1.1.4 Funzione quantile empirica

Se dal quantile singolo si passa a considerare l'intero sistema dei quantili definibile per un dato campione si costituisce la cosiddetta funzione quantile empirica. Per variabili discrete si adotta, in genere, la formulazione

$$Q_n(p) = F_n^{-1}(p) = \inf\{x \,|F_n(x) \geq p\} \quad \text{per } 0 \leq p \leq 1 \tag{1.16}$$

La funzione quantile empirica per variabili discrete è una funzione a gradini con valori corrispondenti alle statistiche d'ordine. Più precisamente

$$Q_n(p) = x_{i:n} \qquad \text{per} \quad \frac{i-1}{n} < p \leq \frac{i}{n}, \quad i = 1, 2, \cdots, n \tag{1.17}$$

In particolare, per le posizioni estreme si conviene:

$$x_0 = \lim_{p \to 0} Q_n(p)\,; \qquad x_1 = \lim_{p \to 1} Q_n(p) \tag{1.18}$$

Appare immediatamente evidente come la funzione quantile empirica altro non sia che l'inversa della funzione di ripartizione empirica ovvero quiesta, ma rappresentata in un sistema con gli assi ruotati.

La definizione [1.16]-[1.18] è appropriata per variabili discrete o comunque per un campione finito di osservazioni. In questo caso, sia la funzione quantile empirica che la funzione di ripartizione empirica sono funzioni a gradini come è illustrato nella figura 1.4.

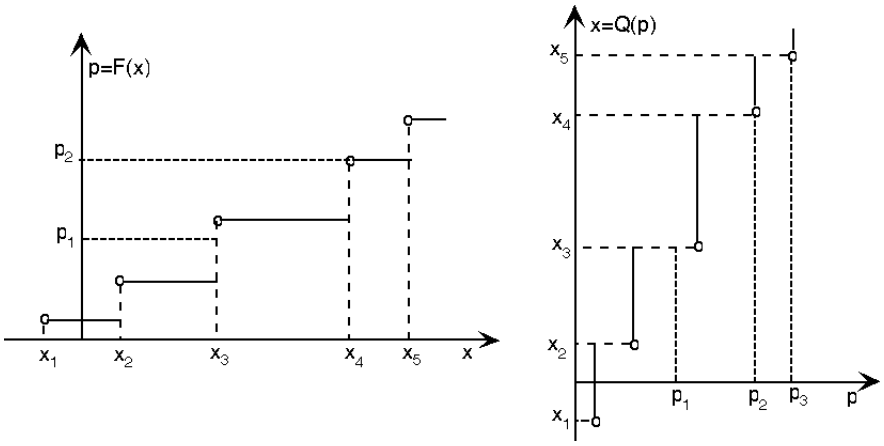

Figura 1.4: Dalla funzione di ripartizione alla funzione quantile

Se $p = p_a$ non esiste alcuna modalità x tale che $F_n(x) = p_a$; in questo caso la definizione [1.16] sceglie come Q_p la modalità più piccola tra tutte quelle che verificano la disuguaglianza $F_n(x) \geq p_a$ e che nel grafico implica X_3. Per $p = p_b$ la funzione di ripartizione è costante ed esiste addirittura un intervallo di valori: $[X_4, X_5)$ per i quali $F_n(x) = p_b$; la [1.16] associa a p_b la modalità più a destra di tale intervallo: $Q_n(p_b) = F_n^{-1}(p_b) = X_5$ infatti, nel punto X_5 si ha

$$Q_n = F_n^{-1}\left[F_n(X_5)\right] > X \text{ per } X \in [X_4, X_5)$$

tranne che nel punto $X = X_{p_b} = X_5$ dove

$$Q_n = F_n^{-1}\left[F_n(X_{p_5})\right] = X_{p_5}$$

Esempio 4. *Durbin e Koopman [?] riportano (Data set [??]) una serie storica di n=100 utenti connessi ad internet attraverso un provider. I primi 30 valori danno luogo al grafico [1.5]. Nel grafico sono presenti dei valori ripetuti.*

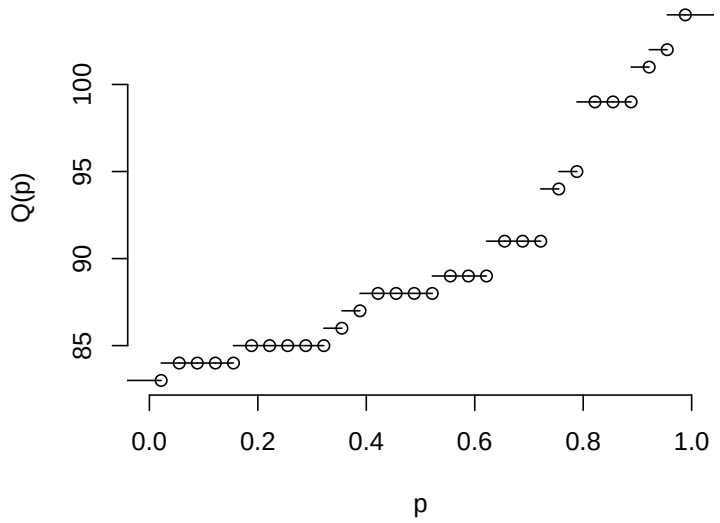

Figura 1.5: Funzione quantile empirica per una variabile discreta

La definizione di funzione quantile empirica cambia se è riferita ad osservazioni su variabili continue.

$$Q_n(p) = n\left(\frac{i}{n} - p\right) x_{i-1:n} + n\left(p - \frac{i-1}{n}\right) x_{i:n} \quad \text{per } \frac{i-1}{n} \leq p \leq \frac{i}{n}; \quad i = 1, 2, \cdots, n \tag{1.19}$$

In questa forma, la funzione $Q_n(p)$ è anche derivabile e permette quindi di ridefinire più rigorosamente la densità quantile empirica

$$\delta_n(p) = Q'_n(p) = n(x_{i:n} - x_{i-1:n}) \quad \text{per} \quad \frac{i-1}{n} < p < \frac{i}{n} \tag{1.20}$$

Le quantità $n(x_{i:n} - x_{i-1:n})$ sono note come gli *spacings* del campione molto utilizzati in statistica matematica perché hanno distribuzione asintotica esponenziale (a prescindere dalla distribuzione da cui provengono) con valore atteso $q(p)$. Ne vedremo una applicazione molto importante nello studio della variabilità campionaria dei quantili. La definizione 1.19 può anche essere estesa alle variabili discrete, ottenendo però dei valori intermedi se $[np]$ non è intero.

