
- $n_2 = 14$.

- $S_1^2 = 1{,}187222$.
- $S_2^2 = 1{,}352308$.

Tenemos ($F_{9,13}$ no sale en la tabla, tomo los valores de $F_{9,12}$):

- $F_{n_1-1,n_2-1,1-\alpha/2} = F_{9,13,0,95} = 0{,}325$.
- $F_{n_1-1,n_2-1,\alpha/2} = F_{9,13,0,05} = 2{,}8$.

Por lo que el intervalo es:

$$\left]0{,}325 \cdot \frac{1{,}189222}{1{,}352308}, 2{,}8 \cdot \frac{1{,}189222}{1{,}352308}\right[=]0{,}285325, 2{,}458184[$$

Ejercicio 2.6.7. Con objeto de estudiar la efectividad de un agente diurético, se eligen al azar 11 pacientes, aplicando dicho fármaco a seis de ellos y un placebo a los cinco restantes. La variable observada en esta experiencia fue la concentración de sodio en la orina a las 24 horas, que se supone tiene una distribución normal en ambos casos. Los resultados observados fueron:

- Diurético: 20,4, 62,5, 61,3, 44,2, 11,1, 23,7.
- Placebo: 1,2, 6,9, 38,7, 20,4, 17,2.

a) Calcular un intervalo de confianza para el cociente de las varianzas al nivel de confianza 0,95.

b) Suponiendo que las varianzas son iguales, calcular un intervalo de confianza para la diferencia de las medias al nivel de confianza 0,9, y una cota inferior de confianza al mismo nivel. Interpretar los resultados.

Solución. Tenemos:

$X \equiv$ "Concentración de `Na` en la orina a las 24h tomando diurético" $\rightsquigarrow \mathcal{N}(\mu_1, \sigma_1^2)$
$Y \equiv$ "Concentración de `Na` en la orina a las 24h tomando placebo" $\rightsquigarrow \mathcal{N}(\mu_2, \sigma_2^2)$

Y tenemos dos m.a.s. de ellas: $(X_1, \ldots, X_{n_1})$ de X y $(Y_1, \ldots, Y_{n_2})$ de Y.

a) Calcular un intervalo de confianza para el cociente de las varianzas al nivel de confianza 0,95.

Buscamos un intervalo de confianza para $\frac{\sigma_1^2}{\sigma_2^2}$ a nivel de confianza $1 - \alpha = 0{,}95 \Longrightarrow \alpha = 0{,}05$. La función pivote en este caso es:

$$T\left(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}; \frac{\sigma_1^2}{\sigma_2^2}\right) = \frac{S_2^2}{S_1^2}\frac{\sigma_1^2}{\sigma_2^2} \rightsquigarrow F(n_2 - 1, n_1 - 1)$$

Que:

- Sale creciente respecto el parámetro.

- Si tomamos λ con:

$$\frac{S_2^2}{S_1^2}\frac{\sigma_1^2}{\sigma_2^2} \rightsquigarrow F(n_2-1, n_1-1) = \lambda \Longrightarrow \frac{\sigma_1^2}{\sigma_2^2} = \frac{S_1^2}{S_1^2}\lambda$$

Por lo que el intervalo general a considerar será:

$$\left]\frac{S_1^2}{S_2^2}\lambda_1, \frac{S_1^2}{S_2^2}\lambda_2\right[$$

Por lo que vimos en teoría, tendremos:

$$\lambda_1 = F_{n_2-1,n_1-1,1-\alpha/2}, \qquad \lambda_2 = F_{n_2-1,n_1-1,\alpha/2}$$

Por lo que el intervalo a considerar será:

$$\left]\frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1,1-\alpha/2}, \frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1,\alpha/2}\right[$$

Calculamos las cuasivarianzas:

$$S_1^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n_1-1} = \frac{(n-1)}{n}Var(X_1,\ldots,X_n)$$

Usando la calculadora, obtenemos:

$$S_1^2 = 483{,}12 \qquad S_2^2 = 208{,}517$$

Y consultando la tabla de la F de Snedecor:

$$F_{n_2-1,n_1-1,0{,}975} = 0{,}107, \qquad F_{n_2-1,n_1-1,0{,}025} = 7{,}39$$

Por lo que el intervalo obtenido es:

$$]0{,}2479, 17{,}1219[$$

Si lo hubiéramos hecho para $\frac{\sigma_2^2}{\sigma_1^2}$ habríamos obtenido:

$$]0{,}0583, 4{,}0398[$$

b) Suponiendo que las varianzas son iguales, calcular un intervalo de confianza para la diferencia de las medias al nivel de confianza 0,9, y una cota inferior de confianza al mismo nivel. Interpretar los resultados.

Como el valor 1 para $\frac{\sigma_1^2}{\sigma_2^2}$ está en los intervalos, no podemos descartar que las varianzas sean iguales, con lo que podemos asumirlo. En teoría vimos que el intervalo a obtener es:

$$\left]\overline{X}-\overline{Y}-t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \overline{X}-\overline{Y}+t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right[$$

Y la cota inferior es:

$$\overline{X}-\overline{Y}-t_{n_1+n_2-2,\alpha}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

donde:

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

Como tenemos los datos:

- $n_1 = 6$.
- $n_2 = 5$.
- $\alpha = 0{,}05$.
- $S_1^2 = 483{,}12$.
- $S_2^2 = 208{,}517$.
- $\overline{x} = 37{,}2$.
- $\overline{y} = 16{,}88$.

Podemos también consultar:

- $t_{n_1+n_2-2,\alpha/2} = t_{9,0{,}025} = 2{,}2622$.
- $t_{n_1+n_2-2,\alpha} = t_{9,0{,}05} = 1{,}8331$.

Por lo que:

$$S_p = \sqrt{\frac{5 \cdot 483{,}12 + 4 \cdot 16{,}88}{6+5-2}} = 16{,}610305$$

y el intervalo y la cota inferior son:

$$\left]37{,}2 - 16{,}88 - 2{,}2622 \cdot 16{,}610305 \cdot \sqrt{\frac{1}{6}+\frac{1}{5}}, 37{,}2 - 16{,}88 + \right.$$
$$\left. + 2{,}2622 \cdot 16{,}610305 \cdot \sqrt{\frac{1}{6}+\frac{1}{5}}\right[$$
$$=]-2{,}433296, 43{,}073269[$$

Vemos que 0 está en dicho intervalo, por lo que no podemos descartar que las dos medias sean iguales.

Ejercicio 2.6.8. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria con distribución $U(0, \theta)$. Dado un nivel de confianza arbitrario, calcular el intervalo de confianza para θ de menor longitud media uniformemente basado en un estadistico suficiente.

Si la muestra es de $X \rightsquigarrow U(0, \theta)$ y consideramos un nivel de confianza $1 - \alpha$ para $\alpha \in]0, 1[$, calculamos un estadístico suficiente para θ:

$$f_\theta^n(x_1, \ldots, x_n) \overset{\text{iid.}}{=} \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{\theta} = \frac{1}{\theta^n} \qquad x_i \in]0, \theta[$$

Por lo que para $0 < X_{(1)} < X_{(n)} < \theta$ tenemos:

$$f_\theta^n(x_1, \ldots, x_n) = I_{\mathbb{R}^+}(X_{(1)}) I_{\mathbb{R}^-}(X_{(n)} - \theta) \frac{1}{\theta^n}$$

Tomando:

$$h(X_1, \ldots, X_n) = I_{\mathbb{R}^+}(X_{(1)}), \qquad T(X_1, \ldots, X_n) = X_{(n)}$$
$$g_\theta(t) = I_{\mathbb{R}^-}(t - \theta) \frac{1}{\theta^n}$$

Tenemos por el Teorema de factorización de Neymann-Fisher que el estadístico $T(X_1, \ldots, X_n) = X_{(n)}$ es suficiente para θ. Calculamos su función de distribución:

$$F_T(t) = (F_X(t))^n = \left(\frac{t}{\theta}\right)^n \qquad t \in [0, \theta]$$

Por lo que una función pivote para aplicar el método de la cantidad pivotal es:

$$T(X_1, \ldots, X_n; \theta) = F_T(T) = \left(\frac{X_{(n)}}{\theta}\right)^n \rightsquigarrow U(0,1)$$

Tenemos que:

- La función es estrictamente decreciente en θ.
- Si tenemos λ de forma que:

$$\lambda = \left(\frac{X_{(n)}}{\theta}\right)^n \Longrightarrow \theta = \frac{X_{(n)}}{\sqrt[n]{\lambda}}$$

Por lo que el intervalo de confianza para θ a nivel de confianza $1-\alpha$ será:

$$\left]\frac{X_{(n)}}{\sqrt[n]{\lambda_2}}, \frac{X_{(n)}}{\sqrt[n]{\lambda_1}}\right[$$

donde λ_1 y λ_2 verifican:

$$1 - \alpha \leqslant P_\theta[\lambda_1 < T < \lambda_2] = \lambda_2 - \lambda_1$$

La longitud esperada del intervalo es:

$$E[L] = E\left[\left(\frac{1}{\sqrt[n]{\lambda_1}} - \frac{1}{\sqrt[n]{\lambda_2}}\right) X_{(n)}\right] = \left(\frac{1}{\sqrt[n]{\lambda_1}} - \frac{1}{\sqrt[n]{\lambda_2}}\right) E[X_{(n)}]$$

Con $E[X_{(n)}] \geqslant 0$, por lo que será suficiente con minimizar el primer término. Por el método de los multiplicadores de Lagrange:

$$F(\lambda_1, \lambda_2) = \left(\frac{1}{\sqrt[n]{\lambda_1}} - \frac{1}{\sqrt[n]{\lambda_2}}\right) - \lambda\left(\lambda_2 - \lambda_1 - (1-\alpha)\right)$$

Si derivamos e igualamos a cero despejando λ:

$$\frac{\partial F}{\partial \lambda_1} = \frac{-1}{n\sqrt[n]{\lambda_1^{n+1}}} + \lambda = 0 \quad \Longrightarrow \quad \lambda = \frac{1}{n\sqrt[n]{\lambda_1^{n+1}}} = \frac{1}{n\lambda_1\sqrt[n]{\lambda_1}}$$
$$\frac{\partial F}{\partial \lambda_2} = \frac{1}{n\sqrt[n]{\lambda_2^{n+1}}} - \lambda = 0 \quad \Longrightarrow \quad \lambda = \frac{1}{n\sqrt[n]{\lambda_2^{n+1}}} = \frac{1}{n\lambda_2\sqrt[n]{\lambda_2}}$$

Ejercicio 2.6.9. Utilizando la desigualdad de Chebychev, dar un intervalo de confianza para p a nivel de confianza arbitrario, basado en una muestra de tamaño arbitrario de una variable aleatoria con distribución $B(1, p)$.

