STATISTICA 1, metodi matematici e statistici

Introduzione al linguaggio R

Esercitazione 5: 19-04-2004

Andrea Tancredi

Università di Roma “La Sapienza”, Rome, Italy

andrea.tancredi@uniroma1.it
http://3w.eco.uniroma1.it/utenti/tancredi

Sufficienza

Le statistiche sufficienti si possono definire nel modo seguente:

Dato un campione casuale $Y = (Y_1, \ldots, Y_n)$ $i.i.d$ da una v.a. avente densità $f(y; \theta)$, una statistica T si dice sufficiente per θ se la distribuzione di $Y|T = t$ non dipende da $\theta \; \forall \, t$.

Consideriamo il seguente esercizio. Dati dei campioni di numerosità 2 provenienti da una distribuzione di Poisson verifichiamo che la somma degli elementi del campione è effettivamente una statistica sufficiente mentre la loro differenza non lo è.

Generiamo 10000 campioni da una Poisson con media 1

```
> nsim <- 10000
> sim <- rpois(nsim * 2, 1)
> sim <- matrix(sim, ncol = 2)
```

Possiamo valutare la distribuzione congiunta di Y_1, Y_2 attraverso

```
> dc <- round(table(sim[, 1], sim[, 2])/nsim, 2)
```

Consideriamo ora solo quei campioni tali che la loro somma vale 3

```
> indici <- c()
> for (i in 1:nsim) {
+     if (sum(sim[i, ]) == 3)
+         indici <- c(indici, i)
+ }
> nsimc <- length(indici)
```

Calcoliamo la distribuzione condizionata di Y1,Y2 dato che la loro somma vale 3

```
> a <- round(table(sim[indici, 1], sim[indici, 2])/nsimc, 2)
```

```
> a

    0    1    2    3
0 0.00 0.00 0.00 0.13
1 0.00 0.00 0.36 0.00
2 0.00 0.40 0.00 0.00
3 0.11 0.00 0.00 0.00
```

Poichè $T = Y_1 + Y_2$ è sufficiente cambiando la media della Poisson e ripetendo tutti i comandi che abbiamo eseguito dovremmo ottenere la stessa distribuzione condizionata per Y_1, Y_2 dato $Y_1 + Y_2 = 3$

```
> nsim <- 10000
> sim <- rpois(nsim * 2, 2)
> sim <- matrix(sim, ncol = 2)
> indici <- c()
> for (i in 1:nsim) {
+     if (sum(sim[i, ]) == 3)
+         indici <- c(indici, i)
+ }
```

```
> nsimc <- length(indici)
> b <- round(table(sim[indici, 1], sim[indici, 2])/nsimc, 2)
> b

     0    1    2    3
  0 0.00 0.00 0.00 0.13
  1 0.00 0.00 0.35 0.00
  2 0.00 0.38 0.00 0.00
  3 0.14 0.00 0.00 0.00
```

Effettivamente le distribuzioni congiunte date dalle tabelle `a` e `b` sono quasi identiche; le differenze osservate diminuiranno se si incremeta il numero di campioni simulati

Provate a ripetere l'esercizio calcolando la distribuzione condizionata di Y_1, Y_2 dato $Y_1 + Y_2 = 4$ considerando sempre diversi valori per la media della Poisson

Consideriamo ora la statistica $Y_2 - Y_1$ e verifichiamo che non è sufficiente

Generiamo 10000 campioni di numerosità 2 da una Poisson di media 1 e calcoliamo la distribuzione condizionata di Y_1, Y_2 dato $Y_2 - Y_1 = 0$

```
> nsim <- 10000
> sim <- rpois(nsim * 2, 1)
> sim <- matrix(sim, ncol = 2)
> indici <- c()
> for (i in 1:nsim) {
+     if (diff(sim[i, ]) == 0)
+         indici <- c(indici, i)
+ }
> nsimc <- length(indici)
> a <- round(table(sim[indici, 1], sim[indici, 2])/nsimc, 2)
```

Il comando *di default* `diff` calcola le differenze tra l'elemento di posto i e l'elemento di posto $i - 1$

Consideriamo ora campioni da una Poisson con media 2

```
> nsim <- 10000
> sim <- rpois(nsim * 2, 2)
> sim <- matrix(sim, ncol = 2)
> indici <- c()
> for (i in 1:nsim) {
+     if (diff(sim[i, ]) == 0)
+         indici <- c(indici, i)
+ }
> nsimc <- length(indici)
> b <- round(table(sim[indici, 1], sim[indici, 2])/nsimc, 2)
```

```
> a

     0    1    2    3    4
0 0.44 0.00 0.00 0.00 0.00
1 0.00 0.44 0.00 0.00 0.00
2 0.00 0.00 0.11 0.00 0.00
3 0.00 0.00 0.00 0.01 0.00
4 0.00 0.00 0.00 0.00 0.00

> b

     0    1    2    3    4    5    6
0 0.09 0.00 0.00 0.00 0.00 0.00 0.00
1 0.00 0.36 0.00 0.00 0.00 0.00 0.00
2 0.00 0.00 0.36 0.00 0.00 0.00 0.00
3 0.00 0.00 0.00 0.15 0.00 0.00 0.00
4 0.00 0.00 0.00 0.00 0.04 0.00 0.00
5 0.00 0.00 0.00 0.00 0.00 0.00 0.00
6 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```

In questo caso le distribuzioni condizionate cambiano quando cambia la media della Poisson per cui $Y_2 - Y_i$ non è sufficiente. Provate a ripetere la simulazione aumentando il numero di simulazioni

Confronto tra due stimatori corretti

Consideriamo ancora dei campioni $i.i.d.$ provenienti da una popolazione avente distribuzione di Poisson. Supponiamo che il campione sia composto da 10 elementi.

