Politecnico di Milano
Appunti delle lezioni del corso di Statistica (2L) per gli allievi INF e TEL, AA 2008/2009*

Inferenza non parametrica

Ilenia Epifani

13 maggio 2009

Indice

Siamo interessati a fare inferenza su una f.d.r. F. Rispetto alle precedenti lezioni, la situazione è cambiata perché consideriamo il caso di completa ignoranza intorno a F e quindi abbiamo bisogno di procedure inferenziali indipendenti dalla forma di F. Queste procedure vanno sotto il nome di *metodi non parametrici*, perché non ci sono parametri di dimensione finita coinvolti nell'indagine. Oppure, questi metodi sono anche detti *metodi distribution-free*, perché l'unica informazione che qualche volta servirà per implementare la procedura riguarderà la natura di F, se è f.d.r. discreta o continua.

1 Funzione di ripartizione empirica

[1] Sia $X_1, \ldots, X_n$ un campione casuale da F. Per stimare la f.d.r. incognita F costruiamo la f.d.r. empirica associata al campione.

Definizione 1.1 La *funzione di ripartizione empirica (o campionaria) associata al campione* $\hat{F}_n$ è una funzione su $\mathbb{R}$ a valori in $[0,1]$ definita da

$$\hat{F}_n(x) = \frac{\#\{j : X_j \leq x\}}{n} \qquad \forall x \in \mathbb{R} \tag{1}$$

Disponiamo le osservazioni del campione $X_1, \ldots, X_n$ in ordine crescente e indichiamo con $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ la sequenza così ottenuta. Le realizzazioni di $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ dipendono solo dal campione osservato, quindi $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ sono statistiche e vanno sotto il nome di *statistiche d'ordine*. $X_{(1)}$ e $X_{(n)}$ sono due statistiche che avete già incontrato: $X_{(1)}$ è il minimo e $X_{(n)}$ è il massimo delle osservazioni.

Possiamo rappresentare $\hat{F}_n$ in termini di $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ nel seguente modo:

$$\hat{F}_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{k}{n} & X_{(k)} \leq x < X_{(k+1)} \quad (k = 1, \ldots, n-1) \\ 1 & x \geq X_{(n)} \end{cases} \tag{2}$$

Osservate che la funzione $\hat{F}$ è aleatoria e dipende soltanto dal campione casuale, quindi è una statistica. Inoltre, qualunque sia la realizzazione campionaria, è una funzione a gradini, compresa fra 0 e 1, monotona crescente e continua da destra. Cioè ogni realizzazione di $\hat{F}_n$ può essere pensata come una f.d.r. discreta.

Esempio 1.2 Supponiamo di aver osservato $1, 0, 1, -1, 3, 2.5, 3, 1$. Allora $\hat{F}_8$ è

$$\hat{F}_8(x) = \begin{cases} 0 & x < -1 \\ 1/8 & -1 \leq x < 0 \\ 2/8 & 0 \leq x < 1 \\ 5/8 & 1 \leq x < 2.5 \\ 6/8 & 2.5 \leq x < 3 \\ 1 & x \geq 3 \end{cases} \tag{3}$$

[1] Riferimenti in Pestman (1998): Sezioni VII.1, VII.3 (solo enunciato del teorema di Glivenko-Cantelli).

Investighiamo ora i valori tipici di sintesi di $\hat{F}_n$. Ritroveremo alcune statistiche viste nelle passate lezioni. Infatti,

- il momento r-esimo di $\hat{F}_n$ è
$$M_r = \frac{1}{n}\sum_{j=1}^{n} X_j^r$$
cioè quello che avevamo chiamato momento campionario r-esimo. In particolare,
- la media di $\hat{F}_n$ è $M_1 = \bar{X}$, cioè la media campionaria;
- la varianza di $\hat{F}_n$ è
$$\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2 = \frac{(n-1)S^2}{n}$$
ed è proporzionale alla statistica che abbiamo chiamato varianza campionaria;
- la *mediana*[2] di $\hat{F}_n$ è
$$\hat{q}_{1/2} = \begin{cases} X_{(n+1)/2} & \text{se } n \text{ è dispari} \\ \frac{X_{(n/2)}+X_{(n/2+1)}}{2} & \text{se } n \text{ è pari} \end{cases}$$

Analizziamo ora le proprietà probabilistiche della statistica $\hat{F}_n(x)$.

Fissato x, introduciamo le v.a. $Y_1, \ldots, Y_n$ definite da $Y_j = \mathbf{1}_{(-\infty,x]}(X_j)$ per $j = 1, \ldots, n$. Le v.a. $Y_1, \ldots, Y_n$ costituiscono un campione casuale estratto dalla densità bernoulliana di parametro $F(x)$ e $\hat{F}_n(x)$ può essere interpretata come la media campionaria delle Y_j, cioè $\hat{F}_n(x) = \bar{Y}$. Seguono da questa rappresentazione probabilistica di $\hat{F}_n$ le seguenti proprietà:

1. Per ogni $x \in \mathbb{R}$ fissato, $n\hat{F}_n(x)$ rappresenta il numero di osservazioni di valore al più pari a x e ha distribuzione binomiale di parametri $n\, F(x)$);

2. $\hat{F}_n(x)$ è stimatore non distorto e consistente in media quadratica di $F(x)$; infatti $\mathrm{E}_F(\hat{F}_n(x)) = F(x)$ e $\mathrm{Var}_F(\hat{F}_n(x)) = F(x)(1-F(x))/n \to 0$, per $n \to \infty \ \forall F, \forall x$.

In realtà vale un risultato più forte di convergenza di $\hat{F}_n$ a F uniforme in x, fornito dal seguente teorema di Glivenko-Cantelli:

3. *Sia $X_1, X_2, \ldots$ una sequenza di v.a. i.i.d. con comune f.d.r. F. Allora*
$$P\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| = 0\right) = 1$$

Quindi, la "funzione aleatoria" $\hat{F}_n$ è uno stimatore consistente "globalmente" *in senso forte* per F.

6. La successione $\{\hat{F}_n(x)\}_n$ è asintoticamente gaussiana, cioè
$$\lim_{n\to\infty} P_F\left(\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)[1-F(x)]}} \le z\right) = \Phi(z), \quad \forall z \in \mathbb{R}, \quad \forall x \in \mathbb{R} \text{ t.c. } 0 < F(x) < 1$$

Per convincersi di ciò è sufficiente applicare il teorema centrale del limite alla media campionaria di v.a. bernoulliane i.i.d.

[2]Nel caso di una f.d.r. discreta, per *mediana* intendiamo un qualunque valore che lasci alla sua sinistra e alla sua destra almeno metà della massa di probabilità. Nel caso di una f.d.r. F strettamente crescente, la mediana coincide con il quantile di ordine $1/2$ di F.

2 Problemi ipotetici per un singolo campione. Test di buon adattamento

Consideriamo ora qualche esempio di problema ipotetico che si può affrontare avendo a disposizione un campione di n osservazioni i.i.d.

Sia $X_1, \ldots, X_n$ un campione casuale da F. Vogliamo stabilire se la comune F sottostante al campione sia un'assegnata F_0 completamente specificata, contro l'alternativa che non lo sia, cioè vogliamo costruire una procedura di verifica dell'ipotesi nulla semplice $H_0 : F = F_0$ contro l'alternativa composta $H_1 : F \neq F_0$. Per esempio:

$$H_0 : F = \mathcal{N}(0,1) \quad \text{contro} \quad H_1 : F \neq \mathcal{N}(0,1),$$

oppure

$$H_0 : F = Poisson(2) \quad \text{contro} \quad H_1 : F \neq Poisson(2).$$

Ancora, potremmo essere interessati a verificare l'ipotesi nulla composta che F appartenga ad una famiglia di f.d.r. $\mathcal{F}_0$ specificata a meno di qualche parametro m-dimensionale, cioè $\mathcal{F}_0 = \{F(\cdot, \theta), \theta \in \Theta \subset \mathbb{R}^m\}$, contro l'alternativa che F non appartenga a $\mathcal{F}_0$. Per esempio:

$$H_0 : F \text{ è gaussiana} \quad \text{contro} \quad H_1 : F \text{ non è gaussiana}$$

oppure

$$H_0 : F \text{ è Poisson} \quad \text{contro} \quad H_1 : F \text{ non è Poisson}$$

Un test usato per verificare ipotesi su F (sia che dette ipotesi specifichino completamente F sia che ne identifichino solo la forma) è detto *test di buon adattamento* ("*test on goodness of fit*") o *test di verifica del modello.*

I test di buon adattamento che vedremo sono basati su statistiche test che misurano lo scostamento fra f.d.r. empirica $\hat{F}_n$ associata al campione e f.d.r. F specificata dall'ipotesi nulla. La distribuzione esatta della statistica test sotto l'ipotesi nulla è ottenuta mediante simulazioni Monte Carlo. Mentre, per grandi campioni, sono disponibili espressioni esplicite di essa.