Sea $(X_1, \ldots, X_n)$ una m.a.s. de $X \rightsquigarrow B(1,p)$, buscamos por el método de Chebyshev un intervalo de confianza para p a nivel de confianza $1-\alpha$. Para ello, es necesario tener un estimador insesgado de p. Si consideramos:

$$T(X_1, \ldots, X_n) = \overline{X}$$

tenemos que $\overline{X} \in [0,1]$, así como que $\overline{X}$ es insesgado, puesto que:

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} p = \frac{np}{n} = p$$

Usando la reproductividad de la binomial vemos que:

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Y si tomamos c una constante de forma que:

$$\frac{p(1-p)}{n} \leqslant c \qquad \forall p \in]0,1[$$

tenemos entonces que el intervalo:

$$\left]\overline{X} - \sqrt{\frac{c}{\alpha}}, \overline{X} + \sqrt{\frac{c}{\alpha}}\right[$$

Es un intervalo de confianza para p a nivel de confianza $1-\alpha$.

Ejercicio 2.6.10. Para una muestra de tamaño n de una variable aleatoria con función de densidad

$$f_\theta(x) = \frac{2x}{\theta^2}, \qquad 0 < x < \theta$$

encontrar el intervalo de confianza para θ de menor longitud media uniformemente a nivel de confianza $1-\alpha$, basado en un estadístico suficiente.

Buscamos un estadístico suficiente de θ:

$$f_\theta^n(X_1, \ldots, X_n) \overset{\text{iid.}}{=} \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \frac{2X_i}{\theta^2} = \frac{2^n}{\theta^{2n}} \prod_{i=1}^{n} X_i \qquad 0 < X_i < \theta$$

Por lo que tomando $0 < X_{(1)} < X_{(n)} < \theta$ tenemos:

$$f_\theta^n(X_1, \ldots, X_n) = \frac{2^n}{\theta^{2n}} I_{\mathbb{R}^+}(X_{(1)}) I_{\mathbb{R}^-}(X_{(n)} - \theta) \prod_{i=1}^{n} X_i$$

Si tomamos:

$$h(X_1, \ldots, X_n) = I_{\mathbb{R}^+}(X_{(1)}) \prod_{i=1}^{n} X_i, \qquad T(X_1, \ldots, X_n) = X_{(n)}$$

$$g_\theta(t) = \frac{2^n}{\theta^{2n}} I_{\mathbb{R}^-}(t - \theta)$$

Tenemos por el Teorema de factorización de Neymann-Fisher que $X_{(n)}$ es un estadístico suficiente para θ. Calculamos su función de distribución:

$$F_T(t) = (F_X(t))^n$$

Y para ello vemos que tenemos que calcular F_X:

$$F_X(t) = \int_0^t \frac{2x}{\theta^2}\, dx = \left[\frac{x^2}{\theta^2}\right]_0^t = \frac{t^2}{\theta^2}$$

Por lo que:

$$F_T(t) = (F_X(t))^n = \left(\frac{t}{\theta}\right)^{2n}$$

La función pivote que consideramos para aplicar el método de la cantidad pivotal es:

$$T(X_1, \ldots, X_n; \theta) = F_T(T) = \left(\frac{X_{(n)}}{\theta}\right)^{2n} \rightsquigarrow U(0,1)$$

Tenemos que:

- Es estrictamente decreciente respecto θ.
- Si tenemos λ de forma que:

$$\lambda = \left(\frac{X_{(n)}}{\theta}\right)^{2n} \quad \Longrightarrow \quad \theta = \frac{X_{(n)}}{\sqrt[2n]{\lambda}}$$

Tenemos por tanto que el intervalo es:

$$\left]\frac{X_{(n)}}{\sqrt[2n]{\lambda_2}}, \frac{X_{(n)}}{\sqrt[2n]{\lambda_1}}\right[$$

donde λ_1 y λ_2 verifican:

$$1 - \alpha \leqslant P_\theta[\lambda_1 < T < \lambda_2] = F_T(\lambda_2) - F_T(\lambda_1) = \lambda_2 - \lambda_1$$

La longitud esperada del intervalo es:

$$E[L] = E\left[\left(\frac{1}{\sqrt[2n]{\lambda_1}} - \frac{1}{\sqrt[2n]{\lambda_2}}\right) X_{(n)}\right] = \left(\frac{1}{\sqrt[2n]{\lambda_1}} - \frac{1}{\sqrt[2n]{\lambda_2}}\right) E[X_{(n)}]$$

Con $E[X_{(n)}] \geqslant 0$, por lo que será suficiente con minimizar el primer término. Si despejamos λ_2:

$$\lambda_2 = 1 - \alpha + \lambda_1$$

Tenemos que la función a minimizar es:

$$F(\lambda_1) 0 \lambda_1^{\frac{-1}{2n}} - (1 - \alpha + \lambda_1)^{\frac{-1}{2n}}$$

Calculamos su derivada e igualdando a cero:

$$F'(\lambda_1) = \frac{-1}{2n}\lambda_1^{\frac{-1}{2n}-1} + \frac{1}{2n}(1 - \alpha + \lambda_1)^{\frac{-1}{2n}-1} = 0$$

Tenemos que $F'(\lambda_1) = 0 \iff \lambda_1 = 1 - \alpha + \lambda_1 \iff \alpha = 1$. Por lo que no podemos encontrar una solución, luego F es estrictamente monótona:

- Supuesto que F es estrictamente crecientes, tenemos entonces que $\lambda_1 = 0$ y $\lambda_2 = 1 - \alpha$, por lo que el intervalo es:

$$\left] \frac{X_{(n)}}{\sqrt[2n]{1-\alpha}}, +\infty \right[$$

- Supuesto que F es estrictamente decreciente, tenemos que $\lambda_2 = 1$ y $\lambda_1 = \alpha$, por lo que el intervalo es:

$$\left] X_{(n)}, \frac{X_{(n)}}{\sqrt[2n]{\alpha}} \right[$$

Como en la segunda opción nos sale un intervalo acotado, este es el de menor longitud esperada, por lo que nos quedamos con:

$$\left] X_{(n)}, \frac{X_{(n)}}{\sqrt[2n]{\alpha}} \right[$$

Ejercicio 2.6.11. Para una muestra de tamaño n de una variable aleatoria con función de densidad

$$f_\theta(x) = \frac{\theta}{x^2}, \qquad x > \theta$$

encontrar el intervalo de confianza para θ de menor longitud media uniformemente a nivel de confianza $1 - \alpha$, basado en el estimador máximo verosímil de θ.

La función de verosimilitud es:

$$L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^2} \qquad \forall x_i > \theta > 0$$

Por lo que:

$$L_{x_1,\ldots,x_n}(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{\theta}{x_i^2} & \text{si } \theta < x_{(1)} \\ 0 & \text{en otro caso} \end{cases}$$

Y vemos que $L_{x_1,\ldots,x_n}(\theta)$ es creciente, con lo que alcanza su máximo en $\theta = x_{(1)}$. En definitiva, el EMV es $\hat{\theta} = X_{(1)}$, por lo que buscamos un intervalo de confianza basado en $X_{(1)}$.

La función de distribución del mínimo es:

$$F_{X_{(1)}}(t) = 1 - (1 - F_X(t))^n$$

Por lo que calculamos:

$$F_X(t) = \int_\theta^t \frac{\theta}{x^2}\, dx = \left[\frac{-\theta}{x}\right]_\theta^t = \frac{-\theta}{t} + 1 \qquad t > \theta$$

de donde:

$$F_{X_{(1)}}(t) = 1 - (1 - F_X(t))^n = 1 - \left(1 - \left(1 - \frac{\theta}{t}\right)\right)^n = 1 - \left(\frac{\theta}{t}\right)^n \qquad t > \theta$$

Tomamos por tanto como función pivote:

$$T(X_1, \dots, X_n; \theta) = 1 - \left(\frac{\theta}{X_{(1)}}\right)^n \rightsquigarrow U(0,1)$$

Comprobamos las condiciones del método de la cantidad pivotal:

- Si derivamos:

$$\frac{\partial T}{\partial \theta} = -n\left(\frac{\theta}{X_{(1)}}\right)^{n-1} \frac{1}{X_{(1)}} = \frac{-n\theta^{n-1}}{X_{(1)}^n} < 0 \qquad \forall \theta$$

  Por lo que $T(X_1, \dots, X_n; \theta)$ es estrictamente decreciente en función de θ.

- Si tratamos de despejar el parámetro para cierto λ:

$$1 - \left(\frac{\theta}{X_{(1)}}\right)^n = \lambda \Longrightarrow \theta = \sqrt[n]{(1-\lambda)X_{(1)}^n} = X_{(1)}\sqrt[n]{1-\lambda}$$

Por lo que hemos obtenido como intervalo de confianza (donde los índices de λ son debidos a que es estrictamente decreciente):

$$\left]X_{(1)}\sqrt[n]{1-\lambda_2}, X_{(1)}\sqrt[n]{1-\lambda_2}\right[$$

Tratamos ahora de minimizar la longitud del intervalo, que es:

$$L = X_{(1)}\left(\sqrt[n]{1-\lambda_1} - \sqrt[n]{1-\lambda_2}\right)$$

de donde:

$$E_\theta[L] = E_\theta[X_{(1)}]\left(\sqrt[n]{1-\lambda_1} - \sqrt[n]{1-\lambda_2}\right)$$

Y como $E_\theta(X_{(1)})$ es una constante positiva, podemos obviarla a la hora de minimizar, por lo que buscamos minimizar el segundo trozo, sujeto a la restricción:

$$P_\theta[\lambda_1 < T < \lambda_2] = 1 - \alpha$$

Y como tenemos:

$$P[\lambda_1 < T < \lambda_2] = F_T(\lambda_2) - F_T(\lambda_1) = \lambda_2 - \lambda_1$$

Por lo que tenemos la restricción $\lambda_2 - \lambda_1 = 1 - \alpha$. Para meterla en la función a minimizar:

- Bien despejamos λ_1 o λ_2, sustituimos en la función a minimizar y obtenemos una función de una variable, que sabemos minimizar.

