Lezione n. 3

3.1 Grafici e distribuzioni

Esempio 3.1 *[Leggi Gamma]*
Come visto in precedenza, R contiene funzioni precostituite che restituiscono la densità di tutte le più comuni distribuzioni. Come già visto la legge Gamma(α, λ) ha densità :

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\lambda x\}, \qquad x > 0,\ \alpha > 0,\ \lambda > 0. \tag{3.1}$$

La parametrizzazione adottata, in funzione di un parametro di forma (α) e di un tasso (λ), è richiamata in R con il comando `dgamma(x,alpha,rate=lambda)`.

Disegniamo la densità Gamma per diversi valori del parametro α e per $\lambda = 1$ utilizzando il comando `curve`:

```
curve(dgamma(x,1,rate=1),from=0,to=25,n=500,ylim=c(0,0.4))
```

sottolineiamo che, come già visto, per $\alpha = 1$ la densità 3.1 coincide con la densità esponenziale (che è appunto una Gamma(1, λ)); in questo caso abbiamo disegnato la densità di un'esponenziale unitaria;

```
curve(dgamma(x,2,rate=1),add=T, lty=2)
curve(dgamma(x,3,rate=1),add=T,lty=3)
curve(dgamma(x,13,rate=1),add=T,lty=4)
```

notiamo che la densità è sempre asimmetrica, e che tale asimmetria tende a diminuire all'aumentare di α.

Esempio 3.2 *[Leggi Beta]*
Come si è visto, la legge Gamma presenta densità sempre asimmetriche. Una distribuzione che permette una maggiore flessibilità , ma che è ristretta all'intervallo $[0, 1]$, è la legge Beta.

Una v.a. X definita sull'intervallo $[0,1]$ e con densità

$$f_X(x) = \frac{x^{1-\alpha}(1-x)^{1-\beta}}{B(\alpha,\beta)}, \qquad x \in [0,1], \quad \alpha > 0,\ \beta > 0$$

è detta v.a. Beta con parametri (α, β). La costante di normalizzazione al denominatore è la cosiddetta funzione Beta: $B(\alpha,\beta) = \int_0^1 x^{1-\alpha}(1-x)^{1-\beta}\,dx$. Vale la seguente relazione:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Utilizzando la relazione $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, determinare i momenti della distribuzione è particolarmente semplice:

$$\mathrm{E}(X) = \frac{\int_0^1 x^{\alpha}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,dx = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}.$$

Analogamente per il momento secondo

$$\mathrm{E}(X^2) = \frac{\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,dx = \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

[In generale, $\mathrm{E}(X^k) = B(\alpha+k,\beta)/B(\alpha,\beta)$].
Di conseguenza, sfruttando la relazione $\mathrm{V}(X) = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2$ otteniamo per la varianza:

$$\mathrm{V}(X) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

A seconda dei parametri, la densità Beta può assumere andamenti molto diversificati: notiamo innanzi tutto che per $\alpha = \beta = 1$ si ottiene la distribuzione uniforme in [0,1]. Disegniamo la densità utilizzando il comando `curve` e richiamando la funzione R `dbeta`, che restituisce la densità Beta:

```
curve(dbeta(x,1,1), from=0,to=1, n=500,ylim=c(0,3))
```

Per $\alpha = \beta$, la densità risulta simmetrica, e concava o convessa a seconda che α ($=\beta$) sia minore o maggiore di 1. Aggiungiamo al grafico una densità Beta(2,2) (colore rosso) e una densità Beta(0.1,0.1) (colore verde):

```
curve(dbeta(x,2,2), n=500, add=T, col=2)
curve(dbeta(x,.1,.1), n=500, add=T, col=3)
```

come si può notare, si tratta di due curve simmetriche. Utilizziamo ora parametri diversi tra loro:

```
curve(dbeta(x,2,6), n=500, add=T, col=6)
curve(dbeta(x,6,2),  n=500, add=T, col=7, lty=2)
```

come si può notare, la curva è ora asimmetrica, con massa spostata verso destra o verso sinistra a seconda del segno di $\alpha - \beta$.