2.1 Test di Kolmogorov-Smirnov

[3] Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim F$. Per affrontare il problema ipotetico $H_0 : F = F_0$ contro $H_1 : F \neq F_0$, introduciamo la statistica test

$$D_n := \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

Se H_0 è vera, sono verosimili valori piccoli di D_n. Inoltre, il Teorema di Glivenko-Cantelli garantisce che $D_n \to 0$ con probabilità 1 al divergere di n. Sulla base di questa osservazione, adottiamo la regola decisionale di *"rifiutare H_0 se D_n è "grande"*. Al solito, quantifichiamo "grande, piccolo" in termini di livello di significatività α, risolvendo l'equazione

$$P_0(D_n > k) = \alpha$$

Da qui la necessità di investigare la distribuzione di D_n sotto H_0. Vale il risultato che segue.

[3] Sezione VII.4 del Pestman (1998)

Proposizione 2.1 *Se $X_1, \ldots, X_n$ è un campione casuale estratto da F_0 e F_0 è continua, allora la distribuzione di D_n non dipende da F_0 (cioè la distribuzione di D_n è la stessa nella classe delle f.d.r. F_0 continue).*

Dimostrazione Per maggiore semplicità, dimostriamo la Proposizione 2.1 nell'ipotesi che F_0 sia strettamente crescente. Se $X \sim F_0$ allora $U := F_0(X) \sim \mathcal{U}(0,1)$ poichè $F_U(u) = 0$ se $u \leq 0$, $F_U(u) = 1$ se $u \geq 1$, mentre, per $0 < u < 1$ abbiamo

$$F_U(u) = P(F_0(X) \leq u) = P(X \leq F_0^{-1}(u)) = F_0(F_0^{-1}(u)) = u$$

(l'inversa F_0^{-1} esiste perché F_0 è strettamente crescente). Inoltre

$$\hat{F}_n(x) = \frac{\#\{j : X_j \leq x\}}{n} = \frac{\#\{j : F_0(X_j) \leq F_0(x)\}}{n} = \frac{\#\{j : U_j \leq F_0(x)\}}{n}$$

dove $U_j := F_0(X_j)$, $j = 1, \ldots, n$. Infine,

$$D_n = \sup_{x \in \mathbb{R}} \left| \frac{\#\{j : U_j \leq F_0(x)\}}{n} - F_0(x) \right| = \sup_{u \in [0,1]} \left| \frac{\#\{j : U_j \leq u\}}{n} - u \right|$$

Segue che la distribuzione di D_n sotto H_0 coincide con quella della statistica D_n che si otterrebbe se il campione fosse estratto dalla f.d.r. $\mathcal{U}(0,1)$. ∎

La statistica test D_n è detta *statistica di Kolmogorov-Smirnov* e il test basato su D_n è il *test di Kolomogorov-Smirnov*.

I quantili di D_n sotto H_0 sono tabulati per vari valori di n e di α. Inoltre, per $n \to \infty$, è nota anche l'espressione chiusa della f.d.r. limite di D_n, nel senso che

$$\lim_{n \to \infty} P_0\left(\sqrt{n} D_n \leq z\right) = H(z) \qquad \forall z \in \mathbb{R}$$

dove

$$H(z) = 1 - 2 \sum_{j=0}^{\infty} (-1)^{j-1} \mathrm{e}^{-2j^2 z^2}$$

Esercizio 2.2 (MPSPS 15 giugno 2000) Dato il campione di ampiezza 4:

$$1.126,\ 3.104,\ 2.577,\ 2.372$$

verificate l'ipotesi H_0 che il campione sia generato da una f.d.r. esponenziale di parametro 1 ai livelli di significatività 1% e 10%, mediante il test di Kolmogorov-Smirnov.

Soluzione Procediamo a determinare la statistica di Kolmogorov-Smirnov D_4:

$x_j =$	1.126	2.372	2.577	3.104
$F_0(x_j) =$	$1 - \mathrm{e}^{-1.126}$ $\simeq 0.6757$	$1 - \mathrm{e}^{-2.372}$ $\simeq 0.9067$	$1 - \mathrm{e}^{-2.577}$ $\simeq 0.9240$	$1 - \mathrm{e}^{-3.104}$ $\simeq 0.9551$
$\hat{F}_4(x_j) =$	0.25	0.5	0.75	1
$\lvert\hat{F}_4(x_j) - F_0(x_j)\rvert =$	0.4257	0.4067	0.174	0.0449
$\lvert\hat{F}_4(x_{j-1}) - F_0(x_j)\rvert =$	$\boxed{0.6757}$	0.6567	0.4240	0.2051

da cui $D_4 = 0.6757$. Il p-value è 0.05186, così a livello 1% accetto H_0 mentre a livello 10% rifiuto H_0.[4] ∎

[4] I dati sono stati generati dalla f.d.r. $gamma(3,1)$). Usando il pacchetto *ctest* del software *R* (`http://cran.r-project.org`) otteniamo:

```
ks.test(c(1.126, 3.104, 2.577, 2.372), pgamma,1, 1)
One-sample Kolmogorov-Smirnov test
data:  c(1.126, 3.104, 2.577, 2.372)
D = 0.6757, p-value = 0.05186
alternative hypothesis:  two.sided
```

Bande di confidenza per F Il fatto che la distribuzione di $D_n = \sup_x |\hat{F}_n(x) - F(x)|$ sia la stessa nella classe delle F continue, permette di costruire una "banda di confidenza per F" nel seguente modo: fissiamo γ e n e calcoliamo, con l'uso delle tavole (o di un software statistico che li fornisce), il quantile di ordine γ della f.d.r. di Kolmogorov-Smirnov, $q_n(\gamma)$, cioè risolviamo l'equazione $P(D_n \leq q_n(\gamma)) = \gamma$. L'ultima eguaglianza è equivalente a

$$P_F\left(\hat{F}_n - q_n(\gamma) \leq F(x) \leq \hat{F}_n + q_n(\gamma), \quad \forall x \in \mathbb{R}\right) = \gamma$$

Per esempio: con $\gamma = 0.95$ e $n = 8$ si ha $q_\gamma \simeq 0.4543$. Se ora abbiamo il valore della "funzione aleatoria" $\hat{F}_8$ per una realizzazione $(x_1, \ldots, x_8)$, (per esempio la f.d.r. empirica (3) dell'esempio 1.2), considerato che $0 \leq F(x) \leq 1\ \forall x$, allora

$$\left\{F \text{ f.d.r. continue t.c. } \max\{0, \hat{F}_8(x) - 0.4543\} \leq F(x) \leq \min\{1, \hat{F}_8(x) + 0.4543\}, \quad \forall x \in \mathbb{R}\right\}$$

è una *banda di confidenza di livello* 95% *per* F.

2.2 Test χ^2 per dati categorici o discreti

[5] A differenza del test di Kolmogorov-Smirnov che può essere usato solo se i dati sono continui e l'ipotesi nulla è semplice, il test di buon adattamento χ^2 permette di affrontare anche i problemi ipotetici: *a*) $H_0 : F = F_0$ contro $H_1 : F \neq F_0$ e *b*) $H_0 : F \in \mathcal{F}_0$ contro $H_1 : F \notin \mathcal{F}_0$ per qualunque tipo di dati discreti e continui.

