1 Modello di una grandezza aleatoria

1.1 Operazioni con Excel

Partiamo dal file *test_medicina* di tipo Excel. In esso sono riportati, relativamente al 2006, i punteggi del test di ammissione al corso di laurea in Medicina e Chirurgia di Milano statale, reperibili sul sito

$$http://www.accessoprogrammato.miur.it/2006/$$
$$page.php?cdl = ME\&pag = TA$$

Si tratta del punteggio di 1520 concorrenti (voto max. 80).

Immaginiamo di introdurre la v.a. X = punteggio di un partecipante al test sopra descritto, per Milano statale. Possiamo stabilire che distribuzione di probabilità ha X?

Come prima operazione, tracciamo un istogramma. Con Excel dobbiamo aprire, sotto Strumenti, Analisi dati, dove si trova Istogramma. Basta inserire la colonna di dati

$$Foglio1!\$B\$1 : \$B\$1520$$

nell'Intervallo di Input, e specificare la casella di partenza dell'intervallo di output (\$A\$1 del Foglio 2). I risultati sono riportati sul Foglio 2 del file. Excel sceglie automaticamente una partizione dei valori di X e riporta il numero di concorrenti con il voto relativo a ciascuna classe della partizione. Possiamo ora tracciare l'istogramma usando un grafico di dispersione.

L'istogramma ha una forma sorprendentemente gaussiana. Tra poco, con R, analizzaremo la gaussianità (qualcosa si può fare con Excel ma non è automatico). Invece, possiamo svolgere una verifica indiretta della gaussianità, con Excel, risolvendo il seguente esercizio.

Exercise 1 *Il voto minimo di ammissione, per ciascuna sede, è riportato sul sito*

$$http://www.ammissione.it/forum/$$
$$viewtopic-nc-t-11585.html$$

Potremmo congetturare il voto minimo conoscendo solo, di una sede, il voto medio, la deviazione standard, il numero di posti a disposizione ed il numero di concorrenti? Nel caso di Milano statale il voto minimo è stato 44,5. Era prevedibile conoscendo quei soli dati e non tutti i voti degli studenti?

Solution 2 *Ipotizziamo che X sia gaussiana. Media e deviazione standard dei punteggi sono*

$$\overline{x} = 34.98, \quad s = 11.1$$

mentre numero di posti e numeri di partecipanti sono

$$300, \quad 1520.$$

La probabilità di vittoria è quindi

$$\frac{300}{1520} = 0.197\,37$$

che va vista, graficamente, come l'area di una fettina a destra nell'istogramma. Il punto che delimita la fettina è la soglia di ammissione, cioè il voto minimo con cui si è ammessi. Nell'ipotesi di gaussianità, tale soglia è data dal quantile gaussiano di ordine $1 - 0.197\,37$. *Esso si trova con Excel come INV.NORM. Il risultato è 44.437, molto vicino al valore effettivamente accaduto. I calcoli si trovano sul Foglio 3 del file.*

1.2 Operazioni con R

Trasportiamo i dati su R. Salviamoli come testo. Prendiamo il Foglio 1 del file, salviamolo come testo delimitato da tabulazione, col nome test_medicina.txt. Chiudiamo Excel ed apriamo Blocco note per verificare il file. Apriamo poi R e cariachiamo il file, con il comando

$$Med < -read.table(file = \\ \text{“}C : /...percorso.../test_medicina.txt\text{”})$$

verificando poi il caricamento scrivendo Med.

Il semplice comando $hist(Med)$ che produrrebbe l'istogramma, non funziona perche' i nostri dati sono due colonne e non una. Se abbiamo avuto l'accortezza di salvare solo la colonna giusta, funziona. Altrimenti bisogna estrarla. Ecco una possibilità:

$$Med2 < -Med[, 2]$$

che poi verifichiamo chiedendo $Med2$.

$$hist(Med2)$$

produce ora l'istogramma desiderato. Se si vuole una maggior suddivisione in classi, si può aprire l'*help*, chiedere *hist* per scoprire in che pacchetto sta. Dopo aver visto che sta in *graphics*, si apre *guida Html* ed in essa *Packages*, poi il pacchetto *graphics*, ed infine si trova la funzione *hist*. Dalle istruzioni proviamo

$$hist(Med2, breaks = 40)$$

che pressapoco è ciò che ha fatto Excel. Il risultato è piuttosto buono.

Qui possiamo eseguire alcuni controlli sulla gaussianità. Uno di tipo grafico (oltre all'istogramma stesso) è il $Q-Q-plot$ *gaussiano*, che si ottiene con

$$qqnorm(Med2)$$

Il disegno è quasi rettilineo, segno di notevole gaussianità. Cos'è il $Q-Q-plot$ gaussiano?