ESEMPIO 3.3 *[T di Student–trasformazioni]*

Al fine di visualizzare la forma della densità e le caratteristiche della distribuzione t di Student, in questo esempio genereremo un numero cospicuo di realizzazioni da una legge T(ν), ne disegneremo l'istogramma (il quale ci fornisce un'approssimazione della densità di riferimento) e sovrapporremo a quest'ultimo la densità T(ν).

Per generare realizzazioni di una v.a. t di Student, R dispone del comando `rt(n,df)` Tuttavia, per generare realizzazioni da questa legge, sfrutteremo *anche* alcune trasformazioni note. Il modo di procedere è valido in generale: se non sappiamo (o non vogliamo) generare direttamente realizzazioni di una v.a. Y, è a volte noto che se X ha una data distribuzione (e siamo in grado di generare realizzazioni di X), una data trasformazione $h(X)$ ha la stessa distribuzione di Y. In tal caso possiamo simulare realizzazioni di X, $(x_1, \ldots, x_n)$ per poi calcolare $(y_1 = h(x_1), \ldots, y_n = h(x_n))$. Il vettore $(y_1, \ldots, y_n)$ così determinato fornisce n realizzazioni della v.a. Y che volevamo simulare. Questo approccio è già stato utilizzato per generare campioni da una legge Chi Quadrato: si è infatti utilizzata la seguente trasformazione:

$$Y = \sum_{i=1}^{\nu} X_i^2 \text{ con } X_i \sim N(0,1);$$

di conseguenza si sono generati ν vettori di lunghezza n di realizzazioni di v.a. normali standard, si sono poi elevati al quadrato questi valori e li si sono sommati, dando luogo a un vettore di n realizzazioni di una variabile $Y \sim \chi^2_\nu$.

È noto che se $Z \sim$ N(0,1) e $Y \sim \chi^2_\nu$ con Z e Y indipendenti, allora la trasformata $T = \frac{Z}{\sqrt{Y/\nu}} \sim t_\nu$, cioè T ha una distribuzione t di Student con i gradi di libertà dati da quelli della variabile chi quadrato utilizzata al denominatore. Avendo già costruito la funzione `chiq`, che genera realizzazioni di variabili Chi quadrato con un numero di gradi di libertà a scelta (cfr. Esempio 3.1), possiamo pensare di generare realizzazioni di una t di Student con gli stessi gradi di libertà . Per costruire un generatore di realizzazioni da una legge t con g.d.l. assegnati scriveremo quindi

```
# 1) riprendiamo la funzione chisq, commentando
#    la parte in cui viene prodotto l'istogramma
#    del campione da una legge Chi quadro:
#
chiq_function (k=2,n=5000)
{
```

```
ck_rep(0,n)
for (i in 1:k){ck_ck+(rnorm(n))^2}
#ak_sort(ck)
#hist(ak,prob=T,nclass=40)
#lines(ak,dchisq(ak,k),col=3)
return(ck)
}
#
# 2) creiamo la funzione per generare campioni da una t:
#    (richiama la funzione chiq)
#
t.student<- function(num=5000, gdl)
{
numer<- rnorm(num) #genera num realizzazioni di normali standard
denom<-sqrt(chiq(k=gdl, n=num)/gdl)
t <- numer/denom
return(t)
}
```

In alternativa potremmo utilizzare la funzione interna a R `rchisq` e scrivere:

```
t.student.1<- function(num=5000, gdl)
{
numer<- rnorm(num) #genera num realizzazioni di normali standard
denom<-sqrt(rchisq(n=num,df=gdl)/gdl)
t <- numer/denom
return(t)
}
```