Sia F una f.d.r. discreta a k salti in $a_1, \ldots, a_k$ e siano $a_1, \ldots, a_k$ noti. In questo caso, X assume con probabilità strettamente positiva solo i valori $a_1, \ldots, a_k$, ma sono incognite le probabilità $P(X = a_k)$. Il problema di ipotesi $H_0 : F = F_0$ contro $H_0 : F \neq F_0$ è un problema falsamente non parametrico[6] in quanto i parametri incogniti da cui F dipende sono $k - 1$, tanti quante le ampiezze del salto di F

$$p_1 = P_F(X_1 = a_1), \ldots, p_{k-1} = P_F(X_1 = a_{k-1})$$

(l'ultimo salto $p_k = 1 - \sum_{i=1}^{k-1} p_i$ è noto una volta noti i primi $k - 1$). Siano ora

$$p_{01} := P_{F_0}(X_1 = a_1), \ldots, p_{0k} := P_{F_0}(X_1 = a_k)$$

rispettivamente i valori per $p_1, \ldots, p_k$ derivanti dall'ipotesi nulla e procediamo a verificare

$$H_0 : p_i = p_{0i}\ \forall i = 1, \ldots, k \quad \text{contro} \quad H_1 : p_i \neq p_{0i} \text{ per qualche } i.$$

Dato un campione casuale $X_1, \ldots, X_n$ estratto da F calcoliamo la *frequenza assoluta campionaria* di ogni modalità a_i, cioè quante osservazioni assumono valore a_i:

$$N_i = \#\{j : X_j = a_i\} \qquad \forall i = 1, \ldots, k$$

e misuriamo lo scostamento fra i dati e il modello specificato in H_0 mediante la *statistica di Pearson*

$$Q := \sum_{i=1}^{k} \frac{(N_i - np_{0i})^2}{np_{0i}} \tag{4}$$

[5] Sezione III.3 del Pestman (1998)

[6] non a caso nel Pestman (1998) è trattato nel Capitolo III

Se H_0 è vera, allora N_i ha distribuzione binomiale di parametri n, p_{0i} e quindi $\mathrm{E}_0(N_i) = np_{0i}$ per ogni $i = 1, \ldots, k$. Così, sotto H_0, sono verosimili valori "piccoli" di Q.

Sulla base di questa osservazione, adottiamo la seguente regola decisionale:

$$\boxed{\text{se } Q \text{ in (4) è grande, si rifiuti } H_0}$$

Per grandi campioni siamo in grado di determinare approssimativamente il livello critico della statistica test usando il seguente risultato asintotico:

Proposizione 2.3 *Sia $X_1, X_2, \ldots$ una sequenza di v.a. i.i.d. con comune f.d.r. F_0. Allora, per $n \to \infty$ la f.d.r. di $\sum_{i=1}^{k}(N_i - np_{0i})^2/(np_{0i})$ converge alla f.d.r. chiquadro con $k-1$ gradi di liberà.*

Pertanto, per n grande, a livello α

$$\boxed{\text{rifiutiamo } H_0 \text{ se } Q > \chi^2_{k-1}(1-\alpha)}$$

Osservazione 2.4 (Regola per praticoni) Per stabilire quanto grande deve essere n, potremmo usare la stessa regola adottata per l'approssimazione della f.d.r. binomiale con la f.d.r. gaussiana: $n\theta_{0i} > 5$ per ogni $i = 1, \ldots, k$.

Osservazione 2.5 Per semplificare il calcolo della statistica di Pearson Q, osserviamo che

$$\sum_{i=1}^{k}\frac{(N_i - np_{0i})^2}{np_{0i}} = \sum_{i=1}^{k}\frac{N_i^2}{np_{0i}} + \sum_{i=1}^{k}\frac{(np_{0i})^2}{np_{0i}} - 2\sum_{i=1}^{k}\frac{N_i np_{0i}}{np_{0i}} =$$
$$= \sum_{i=1}^{k}\frac{N_i^2}{np_{0i}} + n\sum_{i=1}^{k}p_{0i} - 2\sum_{i=1}^{k}N_i = \sum_{i=1}^{k}\frac{N_i^2}{np_{0i}} - n,$$

dove l'ultima eguaglianza deriva dal fatto che $\sum_{i=1}^{p} p_{0i} = 1$ e $\sum_{i=1}^{k} N_i = n$.

Osservazione 2.6 Le *frequenze relative campionarie*

$$\hat{p}_{ni} = \frac{N_i}{n} \qquad \forall i = 1, \ldots, k$$

possono essere lette come le ampiezze dei salti della f.d.r. empirica $\hat{F}_n$. Espressa in termini di $\hat{p}_{ni}$, la statistica di Pearson diventa

$$Q = n\sum_{i=1}^{k}\frac{(\hat{p}_{ni} - p_{0i})^2}{p_{0i}}$$

In altri termini, il test χ^2 calcola lo scostamento fra f.d.r. empirica $\hat{F}_n$ e teorica F_0 in termini di scostamento fra frequenze relative campionarie $(\hat{p}_{n1}, \ldots, \hat{p}_{nk})$ e densità teoriche $p_{01}, \ldots, p_{0k}$.

D'altro canto, densità di probabilità e frequenze relative campionarie si possono definire anche quando i dati non sono *ordinali* ma *categorici*, cioè quando ogni osservazione è classificata come appartenente a una categoria (cioè di essere di un certo tipo) e fra tipi diversi

non possiamo stabilire nessun ordinamento. Se, per esempio, in un'indagine sociologica, sono interessata a verificare ipotesi statistiche sulla distribuzione delle religioni in Italia, posso applicare il test χ^2, ponendo $p-1=$ numero delle diverse religioni presenti in Italia (una categoria è riservata a tutti gli altri che non ne professano nessuna) e interpretando p_{0i} come la probabilità che un soggetto scelto a caso fra quelli che vivono in Italia professi la religione i.

Esercizio 2.7 (MPSPS 3 giugno 1999, forse) Sulla base delle dimensioni, i biologi marini classificano i granchi blu come *giovani*, *adulti* e *anziani*. In una popolazione *sana* le proporzioni ideali sono: 50% giovani, 30% adulti, 20% anziani. Un discostamento da tali proporzioni indica squilibrio dell'ecosistema. In una piccola baia vengono pescati 58 granchi giovani, 33 adulti e 39 anziani. Si può ritenere che la popolazione sia sana?
Soluzione Verifichiamo l'ipotesi

$$H_0 : p_1 = 0.5,\ p_2 = 0.3,\ p_3 = 0.2$$

sapendo che $N_1 = 58$, $N_2 = 33$, $N_3 = 39$ e $n = 58 + 33 + 39 = 130$. Allora

$$Q = \frac{\frac{58^2}{0.5} + \frac{33^2}{0.3} + \frac{39^2}{0.2}}{130} - 130 \simeq 8.177$$

Il p-value del test è $1 - F_{\chi^2_{3-1}}(8.177) = 1 - F_{\mathcal{E}(2)}(8.177) = \mathrm{e}^{-8.177/2} \simeq 0.017 = 1.7\%$. Così, per esempio, a livello 1% accettiamo l'ipotesi che la popolazione sia sana, ma a livello 5% la rifiutiamo[7]. ∎

2.3 Test χ^2 per dati qualunque

Il test χ^2 di buon adattamento può essere implementato anche per la verifica di un modello discreto numerabile o continuo.

Sia nel caso di ipotesi nulla semplice che in quello di ipotesi nulla composta, dobbiamo raggruppare i dati in k classi e confrontare le frequenze osservate di queste classi con le corrispondenti frequenze attese sotto H_0.

