Universidad de Cantabria
Curso 2010–2011

Práctica 3: Papel probabilístico

Los métodos gráficos han sido tradicionalmente utilizados en ingeniería debido a su sencillez y utilidad para analizar datos y comprender su estructura. Con la popularización de los ordenadores como herramienta habitual de trabajo, la representación gráfica ha pasado de realizarse manualmente a ser hecha en el ordenador. Uno de los métodos gráficos tradicionales que puede ser fácilmente implementado en R es el basado en el concepto de papel probabilístico.

La idea básica del papel probabilístico, de una familia biparamétrica de funciones de distribución, es modificar las escalas del dibujo de la variable aleatoria y de la probabilidad acumulada, de tal forma que dicha familia se convierta en una familia de rectas. De esta manera, cuando se representa la función de distribución en dicho papel, el aspecto de la gráfica (rectilinea o no) sirve para decidir si los datos proceden de esa familia de distribuciones o no. Además, de la representación pueden obtenerse también los parámetros de la distribución concreta que mejor se ajusta.

1. Función de distribución empírica

Si tomamos una muestra de una población cuya función de distribución desconocemos, podemos construir una "función de distribución muestral" conocida como **función de distribución empírica**[1]. Esta función nos dará la probabilidad de obtener un valor menor o igual que uno dado a la vista de la muestra que hemos obtenido. Para construirla ordenamos los valores x_i que ha tomado la variable en nuestra muestra de tamaño n. Al elemento que ocupa el lugar i-ésimo una vez ordenada la muestra se le conoce como estadístico de orden i, y lo denotaremos por $x_{i:n}$. La función de distribución empírica asigna a cada valor obtenido en la muestra una probabilidad acumulada i/n:

$$S_n(x_{i:n}) = \frac{i}{n}$$

En realidad, esta función está definida en toda la recta real y tiene forma escalonada (Figura 1). En R, se puede obtener mediante la función `ecdf`:

```
x.i <- rnorm(50)       # Tomamos una muestra de tamaño 50 de la normal
Sn <- ecdf(x.i)        # Construimos su ECDF
Sn(1.2)                # Podemos usarla como una funcion...
plot(Sn)               # ... y la interpreta de forma especial el comando plot
curve(pnorm(x),add=T)  # Podemos compararla con la distribucion de la poblacion
```

Esta función de distribución empírica es la que queremos representar en papel probabilístico, es decir, en un gráfico con nuevas escalas en los ejes, de tal forma que sus puntos queden formando una línea recta en lugar de la curva que describen en la Figura 1. Sin embargo, en el caso de muchas familias de distribuciones, al valor 1 (que alcanza $S_n(x_{n:n})$), cuando se le aplica la transformación de escala se transforman en ∞. Por ello, es imposible dibujarlo. Una solución a este problema consiste en usar otras **fórmulas de punteo**, diferentes de i/n para asignar la probabilidad acumulada hasta cada

[1] en inglés *Empirical Cumulative Distribution Function* (ECDF)

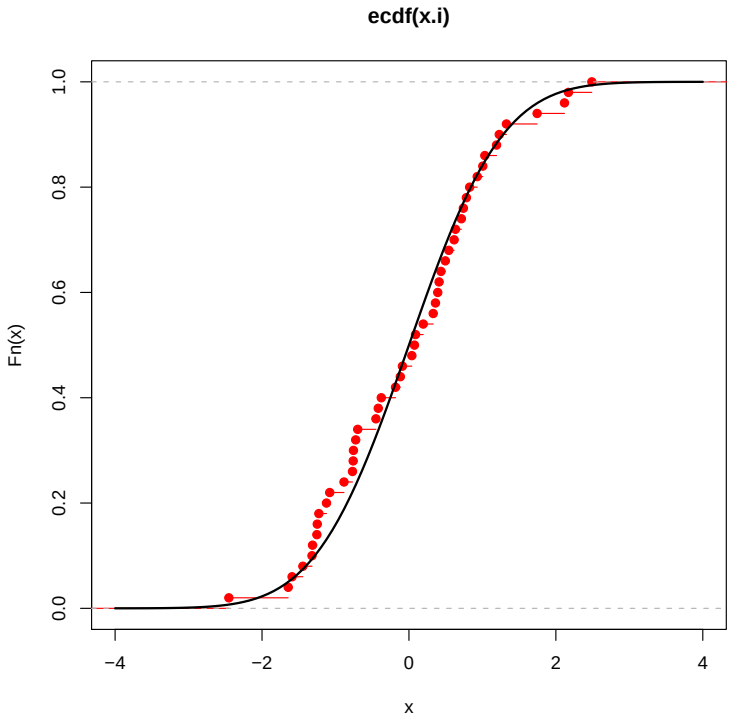

Figura 1: ECDF de una muestra de tamaño 50 tomada de una distribución normal tipificada.

