RAPPRESENTAZIONE ANALITICA DELLE DISTRIBUZIONI STATISTICHE CON R

Versione 0.4- 21 febbraio 2005

Vito Ricci
vito_ricci@yahoo.com

INDICE

1.0 Introduzione

Un'analisi statistica che può essere condotta su dei dati è la verifica che questi siano conformi ad un certo modello teorico. Tra le analisi di questo genere possiamo annoverare la rappresentazione analitica delle distribuzioni statistiche[1] (*distribution fitting*). Essa consiste nel trovare una funzione matematica interpolante che rappresenti opportunamente un fenomeno statistico osservato. Un problema che lo statistico spesso si trova ad affrontare è il seguente: si ha una serie di osservazioni di un carattere quantitativo che denotiamo con x_1, x_2,… x_n e si vuole verificare se tali osservazioni – che costituiscono un campione di una popolazione non nota a priori – provengono da una determinata popolazione caratterizzata da una funzione di densità di frequenza della quale si conosce l'espressione in termini analitici. Indichiamo con $f(x,\boldsymbol{\theta})$ tale funzione, che rappresenta la distribuzione del carattere, dove $\boldsymbol{\theta}$ è un vettore di parametri – valori caratteristici della distribuzione – da stimare sulla base dei dati disponibili.
Gli scopi della rappresentazione analitica possono essere:

a) descrittivi: perequazione di dati, interpolazione di dati mancanti, etc.;
b) investigativi: rientrano nel campo della statistica inferenziale; in particolare, si va alla ricerca di un modello teorico partendo da un campione di osservazioni.

Nella rappresentazione analitica possiamo individuare 4 fasi:

1) scelta della funzione (modello) che si adatta meglio alle caratteristiche della distribuzione dei dati;
2) stima dei parametri della funzione (modello) scelta;
3) calcolo del grado di accostamento delle frequenze osservate rispetto a quelle ottenute con il modello teorico;
4) applicazione di test statistici per saggiare la bontà del modello (*goodness of fit*) in una logica inferenziale; questa fase è importante soprattutto se le finalità della rappresentazione analitica sono di tipo investigativo.

Il presente lavoro si prefigge di affrontare il problema della rappresentazione analitica cogliendo alcuni aspetti teorici e soprattutto quelli pratici e di calcolo con l'ausilio del software statistico R[2] di cui ci siamo occupati in un intervento precedente[3].
R è un ambiente per l'analisi statistica dei dati molto versatile e potente del quale verranno trattati alcuni comandi in una serie di esemplificazioni di carattere pratico. In particolare si affronterà la rappresentazione grafica dei dati (§ 2.0), la scelta della funzione (modello) analitica (§ 3.0), la stima dei parametri di tale funzione (§ 4.0), la misura del grado di accostamento (§ 5.0) ed alcuni test statistici per la verifica in termini inferenziali del modello scelto (§ 6.0).
Per la piena comprensione di questo contributo si presuppone una conoscenza di base del software R acquisita dopo la lettura di "*An introduction to R*"[4]. I comandi di R riportati, se non specificato diversamente, sono compresi nel package denominato `stats` presente nella versione base del software.

2.0 Rappresentazione grafica

Un primo approccio esplorativo con i dati può essere quello grafico. L'uso dello strumento grafico, infatti, può aiutare notevolmente nel prosieguo della rappresentazione analitica. Attraverso un istogramma, ad esempio, si può avere una prima idea del tipo di funzione (o di modello) da scegliere.
Utilizziamo la possibilità offerta da R di simulare campionamenti da popolazioni con funzione di densità di frequenza (o funzione di frequenza) nota (normale, Poisson, gamma, Weibull, etc.) e andiamo a tracciare dei

[1] G. Girone, T. Salvemini, Lezioni di statistica, 1990, vol. I, pag. 243 e segg.
[2] R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.r-project.org.
[3] V. Ricci, "R: un ambiente open source per l'analisi statistica dei dati", Economia e Commercio, n.1, 2004, pagg. 69-82
[4] R Core Team, An introdution to R, versione 2.0.1, novembre 2004

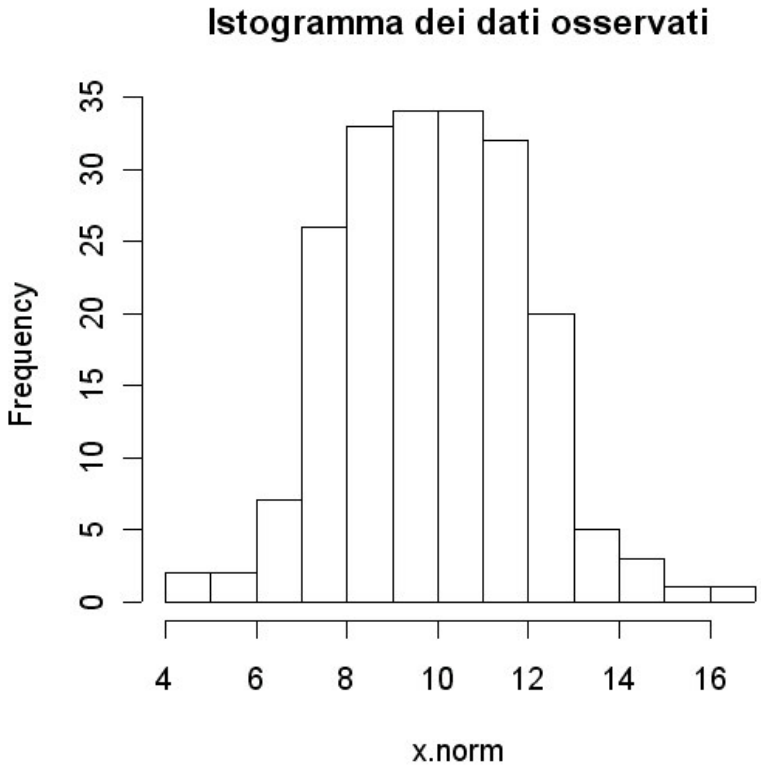
grafici. Simuliamo un campione di n=200 osservazioni provenienti da una popolazione normale N(10,2) con media aritmetica pari a 10 e scarto quadratico medio pari a 2:

```
x.norm<-rnorm(n=200,m=10,sd=2)
```

Possiamo ottenere l'istogramma di questi dati con il comando `hist()` (Fig. 1):

```
hist(x.norm,main="Istogramma dei dati osservati")
```

[Fig. 1]



Un ulteriore opportunità offerta da R è quella di stimare la funzione di densità della frequenza dei dati con il comando `density()` e di stamparne il grafico mediante il comando `plot()` (Fig. 2):

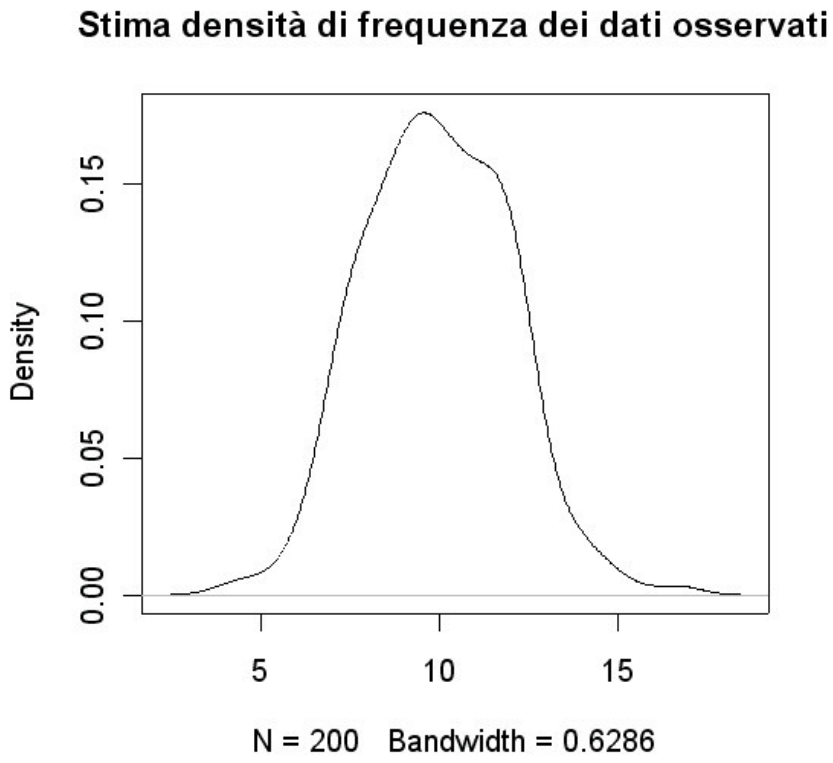
```
plot(density(x.norm),main="Stima densità di frequenza dei dati osservati")
```

