Seminario 1

Técnicas gráficas de ajuste y Método de Montecarlo.

Grau en Matemàtiques i Matemàtiques-Física (UAB)

1. Técnicas gráficas de ajuste.

Dada una muestra de la Uniforme en [0,1], $X_1, \dots X_n$, consideremos los estadísticos de orden (es decir, la muestra ordenada)

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

Las esperanzas de los estadísticos de orden

$$\mathbf{E}\left(X_{(1)}\right), \mathbf{E}\left(X_{(2)}\right), \ldots, \mathbf{E}(X_{(n)})$$

parten el intervalo $(0,1)$ en intervalos equiprobables, cada uno de longitud $\frac{1}{n+1}$: $\mathbf{E}\left(X_{(k)}\right) = k/(n+1)$.

Esto nos permitirá construir los llamados *probability plots*, que usaremos para responder a la pregunta: *¿Es plausible que los datos provengan de una distribución con pdf* $F(x)$? donde F pertenece a una familia de localización y escala.

Para construir el gráfico se procede de esta manera:

(i) Ordenamos los datos $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

Si la suposición es correcta, $F(x_{(1)}), F(x_{(2)}), \ldots, F(x_{(n)})$ es una muestra con la distribución de los estadísticos de orden de la $U(0,1)$.

(ii) Se resuelve (para $y_{(k)}$)

$$F(y_{(k)}) = \frac{k}{n+1}$$

Nota: recordar que si $X \sim F$, entonces $F(X) \sim$ U(0,1), y entonces $\mathbf{E}\left(F(X_{(k)}\right) = \frac{k}{n+1}$.

(iii) Se grafican los pares ordenados $\left(x_{(k)}, y_{(k)}\right)$.

Si la suposición $X_k \sim F$ es razonable, deberíamos tener aproximadamente una línea recta.

Si hay parámetros desconocidos hay que adaptar el método.

Ejemplo 1- Exponenciales: Si queremos ver si una distribución es exponencial con media μ, como $F(x) = 1 - e^{-x/\mu}$, entonces

$$\frac{k}{n+1} \approx 1 - \mathrm{e}^{-y_{(k)}/\mu}.$$

Reacomodando

$$y_{(k)} \approx -\mu \log\left(1 - \frac{k}{n+1}\right).$$

Entonces si la distribución es plausiblemente exponencial, al graficar $\left(x_{(k)}, \log\left(1 - \frac{k}{n+1}\right)\right)$ deberíamos obtener una recta, aproximadamente.

La pendiente será un estimador de μ.

Ejemplo 2: Distribución de Pareto,

$$F(x) = 1 - \left(\frac{x_m}{x}\right)^{\alpha}, \quad x > x_m,$$

donde x_m el extremo izquierdo del soporte de la distribución, y tanto α como x_m son parámetros desconocidos.

Podemos escribir

$$\frac{k}{n+1} \approx 1 - \left(\frac{x_m}{y_{(k)}}\right)^{\alpha}$$

y reacomodando los términos

$$y_{(k)} \approx x_m \left(1 - \frac{k}{n+1}\right)^{-1/\alpha}.$$

No podemos hacer el plot porque no conocemos el valor de α, pero tomando logaritmos a ambos lados tendremos

$$\log y_{(k)} \approx \log x_m - \frac{1}{\alpha} \log\left(1 - \frac{k}{n+1}\right),$$

y podemos graficar los pares

$$\left(\log x_{(k)}, \log\left(1 - \frac{k}{n+1}\right)\right)$$

que nos deberían producir una recta si el modelo es plausible.

1. Generar muestras aleatorias de tamaños $n = 15$, $n = 50$, $n = 100$ con distribución exponencial de parámetro $\lambda = 5$. Verificar el modelo con la ayuda de los plots anteriores..

2. En la librería `evir` hay un conjunto de datos llamado `danish` que corresponde a los montos de reclamaciones por incendios hechas a aseguradoras en Dimanarca entre el 3 de enero de 1980 y el 31 de diciembre de 1990. ¿Es razonable pensar que provienen de una distribución de Pareto?

   Ejemplo 3 - Distribución Normal:

   En la práctica, la distribución que más frecuentemente querremos ver si se adecúa a los datos es la normal, que tiene dos parámetros desconocidos.

   A los números

   $$z_{(k)} = \Phi^{-1}\left(\frac{k}{n+1}\right),$$

   donde Φ es la c.d.f. de $Z \sim N(0,1)$, se los conoce como *normal scores*. Son los $\frac{k}{n+1}$ cuantiles de la distribución normal estándar.

   Usamos los normal scores para construir el *normal probability plot*.

   Supongamos que queremos saber si los datos proceden o no de una distribución $N\left(\mu, \sigma^2\right)$.

   Como antes, ordenamos la muestra $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

   Si $X \sim N\left(\mu, \sigma^2\right)$, entonces

   $$\frac{X - \mu}{\sigma} \sim N(0,1).$$

   Graficamos los pares $\left(x_{(k)}, z_{(k)}\right)$.

   Si los datos son razonablemente normales, tendremos una recta con pendiente $1/\sigma$ y término independiente μ/σ.

   Las instrucciones `qqnorm(x)` y `qqline(x)` de `R` hacen el gráfico anteriormente descrito: en el eje horizontal se grafican los **cuantiles** teóricos, y en el vertical los cuantiles de la muestra.

