FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

EYP1113 - Probabilidad y Estadística

Capítulo 7: Determinación de Modelos de Probabilidad

Ricardo Aravena C. - Cristian Capetillo C. - Ingrid Guevara R.
Bladimir Morales T. - Ricardo Olea O. - Daniel Saavedra M.

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2024

Contenido I

Introducción

El modelo de distribución de probabilidad apropiado para describir un fenómeno es generalmente desconocido.

Bajo ciertas circunstancias, las propiedades básicas del proceso físico subyacente del fenómeno aleatorio sugiere la forma de la distribución de probabilidades

Ejemplos

- Cumple vs. No cumple $\rightarrow$ Bernoulli.
- Número de "eventos" en periodos $\rightarrow$ Poisson.
- Tiempos de duración o espera $\rightarrow$ Exponencial.
- Suma de eventos individuales $\rightarrow$ Normal.
- Condiciones extremas de un proceso $\rightarrow$ Valor Extremo.

Introducción

En muchas situaciones, la distribución de probabilidad debe ser determinada empíricamente a partir de los datos.

Inicialmente, aproximaciones gráficas (Histograma v/s Densidad) nos pueden ayudar a inferir "visualmente" sobre la distribución.

También, con datos disponibles, pueden obtenerse los gráficos de probabilidad (Probability Papers) para distribuciones dadas (si los puntos están en línea recta, la distribución es apropiada).

Por último, dada una distribución a priori puede evaluarse la "bondad de ajuste" (Test χ^2, Test de Kolmogorov-Smirnov o el Test de Anderson-Darling, entre otros).

En esta sección nos enfocaremos en la construcción de un gráfico de probabilidad y test de bondad de ajuste chi-cuadrado

Determinación de Modelos Probabilísticos

Gráficos de Probabilidad

Es la representación gráfica de los datos observados y sus correspondientes frecuencias acumuladas.

Para un conjunto de N observaciones, $x_1, \ldots, x_N$, ordenados de menor a mayor, el m-ésimo valor es graficado contra la probabilidad acumulada de $m/(N+1)$.

La utilidad del "papel" de probabilidad es reflejar "el ajuste" que presentan los datos con respecto a la distribución subyacente.

La linealidad o falta de esta nos indica lo adecuado o inadecuado de la distribución.

Determinación de Modelos Probabilísticos

Gráficos de Probabilidad (Distribución Normal)

Sean $x_{(1)},\ldots,x_{(N)}$ observaciones ordenadas de menor a mayor y $p_1 = \frac{1}{N+1},\ldots,p_N = \frac{N}{N+1}$ sus respectivas probabilidades empíricas.

Calculemos los percentiles teóricos, $\Phi^{-1}(p_i)$, de una distribución Normal Estándar para cada p_i, con $i = 1,\ldots,N$.

Si los x's distribuyen Normal(μ, σ), entonces la siguiente relación lineal se debe cumplir

$$x_{(q)} = \mu + \sigma \cdot \Phi^{-1}(p_q)$$

Determinación de Modelos Probabilísticos

Gráficos de Probabilidad (Distribución Log-Normal)

Sean $x_{(1)},\ldots,x_{(N)}$ observaciones ordenadas de menor a mayor y $p_1 = \frac{1}{N+1},\ldots,p_N = \frac{N}{N+1}$ sus respectivas probabilidades empíricas.

Calculemos los percentiles teóricos, $\Phi^{-1}(p_i)$, de una distribución Normal Estándar para cada p_i, con $i = 1,\ldots,N$.

Si los x's distribuyen log-Normal(λ, ζ), entonces la siguiente relación lineal se debe cumplir

$$\ln x_{(q)} = \lambda + \zeta \cdot \Phi^{-1}(p_q)$$

Determinación de Modelos Probabilísticos

Gráficos de Probabilidad (Distribución Exponencial)

Sean $x_{(1)},\ldots,x_{(N)}$ observaciones ordenadas de menor a mayor y $p_1 = \frac{1}{N+1},\ldots,p_N = \frac{N}{N+1}$ sus respectivas probabilidades empíricas.

Calculemos los percentiles teóricos, $-\ln(1-p_i)$, de una distribución Exponencial(1) para cada p_i, con $i = 1,\ldots,N$.

Si los x's distribuyen Exponencial(ν) trasladada en α, entonces la siguiente relación lineal se debe cumplir

$$x_{(q)} = \alpha + \frac{1}{\nu} \cdot [-\ln(1-p_q)]$$

Test de Bondad de Ajuste

Test de Kolmogorov-Smirnov

Supongamos que queremos evaluar la calidad de ajuste del modelo $f_0(y)$.

$$H_0 : f_Y(y) = f_0(y) \quad \text{vs} \quad H_0 : f_Y(y) \neq f_0(y)$$

$$d = \max\{|F_n(y) - F_0(y)|\}$$

con F_n función de distribución acumulada empírica y F_0 función de distribución acumulada teórica del modelo que se quiere ajustar. Los valores p se obtienen como se describe en Marsaglia, Tsang y Wang (2003).

En R la función `ks.test()` realiza la comparación y calcula el valor-p.

Test de Bondad de Ajuste

Test Chi-Cuadrado

Considere una muestra de n valores observados de una variable aleatoria y suponga una distribución de probabilidad subyacente.

El test χ^2 de bondad de ajuste compara las frecuencias observadas O_1, O_2,..., O_k de k valores (o k intervalos) de la variable con sus correspondientes frecuencias teóricas E_1, E_2,..., E_k que calculados suponiendo la distribución teórica.

Para evaluar la calidad del ajuste se usa el siguiente estadístico de prueba:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

cuya distribución se aproxima por una Chi-Cuadrado$(k-1)$.

Test de Bondad de Ajuste

Test Chi-Cuadrado

Si los parámetros de la distribución son desconocidos, estos deben ser estimados a partir de los datos y debe ser descontado de los grados de libertad de la distribución (por cada parámetro estimado).

Si el estadístico de prueba $X^2 > c_{1-\alpha}(f)$, la hipótesis nula que los datos provienen de la distribución escogida es rechazada.

El parámetro $f = (k-1) - \nu$, con ν el número de estadísticos necesarios para estimar los parámetros.

Se recomienda aplicar este test cuando $k \geq 5$ y $E_i \geq 5$.

En R la función `chisq.test()` realiza la comparación y calcula el valor-p para el caso $\chi^2(k-1)$.