Già da queste rappresentazioni grafiche possiamo iniziare a farci un'idea del modello teorico che meglio si adatta ai nostri dati. R permette anche di calcolare la funzione di ripartizione empirica[5] dei dati tramite il comando `ecdf()` - *empirical cumulative distribution function* - (Fig. 3):

```
plot(ecdf(x.norm),main="Funzione di ripartizione empirica")
```

[5] G. Girone, T. Salvemini, op. cit., vol. II, pag. 308

[Fig. 2]

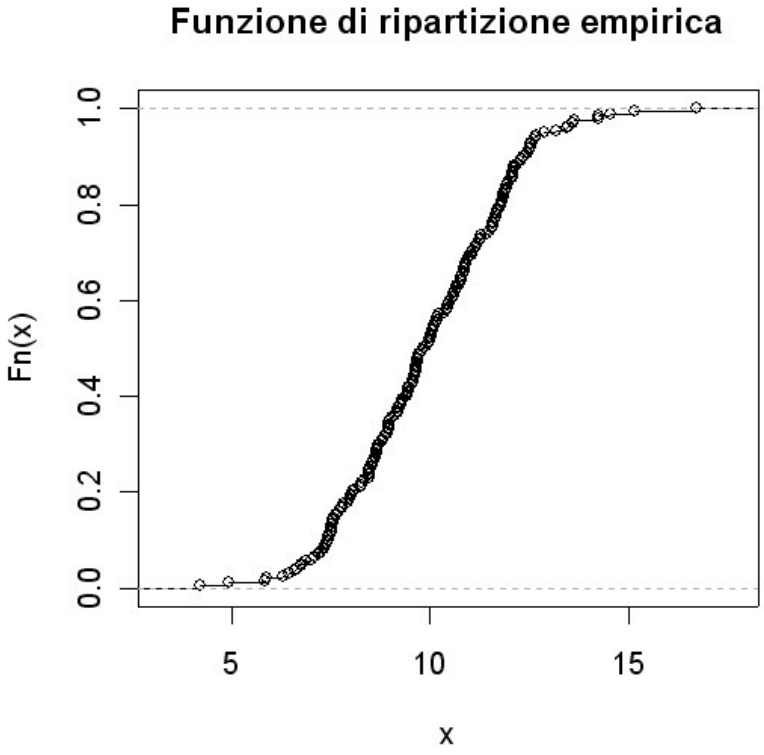


Altro strumento grafico che può aiutarci in questa fase è il *QQ plot*[6] tramite il quale si traccia il grafico con in ordinata i quantili[7] dei dati osservati e in ascissa quelli corrispondenti ottenuti con il modello teorico. Ciò in R può essere ottenuto tramite la funzione `qqnorm()`, nel caso si tratti di verificare la conformità alla distribuzione normale, oppure `qqplot()` o `qqline()`, nel caso si tratti di una qualsiasi distribuzione teorica. Nel nostro esempio abbiamo (Fig. 4):

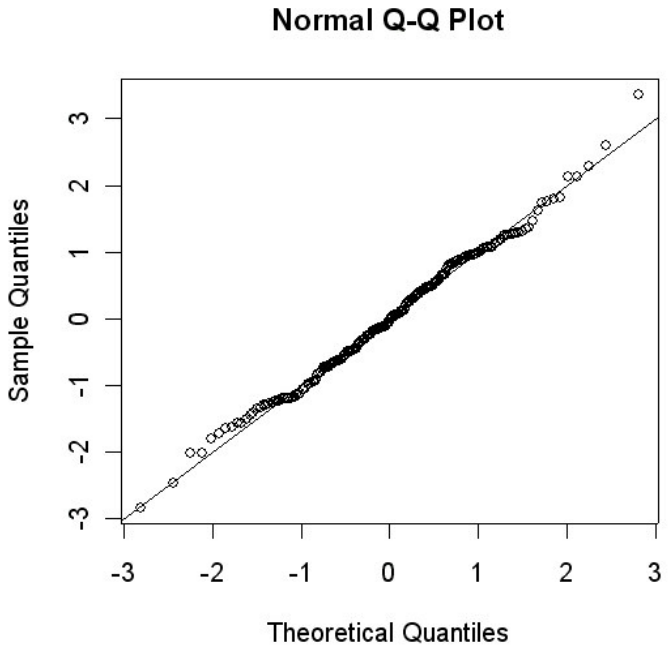
```
z.norm<-(x.norm-mean(x.norm))/sd(x.norm) ## standardizziamo i dati
qqnorm(z.norm) ## tracciamo il QQ plot
abline(0,1) ## tracciamo la diagonale
```

[6] Si veda http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm [consultato in data 2005-01-11]
[7] Per quantile si intende la frazione (o percentuale) di osservazioni inferiori ad un dato valore. Per esempio il quantile 0.3 (o 30%) è il punto in corrispondenza del quale il 30% dei dati è al di sotto di questo valore e il 70% è superiore.

[Fig. 3]


Funzione di ripartizione empirica
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
x

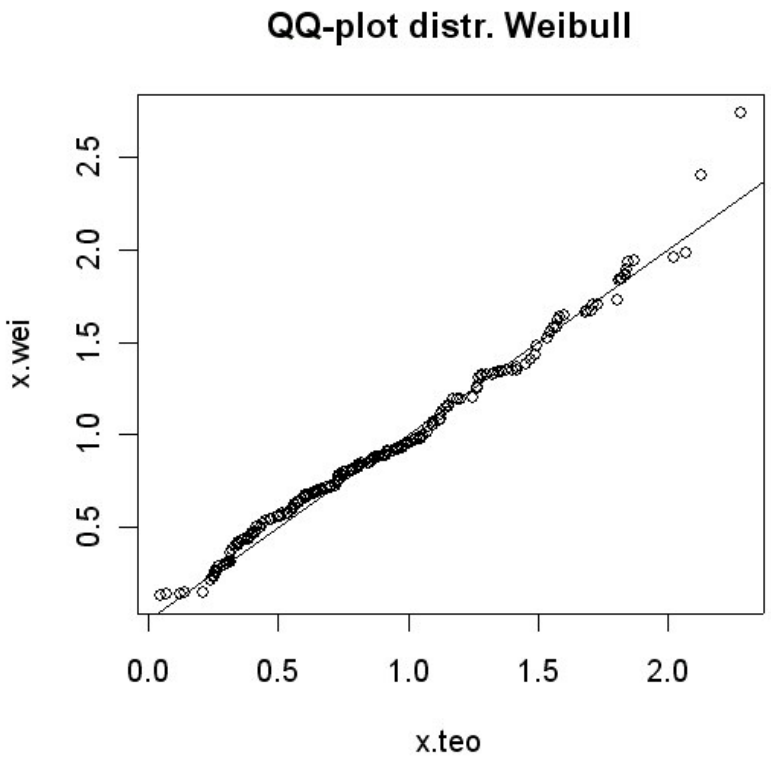
[Fig. 4]


Normal Q-Q Plot
Sample Quantiles
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Se i punti del plot tendono a disporsi sulla diagonale vuol dire che i dati si adattano bene al modello gaussiano. Nel caso di osservazioni di tipo differente da quello normale (ad esempio una funzione di densità della frequenza Weibull) utilizziamo `qqplot()` nel seguente modo (Fig. 5):

```
x.wei<-rweibull(n=200,shape=2.1,scale=1.1)      ##      simulazione      di
campionamento da una distribuzione di Weibull con parametri shape=2.1 e
scale=1.1
x.teo<-rweibull(n=200,shape=2, scale=1) ## calcolo dei valori teorici con
parametri della popolazione supposti noti shape=2 e scale=1
qqplot(x.teo,x.wei,main="QQ-plot distr. Weibull") ## QQ-plot
abline(0,1) ## tracciamo la diagonale
```

[Fig. 5]



dove `x.wei` sono i dati empirici di cui si dispone e `x.teo` sono quelli stimati con la funzione teorica.

