Recenti sviluppi nella teoria delle distribuzioni e loro applicazioni

Adelchi Azzalini

Università di Padova

XLIII Riunione Scientifica della Società Italiana di Statistica

Torino, 14–16 giugno 2006

Abbiamo bisogno di altre distribuzioni?

Contesto:
famiglie parametriche di distribuzioni per variabili continue

- una pletora di distribuzioni, almeno per $d = 1$
- per $d > 1$, la normale multipla predomina ancora (sì, ci sono anche le copule...)
- Esigenze:
  - flessibilità di adattamento ai dati
  - trattabilità matematica, proprietà formali
  - interpretabilità rispetto al problema sostanziale

  l'insieme intersezione è ristretto, soprattutto se $d > 1$

Un approccio al problema delle distribuzioni

Questioni/obiettivi:

- la normale è un caso-limite o un caso-centrale?
  …e operiamo di conseguenza
- possiamo estendere la classe normale e renderla ‘più flessibile’?
- possiamo poi applicare il procedimento ad altre distribuzioni?
- possiamo combinare flessibilità e trattabilità matematica?
- possiamo combinare flessibilità e interpretabilità?

Tratteggiamo un approccio che si muove lungo queste linee-guida.

Una storia con molti attori italiani

i protagonisti . . .

Antonella Capitanio
Monica Chiogna
Alessandra Dalla Valle
Brunero Liseo
Nicola Loperfido
Anna Clara Monti

e con la partecipazione di:

Attilio Meucci	Claudio Agostinelli
Nicola Sartori	Alessandra Salvan
Ruggero Bellio	Alessandra Brazzale
Luca Greco	Laura Ventura
Davide Raggi	Nunzio Cappuccio
Giovanni De Luca	Valeria Ardito
Leonardo Grilli	Carla Rampichini
Andrea Tancredi	etc. etc.

(. . . perdonate omissioni)

. . . e un precursore italiano

Fernando de Helguero
(1880-1908)

Fernando de Helguero (1908):

Il compito della statistica (...) non consiste solo nel determinare la legge di dipendenza dei diversi valori ed esprimerla con pochi numeri, ma anche nel fornire un aiuto allo studioso che vuole cercare le cause della variazione.

...le curve teoriche studiate dal PEARSON e dall'EDGEWORTH (...) mentre danno con molta approssimazione la legge di variazione, a mio avviso sono difettose in quanto (...) nulla ci fanno sapere sulla legge di dipendenza, quasi nulla sulla relazione con la curva normale.

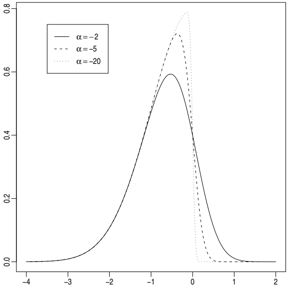
Io penso che miglior aiuto per lo studioso potrebbero essere delle equazioni che supponessero una perturbazione della variabilità normale per opera di cause esterne.

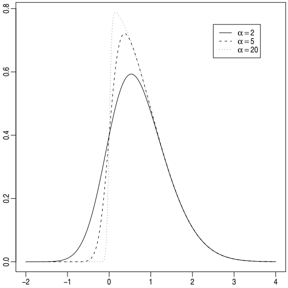
Caso-base: distribuzione normale asimmetrica (SN)

Poniamo $\phi(\cdot)$ densità N(0,1), $\Phi = \int \phi$.
Allora

$$f(x) = \phi(x)\,\{2\,\Phi(\alpha x)\}\,, \qquad (x \in \mathbb{R})$$

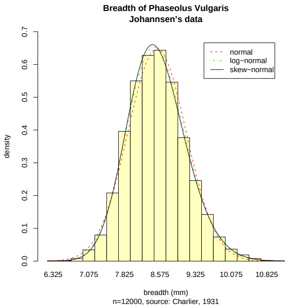
è una densità propria per ogni $\alpha \in \mathbb{R}$.





Un esempio apparentemente innocuo

Dati di Johannsen



famiglia	X^2	gdl	valore p
Normale	196.5	13	< 0.001
Edgeworth, 1° ordine	34.3	12	< 0.001
Edgeworth, 2° ordine	14.9	11	0.19
Log-normale	35.5	17	0.005
SN, 16 celle	14.2	12	0.29
SN, 20 celle	21.2	16	0.17

$$\hat{\gamma}_1 = 0.288, \quad \hat{\gamma}_1\sqrt{n/6} = 12.9$$

Un generatore di distribuzioni

Ingredienti:

$f_0(x)$	f. densità in $\mathbb{R}^d$,	$f_0(x) = f_0(-x)$
$G(y)$	f. ripartizione in $\mathbb{R}$,	$G'(y) = G'(-y)$
$w(x)$	f. dispari, $\mathbb{R}^d \to \mathbb{R}$,	$w(-x) = -w(x)$