Test χ^2 per H_0 semplice. Consideriamo prima di tutto il problema di verifica delle ipotesi $H_0 : F = F_0$ contro $H_1 : F \neq F_0$ con F_0 completamente specificata. Sia $X_1, \ldots, X_n$ un campione casuale e $A_1, \ldots, A_k$ k intervalli disgiunti di $\mathbb{R}$. Per ogni $i = 1, \ldots, k$ calcoliamo

a) il numero N_i di osservazioni che cadono in A_i;

b) la probabilità teorica sotto H_0 che X cada in A_i cioè $p_{0i} = P_{F_0}(X \in A_i)$ (osserviamo che se H_0 è vera, il numero medio delle osservazioni che cadono in A_i è np_{0i});

c) lo scostamento fra $\hat{F}_n$ e F_0 in termini di scostamento fra N_i e np_{0i} mediante la statistica di Pearson $Q := \sum_{i=1}^{k} (N_i - np_{0i})^2/(np_{0i})$. Ad un livello di significatività α e con un campione numeroso,

[7] Implementando il test chiquadrato col software R: `chisq.test(c(58,33,39),p=c(0.5,0.3,0.2))`

```
Chi-squared test for given probabilities
data:  c(58, 33, 39)
X-squared = 8.1769, df = 2, p-value = 0.01677
```

$$\boxed{\text{rifiutiamo } H_0 \text{ se } Q \geq \chi^2_{k-1}(1-\alpha)}$$

Il livello critico $\chi^2_{k-1}(1-\alpha)$ del test χ^2 trova giustificazione nel fatto che la f.d.r. limite di $\sum_{i=1}^{k}(N_i - np_{0i})^2/(np_{0i})$ è χ^2_{k-1}.

Esercizio 2.8 Per testare la bontà di un generatore di numeri pseudo-casuali, genero 250 numeri dalla f.d.r. uniforme sull'intervallo [0,1], ottenendo:

valore di X	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1]$
frequenza	45	53	59	43	50

Sulla base dei dati, cosa concludete circa la bontà del programma di generazione?

Soluzione

$A_i =$	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1]$
$N_i =$	45	53	59	43	50
sotto $H_0 : F = \mathcal{U}(0,1)$: $n \times p_{0i} =$	$250 \times 0.2 = 50$	50	50	50	50

Il valore della statistica di Pearson è:

$$Q = \frac{45^2 + 53^2 + 59^2 + 43^2 + 50^2}{50} - 250 = 3.28$$

I dati sono stati ripartiti in 5 classi; quindi asintoticamente $Q \sim \chi_4$ e il p-value è $\simeq 1 - F_{\chi_4}(3.28) \simeq 1 - 0.4879 = 0.5121$: essendo il p-value molto alto, ai consueti livelli di significatività siamo praticamente certi della bontà del generatore del programma[8]. ∎

Osservazione 2.9 Uno dei problemi più grossi nell'implementazione del test χ^2 per dati raggruppati è la scelta del numero k di intervalli $A_1, \ldots, A_k$ disgiunti e la loro locazione sulla retta. Negli ultimi 80 anni (a partire dai lavori di Fisher degli anni '20, passando per Mann e Wald 1942) sono state elaborate numerose regole per scegliere k in modo tale da non ridurre la potenza del test. Ancora oggi, seppur con qualche modifica, la regola più comune per fissare le classi è quella di Mann e Wald (1942) che proposero di scegliere k in funzione della dimensione del campione n e del livello di significatività α come $k \simeq 4(2n^2/z_{1-\alpha})^{1/5}$ e di scegliere intervalli $A_1, \ldots, A_k$ equiprobabili sotto H_0, cioè tali che $P_{F_0}(A_i) = 1/k$. In alcuni lavori di simulazione dei primi anni 90 (cfr. Del Barrio *et alii* 2000) si è proposto di aggiustare la regola di Mann-Wald dividendo per 4 e pensando $\alpha = 5\%$ (ovvero $z_{1-\alpha} \simeq 1.96$), cosìcché qualcuno sceglie $k =$ parte intera di $n^{2/5}$.

Osservazione 2.10 Per implementare il test di buon adattamento χ^2 dobbiamo "discretizzare" i dati. Se F è continua, ciò produce una perdita di informazione e una conseguente riduzione della potenza del test (leggi aumento della probabilità di errore di seconda specie) rispetto ad altri test di buon adattamento, come per esempio quello di Kolmogorov-Smirnov. Altra critica: a parità di classi $A_1, \ldots, A_k$, la statistica di Pearson non discrimina fra diverse f.d.r. che assegnano a quelle classi stessa probabilità.

D'altro canto, la discretizzazione è anche un pregio del test χ^2 di Pearson, infatti, la sua implementazione richiede la sola conoscenza dei dati raggruppati e non di quelli grezzi.

[8] Usando R:

```
chisq.test(c(45 , 53 , 59 ,43 ,50),p=c(0.2,0.2,0.2,0.2,0.2))
Chi-squared test for given probabilities
data:  c(45, 53, 59, 43, 50)
X-squared = 3.28, df = 4, p-value = 0.5121
```

Test χ^2 per H_0 composta. Effettuiamo ora un test di buon adattamento χ^2 di un modello specificato a meno di qualche parametro incognito e abbiamo raggruppato i dati in k intervalli $(a_0, a_1], (a_1, a_2], \ldots, (a_{k-1}, a_k]$.

Per esempio, vogliamo verificare se

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0 \quad \text{contro} \quad H_1 : X \text{ non è gaussiana}$$

Per implementare il test calcoliamo per ogni $i = 1, \ldots, k$, il numero N_i di osservazioni a valori in $(a_{i-1}, a_i]$ e la probabilità che X cada in $(a_{i-1}, a_i]$ cioè

$$p_i(\mu, \sigma^2) = \Phi\left(\frac{a_i - \mu}{\sigma}\right) - \Phi\left(\frac{a_{i-1} - \mu}{\sigma}\right) \tag{5}$$

Quindi stimiamo i parametri μ, σ^2 e calcoliamo le p_i in (5) usando degli stimatori $\hat{\mu}, \hat{\sigma}^2$. Infine calcoliamo la statistica di Pearson

$$Q^* = \sum_{i=1}^{k} \frac{[N_i - np_i(\hat{\mu}, \hat{\sigma}^2)]^2}{np_i(\hat{\mu}, \hat{\sigma}^2)} \tag{6}$$

La f.d.r. asintotica di Q^* in (6) sotto H_0 è ancora χ^2 ma con diversi gradi di libertà rispetto al caso di ipotesi nulla semplice: se μ e σ^2 sono stimati *"in modo opportuno"*, perdiamo un grado di libertà per ogni parametro stimato e quindi, asintoticamente $Q^* \sim \chi^2_{k-1-2}$.

Ma come possiamo stimare μ e σ^2?

Se abbiamo i dati grezzi usiamo gli stimatori ML: $\hat{\mu}_{ML} = \bar{X}$ e $\hat{\sigma}^2_{ML} = \sum_{j=1}^{n}(X_j - \bar{X})^2/n$. Se, invece, disponiamo solo dei dati raggruppati, una ricetta semplice per stimare μ e σ^2 è la seguente: calcoliamo il valore centrale di ogni intervallo $(a_{i-1}, a_i]$ (di lunghezza finita), cioè

$$c_i = \frac{a_{i-1} + a_i}{2}, \quad i = 1, \ldots, k$$

e lo pesiamo con la numerosità campionaria N_i dell'intervallo. Poi, calcoliamo media campionaria e momento secondo campionario di questi dati, cioè

$$M_1 = \frac{\sum_{i=1}^{k} c_i N_i}{n}, \quad M_2 = \frac{\sum_{i=1}^{k} c_i^2 N_i}{n}$$

e applichiamo il metodo dei momenti. Per un campione casuale gaussiano otteniamo

$$\hat{\mu} = M_1, \qquad \hat{\sigma}^2 = M_2 - M_1^2$$

Il test di buon adattamento χ^2 esemplificato per il modello gaussiano può essere in generale usato ogni qualvolta l'ipotesi nulla specifichi una f.d.r. dipendente da m parametri incogniti, $\theta_1, \ldots, \theta_m$: si stimano $\theta_1, \ldots, \theta_m$ usando il metodo dei momenti con i valori centrali delle classi, ciascuno pesato per la numerosità campionaria della classe, e si usano le stime ottenute $\hat{\theta}_1, \ldots, \hat{\theta}_m$ per calcolare le probabilità $p_1, \ldots, p_k$ specificate da H_0. Si può dimostrare che sotto H_0 la statistica

$$Q^* = \sum_{i=1}^{k} \frac{[N_i - np_i(\hat{\theta}_1, \ldots, \hat{\theta}_m)]^2}{np_i(\hat{\theta}_1, \ldots, \hat{\theta}_m)}$$

ha f.d.r. asintotica χ^2_{k-1-m}: perdiamo un grado di libertà per ogni parametro stimato. A questo punto, la regola di rifiuto da adottare è la seguente:

Per n grande, a livello α,

rifiutiamo $H_0: X \sim F(x;\theta_1,\ldots,\theta_m)$ a livello α se $Q^* > \chi^2_{k-1-m}(1-\alpha)$

Il p-value di questo test asintotico è dato da $1 - F_{\chi^2_{k-1-m}}(Q^*)$.