dato de la muestra. La siguiente tabla resume diferentes fórmulas de punteo propuestas por diferentes autores:

Fórmula de punteo	Autor
$\left(x_{i:n}, \frac{i}{n+1}\right)$	-
$\left(x_{i:n}, \frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$	Blom
$\left(x_{i:n}, \frac{i-\frac{1}{2}}{n}\right)$	Hazen
$\left(x_{i:n}, \frac{i-0{,}44}{n+0{,}12}\right)$	Gringorten

En lo que sigue utilizaremos la fórmula de Hazen, aunque en R podemos utilizar cualquier otra sin mayor dificultad. Si cargamos el fichero `papeles.R` de la página web de la práctica, estas fórmulas se pueden obtener de la siguiente manera:

```
source("papeles.R")     # Cargamos el fichero con funciones de esta practica
punteo.hazen(50)        # Nos devuelve las probabilidades asignadas por este punteo
punteo.gringorten(50)   #   a una muestra de tamaño 50
```

2. Fundamentos del papel probabilistico

El papel probabilístico es simplemente un papel en el que se han cambiado las escalas de tal manera que las funciones de distribución de una cierta familia, cuando se dibujan en él, se convierten en una familia de lineas rectas[2].

[2]Para más detalles, consultar el tema 10 de E. Castillo y R.E. Pruneda "Estadística aplicada", Ed. Moralea, 2001 (Signatura CAM/A62-19)

Sea $F_X(x;\theta_1,\theta_2)$ una familia biparamétrica de funciones de distribución, donde θ_1 y θ_2 son los parámetros. Se busca una transformación

$$\xi = g(x) \quad ; \quad \eta = h(y) \tag{1}$$

tal que la familia de curvas

$$y = F_X(x;\theta_1,\theta_2) \tag{2}$$

cuando se transforma por (1) se convierte en una familia de lineas rectas, es decir:

$$h(y) = h[F_X(x;\theta_1,\theta_2)] = ag(x) + b \qquad \Leftrightarrow \qquad \eta = a\xi + b \tag{3}$$

donde la variable η se llama variable reducida y $a = a(\theta_1,\theta_2)$ y $b = b(\theta_1,\theta_2)$ son, respectivamente, la pendiente y la ordenada en el origen de la recta en la que se transforma $F_X(x;\theta_1,\theta_2)$.

Por tanto, para que exista un papel probabilístico asociado a una cierta familia de funciones de distribución $F_X(x;\theta_1,\theta_2)$ se necesita que

$$F_X(x;\theta_1,\theta_2) = h^{-1}[ag(x) + b] \tag{4}$$

Dada una familia de distribuciones conocida, tendremos que buscar la forma de las transformaciones h y g que hacen posible el papel probabilístico para esa familia. En esta práctica veremos las transformaciones que hacen posible el papel normal y el papel exponencial.

3. Papel normal

Si $F_X(x;\mu,\sigma)$ es la función de distribución de una variable normal, sabemos que puede ser escrita como

$$F_X(x;\mu,\sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{5}$$

donde μ y σ son la media y la desviación típica, respectivamente, y $\Phi(x)$ es la función de distribución de la variable normal estándar $F_{N(0,1)}(x)$.

Entonces, según (1) y (3), la expresión (5) da

$$\xi = g(x) = x \quad ; \quad \eta = h(y) = \Phi^{-1}(y) \quad ; \quad a = \frac{1}{\sigma} \quad ; \quad b = -\frac{\mu}{\sigma} \tag{6}$$

y la familia de lineas rectas es

$$\eta = a\xi + b = \frac{\xi - \mu}{\sigma} \tag{7}$$

Si los puntos $\left(g(x_{i:n}), h\left(\frac{i-0,5}{n}\right)\right) = \left(x_{i:n}, \Phi^{-1}\left(\frac{i-0,5}{n}\right)\right)$ se alinean formando una recta, se acepta la hipótesis de normalidad y la estimación de los parámetros μ y σ puede hacerse tras ajustar una recta a los mismos, notando que haciendo $\eta = 0$ y $\eta = 1$ en (7) se obtiene:

$$\begin{array}{lllll} \eta = 0 & \Rightarrow & 0 = \frac{\xi-\mu}{\sigma} & \Rightarrow & \xi_{\eta=0} = \mu \\ \eta = 1 & \Rightarrow & 1 = \frac{\xi-\mu}{\sigma} & \Rightarrow & \xi_{\eta=1} = \mu + \sigma \end{array} \tag{8}$$

La figura 2 muestra un papel probabilístico normal, donde el eje de las ordenadas ha sido transformado por $\eta = \Phi^{-1}(y)$ y el eje de abscisas no ha sufrido transformación alguna. En este papel se representan directamente los puntos de la función de distribución empírica: $\left(x_{i:n}, \frac{i-0,5}{n}\right)$.