3.0 Scelta del tipo di funzione

La prima fase della rappresentazione analitica altro non è che un problema di specificazione: si tratta di formulare un modello matematico che descriva in maniera soddisfacente il fenomeno oggetto di indagine. Questa descrizione può essere semplicemente fenomenica o empirica se le finalità della rappresentazione analitica sono descrittive. Se gli scopi sono, invece, investigativi la formulazione del modello teorico deve basarsi su un approfondimento della struttura del fenomeno e dovrebbe precedere la fase di raccolta dei dati.
In alcune circostanze il tipo di funzione (o modello) può essere dedotto da plausibili ipotesi sulla natura o sulla struttura del fenomeno. Si sceglie un determinato schema matematico-teorico e successivamente si verifica se è conforme ai dati rilevati. In altri casi può essere d'aiuto il ricorso alla rappresentazione grafica (§ 2.0): dalla forma dell'istogramma è possibile stabilire approssimativamente la funzione che meglio si presta a rappresentare il carattere. Tale metodo, tuttavia, può essere alquanto soggettivo. Un metodo obiettivo per scegliere il tipo di curva teorica da usare nella rappresentazione analitica è il criterio K di

Pearson[8]. Dalla risoluzione di una certa equazione differenziale si perviene ad una famiglia di funzioni dei tipi più disparati atte a rappresentare quasi tutte le distribuzioni empiriche. Tali curve dipendono esclusivamente da 4 caratteristiche: media, variabilità, asimmetria e curtosi. Standardizzando la distribuzione, il tipo di curva dipende solo dalla misura di asimmetria e da quella di curtosi[9] sintetizzate nel criterio K:

$$K = \frac{\gamma_1^2(\gamma_2 + 6)^2}{4(4\gamma_2 - 3\gamma_1^2 + 12)(2\gamma_2 - 3\gamma_1^2)}$$

dove:

$\gamma_1 = \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{n\sigma^3}$ è l'indice di asimmetria

$\gamma_2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^4}{n\sigma^4} - 3$ è l'indice di curtosi.

A seconda del valore di K, che si ottiene in base alle osservazioni disponibili, corrisponde un tipo particolare di curve.
Riportiamo di seguito alcune funzioni analitiche molto diffuse e che verranno trattate nel presente lavoro; per ciascuna sono riportati la forma grafica e i comandi di R per ottenere tali grafici. Esiste in letteratura una notevole varietà di funzioni e modelli teorici che possono essere adoperati[10]. Nel caso di caratteri quantitativi discreti possiamo avere la distribuzione di Poisson (Fig. 6):

$f(x,\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$ con x=0,1,2,…

```
x.poi<-rpois(n=200,lambda=2.5)
hist(x.poi,main="Distribuzione di Poisson")
```

Nel caso di caratteri quantitativi continui abbiamo:

distribuzione normale (gaussiana)[11] (Fig. 7): $f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$ con $x \in R$

```
curve(dnorm(x,m=10,sd=2),from=0,to=20,main="Distribuzione normale")
```

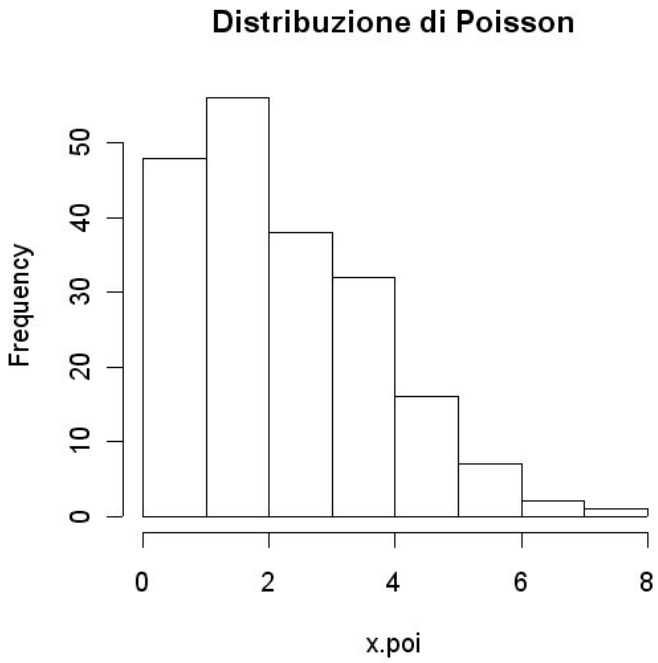
distribuzione gamma[12] (Fig. 8): $f(x,\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$ con $x \in R^+$

```
curve(dgamma(x, scale=1.5, shape=2),from=0, to=15, main="Distribuzione
Gamma")
```

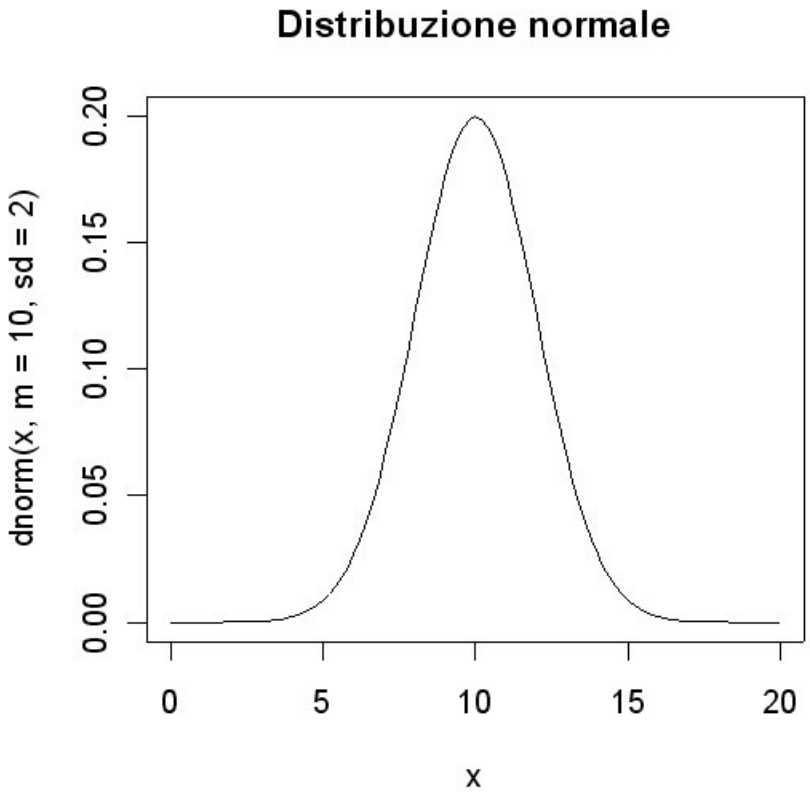
---

[8] G. Girone, T. Salvemini, op. cit., vol. I, pag. 256
[9] G. Girone, T. Salvemini, op. cit., vol. I, pag. 224 e segg. e http://www.cisi.unito.it/progetti/leda/cap8.htm [consultato in data 2005-01-11]
[10] Per farsi un'idea si consultino i seguenti siti: http://www.xycoon.com/continuousdistributions.htm, http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm e http://www.statsoft.com/textbook/stdisfit.html [consultati in data 2005-01-11]
[11] http://www.xycoon.com/normal.htm [consultato in data 2005-01-12]
[12] http://www.xycoon.com/gamma.htm [consultato in data 2005-01-12]

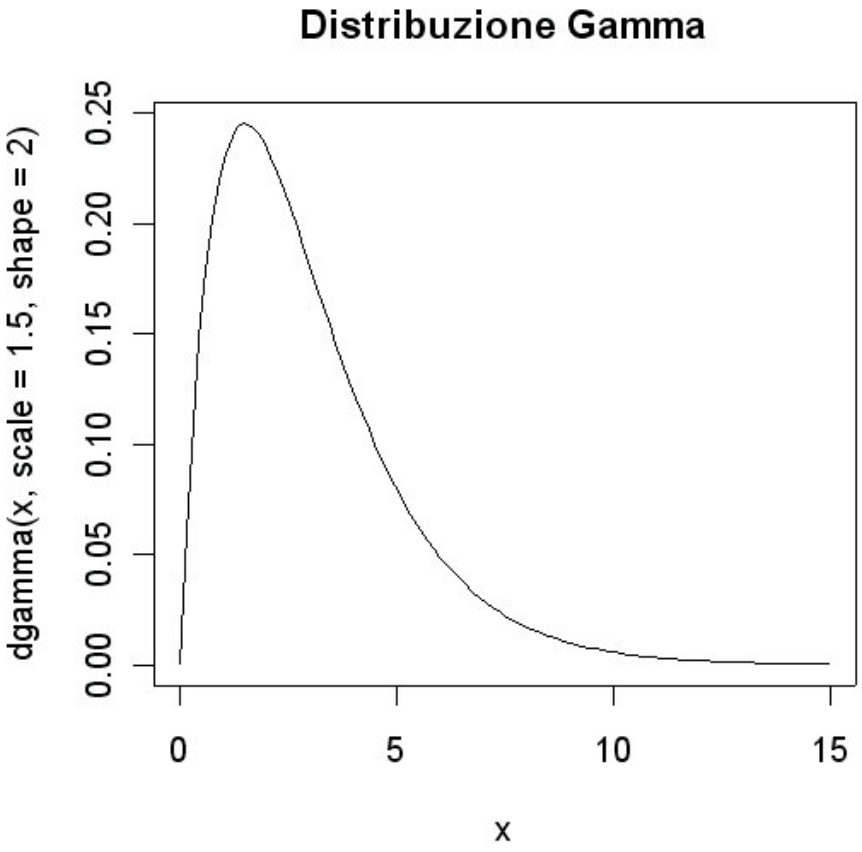
[Fig. 6]


