Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Saucedo

Seminari 2: Tècniques gràfiques d'ajust, estimadors i mètode de Montecarlo

A la primera part d'aquest seminari començarem veient algunes tècniques gràfiques que ens ajuden a verificar de quina distribució provenen les nostres dades. A la segona part, farem algunes simulacions de Montecarlo i calcularem alguns estimadors estadístics.

1 Tècniques gràfiques d'ajust

Donada una mostra Uniforme a $[0,1]$, $X_1, \dots X_n$, considerem els estadístics d'ordre, és a dir, la mostra ordenada

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

Les esperances dels estadístics d'ordre,

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

divideixen l'interval $(0,1)$ en intervals equiprobables, cadascun de longitud $\frac{1}{n+1}$: $\mathbf{E}\left(X_{(k)}\right) = k/(n+1)$ (penseu-ho!).

Aixo ens permetrà construir els anomenats *probability plots* que farem servir per respondre a la pregunta: *és possible que les dades provinguin d'una distribució amb cdf $F(x)$?*

Per a construir el gràfic procedim de la següent manera:

1. Ordenem les dades $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Si la suposició és correcta, $F(x_{(1)}), F(x_{(2)}), \dots, F(x_{(n)})$ és una mostra amb la distribució dels estadístics d'ordre de la $U(0,1)$.

2. Es resol (per $y_{(k)}$)
$$F(y_{(k)}) = \frac{k}{n+1}$$
Nota: recordeu que si $X \sim F$, aleshores $F(X) \sim U(0,1)$ (penseu-ho!), i llavors, $\mathbf{E}\left(F(X_{(k)})\right) = \frac{k}{n+1}$.

3. Es dibuixen els parells ordenats $\left(x_{(k)}, y_{(k)}\right)$.
Si la suposició $X_k \sim F$ és raonable, hauríem d'obtenir aproximadament una línia recta. Si hi ha paràmetres desconeguts, aleshores s'ha d'adaptar el mètode. Veiem alguns exemples.

Exemple 1 - Exponencials: si volem veure si una distribució és exponencial amb mitjana μ, com $F(x) = 1 - e^{-x/\mu}$, aleshores

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

Reescrivint,

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

Aleshores, si la distribució és plausiblement exponencial, al dibuixar $\left(x_{(k)}, \log\left(1-\frac{k}{n+1}\right)\right)$ hauríem d'obtenir aproximadament una recta.

El pendent d'aquesta recta serà un estimador de $-\mu$.

Exemple 2 - Distribució de Pareto: en aquest cas, tenim que

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

on x_m és l'extrem esquerre del suport de la distribució i, per tant, α i x_m són paràmetres desconeguts.

Aleshores, podem escriure,

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

que reescrivint,

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

En aquest cas, no podem fer el plot perquè desconeixem el valor de α, però prenent logaritmes a banda i banda, obtenim

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

i podem dibuixar els parells

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

que ens haurien de reproduir una recta si el model és raonable.

Exercici 1 Genereu mostres aleatòries de mida $n = 15, 20, 100$ que segueixin una distribució exponencial amb paràmetre $\lambda = 5$. Verifiqueu el model amb l'ajuda dels plots anteriors.

Exercici 2 A la llibreria `evir` (per instal·lar-ho fem `install.packages("evir")`) hi ha un conjunt de dades anomenades `danish` que es corresponen amb la quantitat en reclamacions per incendis fetes a asseguradores a Dinamarca entre el 3 de gener de 1980 i el 31 de desembre de 1990. És raonable pensar que provenen d'una distribució de Pareto?

Exemple 3 - Distribució Normal: A la pràctica, la distribució que més freqüentment voldrem veure si s'adequa o no a les nostres dades és la distribució normal que té dos paràmetres desconeguts. En aquest cas, als valors

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

on Φ és la c.d.f. de $Z \sim N(0,1)$, se'ls coneix com *normal scores*. Són els $\frac{k}{n+1}$ quantils de la distribució normal estàndard.

Utilitzarem els normal scores per construir el *normal probability plot*.

Suposem que volem saber si les dades provenen o no d'una distribució $N(\mu, \sigma^2)$. Com abans, ordenarem la nostra mostra, $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

Si $X \sim N(\mu, \sigma^2)$, aleshores

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

i dibuixem els parells $\left(x_{(k)}, z_{(k)}\right)$.

La funció Φ^{-1} es calcula mitjançant `qnorm(x)`. També disposem de dues funcions `qqnorm(x)` i `qqline(x)` que fan una feina molt semblant. Quina? Investigueu l'argument `qtype` de `qqline(x)`.

