Lezione n. 5

5.1 Grafici e distribuzioni

ESEMPIO 5.1 *Legame tra Weibull ed esponenziale; TLC per v.a. esponenziali*
Supponiamo che $X \sim \text{Weibull}(\alpha, \beta)$.

(i) Si consideri la distribuzione di $Y = X^\beta$.

(ii) Fissato $\alpha = 1$, si analizzi la distribuzione esatta di $S_n = \sum_{i=1}^n Y_i$, dove $Y_i, i = 1, \cdots, n$ sono indipendenti e somiglianti e con la stessa distribuzione di Y.

(iii) Applicando il teorema limite centrale, si analizzi la distribuzione limite di S_n. Che cosa possiamo dire di $\Pr(S_{10} > z)$? Per un valore di z a scelta, confrontare i risultati ottenuti basandosi rispettivamente sulla distribuzione esatta e su quella approssimata di S_{10} e verificare l'errore commesso. Produrre una serie di rappresentazioni grafiche delle distribuzioni di S_n al variare di n.

SOLUZIONE.
Se X è una variabile casuale con densità $f_X(\cdot)$, la trasformata $Y = g(X)$ con $g(\cdot)$ funzione strettamente monotona *ha densità*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d(g^{-1}(y))}{dy} \right|.$$

Sapendo che la densità di X è $f_X(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}, \quad x > 0,\ \alpha > 0, \beta > 0$, la densità di $Y = X^\beta$ si scrive come:

$$f_Y(y) = \alpha\beta y^{\frac{\beta-1}{\beta}} e^{-\alpha y^{\frac{\beta}{\beta}}} \frac{1}{\beta} y^{\frac{1-\beta}{\beta}} = \alpha e^{-\alpha y}, \qquad y > 0$$

cioè Y ha una distribuzione esponenziale negativa di parametro α.

Utilizzando la funzione generatrice dei momenti di un'esponenziale negativa di parametro α, $G_Y(t) = \frac{\alpha}{\alpha - t}, \quad t < \alpha$, *e sfruttando l'indipendenza e la somiglianza delle* Y_i, *la funzione generatrice dei momenti di* $\sum_{i=1}^{n} Y_i$ *con* Y_i *i.i.d.* $\sim \text{Exp}(\alpha)$ *è pari a:*

$$G_Y(t) = \{\frac{\alpha}{\alpha - t}\}^n, \quad t < \alpha,$$

che è anche la f.g.m. di una $\text{Gamma}(n, \alpha)$. *Pertanto, ricordando che si considera* $\alpha = 1$, *la distribuzione di* S_{10} *è di tipo* $\text{Gamma}(10, 1)$.

Consideriamo ora la distribuzione limite di S_n. *Ricordando che ogni addendo in* S_n *è distribuito come un'esponenziale negativa di parametro 1 (e quindi anche media 1), si ottiene* $\text{E}(S_n) = n$, *e* $V(S_n) = \sum_i V(Y_i) = n$,*per cui, in base al teorema limite centrale si ha*

$$\frac{S_n - n}{\sqrt{n}} \sim_n N(0, 1) \tag{5.1}$$

In base al risultato limite appena riportato, possiamo approssimare $\Pr(S_{10} > z) = \Pr(\frac{S_{10}-10}{\sqrt{10}} > \frac{z-10}{\sqrt{10}})$ *con* $\Phi(\frac{z-10}{\sqrt{10}})$. *Conoscendo la distribuzione esatta di* S_n, *possiamo valutare la bontà dell'approssimazione (5.1).*

I comandi R per ottenere la probabilità cercata sono:
`pgamma(z,10,1,lower.tail=F)` *oppure* `1-pgamma(z,10,1)` *e*
`pnorm(z,10,sqrt(10),lower.tail=F)` *o ancora* `1-pnorm(z,10,sqrt(10))`,*oppure* `pnorm((z-10)/sqrt(10),lower.tail=F)`.