Distribuzione di Poisson
Frequency
0
10
20
30
40
50
0
2
4
6
8
x.poi

[Fig. 7]


Distribuzione normale
dnorm(x, m = 10, sd = 2)
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
x

[Fig. 8]



distribuzione Weibull[13] (Fig. 9): $f(x,\alpha,\beta) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{[-(\frac{x}{\beta})^{\alpha}]}$ con $x \in R^{+}$

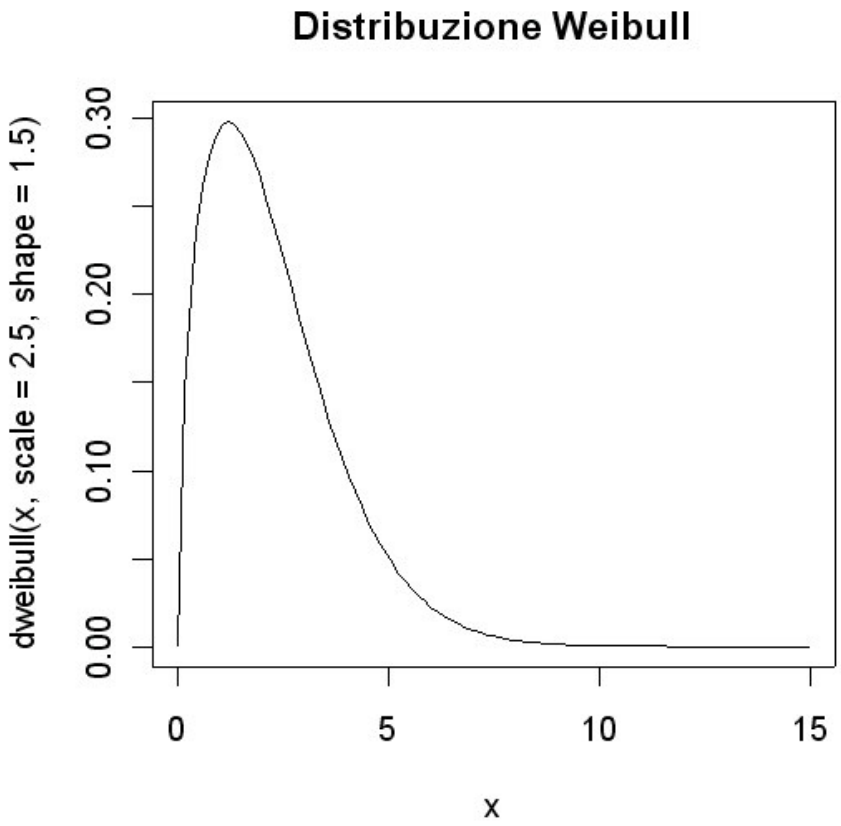
```
curve(dweibull(x, scale=2.5, shape=1.5),from=0, to=15,
main="Distribuzione Weibull")
```

Per calcolare le misure di asimmetria e curtosi di una distribuzione si può ricorre alle funzioni `skewness()` e `kurtosis()` contenute nel package `fBasics` (che occorre scaricare dal sito in quanto non compreso nella versione base di R):

```
library(fBasics) ## caricamento del package
skewness(x.norm) ## asimmetria dist. normale
[1] 0.1242952
kurtosis(x.norm) ## curtosi dist. normale
[1] 0.01372539

skewness(x.wei) asimmetria dist. Weibull
[1] 0.7788843
kurtosis(x.wei) ## curtosi dist. Weibull
[1] 0.4331281
```

[13] http://www.xycoon.com/Weibull.htm [consultato in data 2005-01-12]

[Fig. 9]



4.0 Stima dei parametri

Una volta scelta la funzione che meglio si adatta al fenomeno da rappresentare è necessario stimare i parametri che caratterizzano tale modello sulla base dei dati disponibili. Tra i metodi utilizzati più frequentemente ricordiamo[14]: il metodo delle ordinate fisse, il metodo delle somme, il metodo delle aree e il metodo dei minimi quadrati. Nel presente contesto tratteremo i seguenti metodi di stima:

1) analogico
2) dei momenti
3) della massima verosimiglianza

Il metodo analogico è abbastanza intuitivo e consiste nello stimare i parametri del modello applicando la formula degli stessi ai dati osservati. Ad esempio, si stima la media (non nota) di una popolazione normale con la media delle osservazioni campionarie.

```
stima.media<-mean(x.norm)
stima.media
[1] 9.935537
```

Il metodo dei momenti consiste nell'eguagliare i momenti empirici calcolati con i dati con quelli teorici determinati in base alla funzione scelta e al numero dei parametri da stimare. Definiamo i momenti empirici in questo modo:

[14] G. Girone, T. Salvemini, op. cit., vol. I, pag. 258 e segg.

- *momento empirico di ordine t dall'origine*: $m_t = \sum_{i=1}^{n} x_i^t y_i$ t=0,1,2…
- *momento empirico centrale (o dalla media) di ordine t*: $m_t^{'} = \sum_{i=1}^{n} (x_i - \mu)^t y_i$ t=0,1,2…

mentre quelli teorici:

- *momento teorico di ordine t dall'origine*: $m_t^{*} = \int_{\beta}^{\alpha} x^t f(x,\theta)\,\mathrm{dx}$ t=0,1,2…
- *momento teorico centrale (o dalla media) di ordine t*: $m_t^{'*} = \int_{\beta}^{\alpha} (x-\mu)^t f(x,\theta)\,\mathrm{dx}$ t=0,1,2…

dove $\beta-\alpha$ è il campo di variazione teorico in cui è definita f(x,**θ**), μ è la media della distribuzione e y_i sono le frequenze empiriche relative. Ad esempio, stimiamo i parametri di una distribuzione gamma con il metodo dei momenti considerando il momento del prim'ordine dall'origine, ossia la media, e il momento centrale di ordine due, ossia la varianza:

$$\frac{\alpha}{\lambda} = \bar{x}$$

$$\frac{\alpha}{\lambda^2} = s^2$$

dove nel primo membro sono riportate la media e la varianza della distribuzione gamma e nel secondo membro la media e la varianza non corretta delle osservazioni. Risolvendo si ottengono le stime dei parametri:

$$\hat{\lambda} = \frac{\bar{x}}{s^2}$$

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2}$$

```
x.gam<-rgamma(200,rate=0.5,shape=3.5)  ## simuliamo una distribuzione gamma con λ=0.5 (parametro di scala[15]) e α=3.5 (parametro di forma)

med.gam<-mean(x.gam) ## media dati simulati
var.gam<-var(x.gam) ## varianza dati simulati
l.est<-med.gam/var.gam ## stima dei lambda (corrisponde a rate)
a.est<-((med.gam)^2)/var.gam ## stima di alfa

l.est
[1] 0.5625486
a.est
[1] 3.916339
```

Il metodo della massima verosimiglianza (*maximum likelihood*) è un metodo utilizzato nella statistica