```
> n <- 10
> x <- rpois(n, lambda = 2)
> x

 [1] 0 3 2 2 0 4 2 1 1 1
```

Poichè λ costituisce sia la media che la varianza della Poisson possiamo prendere in considerazione sia la media campionaria che la varianza campionaria corretta come stimatori di λ

```
> stima1 <- mean(x)
> stima2 <- var(x)
```

```
> stima1

[1] 1.6

> stima2

[1] 1.6
```

Supponiamo ora di voler valutare in generale il comportamento dei due stimatori che chiameremo T_1 e T_2

$$T_1 = \bar{X} \quad T_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_1 - \bar{X})^2$$

quando abbiamo un campione di 10 elementi da una Poisson con parametro 2. Diremo che T_1 si comporta meglio di T_2 se l'errore quadratico medio di T_1 è più piccolo di quello di T_2 dove

$$MSE(T) = E\left[(T-\lambda)^2\right]$$

In realtà T_1 e T_2 sono due stimatori corretti (provate a dimostrare che T_2 è corretto) quindi gli errori quadratici medi corrispondono alle varianze dei due stimatori.

Il comando `var` di R restituisce direttamente la varianza campionaria corretta... leggiamo infatti dall'help di `var` che *The denominator n - 1 is used which gives an unbiased estimator of the (co)variance for i.i.d. observations.*

Ci costruiamo allora 10000 campioni di numerosità 10 da una Poisson(2) e per ogni campione ci calcoliamo media e varianza

```
> mediap <- 2
> m <- 10000
> dati.simulazione <- rpois(n * m, mediap)
> campioni <- matrix(dati.simulazione, nrow = m, byrow = T)
> T1 <- apply(campioni, FUN = mean, MAR = 1)
> T2 <- apply(campioni, FUN = var, MAR = 1)
```

Prima di valutare gli errori quadratici medi diamo uno sguardo alle medie di T_1 e T_2 e anche alla densità della loro distribuzione

```
> mean(T1)

[1] 1.99667

> mean(T2)

[1] 2.009943

> plot(density(T1), type = "l")
> lines(density(T2), type = "l", col = 2)
```

Qualche idea su chi ha l'errore quadratico medio più alto?

Aggiungiamo una barra verticale in corrispondenza di 2 ovvero in corrispondenza del vero valore del parametro

```
> abline(v = 2, col = 3)
```

```
> mean((T1 - 2)^2)

[1] 0.200841

> mean((T2 - 2)^2)

[1] 1.106101
```

Tra i due lo stimatore ad avere l'errore quadratico medio più basso è la media campionaria, per lo meno quando la vera media della Poisson è due

In generale, per dire che T_1 è meglio di T_2 dobbiamo essere sicuri che l'errore quadratico medio di T_1 sia più piccolo di quello di T_2 quale che sia la media della popolazione

```
> griglia <- seq(0.5, 5, 0.5)
> MSE.poisson <- matrix(nrow = length(griglia), ncol = 3)
> n <- 10
> for (i in 1:length(griglia)) {
+     mediap <- griglia[i]
+     m <- 10000
+     dati.simulazione <- rpois(n * m, mediap)
+     campioni <- matrix(dati.simulazione, nrow = m, byrow = T)
+     medie <- apply(campioni, FUN = mean, MAR = 1)
+     varianze <- apply(campioni, FUN = var, MAR = 1)
+     MSE.poisson[i, 1] <- mediap
+     MSE.poisson[i, 2] <- mean((medie - mediap)^2)
+     MSE.poisson[i, 3] <- mean((varianze - mediap)^2)
+ }
```

Vediamo ora come variano gli errori quadratici medi dei due stimatori al variare di θ

```
> plot(MSE.poisson[, 1], MSE.poisson[, 3], type = "l", col = 2,
+     xlab = expression(lambda), ylab = "MSE")
> lines(MSE.poisson[, 1], MSE.poisson[, 2], col = 3)
> legend(x = 1, y = 5, legend = c("MSE varianza", "MSE media"),
+     col = c(2, 3), lty = c(1, 1))
```

L'errore quadratico medio della media campionaria è più basso dell'errore quadratico medio della varianza

Proviamo infine a confrontare la varianza campionaria corretta con quella non corretta...infatti non sempre uno stimatore corretto ha un errore quadratico medio più piccolo di uno non corretto, come infatti succede nel nostro esempio.

```
> griglia <- seq(0.5, 5, 0.5)
> MSE.poisson <- matrix(nrow = length(griglia), ncol = 3)
> n <- 10
> for (i in 1:length(griglia)) {
+     mediap <- griglia[i]
+     m <- 10000
+     dati.simulazione <- rpois(n * m, mediap)
+     campioni <- matrix(dati.simulazione, nrow = m, byrow = T)
+     medie <- apply(campioni, FUN = mean, MAR = 1)
+     varianze <- apply(campioni, FUN = var, MAR = 1)
+     MSE.poisson[i, 1] <- mediap
+     MSE.poisson[i, 2] <- mean((varianze - mediap)^2)
+     MSE.poisson[i, 3] <- mean((varianze * (n - 1)/n - mediap)
+ }
```

Rispetto al codice precedente abbiamo modificato solo l'ultima riga

Osserviamo infine il grafico degli errori quadratici medi dei due stimatori.

```
> plot(MSE.poisson[, 1], MSE.poisson[, 3], type = "l", col = 2,
+     xlab = expression(lambda), ylab = "MSE")
> lines(MSE.poisson[, 1], MSE.poisson[, 2], col = 3)
> legend(x = 1, y = 5, legend = c("MSE varianza non corretta",
+     "MSE varianza correta"), col = c(2, 3), lty = c(1, 1))
```