Esempio 5. *Ghosh [?] studia il dataset relativo al livello delle piene del fiume Irequois riportato come dataset* Ghoshr *in* `http://www.ecostat.unical.it/tarsitano/datasets`. *Nella figura che segue è stata disegnata la funzione quantile empirica del dataset.*

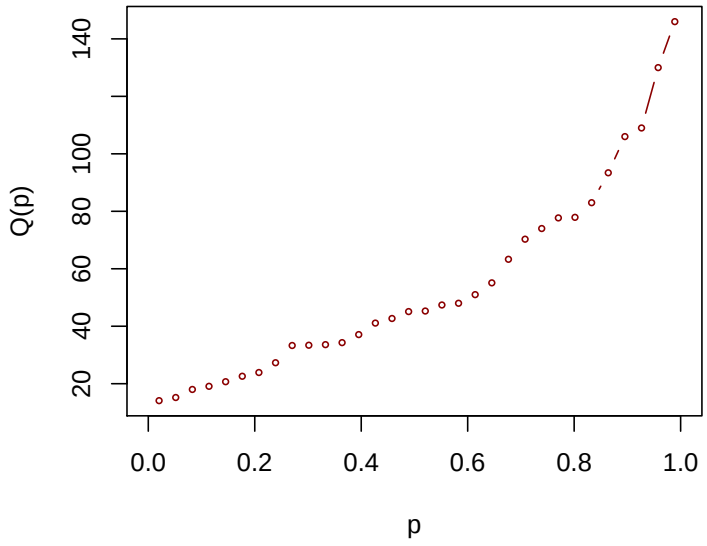

Figura 1.6: Funzione quantile per un campione da variabile continua

La funzione quantile empirica è monotona crescente e mostra diversi punti di svolta cioè valori in cui il ritmo di accrescimento sembra attenuarsi.

1.1.5 Indici descrittivi basati sui quantili

Un concetto si considera fissato quando ne è fornita una definizione operativa cioè si sia stabilito l'insieme di regole mediante le quali effettuarne la classificazione o determinarne la misura o, in generale, per aggangiarlo alla realtà osservabile che fornirà i valori o le categorie della variabile. Dice P.W. Bridgman (Nobel per la fisica nel 1927): $\cdots$ *Se una questione specifica ha senso, deve essere possibile trovare operazioni attraverso le quali ad essa si possa dare una risposta. Se così non fosse allora la questione non ha senso.* Sir Francis Galton, cugino di Darwin, riflette: *Finché i fenomeni di una qualsiasi disciplina del sapere non siano stati sottoposti a misurazioni in base ad una scala, essa non potrà mai assurgere alla dignità di scienza.*

La definizione operativa di un concetto consiste nell'escogitare misurazioni facilmente eseguibili (senza errori ed ambiguità, per quanto possibile) che permettano di collocare l'oggetto osservato in una posizione specifica rispetto ad un ordinamento ed in modo tale che un cambiamento nella misurazione sia legato, in un modo semplice ed univoco, ad un cambiamento in ciò che rappresentano. Un'accusa talvolta rivolta alla statistica (ma generalizzabile a molte altre discipline) è di focalizzarsi sugli elementi maggiormente suscettibili di misurazioni valide, ma non necessariamente le più efficaci per la comprensione del fenomeno. Rispetto a questa critica possiamo tentare di prendere precauzioni ammettendo definizioni più vaghe, ma con la conseguenza di risultati altrettanto vaghi. Un fatto è come un sacco: vuoto non si regge. Perché si regga, bisogna farci entrare dentro la ragione ed i sentimenti che lo hanno determinato. Così Luigi Pirandello fa dire al Padre nei *Sei personaggi in cerca d'autore*. Tuttavia, una volta riconosciuto che un fatto è come un sacco vuoto, si può facilmente comprendere come, a seconda di quello che si mette dentro, si può far assumere al sacco la forma che si vuole".

La statistica, nel corso del tempo, ha focalizzato l'attenzione su alcuni aspetti differenzianti delle distribuzioni quali: la centralità cioè l'esistenza di una modalità -fittizia o reale- che prevalga sulle altre e sia di queste rappresentativa; la variabilità e cioè l'attitudine delle modalità a disperdersi o a concentrarsi su particolari livelli; la simmetria cioè la tendenza all'equilibrio ovvero al prevalere dei valori piccoli o dei valori grandi rispetto agli altri.

La media o indice di centralità è un valore, ipotetico o osservato, rappresentativo della distribuzione. L'unico vincolo nella sua scelta è la condizione di internalità se la scala di misurazione è almeno ordinale: $X_{1:n} \leq \text{Media} \leq X_{n:n}$.

Una classe di indici di centralità molto interessante è la seguente:

$$T_1(\alpha, p) = \alpha \hat{Q}_p + (1-\alpha)\hat{Q}_{1-p} \quad \text{per } 0 < p \leq 0.5, \quad 0 \leq \alpha \leq 1 \tag{1.21}$$

E' evidente che l'indice $T_1(\alpha, p)$ verifica la condizione di internalità di Cauchy. L'espressione [1.21] ha tanti casi particolari, e alcuni di essi sono stati effettivamente utilizzati. Ad esempio, per $\alpha = 0.5$ si definisce il quasiquantile proposto da Reiss [**?**]; la semisomma interquartile si ottiene per $\alpha = 0.5$ e $p = 0.25$; per $p \to 0$ e $\alpha = 0.5$ si perviene al valore centrale della distribuzione; per $p = 0.5$ si ottiene la mediana.

Esempio 6. *Dal dataset* Heart Disease *disponibile su* `http://b-course.hiit.fi/cl_readymade.html` *preleviamo i valori relativi al livello della pressione sanguigna a riposo e calcoliamo la [1.21] per* $\alpha = 0.20, p = 0.10$ *e per* $\alpha = 0.80 ep = 0.10$. *I calcoli danno luogo a*

$$T_1(0.2, 0.1) = 143.6, \; T_1(0.8, 0.1) = 118.4$$

Il primo valore è superiore al secondo in quanto il primo decile vi contribuisce in modo meno incisivo di quanto non faccia nel secondo.

La disponibilità dei quantili campionari è utile quando non è possibile ipotizzare una particolare distribuzione o anche solo assumere che sia simmetrica. In varie occasioni quali la determinazione di soglie di garanzia, intervalli di riferimento e livelli di tolleranza non si è interessati ai parametri della distribuzione; sono piuttosto i quantili che danno le giuste risposte.

Non solo, ma la formula [1.21] consente una distinzione tra la valutazione puntuale della centralità e la sua valutazione globale. Per ogni valore di $p \in (0, 1)$ otteniamo una indicazione sulla centralità della distribuzione che, riunite, formano quella che potremmo chiamare la funzione di centralità. Grazie a questa, potremmo confrontare distribuzioni diverse tra di loro definendo a piacimento il livello di centralità da considerare. Potremo inoltre, comparare una funzione di centralità campionaria con la funzione di centralità di uno o più modelli per asseverarne le somiglianze con ciò si è osservato e trarre spunti per eventuali generalizzazioni.