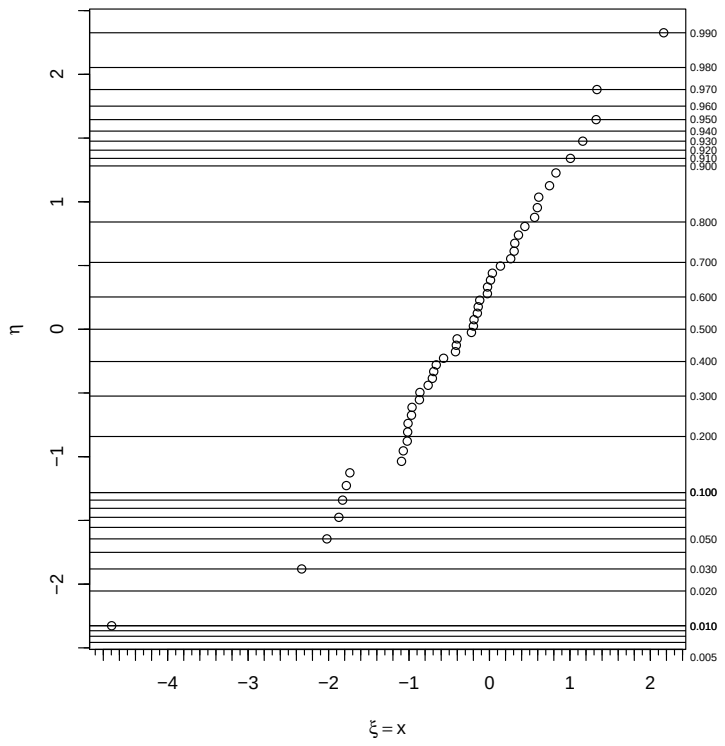

Figura 2: Papel probabilístico normal representando los datos de una muestra de tamaño 50 tomada de una distribución normal tipificada .

3.1. Papel normal en R

En R se pueden dibujar fácilmente las transformaciones de los datos necesarias para representar un conjunto de datos en papel normal. Para ello, basta utilizar la función `qnorm` como $h(y)$ para transformar las probabilidades empíricas. Sin embargo, en las instrucciones del fichero `papeles.R` se define una función `papel.normal` que facilita la representación. Para utilizarla basta hacer:

```
source("papeles.R")
load("normal.rda")   # carga la variable 'data'
papel.normal(data)
```

Como resultado se obtiene la Figura 3. La recta que mejor se ajusta a los datos mediante mínimos cuadrados se puede obtener utilizando el argumento `show.fit`:

```
papel.normal(data, show.fit=TRUE)
```

A menudo, los datos más extremos de una muestra no se ajustan bien a una recta. Esto se debe a que las regiones con poca probabilidad de ocurrencia (las colas de la distribución) necesitan de tamaños de muestra muy grandes para disponer de un número de datos suficiente como para representarlas adecuadamente. Es común que falten valores en rangos que tenian una pequeña probabilidad de ocurrir y que aparezcan valores en otros rangos más improbables. Por esta razón, a veces, es preferible descartar los datos más extremos de la muestra y ajustar la recta únicamente a los valores que realmente se alinean. La función `papel.normal` admite el argumento `trim=N`, para eliminar `N` puntos de cada cola antes de hacer el ajuste lineal. Pruébalo con los datos anteriores y observa el efecto sobre la recta y los parámetros de la ditribución:

```
papel.normal(data, show.fit=TRUE, trim=3)
```