- Usamos los multiplicadores de Lagrange, sumamos a la función la restricción (igualdada a 0) multiplicada por un cierto parámetro, que llamaremos λ, por lo que buscamos minimizar:

$$(1-\lambda_1)^{1/n} - (1-\lambda_2)^{1/n} + \lambda(\lambda_2 - \lambda_1 - 1 + \alpha)$$

Cuando las restricciones son tan sencillas el método de despejar y sustituir es más sencillo, por lo que optamos por dicho método:

$$\lambda_2 = 1 - \alpha + \lambda_1$$

Sustituyendo en la función a minimizar, buscamos la constante λ_1 que minimiza la expresión:

$$f(\lambda_1) = (1-\lambda_1)^{1/n} - (1-(a-\alpha+\lambda_1))^{1/n} = (1-\lambda_1)^{1/n} - (\alpha-\lambda_1)^{1/n}$$

Calculamos la derivada y la igualamos a 0, para despejar λ_1:

$$f'(\lambda_1) = \frac{1}{n}(1-\lambda_1)^{1/n-1}(-1) - \frac{1}{n}(\alpha-\lambda_1)^{1/n-1}(-1) = 0$$

Haciendo cálculos llegamos a que $f'(\lambda_1) = 0$ si y solo si $\alpha = 1$, pero $\alpha \in]0,1[$, por lo que $f'(\lambda_1) \neq 0$, con lo que f es estrictamente monónotona. Sabemos también que $0 < \lambda_1 < \lambda_2 < 1$, buscamos razonar si f es estrictamente creciente o estrictamente decreciente. Para ello, podemos estudiar el signo de la derivada o bien razonarlo de la siguiente forma:

- Si la función es estrictamente creciente, esta alcanza su mínimo en $\lambda_1 = 0$, con lo que usando que $\lambda_2 - \lambda_1 = 1 - \alpha$ tenemos que $\lambda_2 = 1 - \alpha$. Sustituyendo, obtenemos el intervalo:

$$\left]X_{(1)}\sqrt[n]{1-(1-\alpha)}, X_{(1)}\sqrt[n]{1-0}\right[= \left]X_{(1)}\sqrt[n]{\alpha}, X_{(1)}\right[$$

- Si la función es estrictamente decreciente, entonces esta alcana su mínimo cuando $\lambda_2 = 1$, con lo que $\lambda_1 = \alpha$, de donde obtenemos el intervalo:

$$\left]X_{(1)}\sqrt[n]{0}, X_{(1)}\sqrt[n]{1-\alpha}\right[= \left]0, X_{(1)}\sqrt[n]{1-\alpha}\right[$$

El segundo intervalo no está acotado por debajo (ya que el mínimo a considerar en el espacio paramétrico es 0), pero en el primero sí que tenemos un valor acotado por debajo y otro por encima. Preferimos elegir el primer intervalo, puesto que restringe más la longitud del intervalo (es decir, lo acota por ambos lados), aunque puede ocurrir que en función de los valores de $X_{(1)}$ y α el intervalo mínimo sea a veces el primero y otras veces el segundo.

2.7. Contraste de hipótesis

Ejercicio 2.7.1. Se toma una observación de una variable con distribución de Poisson para contrastar que la media vale 1 frente a que vale 2.

a) Construir un test no aleatorizado con nivel de significación 0,05 para el contraste planteado. Calcular las probabilidades de cometer error de tipo 1 y de tipo 2, el tamaño y la potencia del test frente a la hipótesis alternativa.

b) ¿Cómo debe aleatorizarse el test para alcanzar el tamaño 0,05? ¿Cuál es la potencia de este test?

Solución

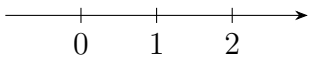
a) Sea $X \rightsquigarrow \mathcal{P}(\lambda)$, queremos resolver el contraste:

$$\begin{cases} H_0 : \lambda = 1 \\ H_1 : \lambda = 2 \end{cases}$$

Trabajaremos por tanto con:

$$\Theta_0 = \{1\}, \qquad \Theta_1 = \{2\}, \qquad \Theta = \{1, 2\}$$

y escribiremos $\lambda_0 = 1$, $\lambda_1 = 2$. Pensamos la forma que ha de tener el test: el espacio muestral de X es $\mathbb{R}_0^+$:



Es claro que cuanto más a la izquierda nos encontremos con dicha observación X, no rechazaremos la hipótesis nula, así como que cuanto más a la derecha estemos la rechazaremos. Por tanto, establecemos un cierto punto $c \in \mathbb{R}_0^+$ que delimite la región crítica del test (que será la región $]c, +\infty[$). Para determinar dicho punto c imponemos un nivel de significación $\alpha = 0{,}05$. Por tanto, el test será de la forma:

$$\varphi(X) = \begin{cases} 1 & \text{si } X > c \\ 0 & \text{si } X \leqslant c \end{cases}$$

Imponemos nivel de significancia α:

$$\alpha \geqslant \sup_{\lambda \in \Theta_0} \beta_\varphi(\lambda) = \sup_{\lambda \in \{\lambda_0\}} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_0) = E_{\lambda_0}[\varphi] = P_{\lambda_0}[X > c] = 1 - P_{\lambda_0}[X \leqslant c]$$

Y lo que hacemos ahora es ir probando con distintos valores de c hasta acercarnos lo máximo posible a $\alpha = 0{,}05$ pero sin pasarnos:

- Para $c = 0$:

$$P_1[X \leqslant 0] = P_1[X = 0] = 0{,}3679 \quad \Longrightarrow \quad 1 - P_1[X \leqslant 0] = 0{,}6321 > \alpha$$

- Para $c = 1$:

$$P_1[X \leqslant 1] = P_1[X = 0] + P_1[X = 1] = 0{,}3679 + 0{,}3679 = 0{,}7358$$
$$\Longrightarrow 1 - P_1[X \leqslant 1] = 0{,}2642 > \alpha$$

- Para $c = 2$:

$$P_1[X \leqslant 2] = P_1[X \leqslant 1] + P_1[X = 2] = 0{,}7358 + 0{,}1839 = 0{,}9197$$
$$\Longrightarrow 1 - P_1[X \leqslant 2] = 0{,}0803 > \alpha$$

- Para $c = 3$:

$$P_1[X \leqslant 3] = P_1[X \leqslant 2] + P_1[X = 3] = 0{,}9197 + 0{,}0613 = 0{,}981$$
$$\Longrightarrow 1 - P_1[X \leqslant 3] = 0{,}019 \leqslant \alpha$$

Por tanto, el valor a tomar es $c = 3$, de donde el test sería:

$$\varphi(X) = \begin{cases} 1 & \text{si } X > 3 \\ 0 & \text{si } X \leqslant 3 \end{cases}$$

Calculamos ahora:

Probabilidad de cometer error tipo 1.

$$P_{\lambda \in \Theta_0}[X \in C] = P_{\lambda_0}[X > 3] = P_1[X > 3] = 0{,}019$$

Probabilidad de cometer error tipo 2.

$$P_{\lambda \in \Theta_1}\left[X \in \overline{C}\right] = P_{\lambda_1}[X \leqslant 3] = P_2[X \leqslant 3] = \sum_{k=0}^{3} P_2[X = k]$$
$$= (0{,}1353 + 0{,}2707 + 0{,}2707 + 0{,}1804) = 0{,}8571$$

Tamaño del test. El tamaño resulta ser igual a la probabilidad encontrada cuando buscábamos el valor de c:

$$\sup_{\lambda \in \Theta_0} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_0) = E_{\lambda_0}[\varphi] = P_{\lambda_0}[X > 3] = 0{,}019$$

Potencia frente a hipótesis alternativa. La potencia del test es:

$$\sup_{\lambda \in \Theta_1} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_1) = E_{\lambda_1}[\varphi] = P_{\lambda_1}[X > 3] = P_2[X > 3] = 1 - P_2[X \leqslant 3]$$
$$= 1 - 0{,}8571 = 0{,}1429$$

b) Para alcanzar un tamaño de 0,05 mediante un test aleatoriezado, lo que hacemos ahora es considerar un test del tipo:

$$\varphi(X) = \begin{cases} 1 & \text{si } X > 3 \\ \gamma & \text{si } X = 3 \\ 0 & \text{si } X < 3 \end{cases}$$

Para cierto $\gamma \in [0, 1]$, de forma que a la hora de imponer un tamaño menor o igual que $\alpha = 0{,}05$ calculemos γ para obtener la igualdad. De esta forma, el tamaño del test es:

$$\sup_{\lambda \in \Theta_0} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_0) = E_{\lambda_0}[\varphi] = P_{\lambda_0}[X > 3] + \gamma P_{\lambda_0}[X = 3] = 0{,}019 + \gamma 0{,}0613$$

Si imponemos que sea igual a α:

$$0{,}05 = \alpha = 0{,}019 + \gamma \cdot 0{,}0613 \quad \Longrightarrow \quad \gamma = \frac{0{,}05 - 0{,}019}{0{,}0613} \approx 0{,}5057096$$

Calculamos la potencia del test:

$$\sup_{\lambda \in \Theta_1} \beta_\varphi(\lambda) = E_{\lambda_1}[\varphi] = P_{\lambda_1}[X > 3] + \gamma P_{\lambda_1}[X = 3] = 0{,}1429 + \gamma 0{,}1804 \approx 0{,}23413$$

Ejercicio 2.7.2. Una urna contiene 10 bolas, blancas y negras. Para contrastar que el número de bolas blancas es 5 frente a que dicho número es 6 o 7, se extraen tres bolas con reemplazamiento y se rechaza H_0 sólo si se obtienen 2 o 3 bolas blancas. Calcular el tamaño de este test y la potencia frente a las alternativas.