La hoja de datos `michelson` de la libreria MASS tiene datos corresponden a un experimento realizado en 1882 por Michelson en el que intenta contrastar el estándar tomado hasta el momento como valor de la velocidad de la luz. Las medidas de la velocidad de la luz están tomadas en km/s - 299000.

Para leerlos tenemos que cargar la librería:

```
library(MASS)
```

y los leemos con la instrucción

`data(michelson);names(michelson)` nos permite ver cómo se llaman las variables.

```
qqnorm(michelson$Speed);qqline(michelson$Speed)
```

2. El método de Monte Carlo

Los métodos de simulación proporcionan un camino sencillo para aproximar probabilidades. Se simula un experimento aleatorio un gran número de veces y la probabilidad de un suceso cualquiera se aproxima mediante la frecuencia relativa del suceso en las repeticiones del experimento. La idea de usar simulaciones para modelizar experimentos aleatorios es muy antigua, durante la II guerra mundial, en Los Álamos se simulaban cascadas de neutrones en diferentes materiales. Como el trabajo era secreto, necesitaban un nombre clave, y le dieron el del famoso casino. Desde ese entonces, el uso de experimentos simulados para entender patrones probabilísticos se conoce como método de Monte Carlo.

Ejemplo

Teóricamente, dos muestras independientes con distribución normal deben tener correlación cero. Sin embargo, en la práctica esto no suele ocurrir. Veamos un ejemplo con dos muestras de tamaño 20 de la distribución normal típica y calculemos la correlación entre ellas.

```
nn =20
muestra.1 = rnorm(nn)
muestra.2 = rnorm(nn)
cor(muestra.1,muestra.2)
plot(muestra.1,muestra.2, pch=16)
abline(h=0, lty=2)
abline(v=0, lty=2)
title(paste("r =", round(cor(muestra.1, muestra.2), 3)))
```

Supongamos ahora que tenemos dos muestras aleatorias de distribuciones normales típicas y queremos decidir si son independientes o no mirando su correlación. Nuestra *hipótesis nula* es que sí lo son. Hemos visto que la correlación puede ser distinta de cero aun cuando las muestras sean independientes. Por lo tanto siempre podemos cometer dos tipos de error:

I- Podemos decir que no son independientes cuando sí lo son. Esto se llama un *error de Tipo I*.

II- Podemos decir que son independientes cuando no lo son. Esto es un *error de Tipo II*.

Ahora parece razonable que si la correlación que obtenemos es pequeña, nos inclinemos por la opción de independencia, mientras que si es grande nos inclinemos más bien por la ausencia de independencia. Esto quiere decir que escogemos un intervalo $(-\eta, \eta)$ y si la correlación empírica de nuestra muestra cae dentro de este intervalo decimos que las muestras son independientes, mientras que si cae fuera decimos que no lo son (o al menos que la evidencia no soporta la hipótesis de independencia). ¿Cómo escogemos η? Un

posible enfoque es tratar de controlar los errores que podemos cometer, es decir, escoger η de modo que la probabilidad de cometer un error de Tipo I esté controlada.

Podemos pedir, por ejemplo, que si tomamos dos muestras de tamaño 20 al azar, la probabilidad de cometer este tipo de error sea menor que 0.9.

Hagámoslo por medio de una simulación:

Crearemos una función mediante a cual simulamos dos muestras de tamaño n de distribuciones normales típicas y nos devuelve su coeficiente de correlación.

```
correlacion=function(n){
x=rnorm(n);y=rnorm(n)
cor(x,y)
}
```

Hagamos ahora 1000 réplicas de la correlación para $n = 20$

```
correla=replicate(1000,correlacion(20))
```

y buscamos el valor $\hat{\eta}$ tal que 10 % de las correlaciones queden fuera del intervalo $(-\hat{\eta}, \hat{\eta})$

```
quantile(correla,c(0.05,0.95))
```

Hacer un histograma de las correlaciones observadas. Agregar líneas verticales en $x = (-\hat{\eta}, \hat{\eta})$ y un punto en la correlación media observada.

```
hist(correla,probability=T,col="cornsilk")
abline(v=quantile(correla,c(0.1,0.9)),col="darkblue")
points(mean(correla),0,pch=8,col="maroon4")
```

Ejemplo: Taxis

Supongamos que los taxis de una ciudad están numerados de 1 a N, y una persona que se aburre ha decidido estimar N a partir de la observación de los números taxis que pasan frente a la terraza en la que pasa la tarde. Supongamos que el observador tiene la misma probabilidad de observar cualquiera de los N taxis de la ciudad en un momento determinado.

Al final del día el observador tiene una muestra $Y_1, \ldots Y_n$, y cada una de las Y_i está distribuida de acuerdo a la uniforme discreta en $\{1, \ldots, N\}$.

Consideremos dos estimadores posibles para N:

$$\hat{N}_1 = \text{máx}\{Y_1, \ldots, Y_n\} \tag{1}$$

$$\hat{N}_2 = 2\bar{Y} \tag{2}$$

Haremos un experimento de simulación para ver cuál de estos estimadores es mejor.

1. Escribir una función de N y n que genere la muestra de taxis observada y devuelva los valores de estos dos estimadores.
2. Hacer un experimento de simulación con 1000 repeticiones.
3. Hacer un histograma de las distribuciones de ambos estimadores.
4. Calcular los sesgos medios (recordar que conocemos N).
5. Cuál de estos estimadores tiene menor error cuadrático medio $\mathbf{E}(N_i - N)^2$, $i = 1, 2$.?