In ciascuno dei comandi z è un oggetto R in cui abbiamo salvato il valore in cui calcolare la probabilità della coda cercata. Ad esempio, ponendo `z_20` *, otteniamo:*

```
> z_20
> pnorm(z,10,sqrt(10),lower.tail=F)
[1] 0.05692315
> pnorm((z-10)/sqrt(10),lower.tail=F)
[1] 0.05692315
> 1-pnorm(z,10,sqrt(10))
[1] 0.05692315
> pgamma(z,10,1,lower.tail=F)
[1] 0.06985366
> zz <- seq(20,25,0.1)
```

come si può constatare ad esempio visualizzando `pgamma(zz,10,1,lower.tail=F)` *e* `pnorm((zz-10)/sqrt(10),lower.tail=F)`, *i valori*

basati sull'approssimazione normale decrescono molto più rapidamente di quelli basati sulla distribuzione esatta; l'approssimazione peggiora al crescere di z.

L'approssimazione alla normale basata sul teorema limite centrale diventa accettabile solo per n molto elevato. Il risultato può essere intuitivamente giustificato dal fatto che passiamo da una distribuzione fortemente asimmetrica (si ricordi che S_1 ha distribuzione esponenziale negativa) ad una distribuzione simmetrica quale è quella limite.

Questo può essere chiarito con una serie di grafici: a questo scopo, apriamo una nuova finestra grafica (comando `win.graph()`*) e suddividiamola ad esempio in 6 aree:* `par(mfrow=c(3,2))` *suddivide la finestra grafica attiva in una matrice di grafici con 3 righe e 2 colonne. N.B.: Se non è aperta alcuna finestra grafica, un comando* `par` *determina comunque l'apertura di una nuova finestra grafica. Plottiamo la distribuzione esponenziale corrispondente a S_1, le distribuzioni di S_n per $n = 1, 2, 5, 10, 30, 35$ e sovrapponiamo le distribuzioni normali con la stessa media e varianza di S_n:*

```
par(mfrow=c(3,2))
curve(dgamma(x,1,1), 0.01,60, main='n=1')
curve(dnorm(x, 1,1),col=2, add=T)
curve(dgamma(x,2,1),0.01,60, main='n=2')
#usato tra gli argomenti di plot(), main='titolo' inserisce un titolo nel plot
curve(dnorm(x, 2,sqrt(2)),col=2,add=T)
curve(dgamma(x,5,1), 0.01,60, main='n=5')
curve(dnorm(x, 5,sqrt(5)),col=2,add=T)
curve(dgamma(x,10,1),0.01,60, main='n=10')
curve(dnorm(x, 10,sqrt(10)),col=2,add=T)
curve(dgamma(x,30,1), 0.01, 60, main='n=30')
curve(dnorm(x, 30,sqrt(30)),col=2,add=T)
curve(dgamma(x,35,1), 0.01,60, main='n=35')
curve(dnorm(x,35,sqrt(35)),col=2,add=T)
```

■

5.2 Rappresentazioni grafiche: gli istogrammi

Dato un campione e quindi una distribuzione empirica, l'istogramma ci fornisce una rappresentazione grafica di tale distribuzione. R dispone del comando `hist`. Se `x` é un vettore in cui é salvato un campione estratto da una certa distribuzione, `hist(x)` restituisce

l'istogramma di x, ossia una rappresentazione grafica della distribuzione empirica dei dati. Il supporto osservato della variabile considerata viene suddiviso in classi e su ciascuna classe viene calcolata la frequenza assoluta (opzione di default, equivale a `prob=F`) o relativa (opzoine `prob=T`) di osservazioni che ricadono in ogni classe. L'ampiezza di ciascun intervallo viene stabilita da R in modo automatico. Si può comunque agire sul numero di classi con l'opzione `nclass=k` dove `k`è un intero, oppure sull'ampiezza degli intervalli, specificando un vettore contenente gli estremi di ciascuna classe, con il comando `breaks=c(a1,a2,...,ak)` dove `a1,...ak` sono gli estremi degli intervalli in cui viene suddiviso il supporto. In generale è raccomandabile usare l'opzione `prob=T` che calcola in ogni classe la frequenza relativa. In questo modo è anche possibile, laddove sia necessario, sovrapporre eventuali distribuzioni teoriche a quelle empiriche (comando `lines`, oppure `curve`, oppure `points`).