Sea:

$$X \equiv \text{“Número de bolas blancas en una extracción”} \rightsquigarrow B(1, p)$$

Se quiere resolver el contraste de hipótesis:

$$\begin{cases} H_0 : p = p_0 = 5/10 \\ H_1 : p \in \{6/10, 7/10\} \end{cases}$$

Por lo que consideraremos:

$$\Theta_0 = \{5/10\}, \qquad \Theta_1 = \{6/10, 7/10\}, \qquad \Theta = \{5/10, 6/10, 7/10\}$$

Y tenemos (X_1, X_2, X_3) una m.a.s. de X. Planteamos el test:

$$\varphi(X_1, X_2, X_3) = \begin{cases} 1 & \text{si } X_1 + X_2 + X_3 \in \{2, 3\} \\ 0 & \text{en otro caso} \end{cases}$$

Por la reproductividad de la binomial tenemos que $X_1 + X_2 + X_3 \rightsquigarrow B(3, p)$. Calculamos:

Tamaño del test.

$$\begin{aligned} \sup_{p \in \Theta_0} \beta_\varphi(p) = \beta_\varphi(p_0) = E_{p_0}[\varphi] &= P_{p_0}[X_1 + X_2 + X_3 = 2] + P_{p_0}[X_1 + X_2 + X_3 = 3] \\ &= P_{1/2}[X_1 + X_2 + X_3 = 2] + P_{1/2}[X_1 + X_2 + X_3 = 3] \\ &= 0{,}375 + 0{,}125 = 0{,}5 \end{aligned}$$

Potencia frente a hipótesis alternativas. Calculamos cada una de ellas:

- Para $p = 6/10$:

$$\begin{aligned} \beta_\varphi(p) = E_p[\varphi] &= P_p[X_1 + X_2 + X_3 = 2] + P_p[X_1 + X_2 + X_3 = 3] \\ &= P_{0{,}6}[X_1 + X_2 + X_3 = 2] + P_{0{,}6}[X_1 + X_2 + X_3 = 3] \end{aligned}$$

  Si consideramos $Y \rightsquigarrow B(3, 1 - p) = B(3,\ 0{,}4)$ tenemos:

$$\begin{aligned} \beta_\varphi(p) &= P_{0{,}6}[X_1 + X_2 + X_3 = 2] + P_{0{,}6}[X_1 + X_2 + X_3 = 3] \\ &= P[Y = 3 - 2] + P[Y = 3 - 3] = P[Y = 1] + P[Y = 0] \\ &= 0{,}432 + 0{,}216 = 0{,}648 \end{aligned}$$

- para $p = {}^{7}/_{10}$:

$$\begin{aligned}\beta_\varphi(p) &= P_p[X_1 + X_2 + X_3 = 2] + P_p[X_1 + X_2 + X_3 = 3] \\ &= P_{0,7}[X_1 + X_2 + X_3 = 2] + P_{0,7}[X_1 + X_2 + X_3 = 3]\end{aligned}$$

Si consideramos $Y \rightsquigarrow B(3, 1-p) = B(3,\ 0{,}3)$ tenemos:

$$\begin{aligned}\beta_\varphi(p) &= P_{0,7}[X_1 + X_2 + X_3 = 2] + P_{0,7}[X_1 + X_2 + X_3 = 3] \\ &= P[Y = 1] + P[Y = 0] = 0{,}441 + 0{,}343 = 0{,}784\end{aligned}$$

Ejercicio 2.7.3. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de una variable aleatoria con distribución de Poisson de parámetro λ. Encontrar el test más potente de tamaño α para resolver el problema de contraste

$$\begin{aligned}H_0 &: \lambda = \lambda_0 \\ H_1 &: \lambda = \lambda_1\end{aligned}$$

Aplicación: En una centralita telefónica el número de llamadas por minuto sigue una distribución de Poisson. Si en cinco minutos se han recibido 12 llamadas, ¿puede aceptarse que el número medio de llamadas por minuto es 1,5, frente a que dicho número es 2, al nivel de significación 0,05? Calcular la potencia del test obtenido.

Sea $X \rightsquigarrow \mathcal{P}(\lambda)$, en este caso tenemos:

$$\Theta_0 = \{\lambda_0\}, \qquad \Theta_1 = \{\lambda_1\}, \qquad \Theta = \{\lambda_0, \lambda_1\}$$

Como nos encontramos ante un contraste simple frente a simple y queremos hayar el test más potente de tamaño α, recurrimos a un test de Neymann-Pearson, que será de la forma:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \lambda(X_1, \ldots, X_n) > k \\ \gamma & \text{si } \lambda(X_1, \ldots, X_n) = k \\ 0 & \text{si } \lambda(X_1, \ldots, X_n) < k \end{cases}$$

Para ciertos valores $\gamma \in [0, 1]$, $k \in \mathbb{R}$ y una cierta función λ que nos disponemos a calcular. Primero observemos que:

$$P_\lambda[X_1 = x_1, \ldots, X_n = x_n] \overset{\text{iid.}}{=} \prod_{i=1}^{n} P_\lambda[X = x_i] = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \cdot \frac{\lambda^{\sum_{k=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} > 0$$

Por lo que:

$$\lambda(x_1, \ldots, x_n) = \frac{P_{\lambda_1}[X_1 = x_1, \ldots, X_n = x_n]}{P_{\lambda_0}[X_1 = x_1, \ldots, X_n = x_n]} = \frac{e^{-n\lambda_1} \cdot \frac{\lambda_1^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}}{e^{-n\lambda_0} \cdot \frac{\lambda_0^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}} = e^{n(\lambda_0 - \lambda_1)} \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=1}^{n} x_i}$$

Y para simplificar el test buscamos uno equivalente, si consideramos el estadístico $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i \rightsquigarrow \mathcal{P}(n\lambda)$, estudiamos el comportamiento de:

$$g(t) = e^{n(\lambda_0 - \lambda_1)} \left(\frac{\lambda_1}{\lambda_0} \right)^t$$

Para ello:

$$\ln g(t) = n(\lambda_0 - \lambda_1) + t \ln \left(\frac{\lambda_1}{\lambda_0} \right)$$

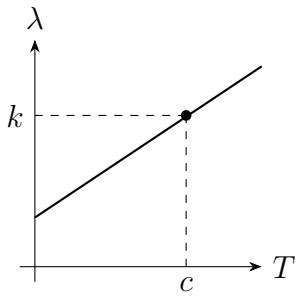
$$g'(t) = \ln \left(\frac{\lambda_1}{\lambda_0} \right)$$

Y observamos que:

$$g'(t) > 0 \iff \ln \left(\frac{\lambda_1}{\lambda_0} \right) > 0 \iff \frac{\lambda_1}{\lambda_0} > 1 \iff \lambda_1 > \lambda_0$$

Por lo que distinguimos casos (sabemos que $\lambda_0 \neq \lambda_1$):

Supuesto que $\lambda_0 < \lambda_1$. Tenemos entonces que g es estrictamente creciente:



Por tanto:

$$\left.\begin{array}{l} \lambda > k \\ \lambda = k \\ \lambda < k \end{array}\right\} \iff \left\{\begin{array}{l} T > k \\ T = k \\ T < k \end{array}\right.$$

De donde el test será:

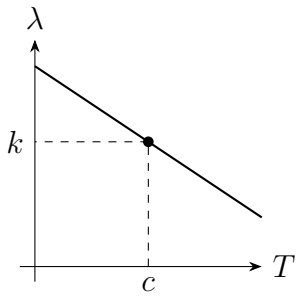
$$\varphi'(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^{n} X_i > c \\ \gamma & \text{si } \sum_{i=1}^{n} X_i = c \\ 0 & \text{si } \sum_{i=1}^{n} X_i < c \end{cases}$$

Imponiendo tamaño α:

$$\alpha = \sup_{\lambda \in \Theta_0} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_0) = E_{\lambda_0}[\varphi] = P_{\lambda_0}[T > c] + \gamma P_{\lambda_0}[T = c]$$

Y lo que haremos será para cierto α calcular los valores de c y γ.

Supuesto que $\lambda_1 < \lambda_0$. En este caso tendremos que g es estrictamente decreciente:



Por lo que obtendremos un test equivalente de la forma:

$$\varphi'(X_1,\dots,X_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^n X_i < c \\ \gamma & \text{si } \sum_{i=1}^n X_i = c \\ 0 & \text{si } \sum_{i=1}^n X_i > c \end{cases}$$

Imponiendo tamaño α:

$$\alpha = \sup_{\lambda\in\Theta_0} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_0) = E_{\lambda_0}[\varphi] = P_{\lambda_0}[T < c] + \gamma P_{\lambda_0}[T = c]$$

Y lo que haremos será para cierto α calcular los valores de c y γ.

Aplicación. Sea:

$$X \equiv \text{“Número de llamadas por minuto”} \rightsquigarrow \mathcal{P}(\lambda)$$

En 5 observaciones $(x_1,\dots,x_5)$ se ha obtenido:

$$\sum_{i=1}^n x_i = 12$$

Y planteamos el contraste de hipótesis:

$$\begin{cases} H_0 : \lambda = 1{,}5 \\ H_1 : \lambda = 2 \end{cases}$$

Sea $\alpha = 0{,}05$, como $\lambda_0 < \lambda_1$ nos encontramos en el primer caso, por lo que usamos el test:

$$\varphi'(X_1,\dots,X_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^n X_i > c \\ \gamma & \text{si } \sum_{i=1}^n X_i = c \\ 0 & \text{si } \sum_{i=1}^n X_i < c \end{cases}$$

Imponiendo tamaño α:

$$0{,}05 = \alpha = \sup_{\lambda \in \Theta_0} \beta_\varphi(\lambda) = P_{\lambda_0}[T > c] + \gamma P_{\lambda_0}[T = c] = P_{1,5}[T > c] + \gamma P_{1,5}[T = c]$$

Recordamos que $T \rightsquigarrow \mathcal{P}(n\lambda)$, y como nos encontramos bajo hipótesis nula, tenemos que $T \rightsquigarrow \mathcal{P}(5 \cdot 1{,}5) \equiv \mathcal{P}(7{,}5)$. Buscamos el primer valor de c de forma que se tenga $1 - P[T \leqslant c] = P[T > c] \leqslant \alpha$:

- Para $c = 12$, se tiene:

$$P[T > c] = 0{,}0427 \leqslant 0{,}05 = \alpha$$

- Para $c = 11$ se tiene que $P[T > c]$ es claramente mayor que 0,05

Por lo que tomamos $c = 12$. Calculamos ahora γ:

$$0{,}05 = 0{,}0427 + \gamma \cdot 0{,}0366 \quad \Longrightarrow \quad \gamma = \frac{0{,}05 - 0{,}0427}{0{,}0366} \approx 0{,}199454$$

De esta forma, el test a usar es:

$$\varphi'(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^{n} X_i < 12 \\ 0{,}199454 & \text{si } \sum_{i=1}^{n} X_i = 12 \\ 0 & \text{si } \sum_{i=1}^{n} X_i > 12 \end{cases}$$

Como $\sum_{i=1}^{5} x_i = 12$, tenemos que:

$$\varphi'(x_1, \ldots, x_5) = 0{,}199454$$

Por lo que rechazamos la hipótesis nula con dicha probabilidad. Lo mejor sería repetir el experimento para obtener una decisión clara. Calculamos ahora la potencia del test:

$$\sup_{\lambda \in \Theta_1} \beta_\varphi(\lambda) = \beta_\varphi(\lambda_1) = E_{\lambda_1}[\varphi] = P_{\lambda_1}[T > 12] + 0{,}199454 \cdot P_{\lambda_1}[T = 12]$$

Recordemos que $T \rightsquigarrow \mathcal{P}(5\lambda_1) \equiv \mathcal{P}(10)$:

$$\sup_{\lambda \in \Theta_1} \beta_\varphi(\lambda) = P_2[T > 12] + 0{,}199454 \cdot P_2[T = 12] = 0{,}227408$$

Ejercicio 2.7.4. Dada una muestra de tamaño n de una variable con distribución $\mathcal{N}(\mu, \sigma_0^2)$, deducir el test más potente de tamaño arbitrario para contrastar hipótesis simples sobre μ.