Infine, possiamo ricorrere alla funzione interna `rt(n,df)`. In ognuno dei casi precedenti, la funzione restituisce un campione, ovvero un vettore di lunghezza n che potremo rappresentare con un istogramma. Assegnando ad oggetti R i campioni di 5000 unità generati con ciascuno dei tre metodi, per rappresentarne l'istogramma e la curva di densità teorica, scriviamo

```
t.samp1 <- t.student(gdl=11)
t.samp2 <- t.student.1(gdl=11)
```

```
t.samp3 <- rt(n=5000,df=11)
par(mfrow=c(2,2))
hist(t.samp1, nclass=50, prob=T, main='primo metodo')
curve(dt(x,11),add=T,col=3)
hist(t.samp2, nclass=50, prob=T, main='secondo metodo')
curve(dt(x,11),add=T,col=3)
hist(t.samp3, nclass=50, prob=T, main='terzo metodo')
curve(dt(x,11),add=T,col=3)
```

Osserviamo che la distribuzione ha forma campanulare centrata sull'origine, che ricorda quella di una normale. Tuttavia, come si può notare dal grafico, la frequenza con cui vengono estratte osservazioni dalle code della distribuzione è più alta di quanto non ci si aspetti per un campione da una normale. Generando infatti una campione da una Normale Standard, con probabilità 0.997 le osservazioni sono concentrate nell'intervallo $[-3.5, 3.5]$, (infatti `1-2*pnorm(-3.5)` è pari a `0.9995347`), mentre generando da una T(11) potremmo frequentemente ottenere valori esterni a tale intervallo. Confrontiamo gli istogrammi e le densità nei due casi:

```
par(mfrow=c(2,1))
set.seed(1224) #imposta il seme della simulazione
norm.sam<-rnorm(5000)
hist(norm.sam,prob=T, nclass=50, main='normale')
set.seed(1224) #imposta il seme della simulazione
t.sam<-rt(5000,11)
hist(t.sam,prob=T, nclass=50, main='t di Student')
```

Come si vede utilizzando il comando `summary` che restituisce tra l'altro minimo e massimo di ogni campione,

```
> summary(t.sam)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-7.201000 -0.701200 -0.017360 -0.004057  0.712800  4.932000
> summary(norm.sam)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.61200 -0.70430 -0.01262 -0.01421  0.67230  3.89100
```

le realizzazioni della v.a. T possono avere valori più estremi di quelli del campione dalla normale. Il grafico delle densità chiarisce ulteriormente questo concetto:

```
par(mfrow=c(1,1))
curve(dt(x,11),add=F,col=3)
curve(dnorm(x),add=T,col=2)
```

le code della normale sono molto più basse di quelle della t. La probabilità di generare un'osservazione proveniente da una coda è dunque più alta sotto una legge t che non sotto una normale; naturalmente al crescere dei gradi di libertà la t tende alla normale e le differenze diventano trascurabili. Per gradi di libertà piccoli le differenze sono più rilevanti. Ad esempio

```
> 2*pt(3,5,lower.tail=FALSE)
[1] 0.03009925
> 2*pnorm(3,lower.tail=FALSE)
[1] 0.002699796
> 2*pt(3,11,lower.tail=FALSE)
[1] 0.01207984
> 2*pt(3,55,lower.tail=FALSE)
[1] 0.004051086
```

Al decrescere dei g.d.l., la presenza di valori estratti dalle code si fa sempre più grande. La situazione più estrema si ha per la legge T(1). Con un solo g.d.l. abbiamo infatti una distribuzione con code molto pesanti: si considerino ad esempio le probabilità assegnate agli insiemi $\mathbb{R} - [-3, 3]$ e $\mathbb{R} - [-30, 30]$:

```
> 2*pt(3,1,lower.tail=FALSE)
[1] 0.2048328
> 2*pt(30,1,lower.tail=FALSE)
[1] 0.02121280
```