[15] Nella funzione `rgamma()` si può specificare l'argomento `rate` oppure l'argomento `scale=1/rate`; `rate` corrisponde al parametro λ

inferenziale per la stima puntuale dei parametri[16]. Abbiamo la funzione di densità della frequenza f(x,**θ**) che descrive il carattere quantitativo a livello di popolazione. Di tale funzione, che si suppone nota nell'espressione analitica, occorre stimare il vettore dei parametri **θ** in base alle osservazioni campionarie (dati osservati) $x_1, x_2, \ldots x_n$. Definiamo funzione di verosimiglianza la seguente espressione:

$$L(x_1, x_2, ...x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

che va vista come funzione del vettore dei parametri **θ**. Il metodo della massima verosimiglianza consiste nello stimare **θ** in modo che venga massimizzata $L(x_1, x_2, ...x_n, \theta)$, oppure, che è lo stesso, per semplificare i calcoli, il suo logaritmo. Nei casi più semplici si ricorre ai metodi dell'analisi matematica uguagliando a zero le derivate parziali; quando la funzione di verosimiglianza assume forme troppo complesse per calcolare le derivate parziali si usano i metodi di calcolo numerico per trovare il punto di massimo di $L(x_1, x_2, ...x_n, \theta)$.
Le stime di massima verosimiglianza (MLE, *maximum likelihood estimates*) godono di una serie di proprietà molto utili da punto di vista statistico e matematico [17].
Ad esempio, nel caso di una funzione gamma , la verosimiglianza assume la seguente espressione[18]:

$$L(x_1, x_2, ...x_n, \alpha, \lambda) = \prod_{i=1}^{n} f(x_i, \alpha, \lambda) = \prod_{i=1}^{n} \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} = (\frac{\lambda^\alpha}{\Gamma(\alpha)})^n (\prod_{i=1}^{n} x_i)^{\alpha-1} e^{-\lambda \sum_{i=1}^{n} x_i}$$

mentre il suo logaritmo è pari a:

$$\log(L) = n\alpha \log(\lambda) - n\log(\Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i$$

Nell'ambiente R si possono ottenere stime di massima verosimiglianza dei parametri di un modello con due comandi:

1) `mle()` compreso nel package `stats4`
2) `fitdistr()` compreso nel package `MASS`

La funzione `mle()` permette di ottenere delle stime di massima verosimiglianza dei parametri utilizzando dei metodi iterativi di calcolo numerico per minimizzare il negativo del logaritmo della funzione di verosimiglianza (che equivale a massimizzare il logaritmo della stessa funzione) specificando tale espressione come argomento e fornendo delle stime iniziali dei parametri. Nel caso di una distribuzione di tipo gamma abbiamo:

```
library(stats4) ## per caricare il package
ll<-function(lambda,alfa) {n<-200
                           x<-x.gam
-n*alfa*log(lambda)+n*log(gamma(alfa))-(alfa-
1)*sum(log(x))+lambda*sum(x)} ## -log(verosimiglianza)

est<-mle(minuslog=ll, start=list(lambda=2,alfa=1))
summary(est)
Maximum likelihood estimation

Call:
mle(minuslogl = ll, start = list(lambda = 2, alfa = 1))
```

[16] G. Girone, T. Salvemini, op. cit., vol. II, pag. 215. Per approfondimenti sul metodo della massima verosimiglianza si rimanda a: http://www.weibull.com/LifeDataWeb/maximum_likelihood_estimation_appendix.htm, http://www.itl.nist.gov/div898/handbook/eda/section3/eda3652.htm [consultati in data 2005-01-13]
[17] G. Cicchitelli, Probabilità e statistica, 1984, pagg. 160-162
[18] http://www-mtl.mit.edu/CIDM/memos/94-13/subsection3.4.1.html [consultato in data 2005-01-12]

```
Coefficients:
        Estimate Std. Error
lambda 0.5290189 0.05430615
alfa   3.6829126 0.35287672

-2 log L: 1044.282
```

Come valori iniziali per il calcolo sono stati scelti dei valori arbitrari, ma, ad esempio, si potevano usare le stime ottenute con il metodo dei momenti. La funzione `mle()` consente di ottenere stime di massima verosimiglianza qualunque sia la funzione di densità della frequenza, l'importante è conoscere l'espressione analitica della funzione di verosimiglianza che va ottimizzata.
Nel package `MASS` è previsto il comando `fitdistr()` che permette di ottenere delle stime di massima verosimiglianza dei parametri delle distribuzioni senza che sia necessario conoscere l'espressione della funzione di verosimiglianza. È sufficiente specificare il vettore dei dati, il tipo di funzione (`densfun`) ed eventualmente i valori iniziali per la procedura iterativa (`start`).

```
library(MASS) ## caricamento del package MASS
fitdistr(x.gam,"gamma") ## stima dei parametri distr. gamma
     shape          rate
  3.68320097   0.52910229
 (0.35290545) (0.05431458)

fitdistr(x.wei,densfun=dweibull,start=list(scale=1,shape=2))##      stima
parametri distr. Weibull
     scale         shape
  1.04721828   2.04959159
 (0.03814184) (0.11080258)

fitdistr(x.norm,"normal") ## stima parametri distr. normale
     mean          sd
  9.9355373   2.0101691
 (0.1421404) (0.1005085)
```

5.0 Misura del grado di accostamento

La determinazione del grado di accostamento[19] serve per verificare e valutare l'approssimazione tra le frequenze osservate del carattere e quelle calcolate con il modello teorico scelto. L'indice di accostamento dovrà essere quindi un'opportuna funzione delle differenze tra le frequenze empiriche e quelle teoriche, dovrà attribuire il medesimo peso a scarti uguali ma di opposto segno, dovrà essere funzione crescente degli scarti, nel senso che deve aumentare al crescere degli scarti. Si possono avere indici assoluti e indici relativi. Tra gli indici assoluti più usati abbiamo:

$\xi = \frac{\sum_{i=1}^{n} | y_i - y_i^* |}{n}$ media aritmetica semplice degli scarti in valore assoluto

$^2\xi = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{n}}$ media quadratica degli scarti

dove le y_i sono le frequenze osservate e le y_i^* sono le frequenze teoriche.

[19] G. Girone, T. Salvemini, op. cit., vol. I, pag. 283 e segg.

Tali misure dipendono dall'ordine di grandezza delle frequenze. Per eliminare tale inconveniente si ricorre agli indici relativi di accostamento che si ottengono rapportando gli indici sopra riportati alla media aritmetica (o quadratica) delle frequenze empiriche. Abbiamo pertanto:

$$\delta = \frac{\xi}{\sum_{i=1}^{n} y_i / n} = \frac{\sum_{i=1}^{n} | y_i - y_i^* |}{\sum_{i=1}^{n} y_i}$$

$${}^2\delta = \frac{{}^2\xi}{\sum_{i=1}^{n} y_i / n} = \frac{\sqrt{\sum_{i=1}^{n} (y_i - y_i^*)^2 / n}}{\sum_{i=1}^{n} y_i / n}$$

$${}^2_2\delta = \frac{{}^2\xi}{\sqrt{\sum_{i=1}^{n} y_i^2 / n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{\sum_{i=1}^{n} y_i^2}}$$

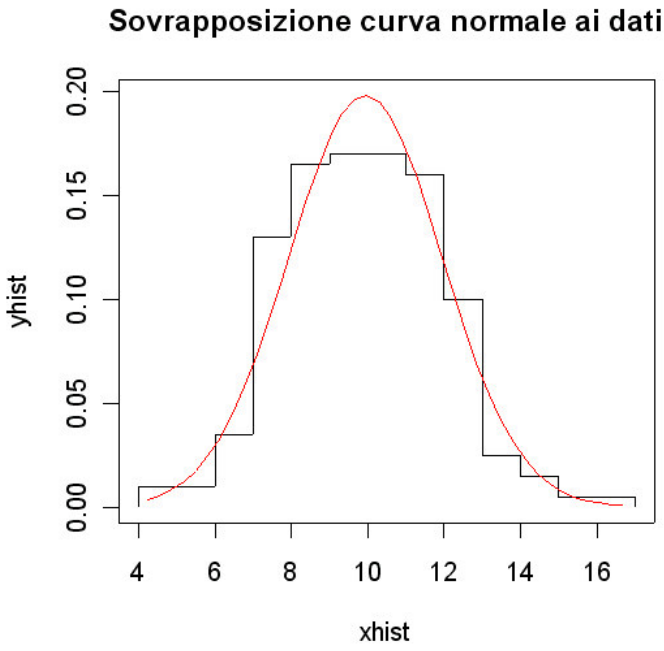
Tali indicatori sono di solito moltiplicati per 100 per esprimere il risultato ottenuto in termini di percentuale della corrispondente media.
Si riporta una esemplificazione di calcolo con R nel caso di funzione di frequenza di Poisson:

```
lambda.est<-mean(x.poi) ## stima del parametro lambda
tab.os<-table(x.poi)## tabella con frequenze empiriche
tab.os
x.poi
 0  1  2  3  4  5  6  7  8
12 36 56 38 32 16  7  2  1

freq.os<-vector()
for(i in 1: length(tab.os)) freq.os[i]<-tab.os[[i]] ## frequenze
empiriche
freq.ex<-(dpois(0:max(x.poi),lambda=lambda.est)*200) ## frequenze
teoriche

freq.os
[1] 12 36 56 38 32 16  7  2  1

freq.ex
[1] 13.850445 36.980688 49.369219 43.938605 29.329019 15.661696  6.969455
[8]  2.658349  0.887224

acc<-mean(abs(freq.os-trunc(freq.ex))) ## indice di accostamento assoluto
acc
[1] 2.111111

acc/mean(freq.os)*100 ## indice di accostamento relativo
[1] 9.5
```