Esempio 7. *Confrontiamo la centralità della pressione sanguigna di donne ed uomini nel dataset* Heart Disease *per* $\alpha = 0.1, 0.2, 0.3, 0.4$. *Quella delle donne è tendenzialmente maggiore di quella degli uomini. Lo scarto è più marcato negli estremi (p vicino a zero e vicino a 0.5). La scelta di* α *non pare essere essenziale.*

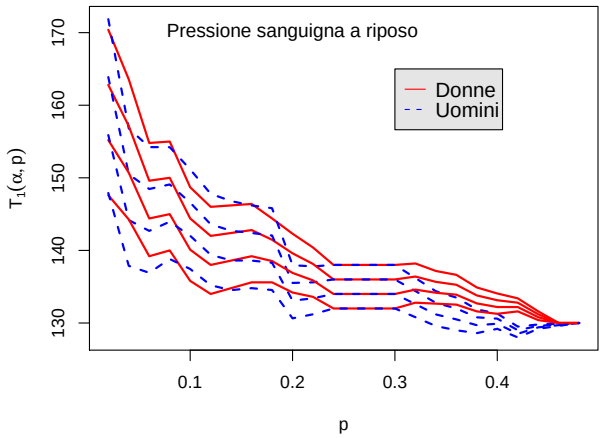

Figura 1.7: Funzione di centralità

Esempio 8. *La centralità del modello di Cauchy è* $T_1(\alpha, p) = c + \alpha \tan[\pi(p - 0.5)] + (1 - \alpha) \tan[\pi(0.5 - p)]$. *Per* $\alpha = 0.25$ *si ha*

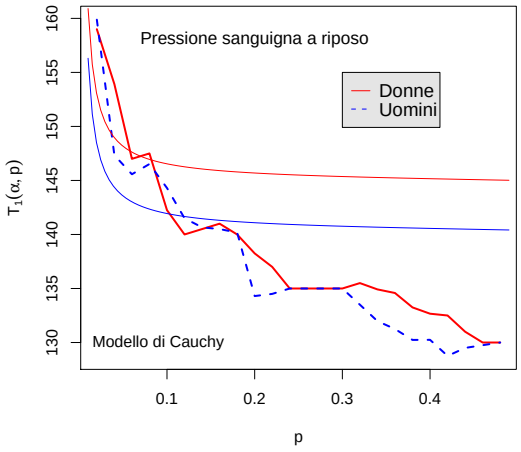

Figura 1.8: Centralità osservata e teorica

La costante c indica il centro della distribuzione e nell'esempio è pari al 9° decile campionario. L'adattamento tra teoria e realtà lascia molto a desiderare. In parte questo è spiegabile con la simmetria del modello che non trova riscontro nei dati.

Se si integra la la [1.21] rispetto a p si ottiene una misura globale di centralità

$$\begin{aligned}\int_0^{0.5} T_1(\alpha, p)\, dp &= \alpha \int_0^{0.5} \hat{Q}_p dp + (1-\alpha) \int_0^{0.5} \hat{Q}_{1-p} dp \qquad (1.22)\\ &= \frac{\alpha}{k} \sum_{i=1}^{k} x_{i:n} + \frac{(1-\alpha)}{n-k} \sum_{i=k+1}^{n} x_{i:n}\end{aligned}$$

dove $k = \lfloor n0.5 \rfloor$. Se $\alpha = 0.5$ ed n è pari la [1.22] diventa la media aritmetica.

I quantili campionari sono spesso coinvolti in misure di variabilità. Ecco una classe di misure di variabilità relativa particolarmente flessibile ed efficace.

$$T_2(\alpha, p) = \frac{\hat{Q}_{1-p} - \hat{Q}_p}{|\hat{Q}_{\alpha+2p}|} \qquad \text{per} \quad 0 < p \leq 0.5,\ |\hat{Q}_{\alpha+2p}| > 0 \qquad (1.23)$$

con $0 \leq \alpha < (1-2p)$. Per ogni p la [1.23] fornisce un indice di variabilità relativa compreso tra zero (assenza di variabilità, $p = 0.5$) ed il range relativo $(x_{n:n} - x_{1:n})/\hat{Q}_\alpha$ raggiunto per $p \to 0$. $p = 0.25$ e $\alpha = 0$ definisce il rapporto tra differenza interquantile e mediana; per $p = 0.1$ e $\alpha = 0.30$ si determina la differenza interdecile rapportata ancora alla mediana.

Esempio 9. *Karagas et al. [?] analizzano il livello di arsenico nelle unghie in* $n = 21$ *soggetti: 0.119, 0.118, 0.099, 0.118, 0.277, 0.358, 0.08, 0.158, 0.31, 0.105, 0.073, 0.832, 0.517, 2.252, 0.851, 0.269, 0.433, 0.141, 0.275, 0.135, 0.175*

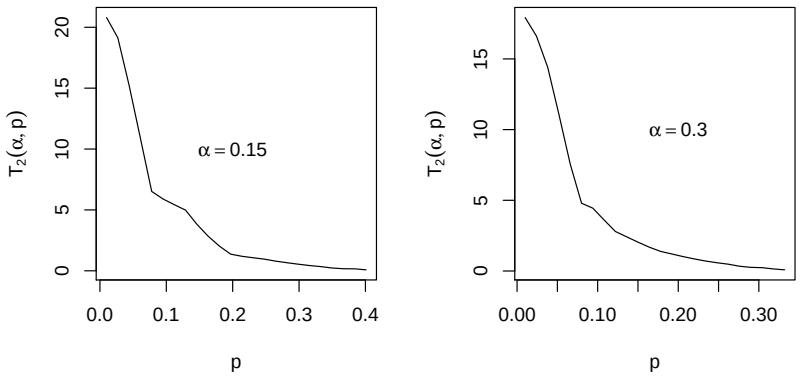

Figura 1.9: Misure della variabilità relativa

La curva di variabilità relativa è discendente dato che per p bassi si confrontano quantili distanti; all'aumentare di p i quantili sono sempre più ravvicinati che tendono a zero. Se il livello d'arsenico fosse costante la curva coinciderebbe con l'asse delle ascisse. Con modalità tutte uguali tranne una che è maggiore delle altre, la curva tenderà a sovrapporsi ai cateti che fanno angolo nell'origine; se, invece, la modalità diversa è inferiore alle altre, la curva si avvicinerà sempre più ai due cateti che si incrontrano nell'angolo opposto all'origine.

Indici di centralità e variabilità non bastano. Se si vuole identificare il modello che è dietro la rilevazione empirica,è necessario estendere il confronto ad altre caratteristiche. In particolare, la simmetria del dataset. Si dirà simmetrica la distribuzione per la quale $\hat{Q}_{1-p} - M_e = M_e - \hat{Q}_p$. La mediana M_e è il polo di simmetria. Esistono varie misure di asimmetria. Ad esempio

$$\frac{\left(\hat{Q}_{1-p} - M_e\right)^2 + \left(M_e - \hat{Q}_p\right)^2}{2M_e^2} - 1 \tag{1.24}$$

Tale misura, proposta da Tukey, è nulla per distribuzioni simmetriche (ma non solo) ed è indipendente dalla unità di misura. Un'altra utile formula è

$$\frac{\hat{Q}_{1-p} + \hat{Q}_p - 2M_e}{\hat{Q}_{1-p} - \hat{Q}_p} \qquad 0 < p < 0.5 \tag{1.25}$$

che ha come caso particolare l'indice di Yule-Bowley per $p = 0.25$. Le misure [1.25] variano nell'intervallo $[-1, 1]$ ed hanno le stesse proprietà delle [1.24].