Allora

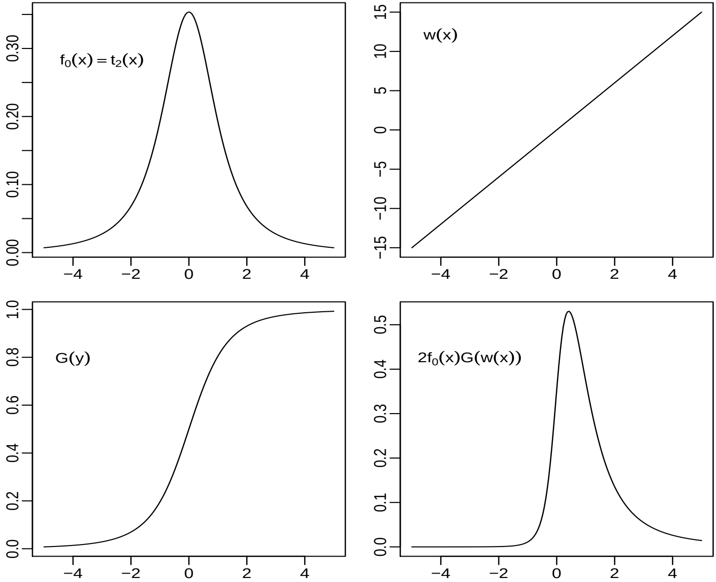
$$f(x) = 2\, f_0(x)\, G\{w(x)\}$$

è una densità propria.

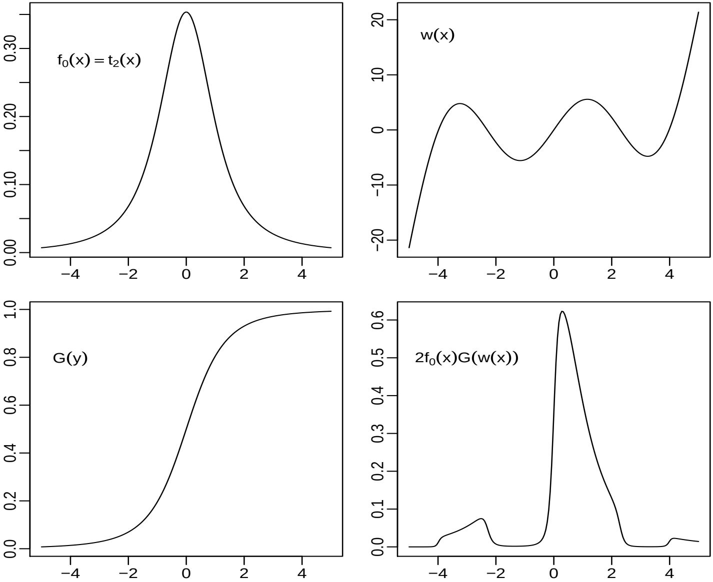
Dimostrazione. Siano $X \sim f_0$, $Y \sim G'$ indipendenti. Allora

$$\tfrac{1}{2} = \mathbb{P}\{Y - w(X) \leq 0\} = \mathbb{E}_X\{Y \leq w(X)|X\} = \mathbb{E}_X\{G\{w(X)\}\}$$

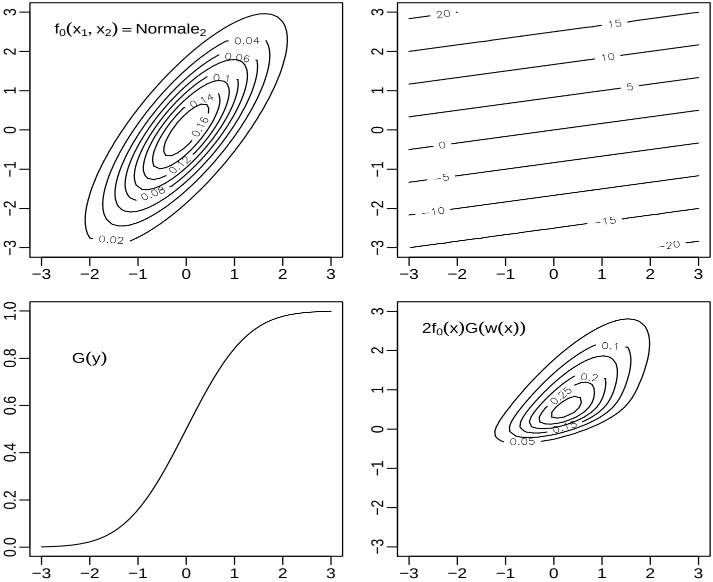
Un esempio semplice ($d = 1$)



Un esempio un po' originale ($d = 1$)



Costruzione della distribuzione SN per $d = 2$



Invarianza rispetto alla perturbazione

Proprietà: se $Y \sim f_0(x)$, $Z \sim f(x) = 2\, f_0(x)\, G\{w(x)\}$, allora

$$t(Y) \stackrel{d}{=} t(Z)$$

se $t(\cdot)$ è una funzione pari.

