

Poiché la distribuzione gamma dipende da due parametri, essa possiede moltissima flessibilità come modello per le distribuzioni reali. Per mostrare questa flessibilità nella figura 1 sono mostrati i grafici di parecchie funzioni di densità  $\chi^2$  corrispondenti quindi come si è già detto a  $\beta = 2$  nella funzione di densità gamma. Facendo osservare che, poiché  $\beta$  funge da parametro di scala, cambiando il valore di  $\beta$  la corrispondente curva di figura 1 semplicemente si stenderà se  $\beta$  aumenta e si comprimerà se  $\beta$  diminuisce, mantenendo però naturalmente uguale a 1 l'area da essa sottesa.

### Inferenza statistica

Finora ci siamo occupati di sviluppare i principi fondamentali della probabilità e di presentare alcune distribuzioni di probabilità che si sono dimostrate particolarmente utili nel risolvere certe classi di problemi reali.

Il motivo è stato quello di costruire modelli per esperimenti di tipo ripetitivo.

Il vantaggio di tali modelli è che ci permettono di studiare le proprietà dell'esperimento e di fare previsioni sui risultati relativi a future esecuzioni dell'esperimento cose che sarebbero entrambe difficili o impossibili da fare senza l'aiuto di un tale modello.

*Il processo di costruire un modello sulla base di dati sperimentali e di trarre conclusioni da esso è un esempio di inferenza induttiva, quando esso si applica a problemi statistici viene chiamato di solito inferenza statistica.*

Gli statistici si occupano principalmente di fare delle inferenze statistiche servendosi di dati sperimentali, molto spesso lo statistico è interessato a costruire un modello per una sola variabile casuale associata ad un esperimento, piuttosto che per l'intero esperimento.

Come conseguenza la maggior parte dei modelli scelti dagli statistici sono funzioni di densità di variabili casuali.

Le inferenze statistiche sono perciò di solito inferenze riguardanti le funzioni di densità.

Come esempio di quanto detto supponiamo che un biologo abbia osservato che su 200 insetti di una data specie ve ne sono 44 che possiedono macchie diverse da quelle del restante insieme.

Supponiamo inoltre che il biologo sospetti che il 25% di tali insetti erediti le macchie meno comuni. Se egli fa l'ipotesi che in questo caso valga la legge dell'eredità e rappresenta con la variabile  $X$  il numero degli insetti e i 200 esaminati che possiedono le macchie meno comuni, allora il modello che egli naturalmente sceglierebbe è la funzione di densità binomiale

$$(I) f(x) = \frac{200!}{x!(200-x)!} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{200-x}$$

Se non ci fosse stata alcuna teoria a suggerire che  $\frac{1}{4}$  di tali insetti dovrebbero possedere le macchie meno comuni, il biologo potrebbe scegliere questa stessa funzione di densità con la probabilità  $\frac{1}{4}$

sostituita dalla frequenza relativa osservata pari a  $\frac{44}{200} = 0.22$  cioè potrebbe mettere

$$\begin{array}{cc} \left(\frac{1}{4}\right)^x & \left(\frac{3}{4}\right)^{200-x} \\ \downarrow & \downarrow \\ 0.22 & 0.78 \end{array}$$

Utilizzando la (I) il biologo poi potrebbe fare previsioni riguardanti i futuri insiemi di 200 osservazioni e scoprire eventuali disaccordi con la sua teoria.

### Dati

*È conveniente nella trattazione statistica chiamare la totalità dei possibili risultati sperimentali come la popolazione di tali risultati, allora un insieme di dati ottenuti eseguendo l'esperimento un certo numero di volte è chiamato un campione della popolazione.*



Secondo questa terminologia l'inferenza statistica consiste allora nel trarre delle conclusioni su di una popolazione servendosi di un campione estratto dalla popolazione stessa.

Un problema fondamentale perciò è come estrarre informazioni dai campioni per utilizzarle nello studiare le popolazioni da cui i campioni sono estratti.

Il tipo di informazione che si dovrebbe ottenere da un insieme di dati dipende dalla natura dei dati e dal modello scelto.

In alcuni problemi si sa da considerazioni teoriche o dall'esperienza con problemi simili quale modello si dovrebbe usare, per esempio la densità binomiale rappresentata dalla (1) è un tale modello.

Tutto ciò che in realtà è necessario dai dati sperimentali per tali modelli è l'informazione atta a fornire buone stime dei parametri implicati, in altri problemi ne teoria, ne esperienza è disponibile per aiutare a scegliere un modello, allora è necessario usare i dati sperimentali per decidere su di un ragionevole tipo di modello prima di poter procedere ulteriormente.

Nel considerare la natura dei dati è importante distinguere fra quegli insiemi di dati, per cui l'ordine in cui le osservazioni sono state ottenute, fornisce una informazione utile, e quegli insiemi per cui non è così, ad esempio se si fosse interessati a studiare i fenomeni atmosferici o la borsa valori di giorno in giorno l'ordine sarebbe molto importante.

L'esperienza industriale indica che l'informazione ottenuta dal considerare l'ordine in cui sono fabbricati gli articoli è indispensabile per una efficiente produzione, se si fosse invece interessati a studiare certe caratteristiche degli studenti di un istituto e si scegliesse un insieme di studenti prendendo un nominativo ogni venti, in una guida dell'istituto difficilmente ci si aspetterebbe che l'ordine in cui vengono presi i nomi fosse di qualche valore nello studio.

Ci occuperemo ora di tecniche che non usano informazioni riguardanti l'ordine.

### **Classificazione dei dati**

Supponiamo ora che siano dati i pesi di 200 individui di un istituto e che si desideri usarli per studiare la distribuzione del peso di quegli individui, ora è molto difficile guardare 200 misure e ottenere contemporaneamente un'idea ragionevolmente accurata di come si distribuiscono quelle misure.

Per ottenere un'idea migliore della distribuzione dei pesi è perciò conveniente condensare i dati classificando le misure in gruppi, sarà allora possibile tracciare il grafico della distribuzione così modificata e imparare di più su come sono distribuiti i pesi.

Questa classificazione sarà utile anche per semplificare i calcoli di alcune medie, particolarmente quando non si dispone di mezzi di calcolo veloci, queste medie forniscono ulteriori informazioni sulla distribuzione, così lo scopo di classificare i dati è quello di aiutare ad estrarre certi tipi di utili informazioni riguardanti distribuzione sottesa.

Nel classificare i dati è conveniente di solito usare da 10 a 20 classi, ma se necessario si può anche scendere fino a 6 classi.

Nella tabella 1 è stato classificato un insieme di misure di 200 diametri di barrette di acciaio i cui valori variano fra 0.431 e 0.503 pollici.

Poiché i diametri sono stati misurati al millesimo di pollice, gli estremi degli intervalli di classe sono stati scelti mezza unità oltre l'accuratezza di questa misura per assicurare che nessuna misura cada su di un estremo, si fa l'ipotesi in questa classificazione che a tutte le misure, che cadono in un dato intervallo di classe, venga assegnato il valore corrispondente al punto di mezzo dell'intervallo, che prende il nome di *marchio* o *centro di classe*.

Dopo che ciascuna misura è stata registrata nella classe appropriata tramite una sbarretta, come mostrato in tabella 1, i risultati della classificazione sono registrati sotto forma di una tabella di frequenza come mostrato nella seconda metà della tabella 1.