Esempio 10. *Glucosio nel sangue di 46 soggetti anziani ([**?**]): 3.52, 3.905, 4.07, 4.07, 4.29, 4.345, 4.4, 4.455, 4.565,4.62, 4.62, 4.675, 4.840, 4.840, 4.895, 4.895, 4.95, 4.95, 5.115, 5.115, 5.225, 5.225, 5.225, 5.335, 5.335, 5.39, 5.39, 5.39, 5.455, 5.555, 5.61, 5.665, 5.72, 5.775, 5.83, 5.83, 5.885, 5.885, 6.215, 7.095, 7.205, 8.14, 9.9, 10.89, 11.605,12.045*

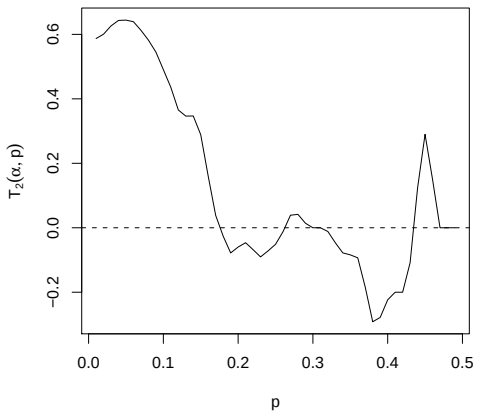

Figura 1.10: Funzione di asimmetria

Se la distribuzione fosse simmetrica il grafico in 1.10 coinciderebbe con l'asse delle ascisse; si nota, invece, una progressiva diminuzione almeno per quantili fino al 40% e subito dopo l'indice comincia a risalire. Questo comportamento evidenzia una forte asimmetria che, in fondo, non deve sorprendere data la possibile presenza di diabetici a vari stadi della malattia nel gruppo.

Un altro aspetto che a volte compare nello studio della morfologia delle distribuzioni è la curtosi cioè la deformazione cui deve essere sottoposta una densità simmetrica e unimodale per sovrapporsi ad una curva gaussiana di pari media e pari scarto quadratico. La ragione di questo concetto e della sua misura è la grande importanza (ora molto discussa) attribuita alla gaussiana.

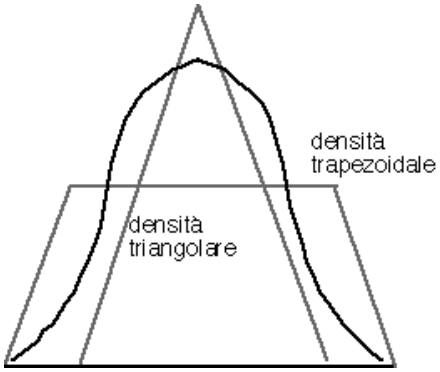

Figura 1.11: Difficoltà nel concetto di curtosi

Il problema con la curtosi è che con essa si tenta di descrivere un effetto composto da due elementi distinti: il grado di appuntimento al centro e lo spessore delle code. Si veda la figura 1.11. Per raggiungere lo stesso grado di appuntimento della densità triangolare, la normale dovrà essere tirata verso l'alto e assottigliata nei fianchi con solo leggere modifiche nelle code. Per portare le code della normale allo spessore della densità trapezoidale si possono allargare i fianchi e rigonfiare le code senza praticamente intervenire sulla sommità. In breve, code e vetta dovrebbero essere trattate separatamente. In questo senso si è tentato di misurarle con indici diversi, ma usati insieme per valutare la forma della distribuzione. Tra gli indici noti in letteratura riteniamo opportuno richiamare la classe di Groeneveld in [**?**] e quella di Wand e Serfling [**?**]

$$\gamma_2(p) = \frac{\hat{Q}[1-0.5p] + \hat{Q}[(1+p)\,0.5] - 2\hat{Q}[0.75]}{\hat{Q}[1-0.5p] - \hat{Q}[(1+p)\,0.5]} \quad \text{per } 0 < p < 0.5 \tag{1.26}$$

(1.27)

$$k(p) = \frac{\hat{Q}(0.75-0.25p) + \hat{Q}(0.75+0.25p) - 2\hat{Q}(0.75)}{\hat{Q}[0.75+0.25p) - \hat{Q}(0.75+0.25p)} \quad \text{per } 0 < p < 1 \tag{1.28}$$

Queste misure vedono la curtosi come il libero spostamento dell'area sottesa alla curva di densità dai fianchi della distribuzione al centro e nelle code.

Esempio 11. *Un dataset molto famoso è quello relativo alle eruzioni del geyser "Old Faithful" nello Yellowstone National Park. Qui si è usata la durata delle eruzioni nella versione di [?] che contempla $n = 299$ osservazioni.*

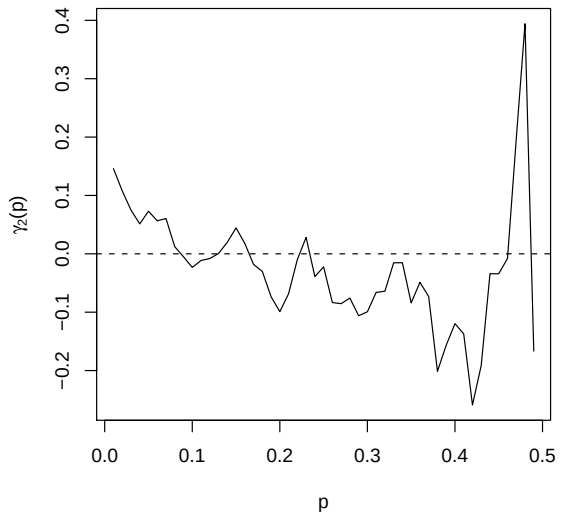

Figura 1.12: Funzioni di curtosi

Il grafico si mantiene prevalentemente al di sotto dello zero segnalando una curtosi accentuata

Gli indici discussi in questo paragrafo, tutti rigorosamente basati sui quantili e quindi robusti rispetto ai valori anomali, possono dare un importante contributo per la descrizione di un dataset e per la comprensione delle relazioni tra più variabili. Essi, infatti, sono suscettibili di un uso puntuale e di uno globale. Quello puntuale si ottiene fissando il valore di p e calcolando il valore della formula. La misura globale si definisce considerando l'intera curva valutando la formula per un insieme fitto e numeroso di valori di p o addirittura integrando rispetto a p il relativo indice. Il vantaggio della misura globale rispetto alle misure puntuali è che per queste sussiste il problema della scarsa sensibilità, vale a dire che può cambiare tutta la rilevazione, ma l'indice puntuale non se ne accorge purché rimangano invariati i tre quantili che compaiono nell'indice. Questo non può succedere per la curva globale che si accorge di ogni cambiamento, anche in un singolo dato.

1.2 Descrizione di una variabile casuale

Le tecniche considerate nel precedente paragrafo hanno illustrato diversi modi di rappresentare graficamente ed analiticamente un certo dataset. Abbiamo inoltre discusso alcune tecniche di sintesi dei vari aspetti che possono interessare nell'analisi descrittiva di una sola variabile. Per andare oltre i fatti osservati c'è bisogno di uno scenario in cui collocare la rilevazione prendendo coscienza dell'aspetto campionario (quindi stocastico) dei dati.

La casualità di una rilevazione empirica è quasi sempre espressa con il modello del campione casuale semplice ovvero estrazioni indipendenti. Poiché tali estrazioni sono governate dalla sorte, il particolare valore che si osserverà in una estrazione non può essere stabilito a priori. Tuttavia, è possibile prevederlo ragionando con una schema di probabilità. Tale schema è la variabile casuale.