Ejercicio 2.7.5. Dada una muestra de tamaño n de una variable aleatoria con distribución $U(-\theta, \theta)$, deducir el test más potente de tamaño α para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ y calcular su potencia. ¿Cuál es el test óptimo fijado un nivel de significación arbitrario?

Ejercicio 2.7.6. Deducir el test más potente de tamaño arbitrario para contrastar $h_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, basándose en una muestra de tamaño n de una variable aleatoria con función de densidad

$$f_\theta(x) = \frac{\theta}{x^2}, \qquad x > \theta$$

Deducir el test óptimo para un nivel de significación arbitrario.

Ejercicio 2.7.7. Construir el test de Neyman-Pearson de tamaño α para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, basándose en una muestra de tamaño n de una variable aleatoria con función de densidad

$$f_\theta(x) = \frac{1}{x \ln \theta}, \qquad 1 < x < \theta$$

Deducir el test óptimo para un nivel de significación arbitrario.

En este caso tenemos:

$$\Theta_0 = \{\theta_0\}, \qquad \Theta_1 = \{\theta_1\}, \qquad \Theta = \{\theta_0, \theta_1\}$$

Sea X una variable aleatoria cuya función de densidad es la enunciada y sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de la misma, el test de Neymann-Pearson de tamaño α para el contraste de hipótesis anunciadas será de la forma:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \lambda(X_1, \ldots, X_n) > k \\ \gamma & \text{si } \lambda(X_1, \ldots, X_n) = k \\ 0 & \text{si } \lambda(X_1, \ldots, X_n) < k \end{cases}$$

Para ciertas constantes γ y k. Calculamos en primer lugar:

$$\begin{aligned} f_\theta^n(x_1, \ldots, x_n) &= \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{I_{\mathbb{R}^+}(x_{(1)} - 1) I_{\mathbb{R}^-}(x_{(n)} - \theta)}{x_i \ln \theta} \\ &= \left((\ln \theta)^n \prod_{i=1}^n x_i \right)^{-1} I_{\mathbb{R}^+}(x_{(1)} - 1) I_{\mathbb{R}^-}(x_{(n)} - \theta) \end{aligned}$$

Y vemos que $f_\theta^n(x_1, \ldots, x_n) \neq 0 \iff x_{(1)} > 1 \wedge x_{(n)} < \theta$. Para dichas realizaciones muestrales calculamos:

$$\lambda(x_1, \ldots, x_n) = \frac{f_{\theta_0}^n(x_1, \ldots, x_n)}{f_{\theta_1}^n(x_1, \ldots, x_n)} = \frac{(\ln \theta_1)^n \prod_{i=1}^n x_i}{(\ln \theta_0)^n \prod_{i=1}^n x_i} = \left(\frac{\ln \theta_1}{\ln \theta_0} \right)^n$$

Ejercicio 2.7.8. Sea X una observación de una variable aleatoria con función de densidad

$$f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}, \qquad x > \theta$$

Construir el test de razón de verosimilitudes de tamaño α arbitrario para contrastar

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta < \theta_1 \end{aligned}$$

Ejercicio 2.7.9. En base a una observación de $X \rightsquigarrow \{B(n,p) : p \in]0,1[\}$, deducir el test de razón de verosimilitudes para contrastar la hipótesis de que el parámetro p no supera un determinado valor, p_0.

Planteamos el contraste de hipótesis:

$$\begin{cases} H_0 : p \leqslant p_0 \\ H_1 : p > p_0 \end{cases}$$

Por lo que tendremos:

$$\Theta_0 =]0, p_0]\,, \qquad \Theta_1 =]p_0, 1[\,, \qquad \Theta =]0,1[$$

Calculamos el EMV de p para $x \in \mathcal{X}$:

$$L_x(p) = P_p[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\ln L_x(p) = \ln \binom{n}{x} + x \ln p + (n-x)\ln(1-p)$$

$$\frac{\partial \ln L_x(p)}{\partial p} = \frac{x}{p} + \frac{x-n}{1-p} = \frac{x - xp + xp - np}{p(1-p)} = \frac{x-np}{p(1-p)} = 0 \iff p = \frac{x}{n}$$

De esta forma:

$$\sup_{p \in \Theta_0} L_x(p) = \begin{cases} L_x(x/n) & \text{si } p_0 \geqslant x/n \\ L_x(p_0) & \text{si } p_0 < x/n \end{cases}$$

Por lo que:

$$\lambda(x) = \frac{\sup\limits_{p \in \Theta_0} L_x(p)}{\sup\limits_{p \in \Theta} L_x(p)} = \begin{cases} 1 & \text{si } p_0 \geqslant x/n \\ \dfrac{L_x(p_0)}{L_x(x/n)} & \text{si } p_0 < x/n \end{cases}$$

Y calculamos para el caso $p_0 < \frac{x}{n}$ dicha fracción:

$$\frac{L_x(p_0)}{L_x(x/n)} = \frac{\binom{n}{x} p_0^x (1-p_0)^{n-x}}{\binom{n}{x}\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} = \left(\frac{np_0}{x}\right)^x \left(\frac{1-p_0}{1-\frac{x}{n}}\right)^{n-x}$$

Y tratamos de ver si es creciente o decreciente en función de x, para lo que consideramos:

$$g(x) = \left(\frac{np_0}{x}\right)^x \left(\frac{1-p_0}{1-\frac{x}{n}}\right)^{n-x}$$

$$\ln g(x) = x\ln(np_0) - x\ln x + (n-x)\ln(1-p_0) - (n-x)\ln\left(1-\frac{x}{n}\right)$$

$$\frac{\partial \ln g(x)}{\partial x} = \ln(np_0) - \ln x \cancel{-1} - \ln(1-p_0) + \ln\left(1-\frac{x}{n}\right) \cancel{+1} = \ln\left(\frac{np_0}{x}\right) + \ln\left(\frac{1-\frac{x}{n}}{1-p_0}\right)$$

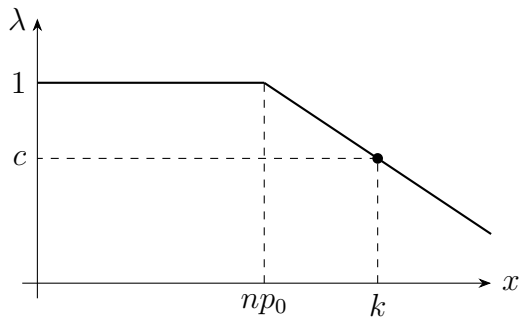
Para el primer sumando:

$$p_0 < \frac{x}{n} \iff np_0 < x \iff \frac{np_0}{x} < 1 \iff \ln\left(\frac{np_0}{x}\right) < 0$$

Para el segundo:

$$p_0 < \frac{x}{n} \iff -p_0 > -\frac{x}{n} \iff 1 - p_0 > 1 - \frac{x}{n} \iff \frac{1 - \frac{x}{n}}{1 - p_0} < 1 \iff \ln\left(\frac{1 - \frac{x}{n}}{1 - p_0}\right) < 0$$

Por tanto tenemos que g es estrictamente decreciente, de donde si consideramos λ como función de x ($p_0 \geqslant {}^{x}/_{n} \iff x \leqslant np_0$):



Tendremos que:

$$\left.\begin{array}{l} \lambda < c \\ \lambda \geqslant c \end{array}\right\} \iff \left\{\begin{array}{ll} x > k & \\ x \leqslant k & \end{array}\right. \qquad \text{con } k \geqslant np_0$$

Por lo que el test a considerar será:

$$\varphi(X) = \left\{\begin{array}{ll} 1 & \text{si } X > k \\ 0 & \text{si } X \leqslant k \end{array}\right.$$

Y determinaremos el valor de k imponiendo un nivel de significación α:

$$\alpha \geqslant \sup_{p \in \Theta_0} \beta_\varphi(p) = \sup_{p \in \Theta_0} E_p[\varphi] = \sup_{p \in \Theta_0} P_p[X > k] = \sup_{p \in \Theta_0} \left(\sum_{t=\lfloor k \rfloor + 1}^{n} P_p[X = t] \right)$$

Veamos ahora que fijado $t \in \{\lfloor k \rfloor + 1, \ldots, n\}$ se tiene que $P_p[X = t]$ es creciente en función de $p \in \Theta_0$. Para ello, si tomamos:

$$h(p) = P_p[X = t] = \binom{n}{x} p^t (1-p)^{n-t} \quad \Longrightarrow \quad \ln h(p) = \ln \binom{n}{x} + t \ln p + (n-t) \ln(1-p)$$

$$\frac{\partial \ln h(p)}{\partial p} = \frac{t - np}{p(1-p)}$$

Observamos ahora que:

$$t \geqslant \lfloor k \rfloor + 1 > k \geqslant np_0 \geqslant np$$

De donde deducimos que $h(p)$ es creciente, por lo que el supremo anterior se alcanza en p_0:

$$\alpha \geqslant \sup_{p \in \Theta_0} \beta_\varphi(p) = \sup_{p \in \Theta_0} \left(\sum_{t=\lfloor k \rfloor + 1}^{n} P_p[X = t] \right) = \sum_{t=\lfloor k \rfloor + 1}^{n} P_{p_0}[X = t]$$

En un caso concreto de esta fórmula se despeja el valor de k.