Exercici 3 El data frame `michelson` de la llibreria `MASS` conté les dades corresponents a un experiment realitzat l'any 1882 per Michelson en el que intentava contrastar l'estàndard pres fins aleshores com a valor de la velocitat de la llum. Les mesures de la velocitat de la llum estan preses en km/s - 299000. Comproveu si realment segueixen una distribució normal.

2 El mètode de Montecarlo

Els mètodes de simulació proporcionen un camí senzill per a aproximar probabilitats. Se simula un experiment aleatori un gran nombre de vegades i la probabilitat d'un esdeveniment qualsevol s'aproxima mitjançant la freqüència relativa d'aquest en cadascuna de les repeticions de l'experiment. La idea d'utilitzar simulacions per modelitzar experiments aleatoris és molt antiga, durant la II Guerra Mundial, a Los Álamos es necessitava simular cascades de neutrons en diferents materials. Com el treball era secret, van escollir com a nom en clau el del famós casino de Mònaco. Des d'aleshores, l'ús d'experiments simulats per entendre patrons probabilístics es coneix com a mètode de Montecarlo.

Exemple

Teòricament, dues mostres independents amb distribució normal han de tenir correlació zero. Tanmateix, a la pràctica, això no succeeix. Veiem un exemple amb dues mostres de mida 20 de la distribució normal típica i calculem la correlació entre elles

```
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)))
```

Suposem ara que tenim dues mostres aleatòries de distribucions normals típiques i volem decidir si són o no independents mirant la seva correlació. La nostra hipòtesi nul·la és que ho són. Hem vist que la correlació pot ser diferent de zero tot i que les mostres siguin independents. Per tant, sempre podem cometre dos tipus d'error.

- Podem dir que no són independents quan sí ho són. Això és un *error de Tipus I*.
- Podem dir que són independents quan no ho són. Això és un *error de Tipus II*.

Ara, sembla raonable dir que si la correlació que hem obtingut és petita, ens inclinem per la opció d'independència, mentre que si és gran ens inclinem per la hipòtesi alternativa. Això vol dir que escollim un interval $(-\eta, \eta)$ i si la correlació empírica de la nostra mostra cau dins d'aquest interval direm que las mostres són independents, mentre que si cau fora direm que no ho són (o almenys que la evidència no suporta la hipòtesi d'independència). Com escollim η? Un possible enfoc és

tractar de controlar els errors que podem cometre, és a dir, escollir η de manera que la probabilitat de cometre un error de Tipus I estigui controlada.

Podem demanar, per exemple, que si prenem dues mostres de mida 20 a l'atzar, la probabilitat de cometre aquest tipus d'error sigui menor que 0.9.

Fem-ho per mitjà d'una simulació:

Comencem creant una funció que ens permeti simular dues mostres de mida n que segueixin una distribució normal típica i ens retorni el seu coeficient de correlació.

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

Fem ara 1000 rèpliques de la correlació per mostres de mida $n = 20$.

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

i busquem el valor $\hat{\eta}$ tal que el 10% de les correlacions quedin fora de l'interval $(-\hat{\eta}, \hat{\eta})$

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

Exercici 4 Feu un histograma de les correlacions observades. Afegiu línies verticals a $x = (-\hat{\eta}, \hat{\eta})$ i un punt a la correlació mitjana observada.

Exemple: Taxis (`https://en.wikipedia.org/wiki/German_tank_problem`)

Suposem que els taxis d'una ciutat estan numerats de $1, \ldots, N$ i que una persona que s'avorreix ha decidit estimar N a partir de l'observació del nombre de taxis que passen per davant d'una terrassa en la que passa la tarda. Suposem que l'observador té la mateixa probabilitat d'observar qualsevol dels N taxis de la ciutat en un moment donat.

Al final del dia, l'observador té una mostra $Y_1, \ldots, Y_n$ i cadascuna de les Y_i està distribuïda seguint una uniforme discreta en $\{1, \ldots, N\}$.

Considerem dos estimadors possibles per N:

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

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

Farem un experiment de simulació per veure quin d'aquests dos estimadors és millor:

1. Escriviu una funció que depengui de N i n que generi la mostra de taxis observada i retorni els valors d'aquests dos estimadors.
2. Feu un experiment de simulació amb 1000 repeticions.
3. Feu un histograma de les distribucions d'ambdós estimadors.
4. Calculeu els biaixos mitjans (recordeu que coneixem N): donat un estimador $\hat{\theta}$ d'un paràmetre mostral θ, aleshores
$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

5. Quin d'aquests dos estimadors té un error quadràtic mitjà més petit?

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2],$$

$i = 1, 2$.

Hint: demostreu i utilitzeu que

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2.$$