Esercizio 5.1. *Legge di Weibull*
Produrre grafici della densità Weibull con differenti scelte dei parametri.

Verificare inoltre tramite simulazione, utilizzando istogrammi, che se $X \sim \text{Weibull}(\alpha, \beta)$ (fissando α, β a piacimento), allora $Y = X^\beta \sim \text{Exp}(\alpha)$.

5.3 Rappresentazioni grafiche: il Q-Q plot per il confronto tra distribuzioni

Per individare, tramite uno strumento grafico, da quale modello sono stati verosimilmente estratti i dati, si può confrontare direttamente l'istogramma relativo alla distribuzione empirica con la densità di riferimento; così facendo, le frequenze relative degli intervalli in cui viene suddiviso il supporto osservato vengono confrontate con le probabilità teoriche di tali intervalli. In alternativa, è possibile confrontare le funzioni di ripartizione empirica e teorica direttamente, oppure attraverso il cosiddetto Q-Q plot, che mette a confronto graficamente i *quantili* della distribuzione empirica con i quantili omotetici (ossia dello stesso livello q) della distribuzione teorica di riferimento.

Il fondamento della tecnica risiede nel fatto che, se la distribuzione empirica si conforma al modello distribuzionale ipotizzato, i quantili empirici dovrebbero essere simili ai quantili "teorici" dello stesso livello q. Pertanto, se il modello da cui sono stati estratti i dati è effettivamente quello ipotizzato, in un grafico a disperisione che rappresenti sulle ascisse i quantili empirici e sulle ordinate i quantili teorici della distribuzione di riferimento, i punti dovrebbero disporsi lungo una retta a 45 gradi.

Se in particolare il modello di riferimento è la normale, abbiamo in R a disposizione la funzione `qqnorm`, che confronta i quantili della distribuzione empirica con quelli di pari livello

della normale *standardizzata*. Il fatto che si utilizzino i quantili della normale *standardizzata* anche se il modello di confronto è normale con media diversa da zero e/o varianza diversa da 1 è semplicemente dovuto al fatto che, se x_q è il quantile di livello q di una distribuzione $N(\mu, \sigma^2)$, si ha

$$q = \Pr(X \leq x_q) = \Pr(\frac{X-\mu}{\sigma} \leq \frac{x_q - \mu}{\sigma}),$$

da cui segue che $\frac{x_q-\mu}{\sigma} = z_q$ è il quantile di livello q di una normale standard. Di conseguenza $x_q = \sigma z_q + \mu$, cioè il quantile di livello q di una normale di media μ e varianza σ^2 si ottiene trasformando opportunamente il quantili omotetico della normale standard.

Questo implica che in generale, se si vuole stabilire se il modello da cui sono stati generati i dati è di tipo normale, è sufficiente calcolare i quantili empirici e confrontarli con i quantili della normale standard. Se i punti rappresentanti le coppie di quantili omotetici si dispongono lungo una retta, la pendenza e l'intercetta di tale retta forniscono inoltre un'informazione sul valore di σ e μ rispettivamente.