Prima di tutto, ricordiamo il concetto di *Funzione di Distribuzione Cumulativa*

$$F(x) = P(X \leq x).$$

La sua inversa è il *quantile*. Poi ricordiamo il concetto di *Cumulativa Emprica* $\widehat{F}(x)$, l'analogo di $F(x)$ ma relativa ad un certo campione sperimentale. $\widehat{F}(x)$

si costruisce nel seguente modo: si prende il campione $x_1, ..., x_n$, lo si ordina in modo crescente, ottenendo la sequenza

$$x'_1 \leq ... \leq x'_n$$

Poi si traccia un grafico costante a tratti che vale 0 tra $-\infty$ e x'_1, vale $\frac{1}{n}$ tra x'_1 e x'_2, vale $\frac{2}{n}$ tra x'_2 e x'_3, $\frac{3}{n}$ tra x'_3 e x'_4, e così via, $\frac{n-1}{n}$ tra x'_{n-1} e x'_n, ed infine vale 1 da x'_n a $+\infty$. L'aspetto è simile a quello di una $F(x)$, ma costante a tratti. Anzi, è identico alla $F(x)$ relativa ad una distribuzione discreta con massa $\frac{1}{n}$ in ciascun punto x_i.

Intuitivamente, per decidere se il campione è sufficientemente gaussiano, basterebbe vedere se $\widehat{F}(x)$ somiglia a $F(x)$. Il problema è che, a parte il dettaglio del valor medio e della deviazione standard, tutti i grafici del tipo $F(x)$ si somigliano tra loro, nel senso che il nostro intuito geometrico non è sufficiente a classificarli (a differenza di quanto avviene per le densità). Ecco allora un trucco grafico che riporta il problema alle rette: per le rette il nostro intuito grafico è molto più sviluppato.

Per la $F(x)$ vale ovviamente la relazione

$$F^{-1}(F(x)) = x$$

cioè $F^{-1}(F(x))$ è una retta, la bisettrice del primo quadrante. Se $\widehat{F}(x)$ è una buona approssimazione di $F(x)$, allora anche la funzione

$$\widehat{F}^{-1}(F(x))$$

dovrebbe somigliare ad una retta. Il $Q-Q-plot$ gaussiano è appunto il grafico di questa funzione. Perciò giudichiamo la gaussianità in base alla somiglianza del $Q-Q-plot$ con una retta.

C'è però un dettaglio. Noi abbiamo a disposizione solo i dati, quindi $\widehat{F}(x)$. Inoltre stiamo esaminando la gaussianità, quindi è implicito nell'uso del comando $qqnorm(Med2)$ che si vuole il confronto con una $F(x)$ gaussiana. Ma quale? Nel senso: con che media e varianza? Ci sono due soluzioni possibili. La prima consiste nello stimare media e varianza dai dati ed usare quelli per trovare $F(x)$. Il SW non fa però questo, come si vede dal fatto che il $Q-Q-plot$ del nostro esempio non è simile alla bisettrice del primo quadrante (sembra tale per via delle particolari unità di misura degli assi). Infatti, ad esempio, per $x=0$ il valore del $Q-Q-plot$ è circa 33 (la media empirica), non zero. Allora il SW segue la seconda strada, speciale delle gaussiane: prende la cumulativa $\Phi(x)$ gaussiana canonica, cioè $F(x)$ relativa a media 0 e varianza 1, e disegna il grafico di

$$\widehat{F}^{-1}(\Phi(x)).$$

Se avessimo a che fare con le cumulative di altre distribuzioni, non potremmo aspettarci nulla di notevole in generale per questo grafico. Ma nel caso gaussiano si dimostra che

$$F^{-1}(\Phi(x)) = \mu + \sigma x$$

cioè è una retta. Quindi $\widehat{F}^{-1}(\Phi(x))$ deve somigliare ad una retta, se il campione è gaussiano.

Infine, svolgiamo il test di Kolmogorov-Smirnov. Il comando è

$$ks.test(Med2, \text{“}pnorm\text{”}, 34.9888, 11.1)$$

dove abbiamo preso i valori empirici di media e varianza. Restituisce un valore p pari a

$$p-value = 0.0778$$

che ad esempio non permetterebbe di rifiutare l'ipotesi

$$X \sim N\left(34.9888, 11.1\right)$$

al 95%. Non descriviamo ora il meccanismo alla base di questo test. Osserviamo solo che nel comando va specificato tutto della distribuzione di confronto.