Osservazione 2.11 (Regola empirica) L'approssimazione χ^2_{k-1-m} della f.d.r. di Q^*_n funziona se $np_i(\hat\theta_1,\ldots,\hat\theta_m) > 5$ per ogni $i = 1,\ldots,k$.

Osservazione 2.12 L'approssimazione asintotica della f.d.r. di Q^* funziona anche quando, disponendo dei dati grezzi, stimiamo $\theta_1,\ldots,\theta_m$ con il metodo di massima verosimiglianza.

Segue qualche esempio.

Esempio 2.13 Una densità esponenziale si adatta ai seguenti dati raggruppati?

classi	N_i
$(0,3]$	40
$(3,4]$	25
$(4,7]$	20
$(7,10]$	15

Abbiamo $40+25+20+15 = 100$ osservazioni ripartite in 4 classi i cui valori centrali sono 1.5, 3.5, 5.5, 8.5; la media campionaria per questi dati raggruppati è 3.85. La densità esponenziale $f(x,\theta) = 1/\theta \mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x)$ ha media θ, quindi la stima di θ è 3.85. Le probabilità attese stimate sotto ipotesi di modello esponenziale sono: $p_1(3.85) = 1 - \mathrm{e}^{3/3.85} \simeq 0.541$, $p_2(3.85) \simeq 0.646 - 0.541 = 0.105$, $p_3(3.85) \simeq 0.838 - 0.646 = 0.192$ e $p_4(3.85) = \mathrm{e}^{-7/3.85} \simeq 0.162$. Inoltre, $100p_i > 5$ per ogni i. La statistica Q^* vale 23.82 e il p-value del test è $1 - F_{\chi^2_{4-1-1}}(23.82) = \mathrm{e}^{-23.82/2} \simeq 6.72 \times 10^{-6}$: vi è fortissima evidenza empirica contro l'ipotesi di dati esponenziali.

Esempio 2.14 Verifichiamo se il numero quotidiano di interruzioni di corrente elettrica da maggio a settembre in una città italiana ha distribuzione di Poisson, basandoci sulle seguenti rilevazioni:

# interruzioni=	0	1	2	3	4	5	6	7	8	9	10	≥ 11
Numero di giorni=	0	5	22	23	32	22	19	13	6	4	4	0

Per un campione casuale proveniente da popolazione poissoniana di parametro θ, lo stimatore ML di θ è dato dalla media campionaria che con i nostri dati vale

$$\hat\theta_{ML} = \frac{1\times 5 + 2\times 22 + 3\times 23 + \cdots + 10\times 4}{5+22+23+32+22+19+13+6+4+4} = \frac{685}{150} = 4.57\ .$$

Le probabilità attese stimate $p_i(4.57) = P(X = i)$ per $i = 0,\ldots,10$ valgono

$$0.010,\ 0.047,\ 0.108,\ 0.165,\ 0.188,\ 0.172,\ 0.131,\ 0.086,\ 0.049,\ 0.025,\ 0.011$$

e $p_{11} = P_0(X \geq 11) = 0.008$, da cui abbiamo che i valori di $150p_i(4.57)$ per $i = 0,\ldots,11$ sono

$$1.50,\ 7.05,\ 16.20,\ 24.75,\ 28.20,\ 25.80,\ 19.65,\ 12.90,\ 7.35,\ 3.75,\ 1.65,\ 1.2$$

Raggruppiamo le prime due e le ultime tre classi, ottenendo

# interruzioni=	≤ 1	2	3	4	5	6	7	8	≥ 9
N_i=	5	22	23	32	22	19	13	6	8
$n \times p_i(4.57)$=	8.55	16.20	24.75	28.20	25.80	19.65	12.90	7.35	6.6

La statistica Q^* ha valore 5.313 ed asintoticamente ha distribuzione χ^2_{9-1-1}; siccome $\chi^2_7(0.9) = 12.02$, accettiamo l'ipotesi nulla di modello di Poisson.

3 Problemi ipotetici per dati accoppiati. Test di indipendenza e concordanza

Siano X, Y due v.a. con f.d.r. congiunta H: $(X, Y) \sim H$ e siano F, G le f.d.r. marginali di X, Y rispettivamente. Cioè:

$$F(x) = \lim_{y\to\infty} H(x, y)$$
$$G(y) = \lim_{x\to\infty} H(x, y)$$

Qui ci poniamo il problema di stabilire se

- i caratteri X e Y sono indipendenti oppure
- se X e Y sono *concordanti*, cioè all'aumentare di X aumenta anche Y (e viceversa) o
- se X, Y sono *discordanti*, cioè all'aumentare dell'uno l'altro diminuisce.

I test che rispondono a questo tipo di problemi sono detti test di indipendenza e test di concordanza.

Affronteremo il problema in ambito non parametrico. Concluderemo la sezione con qualche osservazione per il caso parametrico di f.d.r. congiunta H gaussiana.

3.1 Test χ^2 di indipendenza

In questa sezione costruiamo un test di indipendenza per caratteri X, Y discreti finiti.

Supponiamo che i possibili valori di X siano $x_1, \ldots, x_r$ e che quelli di Y siano $y_1, \ldots, y_s$. Indichiamo con p_{ij} la probabilità congiunta che X assuma valore x_i e Y assuma valore y_j, con p_i la densità marginale di X e con q_j la densità marginale di Y, cioè:

$$p_{ij} = P(X = x_i, Y = y_j), \qquad i = 1, \ldots, r \ ; \ j = 1, \ldots, s,$$
$$p_i = P(X = x_i) = \sum_{j=1}^{s} p_{ij}, \qquad i = 1, \ldots, r$$
$$q_j = P(Y = y_j) = \sum_{i=1}^{r} p_{ij}, \qquad j = 1, \ldots, s \ .$$

Siccome X, Y sono indipendenti se $p_{ij} = p_i q_j \ \forall i = 1, \ldots, r$ e $j = 1, \ldots, s$, allora studieremo l'indipendenza di X, Y con un test dell'ipotesi nulla

$$H_0 : p_{ij} = p_i q_j \ \forall i = 1, \ldots, r, \ j = 1, \ldots, s$$

contro l'alternativa

$$H_1 : p_{ij} \neq p_i q_j \ \text{ per qualche coppia } (i, j).$$

Notate che l'ipotesi H_0 specifica la densità congiunta di X, Y a meno di $r+s-2$ parametri incogniti dati da $p_1, \ldots, p_{r-1}, q_1, \ldots, q_{s-1}$ e l'ipotesi H_1 nega l'ipotesi H_0; quindi, possiamo trattare questo problema di ipotesi con un test chiquadrato di buon adattamento con parametri p_i, q_j da stimare.

Sia $(X_1, Y_1), \ldots, (X_n, Y_n)$ il campione casuale di dati accoppiati. Calcoliamo

1. il numero N_{ij} di coppie del campione $(X_1, Y_1), \ldots, (X_n, Y_n)$ di valore (x_i, y_j);
2. il numero N_{xi} di coppie del campione $(X_1, Y_1), \ldots, (X_n, Y_n)$ con $X = x_i$;
3. il numero N_{yj} di coppie del campione $(X_1, Y_1), \ldots, (X_n, Y_n)$ con $Y = y_j$.