Una variabile casuale X è una regola che permette di associare a ogni evento soggetto a sorteggio un valore numerico. Esistono variabili casuali discrete e continue. Una variabile casuale X è *discreta* se assume un numero finito o un' infinità numerabile di valori $x_1, x_2, \cdots, x_n$, con probabilità rispettivamente $f(x_1), f(x_2), \cdots, f(x_n)$, dove $f(x)$ è detta funzione di probabilità e soddisfa proprietà analoghe alle frequenze relative:

$$f(x_i) \geq 0 \quad i = 1, 2 \cdots, n \qquad \sum_{i=1}^{n} f(x_i) = 1 \tag{1.29}$$

La variabile casuale è *continua* se i valori che essa può assumere sono contenuti in un intervallo reale di numeri reali. In questo caso le probabilità verificano, salvo diversa indicazione, i vincoli

$$f(x) \geq 0 \qquad \int_{-\infty}^{+\infty} f(x)dx = 1 \tag{1.30}$$

Se si ritiene che l'acquisizione dei valori campionari avvenga sostanzialmente nelle medesime condizioni si parlerà di variabili casuali indipendenti e identicamente distribuite (i.i.d.). Se, invece, la struttura delle acquisizioni cambia in maniera percettibile discuteremo di variabili casuali indipendenti non identicamente distribuite (i.n.i.d.). Ritorneremo su questa distinzione nei capitolo dedicati alle applicazioni; al momento, l'importante è dotarsi di tecniche per studiare le variabili casuali, cioè, come analizzare i possibili valori che possono assumere con le associate probabilità.

E. Parzen [**?**] ha introdotto fin dai primi anni '70 diverse chiavi di lettura dei dati che, sebbene avviate da Galton nel 1889 e abbastanza diffuse da molti decenni soprattutto per merito degli statistici italiani, non erano ancora state inquadrate in un progetto unico e coerente: i metodi quantili. Gilchrist [**?**] ne ha seguito le orme operando diversi miglioramenti. Questo paragrafo sviluppa l'idea base di Parzen che la statistica esplorativa e quella inferenziale possono fondarsi sulla trattazione della funzione quantile.

Esistono vari modi per descrivere una variabile casuale, alcuni già discussi nel paragrafo 1.1. In questo paragrafo ne verranno considerati quattro.

1. **La funzione di ripartizione**. Indicata con $F(x)$, è definita come la probabilità che una variabile X sia minore o uguale di un dato valore x:

$$F(x) = Pr[X \leq x] \tag{1.31}$$

2. **La funzione di densità**. Indicata con $f(x)$, è definita da:

$$f(x)\,dx = Pr[x \leq X \leq x + dx] \tag{1.32}$$

dove dx è un intervallo infinitesimo di x. La funzione di densità è la derivata della funzione di ripartizione. Infatti:

$$f(x)\,dx = F(x + dx) - F(x) = dF(x) \longrightarrow f(x) = \frac{dF(x)}{dx} \tag{1.33}$$

3. **La funzione quantile**. Indicata con $Q(p)$. Sia Q_p il valore di x tale che $Pr[X \leq Q_p] = p$; Q_p è detto quantile p-esimo della popolazione. La funzione $Q_p = Q(p)$, che esprime il quantile come una funzione di p, è definita funzione quantile e restituisce valori, non probabilità. La definizione di funzione quantile e di funzione di ripartizione possono essere scritte per coppie di valori(x, p) come $x = Q(p)$ e $p = F(x)$. Queste funzioni sono così l'una l'inversa dell'altra. Possiamo cioè porre:

$$Q(p) = F^{-1}(p); \qquad F(x) = Q^{-1}(x) \tag{1.34}$$

quando, ovviamente, le inverse esistono.

4. **La funzione di sparsità**. Così come derivando a funzione di ripartizione otteniamo la funzione di densità, con la derivata della funzione quantile otteniamo il quarto modo per descrivere una variabile casuale: la funzione di densità quantile o funzione di sparsità. Questa, indicata con $\delta(p)$, è data, infatti, da

$$\delta(p) = \frac{dQ(p)}{dp} \tag{1.35}$$

La funzione di sparsità è il reciproco della funzione di densità per cui è definita nei punti in cui la densità è diversa da zero. Infatti, dalla relazione $F[Q(p)] = p$ si ha

$$\frac{d}{dp}\{F[Q(p)]\} = f[Q(p)]\,\delta(p) = 1 \tag{1.36}$$

da cui consegue

$$\delta(p) = \frac{1}{f[Q(p)]} \tag{1.37}$$

La funzione di sparsità compare come algoritmo di descrizione di una variabile casuale già in Wright [**?**] che la utilizza per generare una famiglia di distribuzioni molto flessibile:

$$\delta(p) = \sqrt{2\pi}\ exp\left\{\frac{1}{2}[Q(p)]^2\right\}$$

che è poi applicata per interpolare le distribuzioni empiriche sia su dati singoli che in classi.

Esempio 12. *Ecco alcuni esempi delle funzioni descritte in precedenza. Da notare che la funzione di densità e la funzione di ripartizione sono definite in un sistema di assi che è invertito rispetto al sistema adoperato per la funzione quantile e la funzione di sparsità (si veda la figura 1.13)*

$$\begin{aligned} F(x) &= \left(\frac{1}{1+x^{-\alpha}}\right)^{\beta} \qquad x > 0 \\ f(x) &= \frac{e^{-|x|}}{2\sqrt{\pi}\,|x|} \qquad -\infty < x < +\infty \\ Q(p) &= p^2(1+p)^{-0.1} \\ \delta(p) &= p^{\lambda} + (1-p)^{\lambda} \end{aligned}$$

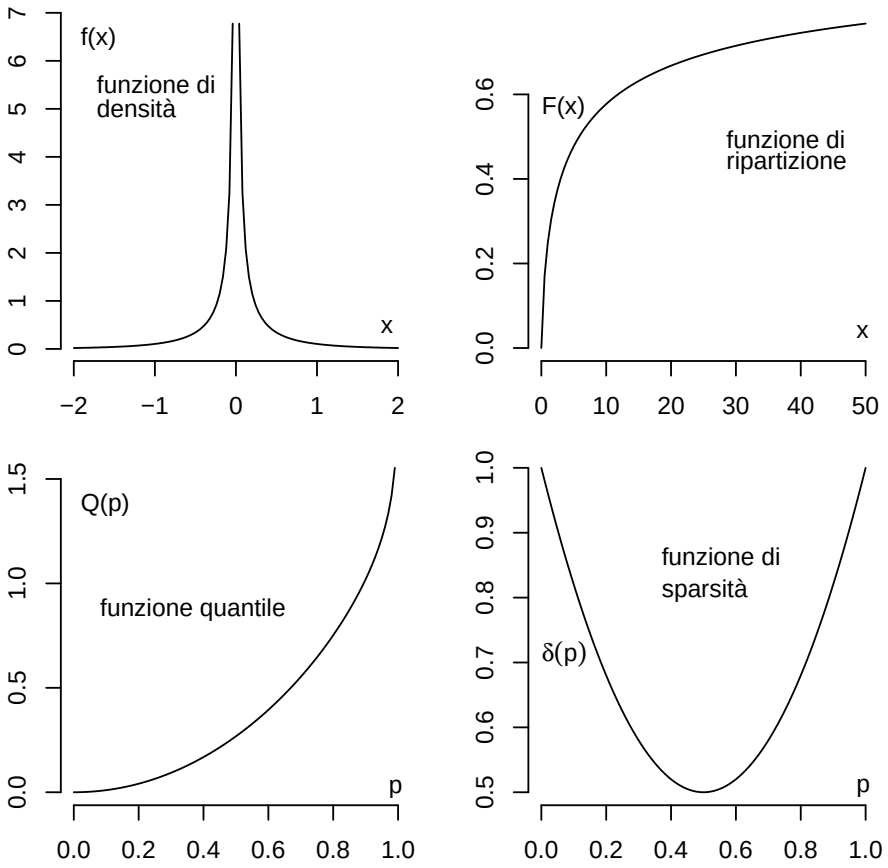

Figura 1.13: Descrizioni di una variabile casuale

Nelle applicazioni di statistica non parametrica interviene anche la derivata della funzione di sparsità nota come funzione punteggio (*quantile score function*) della densità quantile.

$$\delta'(p) = J(p) \tag{1.38}$$

In particolare $J(p)$ compare nella definizione degli indici di rilevanza delle code.