Ejercicio 2.7.10. Sea X una variable con función de densidad

$$f_\theta(x) = \theta x^{\theta-1}, \qquad 0 < x < 1$$

Basándose en una observación de X, deducir el test de razón de verosimilitudes de tamaño arbitrario para contrastar

$$H_0 : \theta \leqslant \theta_0$$
$$H_1 : \theta > \theta_1$$

Tendremos:

$$\Theta_0 =]0, \theta_0]\,, \qquad \Theta_1 =]\theta_0, +\infty[\,, \qquad \Theta = \mathbb{R}^+$$

Calculamos el EMV de θ para $x \in \mathcal{X}$:

$$L_x(\theta) = f_\theta(x) = \theta x^{\theta-1} \quad \Longrightarrow \quad \ln L_x(\theta) = \ln\theta + (\theta-1)\ln x$$
$$\frac{\partial \ln L_x(\theta)}{\partial\theta} = \frac{1}{\theta} + \ln x = 0 \Longleftrightarrow \theta = \frac{-1}{\ln x}$$

Y observemos que como $x \in]0,1[$ tenemos que $\ln x < 0$. El EMV es $\hat\theta = {}^{-1}\!/_{\ln X}$. Por tanto:

$$\sup_{\theta\in\Theta_0} L_x(\theta) = \begin{cases} L_x({}^{-1}\!/_{\ln x}) & \text{si } \theta_0 \geqslant {}^{-1}\!/_{\ln x} \\ L_x(\theta_0) & \text{si } \theta_0 < {}^{-1}\!/_{\ln x} \end{cases}$$

de donde:

$$\lambda(x) = \frac{\sup\limits_{\theta\in\Theta_0} L_x(\theta)}{\sup\limits_{\theta\in\Theta} L_x(\theta)} = \begin{cases} 1 & \text{si } \theta_0 \geqslant {}^{-1}\!/_{\ln x} \\ \dfrac{L_x(\theta_0)}{L_x({}^{-1}\!/_{\ln x})} & \text{si } \theta_0 < {}^{-1}\!/_{\ln x} \end{cases}$$

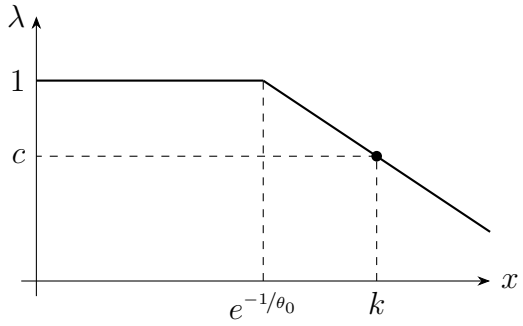
En el caso $\theta_0 < \frac{-1}{\ln x}$:

$$\frac{L_x(\theta_0)}{L_x({}^{-1}\!/_{\ln x})} = \frac{\theta_0 x^{\theta_0-1}}{\frac{-1}{\ln x} x^{\left(\frac{-1}{\ln x}-1\right)}} = -\theta_0 \ln x \cdot x^{\theta_0+\frac{1}{\ln x}}$$

Y veamos si en este caso λ es creciente o decreciente en función de x. Para ello, consideramos:

$$g(x) = -\theta_0 \ln x \cdot x^{\theta_0+\frac{1}{\ln x}}$$
$$\ln g(x) = \ln(-\theta_0 \ln x) + \left(\theta_0 + \frac{1}{\ln x}\right)\ln x = \ln(\theta_0) + \ln(-\ln x) + \theta_0 \ln x + 1$$
$$\frac{\partial \ln g(x)}{\partial x} = \frac{-1}{x\ln x} + \frac{\theta_0}{x} = \frac{-1+\theta_0\ln x}{x\ln x} < 0 \qquad (x \in]0,1[)$$

Por lo que g es estrictamente decreciente, de donde si vemos λ como función de x ($\theta_0 \geqslant {}^{-1}\!/_{\ln x} \Longleftrightarrow \ln x \leqslant {}^{-1}\!/_{\theta_0} \Longleftrightarrow x \leqslant e^{-1/\theta_0}$):



Tenemos entonces que:

$$\left.\begin{array}{l}\lambda < c\\ \lambda \geqslant c\end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} x > k\\ x \leqslant k\end{array}\right. \qquad \text{con} \quad k \geqslant e^{-1/\theta_0}$$

Y el test a considerar será:

$$\varphi(X) = \left\{\begin{array}{ll} 1 & \text{si } X > k\\ 0 & \text{si } X \leqslant k\end{array}\right.$$

y determinamos el valor de k imponiendo tamaño α:

$$\begin{aligned}\alpha &= \sup_{\theta\in\Theta_0} \beta_\varphi(\theta) = \sup_{\theta\in\Theta_0} E_\theta[\varphi] = \sup_{\theta\in\Theta_0} P_\theta[X > k] = \sup_{\theta\in\Theta_0} \left(\int_k^1 \theta t^{\theta-1}\, dt\right)\\ &= \sup_{\theta\in\Theta_0} \left([t^\theta]_k^1\right) = \sup_{\theta\in\Theta_0} (1-k^\theta) \overset{(k\in[0,1])}{=} 1-k^{\theta_0}\end{aligned}$$

Luego:

$$k^{\theta_0} = 1-\alpha \quad \Longrightarrow \quad \ln k = \frac{\ln(1-\alpha)}{\theta_0} \quad \Longrightarrow \quad k = e^{\ln\left((1-\alpha)^{\frac{1}{\theta_0}}\right)} = (1-\alpha)^{\frac{1}{\theta_0}}$$

Ejercicio 2.7.11. Un fabricante de coches asegura que la distancia media recorrida con un galón de gasolina es al menos 30 millas. Probados 9 coches de esta fábrica, la distancia media recorrida con un galón de gasolina ha sido 26 millas, y la suma de los cuadrados 6106 millas al cuadrado.

a) Suponiendo que la distancia recorrida por estos coches con un galón de gasolina tiene distribución normal, contrastar la hipótesis del fabricante a partir de estos datos, a nivel de significación 0,01.

b) ¿Qué conclusión se obtendría de estos mismos datos, al mismo nivel de significación, si se sabe que la desviación típica de la variable considerada es 5,5?

Sale 1 si $36 > \chi^2_{25,;0,02}$, $\chi^2_{25;0,02}$. Los datos no dan evidencia para rechazar.

Ejercicio 2.7.12. Un fabricante de baterías asegura que la desviación típica del tiempo de vida de las mismas es, a lo sumo, 70 horas. Una muestra de 26 baterías tomadas al azar ha dado una cuasidesviación típica de 84 horas. Haciendo las hipótesis adecuadas de normalidad, ¿proporcionan los datos evidencia para rechazar la hipótesis del fabricante al nivel 0,02?

Ejercicio 2.7.13. Un profesor asegura que tiene un nuevo método de enseñanza mejor que el usado tradicionalmente. Para comprobar si tiene razón se selecciona de forma aleatoria e independiente dos grupos de alumnos, A y B, utilizándose el nuevo método con el grupo A y el tradicional con el B. A final de curso se hace un examen a los alumnos, obteniéndose las siguientes puntuaciones:
Grupo A: 6, 5, 4, 7, 3, 5,5, 6, 7, 6.
Grupo B: 5, 4, 5, 6, 4, 6, 5, 3, 7.
Supuesto que las puntuaciones de cada grupo tienen distribución normal, ¿proporcionan estos datos evidencia para rechazar el nuevo método a nivel de significación 0,05?

Si tomamos:

$$X \equiv \text{“Puntuaciones del grupo A (nuevo método)”} \rightsquigarrow \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y \equiv \text{“Puntuaciones del grupo B (método tradicional)”} \rightsquigarrow \mathcal{N}(\mu_2, \sigma_2^2)$$

Consideramos la hipótesis $\mu_1 > \mu_2 \iff \mu_1 - \mu_2 > 0$. Como no sabemos resolver este tipo de contrastes (donde en H_0 no aparece una igualdad), consideramos que esta hipótesis es H_1, por lo que el contraste sería el siguiente:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leqslant 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

En teoría vimos que se rechaza la hipótesis nula si:

$$t_{exp} > t_{n_1+n_2-2;\alpha}$$

donde:

$$t_{exp} = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Calculamos para ello:

- $n_1 = n_2 = 9$.
- $\overline{x} = 5{,}5$.
- $s_1^2 = 1{,}75$.
- $\overline{y} = 5$.
- $s_2^2 = 1{,}5$.

Por lo que:

$$s_p = \sqrt{\frac{8 \cdot 1{,}75 + 8 \cdot 1{,}5}{9+9-2}} \approx 1{,}274755, \qquad t_{exp} = \frac{5{,}5 - 5}{1{,}274755 \cdot \sqrt{\frac{2}{9}}} \approx 0{,}83205$$

Y si calculamos ahora $t_{n_1+n_2-2;\alpha}$:

$$t_{n_1+n_2-2;\alpha} = t_{9+9-2;0,05} = t_{16;0,05} = 1{,}7459$$

Como t_{exp} no es mayor que $t_{n_1+n_2-2;\alpha}$, los datos no proporcionan evidencia suficiente para rechazar el nuevo método a nivel de significación 0,05.

Ejercicio 2.7.14. A partir de las siguientes observaciones de muestras independientes de dos poblaciones normales, contrastar, al nivel de significación 0,01, si la media de la primera población supera en al menos una unidad la media de la segunda.
Muestra 1: 132, 139, 126, 114, 122, 132, 141, 126.
Muestra 2: 124, 141, 118, 116, 114, 132, 145, 123, 121.