Poi, stimiamo p_i e q_j con gli stimatori:

$$\hat{p}_i = \frac{N_{xi}}{n} \ , \qquad \hat{q}_j = \frac{N_{yj}}{n} \ .$$

Infine, usiamo come statistica test la statistica di Pearson T data da

$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}. \tag{7}$$

Infatti, $\mathrm{E}(N_{ij}) = np_iq_j$ se H_0 è soddisfatta e $n\hat{p}_i\hat{q}_j$ è una stima di np_iq_j, quando p_i, q_j sono incogniti. La statistica T coincide con

$$T = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - \frac{N_{xi}N_{yj}}{n})^2}{N_{xi}N_{yj}} \tag{8}$$

e può essere calcolata come

$$T = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{ij}^2}{N_{xi}N_{yj}} - n \ . \tag{9}$$

La statistica di Pearson qui definita ha f.d.r. asintotica χ^2. Per stabilirne i gradi di libertà consideriamo che il numero delle possibili coppie (x, y) sono $r \times s$ e che i parametri incogniti sotto H_0 sono $r+s-2$. Quindi, per quanto discusso nel Paragrafo 2.3, i gradi di libertà risultano:

$$rs - 1 - (r+s-2) = (r-1)(s-1).$$

Riassumendo abbiamo che la regola di rifiuto da adottare è la seguente:

Per n grande, a livello α,

rifiutiamo H_0 : "X, Y sono indipendenti" a livello α se $T > \chi^2_{(r-1)(s-1)}(1-\alpha)$

Il p-value di questo test asintotico è dato da $1 - F_{\chi^2_{(r-1)(s-1)}}(T)$.

Osservazione 3.1 (Regola empirica) L'approssimazione $\chi^2_{(r-1)(s-1)}$ della f.d.r. di T funziona se abbiamo almeno 5 osservazioni per ogni coppia (i, j).

Osservazione 3.2 Sia $(X_1, Y_1), \ldots, (X_n, Y_n)$ un campione casuale bidimensionale da una f.d.r. H. La funzione aleatoria

$$\hat{H}_n(x, y) = \frac{\#\{i : X_i \le x \ e \ Y_i \le y\}}{n} \qquad \forall x \in \mathbb{R}, \ \forall y \in \mathbb{R}$$

è la *funzione di ripartizione empirica associata al campione* $(X_1, Y_1), \ldots, (X_n, Y_n)$. Se siamo in uno stato di completa ignoranza su H, $\hat{H}_n$ è la migliore stima che possiamo produrre per H. Se, invece, l'ipotesi H_0 : "X, Y *sono indipendenti*" è vera, allora useremo come stimatore di H il prodotto delle f.d.r. empiriche $\hat{F}_n \times \hat{G}_n$ associate ai campioni singoli estratti da F e G, rispettivamente. Segue che possiamo interpretare la statistica test di Pearson T per l'indipendenza come una sintesi (quadratica) della distanza fra le f.d.r. $\hat{H}_n$ e $\hat{F}_n \times \hat{G}_n$. Quindi, di nuovo, anche questo test è basato sulle f.d.r. empiriche congiunta e marginali.

3.2 Test di indipendenza e concordanza di Kendall

Poniamoci ora il problema di verificare se i due caratteri X e Y siano dipendenti: dipendenza positiva fra X, Y significa tendenza dei due ad associarsi in modo tale che all'aumentare dell'uno aumenti anche l'altro; viceversa, dipendenza negativa significa che all'aumentare dell'uno l'altro diminuisce.

3.2.1 Coefficiente τ di Kendall

Kendall ha tradotto matematicamente quest'idea nel seguente modo. Siano (X_1, Y_1), (X_2, Y_2) due copie indipendenti del vettore $(X, Y) \sim H$ e definiamo

$$\pi_c := P[\text{“}(X_1 - X_2) \text{ e } (Y_1 - Y_2) \text{ hanno stesso segno”}] = P[(X_1 - X_2)(Y_1 - Y_2) > 0] \tag{10}$$

e

$$\pi_d := P[\text{“}(X_1 - X_2) \text{ e } (Y_1 - Y_2) \text{ hanno segno opposto”}] = P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{11}$$

Definizione 3.3 I caratteri X, Y sono *perfettamente concordanti* se $\pi_c = 1$, cioè X, Y sono perfettamente concordanti se $X_1 < X_2$ implica *essenzialmente* $Y_1 < Y_2$ e $X_1 > X_2$ implica *essenzialmente* $Y_1 > Y_2$. I caratteri X, Y sono *perfettamente discordanti* se $\pi_d = 1$, cioè X, Y sono perfettamente discordanti se $X_1 < X_2$ implica *essenzialmente* $Y_1 > Y_2$ e $X_1 > X_2$ implica *essenzialmente* $Y_1 < Y_2$.

Se F e G sono continue, allora $\pi_d = 1 - \pi_c$.
Infatti, se F e G sono continue, allora $P(X_1 - X_2 = 0) = P(Y_1 - Y_2) = 0$ e quindi

$$1 = P((X_1 - X_2)(Y_1 - Y_2) > 0) + P((X_1 - X_2)(Y_1 - Y_2) < 0) = \pi_c + \pi_d$$

Definizione 3.4 Una misura di associazione fra X, Y è data da

$$\tau := \pi_c - \pi_d$$

ed è detta *coefficiente τ di Kendall*.

Valgono le seguenti proprietà del coefficiente τ di Kendall.

Proposizione 3.5 *1. Per ogni vettore aleatorio (X, Y) abbiamo $-1 \leq \tau \leq 1$;*

2. X, Y sono perfettamente concordanti se e solo se $\tau = 1$;

3. X, Y sono perfettamente discordanti se e solo se $\tau = -1$;

4. se X, Y sono indipendenti allora $\tau = 0$.

Dimostriamo solo il punto 4. Se X, Y sono indipendenti allora

$$\begin{aligned}\pi_c &= P((X_1 - X_2)(Y_1 - Y_2) > 0) \\ &= P(X_1 - X_2 > 0, Y_1 - Y_2 > 0) + P(X_1 - X_2 < 0, Y_1 - Y_2 < 0) \\ &= P(X_1 - X_2 > 0) P_H(Y_1 - Y_2 > 0) + P(X_1 - X_2 < 0) P_H(Y_1 - Y_2 < 0) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}\end{aligned}$$

Quindi $\pi_d = 1 - \pi_c = 1/2$ da cui segue $\tau = 0$.

Osservazione 3.6 Notate che se $\tau = 0$ non è detto che X, Y siano indipendenti, eccezion fatta per i dati gaussiani. Ma ritorneremo nella prossima sezione su questo punto. Se $\tau = 0$ parliamo genericamente di una situazione di assenza di associazione fra X, Y.

Osservazione* 3.7 Si dimostra[9] che fissate le f.d.r. marginali F di X e G di Y, allora

1. X, Y sono perfettamente concordanti se e solo se $(X, Y) \sim H^+(x, y) = \min(F(x), G(y))$
2. X, Y sono perfettamente discordanti se e solo se $(X, Y) \sim H^-(x, y) = \max(F(x) + G(y) - 1, 0)$.
3. Inoltre, le f.d.r. marginali di H^+ e H^- sono F e G.
4. Infine, la classe delle f.d.r. H tali che $\tau = 0$ è stata anche essa completamente caratterizzata in termini delle f.d.r. marginali.

3.3 Test di indipendenza di Kendall

Vediamo ora alcuni possibili problemi ipotetici di dipendenza che si possono affrontare:

$$H_0 : \tau = 0 \quad \text{contro} \quad H_1 : \tau > 0 \tag{12}$$

per verificare se c'è concordanza;

$$H_0 : \tau = 0 \quad \text{contro} \quad H_1 : \tau < 0 \tag{13}$$

per verificare se c'è discordanza;

$$H_0 : \tau = 0 \quad \text{contro} \quad H_1 : \tau \neq 0 \tag{14}$$

per verificare che non ci sia nessun tipo di associazione fra X e Y. Un test per il problema (14), (con H_1 bilatera), può essere letto come un test di indipendenza. Infatti H_1 è compatibile solo con una situazione di non indipendenza.

Motivati dalle precedenti considerazioni, costruiamo il test di indipendenza e concordanza di Kendall.

Sia $(X_1, Y_1), \ldots, (X_n, Y_n)$ un campione casuale bidimensionale estratto da una popolazione con f.d.r. H incognita. Contiamo quante sono le coppie di dati $(X_i, Y_i), (X_j, Y_j)$ $(i < j,$ $i, j = 1, \ldots, n)$ concordanti e quante quelle discordanti su tutte le possibili $n(n-1)/2$ coppie nel seguente modo.

