Ejercicios del tema 1

Estimación paramétrica

Ejercicio 1 El paquete `gapminder` contiene un fichero de datos de población, esperanza de vida y renta per cápita de los países del mundo entre 1952 y 2007. Instala el paquete y lleva a cabo los siguientes gráficos:

1. Un histograma de la esperanza de vida en 2007 de los países de Europa.
2. Diagramas de cajas con las esperanzas de vida de cada continente en el año 1952.
3. Un diagrama de dispersión de la renta per cápita y la esperanza de vida de cada país en el año 2007.
4. Mejora el gráfico anterior representando cada punto con un color diferente en función del continente al que pertenece cada país y representando la renta per cápita en una escala logarítmica (*Indicación:* consulta esta documentación)

Ejercicio 2 Se observa una muestra de una distribución de Poisson de parámetro θ en la que el parámetro θ solo puede tomar alguno de los valores del espacio paramétrico $\Theta = \{1,\ 2,\ 3\}$. Antes de observar la muestra, la opinión de un experto sobre los distintos valores que puede tomar θ queda reflejada en la siguiente distribución de probabilidad a priori sobre Θ:

θ	**1**	**2**	**3**
$\pi(\theta)$	1/2	1/4	1/4

1. Si definimos el estimador bayesiano de θ como el valor del parámetro más probable a posteriori tras observar la muestra, y ésta consiste en dos observaciones independientes x_1, x_2 tales que $x_1 + x_2 = 2$

, ¿cuánto vale el estimador bayesiano de θ?

2. Para una muestra que cumple la condición del apartado anterior, ¿cuál es el estimador de máxima verosimilitud de θ?

Ejercicio 3 Se desea estimar la prevalencia (proporción p de enfermos en la población) de cierto trastorno gástrico. Como la aparición de la enfermedad está relacionada con la edad, se divide la población en dos clases o estratos: (1) menores de 30 años (que constituyen el 40% de la población) y (2) mayores de 30 años (el 60% restante). Se toma una muestra de 60 individuos del estrato (1) y otra de 90 individuos del estrato (2). En total, tenemos lo que se denomina una muestra estratificada de tamaño $n = 150$ individuos. Para cada uno de ellos se observa si tienen o no la enfermedad.

1. A partir de $\hat{p}_1$, la proporción muestral de individuos enfermos en el estrato (1), y $\hat{p}_2$, la proporción muestral de individuos enfermos en el estrato (2), formula un estimador insesgado de la prevalencia p en la población.
2. En función de p_1 y p_2 (las prevalencias correspondientes a cada uno de los dos estratos) calcula la varianza del estimador propuesto en el apartado anterior.
3. Si $p_1 = p_2$, ¿se incrementa la eficiencia por el hecho de usar una muestra estratificada en lugar de una muestra de v.a.i.i.d. de tamaño 150, extraída sin tener en cuenta los estratos (una muestra aleatoria simple)?
4. Supongamos que diez de cada cien personas mayores de 30 años tiene la enfermedad (es decir $p_2 = 0.1$). Representa gráficamente las varianzas de los estimadores correspondientes a la muestra no estratificada y a la muestra estratificada como función de p_1. ¿Para qué valores de p_1 es mejor utilizar muestreo estratificado en lugar de muestreo aleatorio simple?

Ejercicio 4 Para estimar el número N desconocido de ejemplares en un área determinada se suelen emplear los procedimientos llamados de *captura y recaptura*. Lo que sigue es un ejemplo del esquema más sencillo posible.

Se capturan n ejemplares de la población, se identifican con una etiqueta, y se vuelven a soltar. Posteriormente, se seleccionan aleatoriamente con reemplazamiento m ejemplares de la población, y se observa que hay r de ellos que están etiquetados. Con esta información, ¿cuál es el estimador de máxima verosimilitud de N?

Ejercicio 5 Se dispone de una muestra $X_1, \ldots, X_n$ de vaiid con distribución exponencial de parámetro λ (es decir, su función de distribución es $F(t) = 1 - e^{-\lambda t}$) para $t \geq 0$ y se desea estimar el momento de orden 3, $\alpha_3 = \mathrm{E}(X^3)$, de esta distribución.

1. Define algún estimador natural para α_3.
2. Calcula el error cuadrático medio del estimador del apartado anterior.

(*Indicación:* Si $X \sim \exp(\lambda)$, entonces $\mathrm{E}(X^n) = n!/\lambda^n$ para todo entero positivo n.)

Ejercicio 6 Sean $X_1, \ldots, X_{10}$ vaiid de una población normal de media μ y varianza σ^2. Sea S^2 la correspondiente varianza muestral.

1. Calcula $\mathrm{P}(\sigma^2 < S^2 < 1.5\sigma^2)$, la probabilidad de que S^2 no infraestime σ^2, ni la sobreestime en más de 1.5 veces su valor.
2. Repite el ejercicio si tuviéramos una muestra de 50 observaciones en lugar de 10.

Ejercicio 7 El siguiente código genera una muestra de 100 datos de una distribución de Cauchy con parámetro de posición θ:

```r
set.seed(123)
theta <- 10
n <- 100
muestra <- rt(n, 1) + theta
```

Esta distribución tiene la siguiente función de densidad

$$f(x;\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2},$$

que coincide de hecho con la de una distribución t de Student con 1 grado de libertad trasladada una magnitud θ.

1. Calcula el estimador de máxima verosimilitud de θ. ¿Se parece al valor verdadero?
2. Lleva a cabo algún experimento de simulación para aproximar la varianza del estimador de máxima verosimilitud.

(*Indicación*: Busca información de funciones de R que sirvan para optimizar funciones o resolver ecuaciones no lineales: `optimize`, `nlm`, `uniroot`, etc.)

Ejercicio 8 Un fabricante de barras de cierto material afirma que más del 55% de las barras pueden resistir al menos 230 unidades de presión. En un experimento con 20 barras, se obtuvieron resistencias máximas a las siguientes presiones:

```
resistencias <- c(230.0004, 234.2818, 230.4031, 226.2253, 23
  226.8214, 238.9564, 236.7709, 227.0411, 233.1481,
  240.9500, 237.3096, 235.6890, 215.2487, 244.4351,
  211.2601, 235.7786, 227.4435, 227.6600, 226.4903)
```

1. Suponiendo que las resistencias siguen una distribución normal, calcula un intervalo de confianza de nivel 0.95 para la resistencia media.
2. Contrasta, con un nivel de significación 0.1 si los datos disponibles aportan evidencia de que la afirmación del fabricante es cierta.

Ejercicio 9 Tomamos una muestra $X_1, \ldots X_{100}$ de tamaño $n = 100$ de una v.a. X con distribución de Bernoulli de parámetro p. Consideramos el contraste de hipótesis que rechaza $H_0 : p = 0.5$ en favor de

$H_1 : p \neq 0.5$ en la región crítica $R = \{|\hat{p} - 0.5| > 0.1\}$, donde $\hat{p} = \sum_{i=1}^{n} X_i/n$.

1. Determina el tamaño (o nivel de significación) aproximado del test.
2. Calcula la potencia aproximada del test cuando $p = 0.6$.