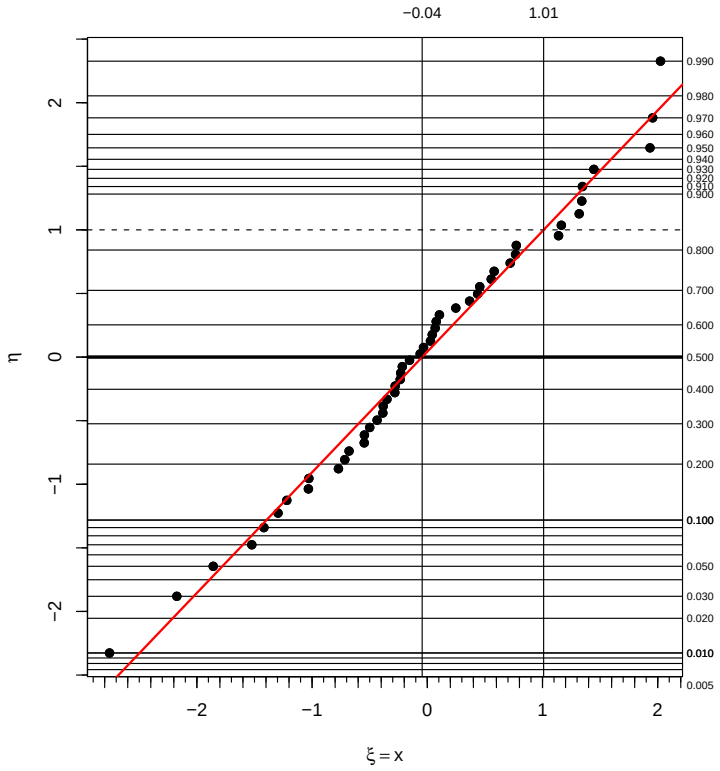

Figura 3: Papel probabilístico normal representando los datos del fichero normal.rda

Practica tú mismo

1) El fichero `resistencias.rda` contiene datos de resistencia a la compresión (en Kg/cm^2) de 40 probetas de hormigón.

1. Comprobar si se puede aceptar la hipótesis de que las resistencias se distribuyen de forma normal.
2. Obtener la resistencia media y la desviación típica a partir del modelo normal ajustado.
3. Si se define la resistencia característica como aquella que es superada por el 95 % de las muestras y damos por bueno el ajuste anterior ¿Cuál es la resistencia característica de este hormigón?

Practica tú mismo

2) El fichero `datos_papeles.rda` contiene 4 muestras de 30 datos cada una. Al cargar el fichero, están disponibles en las variables: `data1`, `data2`, `data3` y `data4`. Representar las muestras sobre papel normal y en caso de que sea aceptable este modelo, obtener los parámetros de la distribución.

4. Papel exponencial

El papel probabilístico exponencial transforma la función de distribución exponencial en una línea recta. Vamos a generalizar la variable aleatoria exponencial respecto a la que hemos visto en teoría, añadiéndole un parámetro de localización x_0, es decir:

$$F_{Ex(\alpha,x_0)}(x) = 1 - e^{-\alpha(x-x_0)} \qquad \forall x \geq x_0$$

Esta distribución se conoce como distribución exponencial desplazada. Haciendo $x_0 = 0$ se recupera la distribución exponencial ordinaria. Si hacemos las transformaciones

$$\eta = h(y) = -\ln(1-y) \qquad \xi = g(x) = x$$

podemos escribir la exponencial desplazada como

$$y = 1 - e^{-\alpha(x-x_0)} \;\Rightarrow\; \eta = h(y) = -\ln(1-y) = \alpha(x-x_0) = \alpha(\xi - x_0)$$

es decir

$$\eta = a\xi + b$$

con $a = \alpha$ y $b = -\alpha x_0$.

De nuevo, los cortes con las rectas $\eta = 0$ y $\eta = 1$, nos permiten obtener los parámetros:

$$\eta = 0 \quad \Rightarrow \quad 0 = \alpha(\xi - x_0) \;\Rightarrow\; \xi_{\eta=0} = x_0 \tag{9}$$

$$\eta = 1 \quad \Rightarrow \quad 1 = \alpha(\xi - x_0) \;\Rightarrow\; \xi_{\eta=1} = x_0 + \alpha^{-1} \tag{10}$$

4.1. Papel exponencial en R

El fichero `papeles.R` incluye la función `papel.exponencial` para dibujar fácilmente las transformaciones de los datos necesarias para representar un conjunto de datos en papel exponencial. Para ello, basta hacer:

```
source("papeles.R")
data <- rexp(40, 0.5)  # 40 datos aleatorios de una dist. Ex(0.5)
papel.exponencial(data)
```

Al igual que en el caso de la normal, esta función admite los argumentos `show.fit` y `trim`. Además tiene un argumento `allow.shift=FALSE`, que ajusta la versión no desplazada de la distribución expo-

nencial (es decir, obliga a tener $x_0 = 0$).

Practica tú mismo

---

3) El fichero `T1007_Llamadas_telefonicas.txt` contiene los tiempos en segundos entre 36 llamadas consecutivas realizadas en una central telefónica automática (Castillo y Pruneda, 2001)

1. Cargar estos datos en R.
2. Representar estos datos en papel probabilístico exponencial ¿Puede decirse que la ocurrencia de llamadas a esta centralita sigue un proceso de Poisson homogéneo?
3. Obtener la tasa de ocurrencia de llamadas en la centralita.

Practica tú mismo

---

4) Comprobar si alguna de las muestras del fichero `datos_papeles.rda` responde a una distribución exponencial y, en tal caso, obtener sus parámetros.