[9] Se interessati cfr Cifarelli, Conti e Regazzini (1996)

Primo Passo Ordiniamo le coppie del campione guardando al solo carattere X. Formalmente: se $X_{(1)}, \ldots, X_{(n)}$ sono le statistiche d'ordine di $X_1, \ldots, X_n$ e $Y_{[k]}$ è il valore dell'osservazione su Y accoppiata (*concomitante*) a $X_{(k)}$, per $k = 1, \ldots, n$, allora il campione sarà riordinato nel seguente modo: $(X_{(1)}, Y_{[1]}), \ldots, (X_{(n)}, Y_{[n]})$.

Esempio 3.8 (Example 2 in Pestman 1998 pag. 245)
Se ho il campione di 5 dati: $(3.7, 5.4), (2.1, 3.6), (4.2, 1.1), (3.2, 1.9), (2.3, 4.8)$, il campione ordinato come spiegato prima è: $(2.1, 3.6), (2.3, 4.8), (3.2, 1.9), (3.7, 5.4), (4.2, 1.1)$

Secondo Passo Contiamo quante volte sulle $n(n-1)/2$ coppie i segni sono concordanti:

$$C = \#\{(i,j) \text{ t.c. } i, j = 1, \ldots, n \text{ con } i < j \text{ e } Y_{[i]} < Y_{[j]}\}$$

e quante volte i segni sono discordanti:

$$D = \#\{(i,j) \text{ t.c. } i, j = 1, \ldots, n \text{ con } i < j \text{ ma } Y_{[i]} > Y_{[j]}\}$$

Infine definiamo il *coefficiente campionario di concordanza di Kendall*

$$R_K := \frac{2(C-D)}{n(n-1)}$$

È facile dimostrare che

1. R_K è stimatore non distorto di $\tau = \pi_c - \pi_d$
2. se i segni di tutte le coppie concordano, allora $C = n(n-1)/2$, $D = 0$ e $R_K = 1$. Se invece tutte discordano allora $C = 0$, $D = n(n-1)/2$ e $R_K = -1$
3. se $\tau = 0$, R_K è v.a. simmetrica perché ogni ogni coppia ha la stessa probabilità di essere concordante e discordante.

Se $\tau = 0$ la probabilità che la statistica R_K esibisca valori "prossimi" a -1 o a 1 è bassa, perché non c'è nessun tipo di associazione fra X e Y. Quindi, per decidere sull'associazione fra X, Y valutiamo se i valori della statistica $C-D$ siano prossimi a $-n(n-1)/2$ o a $n(n-1)/2$ o lontani da entrambi.

Seguendo questa intuizione, costruiamo le seguenti regioni critiche di ampiezza α:

$\mathcal{G}_1 = \{C - D > q(1-\alpha)\}$ è una regione critica per $H_0 : \tau = 0$ contro $H_1 : \tau > 0$
$\mathcal{G}_2 = \{C - D < -q(1-\alpha)\}$ è una regione critica per $H_0 : \tau = 0$ contro $H_1 : \tau < 0$
$\mathcal{G}_3 = \{|C - D| > q(1 - \frac{\alpha}{2})\}$ è una regione critica per $H_0 : \tau = 0$ contro $H_1 : \tau \neq 0$

Nelle righe precedenti, $q(a)$ è il quantile di ordine a della f.d.r. di $C - D$ sotto H_0 e $q(a) = -q(1-a)$, $\forall a \in (0,1)$ perché $C - D$ è v.a. simmetrica se $\tau = 0$.

Rimane infine da indagare la f.d.r. di $C - D$ quando $\tau = 0$, cioè sotto H_0: per n piccolo i quantili di $C-D$ sono tabulati, per esempio in Conover (1999) pagine 391-392. Per n grande, vale il seguente risultato di approssimazione gaussiana: se $\tau = 0$, allora

$$\lim_{n\to\infty} P\left(3\sqrt{\frac{n(n-1)}{2(2n+5)}} R_K \leq z\right) = \Phi(z) \quad \forall z \in \mathbb{R}$$

Quindi, per grandi campioni approssimiamo il quantile q di $C - D$ (sotto H_0) a partire da quello della f.d.r. gaussiana standard z_a, usando la seguente relazione:

$$q_{C-D}(a) \simeq z_a \sqrt{\frac{n(n-1)(2n+5)}{18}}$$

Esempio 3.9 (Continuazione dell'Esempio 2 del Pestman (1998) pag. 245) Se ho un campione di 5 dati:

$(3.7, 5.4), (2.1, 3.6), (4.2, 1.1), (3.2, 1.9), (2.3, 4.8)$,

il campione ordinato è:

$(2.1, 3.6), (2.3, 4.8), (3.2, 1.9), (3.7, 5.4), (4.2, 1.1)$ e $R_K = -2/10$ perché

	2	3	4	5
$-signum(Y_{[1]} - Y_{[j]}) =$	$+1$	-1	$+1$	-1
$-signum(Y_{[2]} - Y_{[j]}) =$		-1	$+1$	-1
$-signum(Y_{[3]} - Y_{[j]}) =$			$+1$	-1
$-signum(Y_{[4]} - Y_{[j]}) =$				-1

Sia $H_1 : \tau \neq 0$. Con $n = 5$, il p-value è $P(|C - D| \geq 2) \simeq 0.817$ e accettiamo H_0[10].

Esercizio 3.10 Svolgere l'esercizio n. 14 pagina 255 in Pestman (1998), sostituendo alle domande i), ii) le seguenti:

i) Compute the outcome of the sample Kendall's coefficent of concordance (R_K).

ii) Test at a level significance of $\alpha = 0.10$ the null hypothesis $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$.

3.4 Test di indipendenza e concordanza per dati gaussiani

Nel caso di dati congiuntamente gaussiani i test di indipendenza e concordanza si traducono in test sul coefficiente di correlazione lineare ρ. Vale infatti che

Proposizione 3.11 *Se X, Y sono v.a. congiuntamente gaussiane, allora $\tau(X, Y) = 0$ se e solo se $\rho(X, Y) = 0$.*

Dimostrazione Chiedere al docente se interessati. ∎

I test su ρ per dati accoppiati gaussiani sono descritti nella Sezione 8.2 della dispensa su *Verifica di ipotesi, AA 07/08.*

[10] con il software R:

```
x<-c(3.7,2.1,4.2,3.2,2.3);
y<-c(5.4,3.6,1.1,1.9,4.8)
cor.test(x, y, alternative = c(two.sided), method =c(kendall), exact = NULL)
Kendall's rank correlation tau
data:  x and y T = 4, p-value = 0.8167 alternative hypothesis:  true tau is not equal to 0
sample estimates:  tau -0.2
```

3.5 Test di aleatorietà di Kendall (Test of randomness)

Alla base delle procedure inferenziali presentate nel corso è l'ipotesi di casualità, aleatorietà, del campione. Se le n osservazioni $X_1, \ldots, X_n$ sono i.i.d. , allora non importa l'ordine con cui esse "arrivano", cioè un trend crescente o decrescente fra le osservazioni è incompatibile con l'ipotesi di casualità del campione. Vediamo ora come il test di concordanza di Kendall possa essere applicato per verificare l'ipotesi nulla di casualità del campione, in questa particolare accezione di assenza di trend. Quindi costruiamo un test per verificare H_0 : "$X_1, \ldots, X_n$ *sono i.i.d.* " contro un'ipotesi alternativa di trend crescente, o di trend decrescente, o semplicemente contro l'alternativa bilatera: "$H_1 : X_1, \ldots, X_n$ *non sono i.i.d.* ".

Per ogni $i = 1, \ldots, n-1$ siano C_i=numero di osservazioni successive a X_i maggiori di X_i, D_i=numero di osservazioni successive a X_i minori di X_i e T la somma di tutte le C_i meno la somma di tutte le D_i cioè: $C = \sum_{i=1}^{n-1} C_i$, $D = \sum_{i=1}^{n-1} D_i$ e $T = C - D$.