Un'ulteriore verifica del grado di accostamento può essere fatta ricorrendo alla sovrapposizione della funzione di densità della frequenza teorica all'istogramma dei dati (Fig. 10).

```
h<-hist(x.norm,breaks=15)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(x.norm),max(x.norm),length=40)
yfit<-dnorm(xfit,mean=mean(x.norm),sd=sd(x.norm))
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)),
main="Sovrapposizione curva normale ai dati")
lines(xfit,yfit, col="red")
```

[Fig. 10]



6.0 Test statistici per verificare la conformità al modello

La verifica della conformità dei dati osservati ad un modello teorico (*goodness of fit*) può essere effettuata con diversi test statistici. Si tratta di test talvolta definiti *test omnibus*. Essi offrono un approccio globale al problema: la conformità tra i dati campionari e la popolazione viene esaminata in un quadro complessivo che include tutte le caratteristiche del carattere oggetto di studio (media, variabilità, forma della distribuzione, etc.). Tali test sono *distribution free* ossia risultano indipendenti dalla distribuzione del carattere. Particolare attenzione verrà prestata ai test di normalità.
Un primo test per verificare la *goodness of fit* dei dati osservati rispetto ad un modello teorico è il test χ^2 (chi-quadro)[20]. Esso si basa sul confronto delle frequenze empiriche con quelle attese calcolate in base alla funzione di densità della frequenza impiegata. Può essere utilizzato sia nel caso di caratteri discreti che continui, ed anche nel caso di parametri del modello stimati con i dati rilevati. La statistica di riferimento è:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

dove le O_i sono le frequenze assolute osservate, le E_i le frequenze assolute teoriche e k il numero di classi o intervalli in cui è stato diviso il carattere. Tale statistica si distribuisce asintoticamente secondo una variabile casuale χ^2 con k-p-1 gradi di libertà (p è il numero di parametri del modello stimati con i dati). Si accetta l'ipotesi di conformità al modello se il valore della statistica è inferiore al valore soglia, ovvero, se il p-value ottenuto dal test è superiore al livello di significatività prefissato.

[20] L. Soliani,, Statistica univariata e bivariata parametrica e non-parametrica per le discipline ambientali e biologiche, cap. III, pag. 1 e segg. e http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm [consultato in data 2005-01-15]

Le condizioni di validità del test χ^2 sono le seguenti:

- campione sufficientemente grande perché la distribuzione della statistica è asintoticamente χ^2
- il numero di frequenze attese entro ogni classe non deve essere minore di 5
- occorre applicare la correzione di Yates (o di continuità) per campioni di dimensione inferiore a 200 che consiste nell'aumentare di 0,5 gli scarti in valore assoluto $| O_i - E_i|$

Uno degli inconvenienti di questo test è la soggettività della ripartizione delle frequenze tra i vari gruppi.

In R abbiamo tre modalità per effettuare il test del chi-quadro.
Nel caso di carattere discreto si può ricorre alla funzione `goodfit()` presente nel package `vcd` (che va scaricato dal sito):

```
library(vcd)## caricamento del package

gf<-goodfit(x.poi,type= "poisson",method= "MinChisq")
summary(gf)

         Goodness-of-fit test for poisson distribution

              X^2 df P(> X^2)
Pearson 2.426986  7 0.932494

plot(gf,main="Approssimazione distribuzione di Poisson")
```

Nella funzione va specificato il tipo di distribuzione e il metodo di stima dei parametri adoperato.

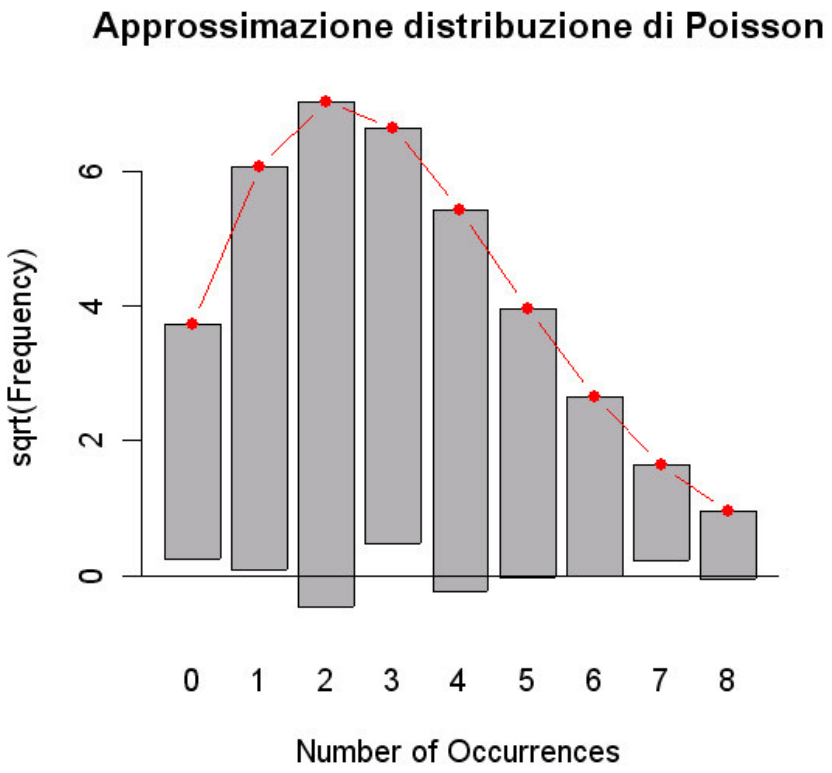
Nel caso di carattere continuo, nella fattispecie di dati provenienti da una distribuzione gamma, con parametri stimati con i dati osservati si procede in questo modo:

```
x.gam.cut<-cut(x.gam,breaks=c(0,3,6,9,12,18)) ##divisione del carattere
in intervalli
table(x.gam.cut) ##distribuzione con carattere diviso in intervalli
x.gam.cut
  (0,3]   (3,6]   (6,9]  (9,12] (12,18]
     26      64      60      27      23

## calcolo delle frequenze assolute teoriche
(pgamma(3,shape=a.est,rate=l.est)-pgamma(0,shape=a.est,rate=l.est))*200
[1] 19.95678

(pgamma(6,shape=a.est,rate=l.est)-pgamma(3,shape=a.est,rate=l.est))*200
[1] 70.82366

(pgamma(9,shape=a.est,rate=l.est)-pgamma(6,shape=a.est,rate=l.est))*200
[1] 60.61188

(pgamma(12,shape=a.est,rate=l.est)-pgamma(9,shape=a.est,rate=l.est))*200
[1] 30.77605

(pgamma(18,shape=a.est,rate=l.est)-pgamma(12,shape=a.est,rate=l.est))*200
[1] 16.12495

f.ex<-c(20,71,61,31,17) ## freq. teoriche distr. gamma
f.os<-vector()
```

```
for(i in 1:5) f.os[i]<- table(x.gam.cut)[[i]] ## freq. empiriche

X2<-sum(((f.os-f.ex)^2)/f.ex) ## statistica chi-quadro

gdl<-5-2-1 ## gradi di liberta
1-pchisq(X2,gdl) ## p-value
[1] 0.07652367
```

Si accetta l'ipotesi nulla di conformità al modello della distribuzione gamma del carattere osservato poiché il p-value ottenuto è superiore ad un livello di significatività pari almeno al 5%

[Fig. 11]



Se il carattere è continuo e la funzione di densità della frequenza è completamente specificata si può usare il comando `chisq.test()`:

```
## calcolo delle frequenze relative teoriche
p<-c((pgamma(3,shape=3.5,rate=0.5)-pgamma(0,shape=3.5,rate=0.5)),
(pgamma(6,shape=3.5,rate=0.5)-pgamma(3,shape=3.5,rate=0.5)),
(pgamma(9,shape=3.5,rate=0.5)-pgamma(6,shape=3.5,rate=0.5)),
(pgamma(12,shape=3.5,rate=0.5)-pgamma(9,shape=3.5,rate=0.5)),
(pgamma(18,shape=3.5,rate=0.5)-pgamma(12,shape=3.5,rate=0.5)))

chisq.test(x=f.os,p=p) ## test del X2

        Chi-squared test for given probabilities

data:  f.os
X-squared = 2.8361, df = 4, p-value = 0.5856
```

Anche il tale circostanza si accetta l'ipotesi di conformità al modello teorico.
Il test di Kolmogorov-Smirnov[21] per la bontà dell'adattamento (*Kolmogorov-Smirnov goodness of fit test*), per la sua ampia utilizzazione è proposto su molti testi di statistica applicata. Esso può essere utilizzato:

- sia per dati misurati su una scala ordinale discreta o dati continui raggruppati in classi, anche se non tutti gli autori sono d'accordo su questo punto[22]

- sia per dati continui, che possono essere misurati con una scala di rapporti oppure a intervalli oppure ordinale.