Una strategia alternativa, che consente di confrontare la distribuzione empirica con una distribuzione di riferimento non necessariamente normale, consiste nellutilizzare il comando `qqplot`. Con questo comando si confrontano le distribuzioni *empiriche* di due campioni per stabilire se essi provengano dalla stessa distribuzione. Nel nostro caso utilizzeremo però come secondo campione un vettore simulato dalla distribuzione di confronto. Di solito, avendo un campione di numerosità non molto elevata, i punti rappresentanti i quantili non si disporranno *esattamente* lungo una retta a 45 gradi, proprio perchè , per effetto della variabilità empirica, i quantili empirici non sarà mai esattamente uguali a quelli teorici. Naturalmente all'aumentare della numerosità campionaria, se la distribuzione ipotizzata corrisponde a quella da cui sono sati estratti i dati, l'effetto visivo di una retta aumenterà . In ogni caso sulle code della distribuzione avremo sempre una maggiore variabilità che nel centro della distribuzione, per cui gli scostamenti tra valori empirici e valori teorici saranno maggiori per valori di q vicini a 0 e 1 rispettivamente. per questo motivo si usa basare le proprie valutazioni sul 50% centrale della distribuzione. Per aiutare le valutazioni grafiche, si usa tracciare una retta a 45 gradi o, meglio, una retta passante per i punti definiti dal primo e terzo quartile empirico e teorico, $(q_{n,0.25}, z_{0.25})$ e $(q_{n,0.75}, z_{0.75})$. Mentre la retta a 45 gradi si traccia semplicemente con il comando `abline(0,1)` in cui 0 e 1 sono intercetta e coefficiente angolare rispettivamente, l'esempio successivo mostra come tracciare la retta passante per i punti corrispondenti ai quantili di cui sopra.

Esempio 5.2 *(segue da Esempio 5.1)*

In riferimento all'Esempio 5.1, costruiamo per simulazione campioni dalla legge esatta di S_n e su questi costruiamo il Q-Q plot per verificare la bontà dell'approssimazione normale fornita

dal teorema limite centrale, per diversi valori di n. Verificheremo che l'approssimazione non é soddisfacente anche per valori di n elevati, a causa dell'asimmetria della distribuzione di S_n, di cui conosciamo la legge esatta, che é di tipo Gamma$(n, 1)$.

I comandi per generare i campioni e costruire istogrammi e Q-Q plot sono i seguenti:

```
x.40 <- rgamma(5000,40,1)
hist(x.40, prob=T, nclass=40)
curve(dnorm(x,40,sqrt(40)),add=T)
qqnorm(x.40)
qqline(x.40)
#seconda opzione: confronto i quantili empirici di due campioni,
#uno dei quali estratto da una normale con parametri n, sqrt(n).
win.graph()
qqplot(rnorm(5000,40,sqrt(40)),x.40)
#comandi per tracciare la retta passante per i punti individuati
#dal primo e terzo quartile in ciascuno dei due campioni
coeff <- diff(quantile(x.40, c(0.25,0.75)))/diff(qnorm(c(0.25,0.75),40,sqrt(40)))
interc <- quantile(x.40, 0.25)-coeff*qnorm(0.25,40,sqrt(40))
abline(interc,coeff)
#aumentiamo n:
win.graph()
x.60 <- rgamma(5000,60,1)
hist(x.60, prob=T, nclass=40)
curve(dnorm(x,60,sqrt(60)),add=T)
qqnorm(x.60)
qqline(x.60)
```

Come si vede sia dal confronto tra l'istogramma e la curva della densità normale di riferimento sia dal Q-Q plot, l'approssimazione normale non é buona, in particolare le distribuzioni differiscono sulle code. Il confronto tra quantili "estremi" indica che infatti che la distribuzione empirica é asimmetrica, per cui sia i quantili sulla coda sinistra che quelli sulla coda destra sono maggiori dei quantili dello stesso livello relativi alla distribuzione normale. A titolo puramente esemplificativo, vediamo che cosa succede confrontando un campione da una normale con la distribuzione normale:

```
#campione da una normale standardizzata
a <- rnorm(5000)
```

```
qqnorm(a)
qqline(a)
#campione da una normale(3,1.5)
a <- rnorm(5000, 3, sqrt(1.5))
qqnorm(a)
qqline(a)
```

come si puó notare, i punti si dispongono molto più marcatamente lungo una retta, segno di una maggiore corrispondenza tra quantili empirici e teorici.