Esempio 13. *La funzione di distribuzione potenza ha una funzione quantile elementare:* $Q(p) = p^{\gamma}$. *Altrettanto semplice è la sua funzione di sparsità* $\delta(p) = \gamma p^{\gamma-1}$. *Si può a questo punto agevolmente determinare la funzione punteggio quantile*

$$\frac{d\ \delta(p)}{dp} = \gamma(\gamma - 1) p^{\gamma-2}$$

1.2.1 La trasformazione dell'integrale di probabilità

Kolmogorov negli anni '30 del secolo scorso si accorse che se X è una variabile casuale con funzione di ripartizione continua $F(.)$ la trasformazione $p = F(x)$ produce una variabile casuale $p \sim U[0,1]$ con distribuzione uniforme uniforme nell'intervallo unitario.Tale relazione è detta trasformazione dell'integrale di probabilità perché se X ha una funzione di densità $f(.)$ la $F(.)$ ne è l'integrale indefinito.

Riprendiamo la funzione quantile: $Q(p) = F^{-1}(p)$. Se la $F(.)$ è monotona, ad ogni $x \in R^+$ corrisponde una sola $p \in [0,1]$ tale che $F(x) = p$ e all'intervallo semiaperto $-\infty < X \leq x$ corrisponde l'intervallo $0 \leq U \leq p = F(x)$. Il primo tipo è probabilizzabile dato che X è una variabile casuale; dobbiamo accertare che lo sia anche l'altro. Notiamo l'uguaglianza tra: $E_1 = \{Q(p) \leq x\}$ e $E_2 = \{p \leq F(x)\}$; infatti, $Q(p) \leq x$ implica che per ogni $\epsilon > 0, p < F(x+\epsilon)$ perché $F(.)$ è continua a destra e, data l'arbitrarietà di ϵ, si ha anche $p \leq F(x)$; ne consegue che $Q(p) \leq x$ e $E_1 = E_2$ ed inoltre:

$$Pr(E_1) = Pr(E_2) \Rightarrow Pr[Q(p) \leq x] = Pr[p \leq F(x)] \tag{1.39}$$

In definitiva, $U = F(x)$ è una variabile casuale; vediamo come determinarne la funzione di densità.

$$\begin{aligned} Pr(U \leq p) = Pr[F(x) \leq p] = Pr\{Q[F(x)] \leq Q(p)\} \\ = Pr[x \leq Q(p)] = F[Q(p)] = p \end{aligned} \tag{1.40}$$

in cui la relazione $F[Q(x)] = x$ scaturisce dalla univocità, nelle condizioni date,della funzione quantile $Q()$. Se con $F_U(p)$ indichiamo la funzione di ripartizione di U allora, in caso di variabile casuale assolutamente continua, si otterrà:

$$F_U(p) = \begin{cases} 0 & se \quad p \leq 0 \\ p & se \quad 0 < p < 1 \\ 1 & se \quad p \geq 1 \end{cases} \tag{1.41}$$

ne consegue che

$$f(p) = \begin{cases} 1 & se \quad 0 \leq p \leq 1 \\ 0 & altrove \end{cases}$$

cioè il modello di densità uniforme sull'intervallo unitario.

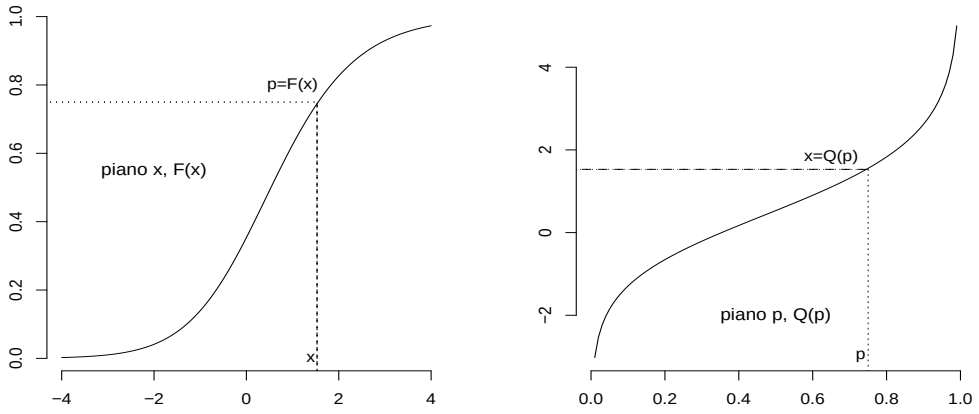

Figura 1.14: Piani di trasformazione di una variabile casuale

Esempio 14. *L'aspetto più interessante di questa trasformazione è però la sua inversa.*

Partiamo dalla variabile casuale uniforme $p \sim U[0,1]$ *e definiamo una nuova variabile casuale con la trasformazione* $X = Q(p)$ *dove* $Q(.)$ *è la funzione quantile della* X *e quindi:*

$$Pr(X \leq x) = Pr[Q(p) \leq x] = Pr[p \leq F(x)] = F(x) \tag{1.42}$$

Per ottenere una variabile casuale che abbia la funzione di ripartizione $F(.)$ *basta disporre della sua inversa in cui inserire come argomento la variabile casuale uniforme.*

Esempio 15. *Se* X *è uniforme sull'intervallo* $[0,1]$*, la trasformazione per arrivare alla ripartizione di Gumbel è:*

$$F(x) = 1 - exp\left\{-e^{-\left(\frac{x-a}{b}\right)}\right\}; \quad x > 0; \quad p = F(x)$$

$$\Rightarrow Q_p = a + b\ln[-\ln(p)]; \quad 0 < p < 1$$

Per ottenere una variabile casuale con la densità di Weibull si adotta la trasformazione:

$$F(x) = 1 - e^{-\left[\frac{x-\mu}{\sigma}\right]^{\gamma}} \Rightarrow X = \mu + \sigma[-\ln(p)]^{\frac{1}{\gamma}}$$

1.2.2 Quantili di variabile casuale

Molte moderne tecniche statistiche che si stanno sempre più diffondendo a causa della loro utilità in molte applicazioni traggono fondamento dai quantili teorici ovvero o della popolazione. Il quantile Q_p di una variabile casuale (continua o discreta) con funzione di ripartizione F può essere definito come

$$F\left(x < Q_p\right) \leq p \leq F\left(x \leq Q_p\right) \tag{1.43}$$

Una soluzione per la [1.43] è sempre disponibile, ma può essere unica solo per alcuni quantili se la variabile casuale è discreta, altrimenti ricade in un intervallo di valori. E' sempre unica, tuttavia, per le variabili casuali continue, almeno quelle dotate di una funzione di ripartizione strettamente crescente in un intorno di Q_p dato che, in virtù di questa condizione, i due estremi della disuguaglianza coincidono.