Il test si basa sul confronto tra la funzione di ripartizione empirica del carattere e quella teorica dedotta dal modello adottato[23]. Date n osservazioni ordinate Y_1, Y_2, ..., Yn, la funzione di ripartizione empirica è così definita:

$$F_n(Y_i)=N(i)/n$$

dove N(i) è il numero di osservazioni minori di Y_i (con i valori Y_i in ordine crescente). La statistica utilizzata per verificare l'ipotesi di conformità al modello teorico è:

$$D_n = \sup_{1 \le i \le n} |F(x_i) - F_n(x_i)|$$

ossia l'estremo superiore delle differenze in valore assoluto tra la funzione di ripartizione teorica e quella empirica. Si accetta l'ipotesi di conformità al modello se il valore della statistica è inferiore al valore soglia, ovvero, se il p-value ottenuto dal test è superiore al livello di significatività prefissato.
Il test di Kolmogorov-Smirnov è più potente del test $\chi 2$ in particolare quando il campione non è grande. Quando la numerosità del campione è grande, i due test hanno potenza simile e forniscono probabilità simili. Con il test di Kolmogorov-Smirnov i parametri della distribuzione teorica non possono essere stimati con i dati osservati, ossia la distribuzione deve essere completamente specificata (a causa di tale limitazione si ricorre al test di Anderson-Darling che, tuttavia, è disponibile solo per alcuni tipi di distribuzione).
Il software R mette a disposizione il comando `ks.test()`. Applichiamo tale test ai dati estratti da una popolazione Weibull, sapendo che i parametri di detta popolazione sono noti e pari a 2, il parametro di forma, e 1, il parametro di scala.

```
ks.test(x.wei,"pweibull", shape=2,scale=1)

        One-sample Kolmogorov-Smirnov test

data:  x.wei
D = 0.0623, p-value = 0.4198
alternative hypothesis: two.sided
```

Si accetta l'ipotesi di conformità al modello scelto poiché il p-value ottenuto è sufficientemente superiore ai livelli di significatività di solito adoperati nei test statistici.
Tracciamo il grafico con la sovrapposizione delle funzioni di ripartizione empirica e teorica dei dati osservati (Fig.12).

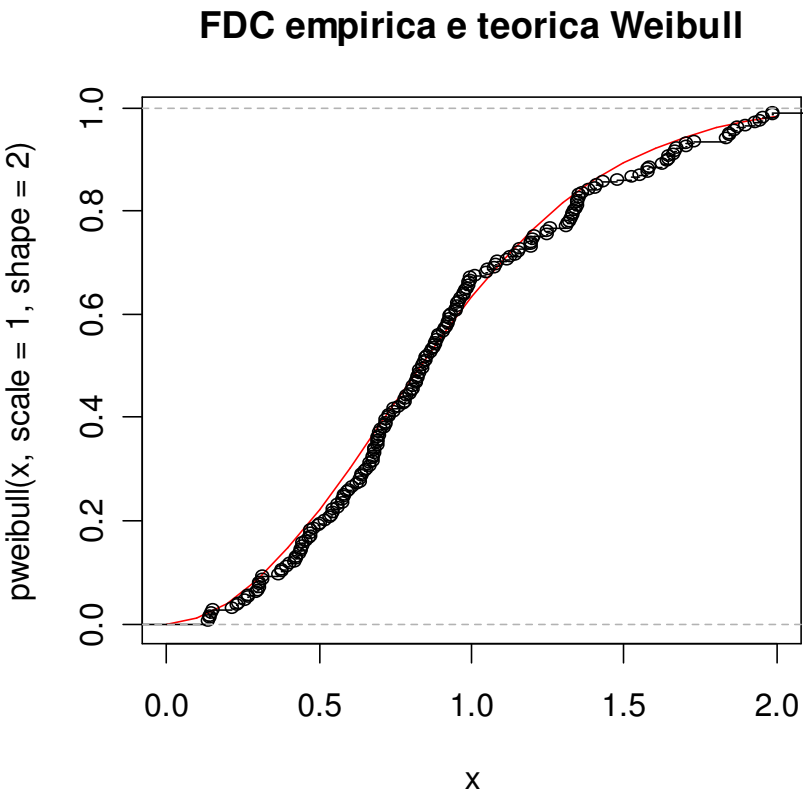
```
x<-seq(0,2,0.1)
```

[21] L. Soliani, op. cit., cap. VII, pag. 86 e segg. e http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm [consultato in data 2005-01-14]
[22] Cfr. G. Girone, T. Salvemini, op. cit., vol. II, pag. 311 e F. Del Vecchio, Statistica per la ricerca sociale, 1992, pag. 257

[23] Si definisce funzione di ripartizione teorica l'espressione $F(x) = \int_{\alpha}^{x} f(y,\theta)dy$ dove α e l'estremo inferiore del campo di variazione della funzione di densità della frequenza.

```
plot(x,pweibull(x,scale=1,shape=2),type="l",col="red", main="FDC empirica
e teorica Weibull")
plot(ecdf(x.wei),add=TRUE)
```

[Fig. 12]



6.1 Test di normalità

Molto spesso lo statistico è chiamato a verificare se i dati raccolti provengono o meno da una popolazione normale, data l'importanza di tale distribuzione nella metodologia statistica. Verranno di seguito esaminati i principali test di normalità[24].

Per completezza ricordiamo che esistono in letteratura dei test per verificare solo la simmetria o solo la curtosi (o entrambi contemporaneamente) di una distribuzione basati sui coefficienti b3 e b4 (o gamma3 e gamma4).[25]

Il test di Shapiro-Wilk[26] è considerato uno dei test più potenti per la verifica della normalità, soprattutto per piccoli campioni. La verifica della normalità avviene confrontando due stimatori alternativi della varianza s^2: uno stimatore non parametrico basato sulla combinazione lineare ottimale della statistica d'ordine di una variabile aleatoria normale al numeratore, e il consueto stimatore parametrico, ossia la varianza campionaria, al denominatore. I pesi per la combinazione lineare (a_i) sono disponibili su apposite tavole. La statistica W può essere interpretata come il quadrato del coefficiente di correlazione in un diagramma quantile-quantile (QQ plot).

[24] L. Soliani, op. cit., cap. XIII, pag. 33 e segg., E. Seier, Testing for normality [consultato in data 2005-02-18] e E. Seier, Comparison of tests for univariate normality [consultato in data 2005-02-18]

[25] Per approfondimenti si rinvia a L. Soliani, op. cit., cap. XIII, pagg. 36 - 46

[26] Si veda: http://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm [consultato in data 2005-01-15]

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Il comando per effettuare il test di normalità in questione in ambiente R è `shapiro.test()`: esso restituisce come risultato il valore della statistica W e il relativo p-value:

```
shapiro.test(x.norm)

        Shapiro-Wilk normality test

data:  x.norm
W = 0.9938, p-value = 0.5659
```

Il p-value è decisamente elevato rispetto ai livelli di significatività a cui di solito si fa riferimento: ciò ci fa propendere per l'ipotesi nulla ovvero la normalità della distribuzione dei valori osservati.

