EYP1113 - Probabilidad y Estadística

Capítulo 7: Determinación de Modelos de Probabilidad

Ricardo Aravena C.
Ricardo Olea O.

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2019

Contenido I

1 Introducción

2 Determinación de Modelos Probabilísticos
- Gráficos de Probabilidad
- Gráficos de Probabilidad (Distribución Normal)
- Gráficos de Probabilidad (Distribución Log-Normal)
- Gráficos de Probabilidad (Distribución Exponencial)
- Aplicación

3 Test de Bondad de Ajuste
- Test Chi-Cuadrado

Introducción

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

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.

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

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

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

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

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

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

1. Construya un gráfico de probabilidad para una distribución Normal, Log-Normal y Exponencial.
2. Simule datos Normales, Log-Normales y Exponenciales para probar los gráficos construidos en (1).
3. Simule datos Gamma($k = 5$, $\nu = 0.3$) trasladados en 10 y ajuste los modelos anteriores en un histograma según la estimación obtenida del gráfico de probabilidad.

```
qqnormal = function(X){
y = sort(X)
n = length(X)
x = (1:n)/(n+1)
p = c(0.01, 0.1, 1,2,5,10,20,30,40,50,60,70,80,90,95,99,99.9,99.99)/100
plot(y~qnorm(x), lwd = 2, bty = "n", las = 1,  ylab = "Values of Y",
xlab = "Cumulative Probability", xaxt = "n", pch = 20,
xlim = c(min(qnorm(p)), max(qnorm(p))), main = expression("QQ-Normal"))
axis(1, at = qnorm(p), label = p)
abline(lm(y~qnorm(x)), lwd = 2, col = "red")
fit = lm(y~qnorm(x))$coef
hat.mu = fit[1]
hat.sigma = fit[2]
list("mu" = as.vector(hat.mu), "sigma" = as.vector(hat.sigma))
}
```

```
qqlognormal = function(X){
y = sort(X)
n = length(X)
x = (1:n)/(n+1)
p = c(0.01, 0.1, 1,2,5,10,20,30,40,50,60,70,80,90,95,99,99.9,99.99)/100
plot(log(y)~qnorm(x), lwd = 2, bty = "n", las = 1,  ylab = "Values of Y"
xlab = "Cumulative Probability", xaxt = "n", pch = 20,
xlim = c(min(qnorm(p)), max(qnorm(p))), main = expression("QQ-LogNormal"
axis(1, at = qnorm(p), label = p)
abline(lm(log(y)~qnorm(x)), lwd = 2, col = "red")
fit = lm(log(y)~qnorm(x))$coef
hat.lambda = fit[1]
hat.zeta = fit[2]
list("lambda" = as.vector(hat.lambda), "zeta" = as.vector(hat.zeta))
}
```

```
qqexp = function(X){
y = sort(X)
n = length(X)
x = (1:n)/(n+1)
p = c(0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.92, 0.95, 0.96, 0.97,
0.98, 0.99, 0.995, 0.999)
plot(y~qexp(x), lwd = 2, bty = "n", las = 1,  ylab = "Values of Y",
xlab = "Cumulative Probability", xaxt = "n", pch = 20,
xlim = c(min(qexp(p)), max(qexp(p))), main = expression("QQ-Exponencial"
axis(1, at = qexp(p), label = p)
abline(lm(y~qexp(x)), lwd = 2, col = "red")
fit = lm(y~qexp(x))$coef
hat.gamma = fit[1]
hat.nu = 1/fit[2]
list("gamma" = as.vector(hat.gamma), "nu" = as.vector(hat.nu))
}
```

Determinación de Modelos Probabilísticos

```
set.seed(1234567890)
x = rgamma(300, shape = 5, rate = 0.3)+10
hist(x, freq = F, xlim = c(0,60), col = "gray",
border = "white", xlab = "", main = "")

par(mfrow = c(1,3))
par1 = qqnormal(x)
par2 = qqlognormal(x)
par3 = qqexp(x)

par(mfrow = c(1,1))
hist(x, freq = F, xlim = c(0,60), ylim = c(0,.15),
col = "gray", border = "white", xlab = "", main = "")
x = seq(0,60,0.01)
lines(dnorm(x, mean = par1$mu, sd = par1$sigma)~x, col = "red")
lines(dlnorm(x, meanlog = par2$lambda, sdlog = par2$zeta)~x, col = "blue
lines(dexp(x-par3$gamma, rate = par3$nu)~x, col = "orange")
legend("topright", c("Normal", "Log-Normal", "Exponencial"),
col = c("red", "blue", "orange"), lty = 1, bty = "n")
```

Test de Bondad de Ajuste

Test de Bondad de Ajuste

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

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 $E_i \geq 5$.