$$Q\left(p\right) = F^{-1}\left(Q_p\right) \tag{1.44}$$

La definizione [1.43] non sempre è univoca per le variabili continue, la cui funzione di densità non sia strettamente positiva. Per ampliarla possiamo ricorrere alla definizione seguente, sviluppata da Gastwirth [**?**]

$$Q\left(p\right) = inf\left\{x\left|F(x) \geq p\right.\right\} \qquad per \quad 0 \leq \ p \ \leq 1 \tag{1.45}$$

Tale definizione implica che, tra tutti i valori della variabile che implicano la verifica della disuguaglianza, si sceglie il valore mi nimo. In questo modo è risolta ogni incertezza sulla univocità del legame tra Q e p. La relazione [1.45] è continua a sinistra (dato che, convenzionalmente, la funzione di ripartizione è continua a destra).

La definizione alternativa del quantile come soluzione di un problema di minimo è possibile anche per i quantili teorici o della popolazione. A questo fine consideriamo la quantità

$$\begin{aligned} \rho\left(Q_p\right) = & \quad pE\left[\left|x - Q_p\right| \left|x > Q_p\right.\right] + \left(1-p\right) E\left[\left|x - Q_p\right| \left|x \leq Q_p\right.\right] \\ = & \quad p\int_{x>Q_p} \left|x - Q_p\right| \ dF\left(x\right) + \left(1-p\right)\int_{x\leq Q_p} \left|x - Q_p\right| \ dF\left(x\right) \\ = & \quad p\int_{Q_p}^{\infty} \left(x - Q_p\right) \ dF\left(x\right) + \left(1-p\right)\int_{-\infty}^{Q_p} \left(Q_p - x\right) dF\left(x\right) \end{aligned} \tag{1.46}$$

Per minimizzare la [1.46] si può calcolare la derivata prima rispetto ad Q_p. Questo comporta la derivazione sotto il segno di integrale

$$\frac{\partial}{\partial x}\left[\int_{-\infty}^{g(x)}\psi(x,y)\,dy\right]=\int_{-\infty}^{g(x)}\frac{\partial\psi(x,y)}{\partial x}dy+\psi[x,g(x)]\frac{\partial g(x)}{\partial x} \tag{1.47}$$

dove $g(x)$ e $\psi(x,y)$ sono funzioni derivabili. La regola [1.47] si applica alla [1.46] per produrre

$$\begin{aligned}\frac{\partial\rho(Q_p)}{\partial Q_p} &= p\left[\frac{\partial\int_{Q_p}^{\infty}(x-Q_p)\,dF(x)}{\partial Q_p}\right]+(1-p)\left[\frac{\int_{-\infty}^{Q_p}(Q_p-x)\,dF(x)}{\partial Q_p}\right]\\ &= -p\int_{Q_p}^{\infty}dF(x)+(1-p)\int_{-\infty}^{Q_p}dF(x)\\ &= -p[1-F(Q_p)]+(1-p)F(Q_p)\\ &= -p+pF(Q_p)+F(Q_p)-pF(Q_p)\\ &= F(Q_p)-p \end{aligned} \tag{1.48}$$

Se ora poniamo la [1.48] pari a zero otteniamo Q_p come radice dell'equazione $F(Q_p)=p$ che è esattamente la definzione del quantile Q_p. Che poi questo corrisponda ad un minimo della [1.48] si può controllare verificando che la sua derivata seconda rispetto ad Q_p è positiva laddove esista la funzione di densità.

Esempio 16. *I quantili teorici possono essere utilizzati per simulare, grazie al computer, moltissime variabili casuali. Tale risultato, di grande importanza nelle applicazioni, deriva direttamente dalla trasformazione dell'integrale di probabilità. Infatti, se p è una variabile casuale con distribuzione uniforme nell'intervallo $[0,1]$ e x è una variabile casuale con funzione di ripartizione F, dotata di inversa, in corrsipondenza di $p\in[0,1]$ si ha*

$$Pr[Q(p)\le x]=Pr\{F[Q(p)]\le F(x)\}=$$
$$Pr[p\le F(x)]=F(x)=Pr[X\le x]$$

Se si deve simulare un valore della x si può procedere in due passi: si simula un valore della uniforme; al risultato si applica la trasformazione quantile $Q(p)=F^{-1}(p)=Q_p$ che è appunto un valore dalla F. Ad esempio, per la variabile casuale esponenziale si usa la trasformazione

$$Q_p=\beta\ln\left(\frac{1}{1-p}\right)\quad per\quad p\sim U(0,1)$$

Naturalmente questo è un vantaggio, perchè le simulazioni dalla uniforme sono più semplici e rapide che le simulazioni da ogni altra distribuzione.

1.2.3 La funzione quantile

E' il meccanismo che lega le modalità osservate in un campione ovvero i valori possibili di una variabile casuale (o popolazione) alle frequenze o probabilità con cui sono osservate o potrebbero essere osservate. Costituisce inoltre la chiave di lettura più importante del presente lavoro. La funzione quantile è detta empirica se modalità e frequenze derivano da una rilevazione effettivamente realizzatesi; è teorica se basata su modalità e frequenze nate da una funzione matematica che simula -per motivi di studio o di controllo- un processo stabile e non caotico di rilevazione; esse non sono osservate in realtà, forse non lo saranno mai e, in fondo, non interessa più di tanto sapere se si sono o non si sono verificate. Rientrano nella problematica dell'inferenza come modelli di ciò che si osserva o si potrebbe osservare se certe condizioni si verificassero.

La funzione quantile esprime, per ogni p, il valore della variabile casuale cui è associata la probabilità p che si realizzi una osservazione ad esso inferiore o uguale e probabilità $(1-p)$ che si realizzi un valore ad essa superiore.

$$Q(p) = F^{-1}(p) = \inf\{x \,|F(x) \geq p\} \quad \text{per } 0 \leq p \leq 1 \tag{1.49}$$

dove $F(.)$ è la funzione di ripartizione della variabile casuale. Ecco alcuni esempi di funzione quantile e le variabili casuali ad esse associate.