Il test di Jarque-Bera[27] è impiegato molto spesso per la verifica dell'ipotesi di normalità in campo econometrico. Esso si basa sulla misura dell'asimmetria e della curtosi di una distribuzione. Si considera in particolare la distribuzione asintotica di una combinazione dei noti coefficienti b3 e b4 (o gamma3 e gamma4) che, sotto l'ipotesi nulla, è di tipo chi-quadro con 2 gradi di libertà.
In R tale test è presente nel package `tseries` (che va scaricato dal sito in quanto non fa parte della versione base del software) ed è richiamabile tramite il comando `jarque.bera.test()` che restituisce il valore della statistica, i gradi di libertà e il p-value:

```
library(tseries) ## caricamento del package
jarque.bera.test(x.norm)

        Jarque Bera Test

data:  x.norm
X-squared = 0.539, df = 2, p-value = 0.7638
```

Il test di Cucconi[28] consente di verificare la normalità superando il problema dei parametri stimati con i dati campionari. Sia $x_1 \geq x_2 \geq \ldots \geq x_n$ un campione tratto da una popolazione continua; si estraggano n numeri casuali (o pseudocasuali) normali standardizzati $\zeta_1, \zeta_2, \ldots \zeta_n$ e posto:

$$r = \zeta_n \quad \text{e} \quad q = \sqrt{\frac{\sum_{i=1}^{n-1} \zeta_i^2}{n-1}}$$

si consideri la trasformata delle x_i: $y_i = q\frac{x_i - \bar{x}}{\hat{\sigma}} + \frac{r}{\sqrt{n}}$ dove $\bar{x}$ è la media del campione e $\hat{\sigma}$ è la radice quadrata della varianza campionaria corretta. Si dimostra che, se le x_i provengono da una popolazione normale, le y_i si distribuiscono secondo una normale standardizzata. Si può usare il test di Kolmogorov per verificare tale ipotesi. Riportiamo una esemplificazione in R:

```
zz<-rnorm(n=200,m=0,sd=1) ## generazione numeri pseudocasuali
r<-zz[200]
q<-sd(zz[-200])
```

[27] Si veda: http://homepages.uel.ac.uk/D.A.C.Boyd/JARQUE-B.PDF [consultato in data 2005-01-14]
[28] F. Del Vecchio, op. cit., pagg. 257-258

```
m<-mean(x.norm)
s<-sqrt(var(x.norm))
y<-q*((x.norm-m)/s)+(r/sqrt(200))
ks.test(y,"pnorm",m=0,sd=1)

        One-sample Kolmogorov-Smirnov test

data:  y
D = 0.0298, p-value = 0.9943
alternative hypothesis: two.sided
```

Esiste un apposito package di R denominato `nortest` (va scaricato dal sito) che permette di effettuare ben 5 diversi tipi di test di normalità:

1) `sf.test()` effettua il test di Shapiro-Francia:

```
library(nortest) ## caricamento del package
sf.test(x.norm)

        Shapiro-Francia normality test

data:  x.norm
W = 0.9926, p-value = 0.3471
```

2) `ad.test()` effettua il test di Anderson-Darling[29]:
questo test è una variante del test di Kolmogorov-Smirnov e può essere usato per verificare la conformità ai seguenti modelli: normale, lognormale, esponenziale, Weibull, valori estremi di I tipo e logistico. Esso si basa sulla statistica:

$$A^2=-n-S$$

con:

$S=\sum_{i=1}^{n}\frac{(2i-1)}{n}[\ln F(x_i)+\ln(1-F(x_{n+i+1}))]$, n l'ampiezza del campione e F(x) la funzione di ripartizione del modello teorico (nella fattispecie è la funzione di ripartizione della distribuzione normale). Nel software R tale test può essere fatto solo per verificare la normalità della distribuzione:

```
library(nortest) ## caricamento del package
ad.test(x.norm)

        Anderson-Darling normality test

data:  x.norm
A = 0.4007, p-value = 0.3581
```

3) `cvm.test()` effettua il test di Cramer-Von Mises:
si basa sulla seguente statistica:

$$W^2=\int_{-\infty}^{+\infty}(F_n(x)-F(x))^2 f(x)dx$$

```
library(nortest) ## caricamento del package
cvm.test(x.norm)

        Cramer-von Mises normality test
```

[29] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm [consultato in data 2005-01-18]

```
data:  x.norm
W = 0.0545, p-value = 0.4449
```

4) `lillie.test()` effettua il test di Lilliefors[30]:
è una variante del test di Kolmogorov-Smirnov particolarmente utile nel caso di campioni di piccole dimensioni. Esso consente di stimare i parametri del modello teorico con i dati campionari restando *distribution-free*.

```
library(nortest) ## caricamento del package
lillie.test(x.norm)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x.norm
D = 0.0414, p-value = 0.5509
```

5) `pearson.test()` effettua il test del chi-quadro di Pearson:
è lo stesso test del χ^2 del quale si è detto in precedenza applicato nel caso in cui si vuole verificare la conformità al modello normale:

```
library(nortest) ## caricamento del package
pearson.test(x.norm)

        Pearson chi-square normality test

data:  x.norm
P = 10.12, p-value = 0.753
```

7.0 Conclusioni

Nel presente lavoro si è affrontato il problema della rappresentazione analitica delle distribuzioni statistiche con l'ausilio del software R che è stato impiegato nelle varie fasi, dalla rappresentazione grafica dei dati ai test statistici per la verifica della conformità degli stessi al modello. Sono stati trattati sommariamente alcuni aspetti teorici, mentre ampio spazio hanno trovato gli aspetti pratici, con delle esemplificazioni, utilizzando i comandi dei vari package di R. Tale ambiente computazionale di analisi statistica si è dimostrato particolarmente utile e potente per la rappresentazione analitica delle distribuzioni statistiche.

[30] L. Soliani, op. cit., cap. XIII, pag. 50

Appendice

Elenco dei comandi di R utili per la rappresentazione analitica delle distribuzioni statistiche. Tra parentesi è riportato il package del quale fa parte il comando.

`ad.test()`: test di Anderson-Darling per la normalità (`nortest`)
`chisq.test()`: test chi-quadro (`stats`)
`cvm.test()`: test di Cramer-Von Mises per la normalità (`nortest`)
`ecdf()`: funzione di ripartizione cumulativa empirica (`stats`)
`fitdistr()`: stima di massima verosimiglianza dei parametri (`MASS`)
`goodfit()`: test di conformità per caratteri discreti (`vcd`)
`hist()`: istogramma (`stats`)
`jarque.bera.test()`: test di Jarque-Bera per la normalità (`tseries`)
`ks.test()`: test di Kolmogorov-Sminorv (`stats`)
`kurtosis()`: coefficiente di curtosi (`fBasics`)
`lillie.test()`: test di Lilliefors per la normalità (`nortest`)
`mle()`: stima di massima verosimiglianza dei parametri (`stats4`)
`pearson.test()`: test chi-quadro per la normalità (`nortest`)
`plot()`: diagramma cartesiano (`stats`)
`qqnorm()`: QQ-plot per distribuzione normale (`stats`)
`qqline()`, `qqplot()`: QQ-plot (`stats`)
`sf.test()`: test di Shapiro-Francia per la normalità (`nortest`)
`shapiro.test()`: test di Shapiro-Wilk per la normalità (`stats`)
`skewness()`: coefficiente di asimmetria (`fBasics`)
`table()`: tabella di contingenza (`stats`)

Riferimenti

D.M. BATES, "Using Open Source Software to Teach Mathematical Statistics", 2001 (scaricabile all'indirizzo: http://www.stat.wisc.edu/~bates/JSM2001.pdf)
G. CICCHITELLI, Probabilità e statistica, 1984
F. DEL VECCHIO, Statistica per la ricerca sociale, 1992
G. GIRONE, T. SALVEMINI, Lezioni di statistica, 1990
R CORE TEAM, An introcution to R, versione 2.0.1, novembre 2004 (scaricabile all'indirizzo: http://cran.r-project.org/doc/manuals/R-intro.pdf)
V. RICCI, "R: un ambiente open source per l'analisi statistica dei dati", Economia e Commercio, n.1, 2004, pagg. 69-82 (scaricabile all'indirizzo: http://www.dsa.unipr.it/soliani/allegato.pdf)
E. SEIER, Testing for normality
(scaricabile all'indirizzo: http://www.etsu.edu/math/seier/2050/normtest.doc)
E. SEIER, Comparison of tests for univariate normality
(scaricabile all'indirizzo: http://interstat.stat.vt.edu/InterStat/ARTICLES/2002/articles/J02001.pdf)
L. SOLIANI Statistica univariata e bivariata parametrica e non-parametrica per le discipline ambientali e biologiche, novembre 2004, (scaricabile all'indirizzo: http://www.dsa.unipr.it/soliani/soliani.html)
SYRACUSE RESEARCH CORPORATION, ENVIRONMENTAL SCIENCE CENTER
Selecting and Parameterizing Data-Rich Distributions PRA Center Short Course October 20-21, 2004
http://esc.syrres.com/pracenter/esf2004/downloads/classnotes/_2_Select%20Parameterize.ppt

Statistics - Econometrics - Forecasting: http://www.xycoon.com/
NIST/SEMATECH e-Handbook of Statistical Methods: http://www.itl.nist.gov/div898/handbook/
Reliability Engineering and Weibull Analysis Resources: http://www.weibull.com/