Planteamos el constrate de hipótesis:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \geqslant \mu_0 = 1 \\ H_1 : \mu_1 - \mu_2 < \mu_0 \end{cases}$$

Sean $X \rightsquigarrow \mathcal{N}(\mu_1, \sigma^2)$, $Y \rightsquigarrow \mathcal{N}(\mu_2, \sigma^2)$, tenemos dos observaciones: $(x_1, \ldots, x_8)$ de X y $(y_1, \ldots, y_9)$ de Y (tomamos $n_1 = 8$ y $n_2 = 9$). Usaremos el estadístico:

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}, \qquad S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

Por lo que calculamos t_{exp}:

$$t_{exp} = \frac{\overline{x} - \overline{y} - \mu_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

para lo que necesitaremos:

$$\overline{x} = 129, \qquad \overline{y} = 126, \qquad s_1^2 = 79{,}142857, \qquad s_2^2 = 121$$

Por lo que:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{7 \cdot 79{,}142857 + 8 \cdot 121}{8+9-2}} \approx 10{,}073066$$

$$t_{exp} = \frac{\overline{x} - \overline{y} - \mu_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{129 - 126 - 1}{10{,}073066\sqrt{\frac{1}{8} + \frac{1}{9}}} \approx 0{,}408611$$

Sea $\alpha = 0{,}01$, vimos en teoría que la región crítica era:

$$t_{exp} < -t_{n_1+n_2-2;\alpha} = -t_{15;0,01} = 2{,}6025$$

Por tanto, como $0{,}408611 < 2{,}6025$ estamos en la región crítica, por lo que podemos rechazar a nivel de significación 0,01 que la media de la primera población supera en al menos una unidad la media de la segunda.

Ejercicio 2.7.15. Una central lechera recibe diariamente leche de dos granjas A y B. Para comparar la calidad de los productos recibidos se ha medido el contenido en grasa en muestras de leche tomadas al azar de cada una de las granjas, con los siguientes resultados:

	Contenido en grasa (%)					
Granja A	14	12	15	15	11	16
Granja B	20	18	18	19	15	

a) ¿Puede suponerse, a nivel de significación 0,05, que el contenido medio en grasa de la leche de las dos granjas es el mismo? Especificar las hipótesis bajo las que se resuelve este problema.

b) Calcular un intervalo de confianza, a nivel de confianza 0,9, para la varianza del contenido en grasa de la leche de la granja B. A partir de dicho intervalo, deducir si puede aceptarse que la varianza de esta población es igual a 3. Especificar el problema de contraste y el test utilizado; calcular su tamaño.

2.8. Regresión lineal y análisis de la varianza

Ejercicio 2.8.1. Una compañía de energía eléctrica pretende desarrollar un modelo lineal para el consumo de energía en función de la temperatura media diaria durante los meses de invierno. En 9 días elegidos al azar se obtuvo la siguiente información:

Temperatura	0	2	4	8	13	-4	-6	-8	-11
Consumo	70	79	67	66	63	97	82	90	107

a) Obtener la recta de regresión estimada e interpretar sus coeficientes.

b) Descomponer la variabilidad de los datos de consumo. Obtener la varianza residual, el coeficiente de determinación y dar su interpretación.

c) Obtener la predicción del consumo de energía para un día con temperatura media 10 grados y estimar el error cuadrático medio de esta predicción.

d) Si se suponen las hipótesis adecuadas de normalidad, ¿se puede considerar, al nivel de significación $\alpha = 0{,}05$, que el consumo no depende linealmente de la temperatura?

Ejercicio 2.8.2. Una compañía de seguros desea establecer una relación lineal, y el grado de dicha relación, para determinar el montante anual del seguro de vida del cabeza de familia en función del ingreso mensual familiar. Observada una muestra aleatoria de 10 familias elegidas de forma independiente, se obtuvo la siguiente información:

Ingreso (cientos de euros)	10	10	15	20	20	25	25	30	30	30
Seguro (decenas de euros)	50	35	55	55	70	65	65	60	75	90

a) ¿Cabe pensar en la citada relación lineal? En caso afirmativo, estimar la recta de regresión, interpretar los coeficientes y obtener la predicción del montante del seguro de vida para un ingreso mensual de 2800 euros.

b) Suponiendo las hipótesis adecuadas de normalidad, realizar el contraste de regresión al nivel de significación $\alpha = 0{,}05$ e interpretar el resultado.

Ejercicio 2.8.3. Los datos de la siguiente tabla representan las calificaciones medias (X) de 7 recién graduados y sus respectivos salarios iniciales (Y) en miles de euros:

X	2,75	2,85	2,95	3,05	3,2	3,4	3,6
Y	16,8	18,8	17,2	17,2	21,2	21,5	22,4

a) Estimar la recta de regresión de Y sobre X e interpretar sus coeficientes.

b) Determinar la varianza residual.

c) Calcular e interpretar el coeficiente de determinación.

d) Predecir el salario inicial de un estudiante con una calificación media de 3,25.

e) Suponiendo las hipótesis adecuadas de normalidad, contrastar la hipótesis de que no existe relación lineal entre ambas variables al nivel de significación $\alpha = 0{,}01$.

Ejercicio 2.8.4. En cierto estudio sobre la relación entre el diámetro de los guisantes (X) y el diámetro medio de sus descendientes (Y), Galton obtuvo los siguientes resultados:

D. Padres	21	20	19	18	17	16	15
D. Descendientes	17,26	17,07	16,37	16,4	16,13	16,17	15,98

a) Determinar el modelo de regresión lineal estimado de Y sobre X e interpretar el valor estimado de la pendiente. Dar la predicción del diámetro de los guisantes cuyos progenitores tienen un diámetro de 18,5. Dar una medida de la bondad del ajuste de los datos a la recta estimada.

b) Suponiendo las hipótesis adecuadas de normalidad, ¿puede deducirse, al nivel 0,05, que no hay relación lineal entre las variables consideradas? Relacionar este resultado con las conclusiones anteriores.

Ejercicio 2.8.5. Una compañía farmacéutica investiga los efectos de 5 compuestos. El experimento consiste en inyectar los compuestos a 13 ratas de características similares y anotar los tiempos de reacción. Los animales se clasifican en 5 grupos de 4, 2, 2, 3 y 2 ratas, respectivamente, y a cada grupo se le administra un compuesto diferente, obteniéndose los resultados de la siguiente tabla:

Grupo	Tiempo de reacción			
1	8,3	7,6	8,4	8,3
2	7,4	7,1		
3	8,1	6,4		
4	7,9	8,5	10,0	
5	7,1	8		

Suponiendo que se verifican las hipótesis de normalidad, aleatoriedad, independencia e igualdad de varianzas, contrastar la hipótesis de que los tiempos medios de reacción coinciden en los cinco grupos y, por tanto, la eficacia de los cinco compuestos es la misma.

Ejercicio 2.8.6. Se quiere estudiar la eficacia de tres fertilizantes, A, B y C, en la producción de cierto fruto. Para ello se aplica el A en 8 parcelas, el B en 6, y el C en 12 parcelas. Las parcelas son de características similares en cuanto a fertilidad, por lo que se considera que las diferencias en la producción, si las hay, serán debidas al tipo de fertilizante. Las toneladas producidas en cada parcela en una determinada temporada son:

A	6	7	5	6	5	8	4	7				
B	10	9	9	10	10	6						
C	3	4	8	3	7	6	3	6	4	7	6	3

Suponiendo que las tres muestras proceden de poblaciones normales con varianzas iguales, contrastar la hipótesis de que los abonos son igualmente eficaces.

Ejercicio 2.8.7. Los siguientes datos corresponden a observaciones del consumo medio (en Kw/h) realizado por 5 tipos de calefactores para mantener una habitación a una temperatura determinada durante todo un día:

Tipo	Consumo (en kw/h)					
1	14,5	14,1	14,6	14,2		
2	13,2	13,4	13,0			
3	13,7	13,6	14,1	13,8	14,0	
4	12,7	13,1	12,8	12,9	13,3	13,2
5	14,6	15,2	14,4	14,8	14,3	

Contrastar la hipótesis de igualdad de los consumos medios de los diferentes tipos de calefactores. ¿Bajo qué hipótesis se puede realizar este contraste?

Ejercicio 2.8.8. En un tratamiento contra la hipertensión se seleccionaron 35 enfermos de características similares. Los enfermos se distribuyeron en cuatro grupos de 10 (P, A, B y AB). El grupo P tomó "placebo" (fármaco inocuo), el grupo A tomó un fármaco "A", el grupo B un fármaco "B" y el grupo AB una asociación entre "A" y "B". Para valorar la eficacia de los tratamientos, se registró el descenso de la presión diastólica desde el inicio del tratamiento hasta después de una semana de tratamiento. Los resultados, después de registrarse algunos abandonos, fueron:

P	10	0	15	-20	0	15	-5			
A	20	25	33	25	30	18	27	0	35	20
B	15	10	25	30	15	35	25	22	11	25
AB	10	5	-5	15	20	20	0	10		

A la vista de estos datos, ¿Puede afirmarse que el descenso de la presión diastólica coincide en los cuatro grupos? ¿Bajo qué hipótesis?

2.9. Constrastes de hipótesis no paramétricos

Ejercicio 2.9.1. A partir de los siguientes datos, que muestran el número de accidentes en un determinado regimiento del ejército durante 200 días elegidos al azar, contrastar si el número de accidentes diarios sigue una distribución de Poisson de parámetro 2.

Nº de accidentes	0	1	2	3	4	5	6	7
Nº de días	22	53	58	39	20	5	2	1

Sea la variable aleatoria:

$$X \equiv \text{“Número de accidentes en un día.”}$$

que se distribuye según una cierta función de distribución F, se plantea el contraste de hipótesis:

$$\begin{cases} H_0 : F = F_{\mathcal{P}(2)} \\ H_1 : F \neq F_{\mathcal{P}(2)} \end{cases}$$

En teoría hemos visto cómo resolver este contraste usando dos tests distintos:

- Test χ^2 de Pearson.
- Test de Kolmogorov-Smirnov.

El segundo solo se puede aplicar para distribuciones continuas, por lo que no podemos usarlo en este caso. Usaremos por tanto el test χ^2 de Pearson, que usa el estadístico:

$$\chi^2(N_0, \dots, N_k) = \sum_{i=0}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} = -n + \sum_{i=0}^{k} \frac{N_i^2}{np_i^0}$$

donde tenemos $n = 200$, $k = 8$ y los N_i dados por la tabla anterior. Calculamos cada una de las $p_i^0 = P_{H_0}[X = i]$:

p_0^0	p_1^0	p_2^0	p_3^0	p_4^0	p_5^0	p_6^0	p_7^0
0,1353	0,2707	0,2707	0,1804	0,0902	0,0361	0,0120	0,0034

de donde:

$$\chi^2_{\text{exp}} = -n + \sum_{i=0}^{k} \frac{N_i^2}{np_i^0} = -n + \frac{1}{n} \sum_{i=0}^{k} \frac{N_i^2}{p_i^0} \approx 2{,}834515$$

Si calculamos ahora el $p-$valor obtenemos que:

$$P[\chi^2(k-1) \geqslant \chi^2_{\text{exp}}] = P[\chi^2(7) \geqslant 2{,}834515] \approx P[\chi^2(7) \geqslant 2{,}8331] = 0{,}9$$

Por lo que no podemos rechazar H_0.