Esempio: $Y^\top AY \stackrel{d}{=} Z^\top AZ$

Campionamento selettivo

- modello-base di Heckman

$$Y_0 = X_0\beta_0 + U_0, \qquad Y_1 = X_1\beta_1 + U_1$$

  dove $(U_0, U_1) \sim N_2(0, \Sigma)$, ma Y_1 osservato solo se $Y_0 > 0$
- la distribuzione di $U_1|U_0 > 0$ è SN
  la distribuzione di $Y_1|Y_0 > 0$ è SN ‘estesa’
- dai risultati precedenti $\longrightarrow$
  - costruire formulazioni più avanzate,
    ad esempio con condizionamento multiplo
  - sostituire ipotesi di normalità con altre più flessibili
    in particolare distribuzioni ellittiche

Modelli a frontiera stocastica

- $$Y = x^\top \beta + W_1 - |W_0|$$
  esprime il prodotto Y realizzato con fattori produttivi x, ma 'inefficienza' $W_0 \sim N(0, \sigma_0^2)$ ed 'errore' $W_1 \sim N(0, \sigma_1^2)$
- riscriviamo Y (usando $V_0, V_1 \sim N(0,1)$ indip.) come
  $$\begin{aligned} Y &= x^\top \beta + \omega(\sqrt{1-\delta^2}\,V_1 - \delta|V_0|) \\ &= x^\top \beta + \omega Z, \qquad Z \sim SN \end{aligned}$$
- dai risultati precedenti $\rightarrow$
  alleggerire ipotesi di normalità di (W_0, W_1), ad esempio
  $$Y = x^\top \beta + \omega Z, \qquad Z \sim t\text{–asimmetrica}$$

Osservazioni ‘contralaterali’ e statistiche ordinate

- In taluni casi, da un soggetto osserviamo coppia ‘identica’ X_1, X_2, ma consideriamo solo

$$Z_1 = \max(X_1, X_2), \quad \text{oppure} \quad Z_2 = \min(X_1, X_2)$$

- Esempi: misurazione acuità visiva massima, misurazioni di altri organi bilaterali
- Se $(X_1, X_2) \sim N_2$ con componenti $N(0,1)$ e correlazione ρ, allora

$$Z_1 \sim SN(\alpha), \quad Z_2 \sim SN(-\alpha) \qquad \alpha = \sqrt{\frac{1-\rho}{1+\rho}}$$

- Estensione al caso di

$$Z = a_1 X_{(1)} + a_2 X_{(2)} \sim (\text{scala}) \times SN(\cdot)$$

Applicazioni nei mercati finanziari

- rilettura di formulazioni economiche (ad es. CAPM) in termini più generali
- impiego in modelli ARCH, GARCH, SV etc termine di errore di maggior flessibilità,
- asimmetria sembra nascere da una diversa reazione dei mercati ai segnali positivi e ai negativi

Piani adattivi in studi clinici

- studio di fase II a dosaggi crescenti (placebo e 2–5 gruppi) $\longrightarrow$ selezione di dose 'ottima'
- studio (esteso) di fase III per la dose prescelta
- obiettivo: combinare i dati dalle due fasi (per limitare i costi devastanti degli studi clinici)
- bisogna tener conto del meccanismo di condizionamento in atto nella fase II
- $\longrightarrow$ distribuzione SUN, estensione di SN al caso di condizionamenti multipli

Rapporti di composizione

- $X_j \in (0,1)$, per $j = 1, \ldots, D$, con vincolo $\sum_{j=1}^{D} X_j = 1$
- rilevanza in vari campi, tra cui chimica, geologia e simili
- trasformazione ALR di Aitchison

$$Z_j = \log(Z_j / Z_D), \qquad (j = 1, \ldots, D-1)$$

  per cui $Z \in \mathbb{R}^{D-1}$
- tipica ipotesi di normalità di Z può essere inadeguata
- ma distribuzioni alternative devono essere chiuse a trasformazioni lineari, per garantire coerenze formali
- famiglia SN è chiusa rispetto a trasformazioni affini $\rightarrow$ OK

Ricapitoliamo

- ci sono motivi per studiare (ancora) famiglie di distribuzioni
- la distribuzione SN rappresenta una semplice forma di estensione della distribuzione normale
- possiamo costruire in modo semplice famiglie di distribuzioni molto articolate, e dotate di interessanti proprietà
- questo approccio presenta forti e 'naturali' legami con varie tematiche di ambito applicativo
- i risultati matematici a disposizione possono offrire strumenti per potenziare il lavoro in ambito applicativo