Disegnando la densità T(1) e confrontandola con quella di una normale

```
curve(dnorm(x),-10,10,col=2)
curve(dt(x,1),add=T,n=1000)
```

notiamo immediatamente le differenze tra le due densità. La T con un grado di libertà è un caso particolare di *distribuzione di Cauchy*. Esprimendola in funzione di un parametro di locazione θ e di un parametro di scala β, essa ha densità

$$f(x; \theta, \beta) = \frac{1}{\pi\beta[1 + ((x - \theta)/\beta)^2]}.$$

La T(1) è una distribuzione di Cauchy localizzata sull'origine (cioè con $\theta = 0$) e con parametro di scala β pari a 1. La caratteristica più rilevante di questa densità è che essa non ha media, una conseguenza della proprietà di avere code molto pesanti. Come conseguenza pratica abbiamo che, simulando da una Cauchy(1) otteniamo con elevata probabilità valori estremi (ossia molto grandi o molto piccoli). L'effetto di queste osservazioni è che ad esempio le rappresentazioni grafiche possono essere molto poco leggibili a causa di queste osservazioni estreme.

Osserviamo che cosa succede simulando da una legge T(1).

Per simulare possiamo utilizzare il comando `rt(n,1)` oppure sfruttare la relazione X_1, X_2 i.i.d $\sim N(0,1) \Rightarrow Y = X_1/X_2 \sim t_1$. Scrivendo `set.seed(1233); t1.sam <- rt(5000,1)`, otteniamo

```
> min(t1.sam)
[1] -1350.661
> max(t1.sam)
[1] 93351.03
```

l'istogramma è condizionato dalla presenza di questi valori estremi, e non risulterà leggibile a meno che non restringiamo il *range* delle osservazioni: ad esempio

```
hist(t1.sam,prob=T,nclass=10000)
hist(t1.sam,prob=T,nclass=10000,xlim=c(-1000,1000))
hist(t1.sam,prob=T,nclass=10000,xlim=c(-200,200))
```

La media campionaria calcolata sul campione "completo" sarà molto instabile (al variare del campione), risentendo delle osservazioni estreme; pertanto non possiamo pensare di adottare la media per stimare il parametro di locazione della distribuzione (che in questo caso è zero). In casi come questo una possibilità è di "potare" il campione delle osservazioni più estreme.

```
y <- sort(t1.sam)
y[1:10]
y[4991:5000]
hist(y[11:4990],prob=T,nclass=50)
mean(t1.sam)
mean(y[11:4990])
```

Si osservi che potevamo anche generare un campione da una legge di Cauchy(0,1) osservando che si tratta di una t di Student con un grado di libertà e utilizzando anche

il fatto che un chi quadrato con un g.d.l. è una normale standard al quadrato; in sostanza si tratta cioè del rapporto tra due normali standard. Potremmo quindi scrivere `t1.sam2<- rnorm(5000)/rnorm(5000)`.

ESEMPIO 3.4 *[Legge F di Fisher]*

La legge F con m_1 e m_2 g.d.l. è introdotta derivando la distribuzione della trasformata

$$Y = \frac{X_1/m_1}{X_2/m_2} \tag{3.2}$$

con X_1, X_2 indipendenti e $X_1 \sim \chi^2_{m_1}$,$X_2 \sim \chi^2_{m_2}$. Per costruire campioni da questa legge procediamo in due modi:

- indirettamente, a partire da campioni indipendenti da chi quadrati utilizzando la trasformata (3.2);
- direttamente usando la funzione R `rf(n,df1,df2)`.

Nel primo caso per generare ad esempio una F(12,16) potremo scrivere

```
effe.sam <- (rchisq(5000,12)/12)/(rchisq(5000,16)/16)
hist(effe.sam,prob=T,nclass=50)
#sovrapponiamo la curva della densita' di riferimento:
curve(df(x,12,16),add=T)
```

Nel secondo caso scriveremo invece

```
effe.sam2 <- rf(5000,12,16)
hist(effe.sam2,prob=T,nclass=50)
#sovrapponiamo la curva della densita' di riferimento:
curve(df(x,12,16),add=T)
```