Modello	*Dominio*	$Q(p)$	*Modello*	*Dominio*	$Q(p)$
Uniforme	$[0,1]$	$a+(b-a)p$	*Logistica*	$(-\infty,\infty)$	$\ln\left(\frac{p}{1-p}\right)$
Normale	$(-\infty,\infty)$	$\Phi^{-1}(p)$	*Pareto*	$(0,\infty)$	$\alpha^{-1}(1-p)^{-\alpha}$
Lognormale	$(0,\infty)$	$\exp\{-\Phi^{-1}(p)\}$	*Burr/3*	$(0,\infty)$	$\left(p^{-b}-1\right)^{-a}$
Esponenziale	$(0,\infty)$	$-\ln(1-p)$	*Cauchy*	$(-\infty,\infty)$	$\tan[\pi(p-0.5)]$
Weibull	$(0,\infty)$	$\beta^{-1}[-\ln(1-p)]^{\beta}$	*Perks*	$(-\infty,\infty)$	$\tan^{-1}(2p-1)$

La funzione quantile $Q(p)$ ha diverse proprietà che discendono dalla sua definizione (si veda comunque [**?**]). Per $-\infty < x < +\infty$ e $0 < p < 1$ si ha $F(x) \geq p$ se e solo se $Q(p) \leq x$ Di conseguenza, $X \sim Q(p)$ se $p \sim U[0,1]$. Per le funzioni composte quantile/ripartizione si ha $Q[F(x)] \leq x$ per $-\infty < x < +\infty$ nonché $F[Q(p)] \geq p$ per $0 < p < 1$. Se $F(.)$ è strettamente crescente in un intorno infinitesimo di x allora $Q[F(x)] = x$ e $F[Q(p)] = p$. Se $S(p)$ è una

funzione quantile allora lo è anche $Q(p) = \mu + \sigma S(p)$ per $\sigma > 0$. Il primo parametro incide sulla posizione della variabile casuale laddove il secondo ne modifica l'unità di misura o scala. Ad esempio, la variabile casuale potenza ha funzione quantile $S(p) = p^{\alpha}$. Un modello di funzione quantile più flessibile si ottiene con $Q(p) = \mu + \sigma p^{\alpha}$. Nel grafico 1.15 sono riportati alcuni casi particolari per $\alpha = 0.25, \mu = -5,5\ \sigma = 2,8$

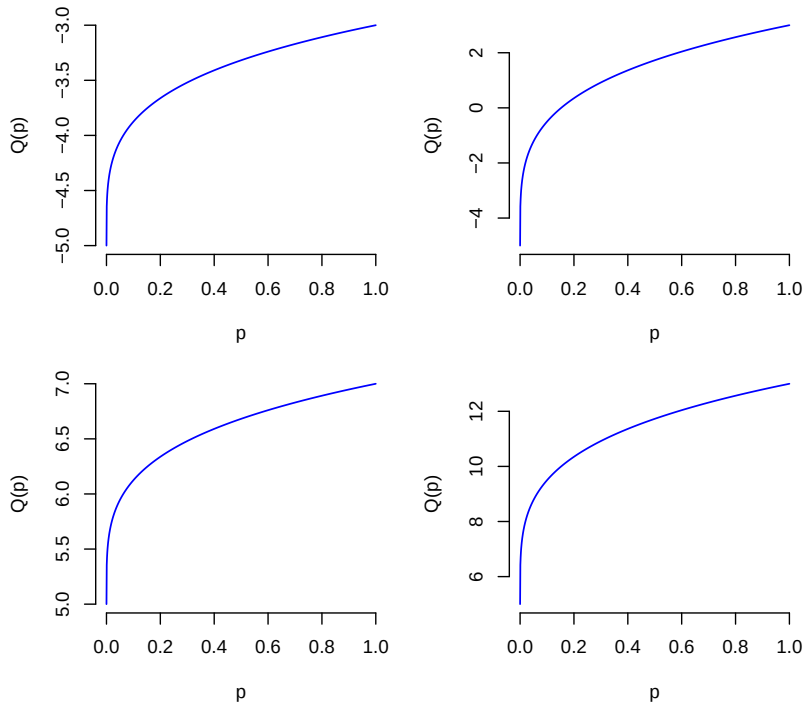

Figura 1.15: Casi particolari della funzione quantile potenza

Se $Q_i(p), i = 1, 2, \cdots, m$ sono delle funzioni quantili allora è anche una funzione quantile ogni loro combinazione lineare finita $Q(p) = \sum_{i=1}^{n} w_i Q_i(p)$. Ad esempio, se si somma la funzione quantile della variabile casuale esponenziale $Q(p) = -\ln(1-p)$ con quella della variabile casuale esponenziale riflessa $Q(p) = \ln(p)$ si ottiene

$$Q(p) = \ln(p) - \ln(1-p) = \ln\left(\frac{p}{1-p}\right)$$

che è la funzione quantile della variabile casuale logistica. La funzione quantile può esistere in forma esplicita come nel caso della Pareto oppure in forma implicita come nella variabile casuale normale. Esistono inoltre funzioni definite solo

nel piano quantile di cui discuteremo più compiutamente alla fine del presente capitolo. Al momento presentiamo solo un esempio illustrativo delle potenzialità di descrizione che si possono ottenere agendo nel piano quantile.

Esempio 17. *Deng e Jiang [?] propongono un modello di funzione quantile per descrivere le* option price *delle imprese legate al settore dell'elettricità.*

$$\frac{Q\left(p;\alpha;\beta;\delta;\mu\right)-\mu}{\delta^{\frac{1}{\alpha}}}=\begin{cases}\left[\log\left(\frac{p^{\beta}}{1-p^{\beta}}\right)\right]^{\frac{1}{\alpha}} & se\ p\geq 0.5^{\frac{1}{\beta}}\\ -\left[-\log\left(\frac{p^{\beta}}{1-p^{\beta}}\right)\right]^{\frac{1}{\alpha}} & se\ p<0.5^{\frac{1}{\beta}}\end{cases} \tag{1.50}$$

dove μ è un parametro di centralità; δ è un parametro di scala; β controlla il bilanciamento delle code: $\beta=1$ significa code bilanciate e $\beta<1\,(>1)$ significa che la coda sinistra (destra) è più pesante di quella opposta (corrisponde al tail coefficient *suggerito da Parzen per classificare i modelli di distribuzione non gaussiani. Infine, il parametro α caratterizza lo spessore delle code nel senso che minore è tale parametro maggiore è lo spessore delle code.*

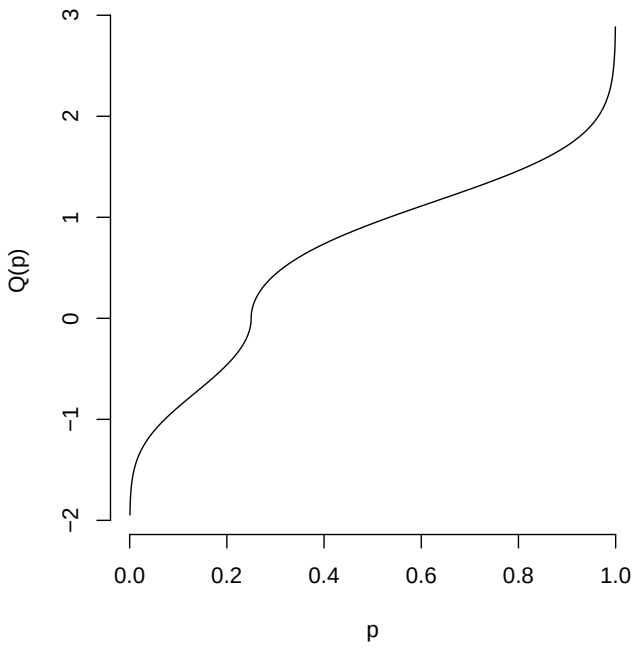

Figura 1.16: Funzione quantile di Deng e Jiang

L'espressione [1.50] ha, inoltre, il merito di definire in forma esplicita anche la funzione di ripartizione, cosicché si è liberi di scegliere il piano su cui proporre la descrizione della variabile casuale.