Ejercicio 2.9.2. Una tela cuadrada tiene 60 defectos de fabricación. Con objeto de analizar la distribución de los defectos en la superficie de la tela, se ha dividido ésta en 9 zonas cuadradas exactamente iguales, observándose los siguientes defectos en cada zona:

8	7	3
5	9	11
6	4	7

Contrastar, a partir de estos datos, si los defectos se distribuyen uniformemente en toda la superficie o, por el contrario, siguen algún patrón de ocurrencia.

Sea la variable aleatoria:

$$X \equiv \text{“Región en la que se localiza un defecto.”}$$

que se distribuye según una función de distribución F, planteamos el contraste:

$$\begin{cases} H_0 : F = F_{U(1,2,\dots,9)} \\ H_1 : F \neq F_{U(1,2,\dots,9)} \end{cases}$$

Usaremos el test χ^2 de Pearson, pues la distribución es discreta. Para ello, calculamos primero $p_i^0 = P_{H_0}[X = i]$, obteniendo en este caso que:

$$p_i^0 = \frac{1}{9} \qquad \forall i \in \{1, \dots, 9\}$$

Por lo que ($k = 9$, $n = 60$):

$$\chi^2_{\text{exp}} = -n + \sum_{i=1}^{k} \frac{N_i^2}{np_i^0} = -n + \frac{1}{np_1^0} \sum_{i=1}^{k} N_i^2 = 7{,}5$$

Calculamos ahora el valor del $p-$valor:

$$P_{H_0}[\chi^2(k-1) \geqslant \chi^2_{\text{exp}}] = P[\chi^2(8) \geqslant 7{,}5] \approx 0{,}475$$

Por lo que los datos no aportan evidencia suficiente para rechazar la hipótesis nula.

Ejercicio 2.9.3. Un modelo genético indica que la distribución de una población de hombres y mujeres, daltónicos o no, se ajusta a probabilidades de la forma:

	Hombres	Mujeres
No daltónicos	$(1-p)/2$	$(1-p^2)/2$
Daltónicos	$p/2$	$p^2/2$

Para comprobar esta teoría se examinaron 2000 individuos de la población, elegidos al azar, obteniéndose los siguientes resultados:

	Hombres	Mujeres
No daltónicos	894	1015
Daltónicos	81	10

Contrastar, mediante el test de la χ^2, si esta muestra concuerda con el modelo teórico.

Ejercicio 2.9.4. Un laboratorio farmacéutico afirma que uno de sus productos confiere inmunidad a la picadura de insectos durante un tiempo exponencial de media 2,5 horas. Probado el producto en 20 sujetos, en un ambiente con gran número de mosquitos, los instantes (en horas) en que recibieron la primera picadura fueron:

0,01	0,02	0,03	0,23	0,51	0,74	0,96	1,17	1,46	1,62
2,18	2,25	2,79	3,45	3,82	3,92	4,27	5,43	5,79	6,34

Usando el test de Kolmogorov-Smirnov, contrastar, a partir de estos datos, si puede aceptarse la afirmación del laboratorio.

Sea la variable aleatoria:

$$X \equiv \text{“Tiempo (en horas) en recibir la primera picadura”}$$

que se distribuye según una función de distribución F, planteamos el contraste:

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$$

donde $F_0 = F_{\exp(2,5)}$. Como la distribución exponencial es continua podemos aplicar el Test de Kolmogorov-Smirnov, por lo que nos disponemos a calcular primero $D_{\exp}$, mediante la fórmula (como las $n = 20$ observaciones son distintas):

$$D_{\exp} = \text{máx}\{\max_{x_i}\{F^*_{X_1,\dots,X_n}(x_i) - F_0(x_i)\}, \max_{x_i}\{F_0(x_i) - F^*_{X_1,\dots,X_n}(x_i^-)\}\}$$

Para ello, usamos la tabla (abreviando con $F^* = F^*_{X_1,\dots,X_n}$):

x_i	$nF^*(x_i)$	$F^*(x_i)$	$F^*(x_i^-)$	$F_0(x_i)$	$F^*(x_i) - F_0(x_i)$	$F_0(x_i) - F^*(x_i^-)$
0,01	1	0.05	0	0.02469	0.02530	0.02469
0,02	2	0.1	0.05	0.04877	0.05122	-0.0012
0,03	3	0.15	0.1	0.07225	0.07774	-0.0277
0,23	4	0.2	0.15	0.43729	-0.2372	0.28729
0,51	5	0.25	0.2	0.72056	-0.4705	0.52056
0,74	6	0.3	0.25	0.84276	-0.5427	0.59276
0,96	7	0.35	0.3	0.90928	-0.5592	0.60928
1,17	8	0.4	0.35	0.94633	-0.5463	0.59633
1,46	9	0.45	0.4	0.98772	-0.5377	0.58772
1,62	10	0.5	0.45	0.98257	-0.4825	0.53257
2,18	11	0.55	0.5	0.99570	-0.4457	0.49570
2,25	12	0.6	0.55	0.99639	-0.3963	0.44639
2,79	13	0.65	0.6	0.99906	-0.3490	0.39906
3,45	14	0.7	0.65	0.99982	-0.2998	0.34982
3,82	15	0.75	0.7	0.99993	-0.2499	0.29993
3,92	16	0.8	0.75	0.99994	-0.1999	0.24994
4,27	17	0.85	0.8	0.99997	-0.1499	0.19997
5,43	18	0.9	0.85	0.99999	-0.0999	0.14999
5,79	19	0.95	0.9	0.99999	-0.0499	0.09999
6,34	20	1	0.95	0.99999	0	0.04999

de donde obtenemos:

$$\left.\begin{array}{l} \max\limits_{x_i}\{F^*_{X_1,\ldots,X_n}(x_i) - F_0(x_i)\} = 0{,}07774 \\ \max\limits_{x_i}\{F_0(x_i) - F^*_{X_1,\ldots,X_n}(x_i^-)\} = 0{,}60928 \end{array}\right\} \implies D_{\text{exp}} = 0{,}60928$$

Calculamos el $p-$valor, usando para ello la distribución Z de Kolmogorov, donde $D(X_1,\ldots,X_{20}) \rightsquigarrow Z(20)$:

$$P_{H_0}[D(X_1,\ldots,X_n) \geqslant D_{\text{exp}}] = P_{H_0}[D(X_1,\ldots,X_n) \geqslant 0{,}60928] < 0{,}001$$

obtenemos un dato mucho menor que 0,001, por lo que podemos rechazar la afirmación del laboratorio.

Ejercicio 2.9.5. Cierta comunidad ha modificado la procedencia del agua destinada al consumo doméstico. Se sabe que, con el antiguo suministro, la distribución de la cantidad de sodio por unidad de volumen de sangre de sus habitantes es simétrica alrededor de 3,24 gr. Tras cierto tiempo, se quiere comprobar si la modificación ha afectado a la concentración de sodio, en el sentido de que su distribución se haya trasladado o no. Para ello, se han realizado 15 análisis con los siguientes resultados (en gr. por unidad):

2,37	2,95	3,4	2,64	3,66	3,18	2,72	3,61	3,87	1,97	1,66	3,72	2,10	
											1,83	3,03	

¿Se puede afirmar, al nivel de significación 0,1, que la distribución de la cantidad de sodio no ha variado?

Ejercicio 2.9.6. La siguiente tabla presenta las presiones sanguíneas sistólicas de 10 individuos antes y después de haber dejado la bebida.

A	140	165	160	160	175	190	170	175	155	160
D	145	150	150	160	170	175	160	165	145	170

¿Se puede afirmar a partir de los datos que el abandono de la bebida no disminuye la presión sanguínea? ¿Bajo qué hipótesis?

Ejercicio 2.9.7. En cierta comunidad de E.E.U.U. se realizó un estudio para investigar si el sueldo anual de las familias influía en los hijos para la elección de los diferentes cursos de enseñanza secundaria (Preparatorio, General y Comercial). Para ello, se hizo una clasificación de los sueldos en cuatro niveles (I, II, III y IV), y se tomó una muestra aleatoria simple de 390 estudiantes, obteniéndose la siguiente tabla de frecuencias:

Sueldo	I	II	III	IV
Preparatorio	23	40	16	2
General	11	75	107	14
Comercial	1	31	60	10

A la vista de los datos, decidir, al nivel de significación 0,01, si se acepta que el nivel económico familiar no influye en la decisión de los estudiantes a la hora de elegir curso.

Ejercicio 2.9.8. En un estudio sociológico sobre la polución atmosférica se entrevistó a 40 residentes de cada una de tres zonas residenciales en Gran Bretaña. La siguiente tabla muestra las respuestas a la pregunta: ¿Hay problema de polución en su barrio?

Zona residencial	No	Sí	No sabe	No contesta
1	5	31	2	2
2	10	21	4	5
3	11	20	7	2

Contrastar si las tres poblaciones de residentes pueden considerarse homogéneas con respecto a su opinión sobre la polución.

Ejercicio 2.9.9. Para determinar si diferentes tipos de profesiones de los individuos activos de cierto colectivo afectan a la tensión arterial, se clasificó a los individuos en cuatro grupos, atendiendo a su profesión, y se midió la tensión a una muestra de individuos elegidos de forma aleatoria. Clasificando la tensión en los niveles "Bajo", "Normal" y "Alto", se obtuvo los siguientes resultados, que muestran el número de individuos de cada tipo de profesión con los distintos niveles de tensión:

Profesión	Bajo	Normal	Alto
I	8	4	3
II	5	7	7
III	4	8	8
IV	5	7	8

¿Qué conclusión acerca del problema planteado se obtiene a la vista de estos datos? Especificar las hipótesis nula y alternativa que se contrastan.

Ejercicio 2.9.10. Para determinar si las calificaciones de los alumnos en selectividad son independientes de las calificaciones en bachiller, se eligió de forma aleatoria una muestra de alumnos, a los que se preguntó ambas calificaciones, obteniendo los siguientes resultados:

Bachiller	Suspenso	Aprobado	Notable	Sobresaliente
Aprobado	10	6	4	5
Notable	7	9	9	4
Sobresaliente	6	10	10	6

¿Qué conclusión acerca del problema planteado se obtiene a la vista de estos datos? Especificar las hipótesis nula y alternativa que se contrastan.