Il valore minimo di T è $-n(n-1)/2$ e T vale $-n(n-1)/2$ se $X_1 > X_2 > \ldots > X_n$. D'altro canto T assume valore massimo $n(n-1)/2$ se e solo se $X_1 < X_2 < \ldots < X_n$. Pertanto $T \approx -n(n-1)/2$ è indice di un trend decrescente e $T \approx n(n-1)/2$ è compatibile soltanto con un trend crescente. Mentre, se il campione è casuale ci aspettiamo che coppie concordanti e discordanti mediamente si compensino e quindi $T \approx 0$. Ma, T è semplicemente la differenza fra il numero di coppie concordanti e quelle discordanti del finto campione bidimensionale $(1, X_1), \ldots, (n, X_n)$. Possiamo usare allora il test di concordanza di Kendall. Le regioni critiche di ampiezza α risultano:

$\mathcal{G}_1 = \{C - D > q(1-\alpha)\}$ per H_0: "$X_1, \ldots, X_n$ è un campione casuale" contro H_1 : *"nei dati c'è un trend crescente"*

$\mathcal{G}_2 = \{C - D < -q(1-\alpha)\}$ per H_0 : "$X_1, \ldots, X_n$ *è un campione casuale*" contro H_1 : *"nei dati c'è un trend decrescente"*

$\mathcal{G}_3 = \{|C - D| > q(1-\alpha/2)\}$ è per H_0: "$X_1, \ldots, X_n$ *è un campione casuale*" contro H_1 : *"i dati non sono indipendenti"*.

con $q(a)$ quantile di ordine a di $C - D$ sotto H_0.

Esempio 3.12 Usando R ho generato 10 numeri casuali dalla $U(0,1)$ e, nell'ordine d'arrivo, i dati ottenuti sono

0.5923 0.6944 0.6956 0.6443 0.6114 0.5073 0.0993 0.1070 0.6701 0.3607

Potete considerare questi numeri come realizzazione di un campione casuale?

Soluzione

i	0.5923	0.6944	0.6956	0.6443	0.6114	0.5073	0.0993	0.1070	0.6701	0.3607
$C-D$	$5-4$	$+1-7$	$+0-7$	$+1-5$	$+1-4$	$+1-3$	$+3-0$	$+2-0$	$+0-1$	$\boxed{-17 = T}$

Se H_1 è bilatera, uso la regione $\mathcal{G}_3$ e il p-value del test è $2(1 - P(T \leq 17))$. Sulle tavole, per $n = 10$, i valori più prossimi a 17 sono 15 e 19 con $P(T \leq 15) = 0.90$ e $P(T \leq 19) = 0.95$. Pertanto il p-value è compreso fra 10% e 20%. Interpolando linearmente $(15, 0.90)$ e $(19, 0.95)$ otteniamo p-value$\simeq$ 15% (R forniva 0.1557). Rifiuterò l'ipotesi di randomness solo a un livello $\geq 15\%$. Quindi concludo che il generatore è un buon generatore. ∎

4 Test di omogeneità

Affrontiamo ora il problema di verificare se due v.a. X, Y sono *omogenee* cioè se sono regolate dallo stesso modello. Altrimenti detto verifichiamo se X e Y hanno la stessa f.d.r. Sia F la f.d.r. di X e G quella di Y e costruiamo un *test di omogeneità* per l'ipotesi nulla

$$H_0 : F(x) = G(x) \qquad \forall x \in \mathbb{R}$$

contro l'alternativa

$$H_1 : F(x) \neq G(x) \qquad \text{per qualche } x \in \mathbb{R} \tag{15}$$

oppure

$$H_1 : F(x) \leq G(x) \ \ \forall x \in \mathbb{R} \text{ e } F(x) < G(x) \text{ per qualche } x \tag{16}$$

oppure

$$H_1 : F(x) \geq G(x) \ \ \forall x \in \mathbb{R} \text{ e } F(x) > G(x) \text{ per qualche } x \tag{17}$$

Le ipotesi alternative unilaterali hanno il seguente significato: se (16) è vera, allora X tende a essere più grande di Y e diciamo che X *domina stocasticamente* Y. Se invece è vera (17), allora è Y che tende a essere più grande di X cioè Y domina stocasticamente X.

I dati a nostra disposizione possono essere di due tipi:

1. due campioni casuali indipendenti $X_1, \ldots, X_m$ *i.i.d.* $\sim F$ e $Y_1, \ldots, Y_n$ *i.i.d.* $\sim G$ oppure

2. un campione casuale di dati accoppiati $(X_1, Y_1), \ldots, (X_n, Y_n)$ generati da una f.d.r. congiunta H con f.d.r. marginali F di X e G di Y.

I test di buon adattamento χ^2 e di Kolmogorov-Smirnov possono essere modificati per verificare $H_0 : F(x) = G(x)\ \forall x \in \mathbb{R}$. Qui noi invece sviluppiamo il *test di omogeneità di Wilcoxon-Mann-Whitney per campioni indipendenti* e il *test dei segni di Wilcoxon per dati accoppiati*.

Per evitare complicazioni tecniche, assumiamo F, G continue: dalla continuità di F e G deriva che *essenzialmente* non ci sono ripetizioni nei campioni in quanto $P(X_{i_1} = X_{i_2}) = P(Y_{j_1} = Y_{j_2}) = P(X_i = Y_j) = 0\ \forall i \neq j, i_1 \neq i_2$ e $j_1 \neq j_2$.

L'allievo interessato è rimandato al Capitolo 5 in Conover 1999, per le varianti ai test necessarie in presenza di ripetizioni (*tail*) nel campione.

4.1 Test di omogeneità di Wilcoxon-Mann-Whitney per due campioni indipendenti

[11] Estraiamo due campioni casuali indipendenti $X_1, \ldots, X_m$ da F e $Y_1, \ldots, Y_n$ da G e riuniamo tutte le osservazioni in un unico campione di ampiezza $m+n$. Registriamo il *rango* ("rank" o *grado*) di ogni osservazione, cioè la posizione che essa occupa nella classifica di tutte le $m+n$ osservazioni dalla più piccola alla più grande, chiamiamo R_i il rango di X_i e sommiamo i ranghi delle X_i: $T_X = \sum_{i=1}^{m} R_i$.

[11] Riferimenti bibliografici: Sezione VI.2 "Wilcoxon's rank-sum test" pagine 233-237 in Pestman (1998).

Esempio 4.1 Se $(x_1, \ldots, x_4) = (23, 10, 21, 5)$ e $(y_1, \ldots, y_5) = (3, 8, 20, 25, 12)$ abbiamo

valori ordinati=	3_y	5_x	8_y	10_x	12_y	20_y	21_x	23_x	25_y
classifica=	1	2	3	4	5	6	7	8	9
ranghi di X=		R_1		R_2			R_3	R_4	
T_X=		2+		4+			7+	8 =	$\boxed{21}$

Se tutte le x_i sono più piccole di ogni y_j, T_X ha valore $m(m+1)/2$; se, invece, tutte le x_i sono più grandi di ogni y_j, allora $T_X = m(2n+m+1)/2$.

Se $F(x) < G(x)$, mi aspetto che un gran numero di x_i siano più grandi delle y_j e quindi che T_X sia "grande". Mentre, se $F(x) > G(x)$, mi aspetto che molte y_j siano più grandi delle x_i e quindi T_X sia "piccolo". Infine, se $F = G$, le x_i e y_j sono mescolate casualmente e quindi con alta probabilità T_X sarà lontano dai valori estremi.

Per costruire il test di omogeneità di Wilcoxon-Mann-Whitney abbiamo bisogno di sapere qualcosa in più sulla f.d.r. di T_X se $F = G$. A tal fine, introduciamo la statistica U che fornisce il numero complessivo di X_i maggiori di Y_j nel campione riunito:

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}_{(Y_j,\infty)}(X_i) \tag{18}$$

Per ottenere U, individuiamo le m statistiche d'ordine $x_{(1)} < \cdots < x_{(m)}$ e contiamo il numero a_1 di y_j superate da $x_{(1)}$, il numero a_2 di y_j superate da $x_{(2)}$,..., e il numero a_m di y_j superate da $x_{(m)}$. Poi sommiamo questi numeri cioè $U = a_1 + a_2 + \cdots + a_m$. Ma $a_1 =$ "$\#y_j \le x_{(1)}$"$= R_1 - 1$, $a_2 =$ "$\#y_j \le x_{(2)}$"$= R_2 - 2, \ldots, a_m =$ "$\#y_j \le x_{(m)}$"$= R_m - m$ da cui deduciamo che

$$U = \sum_{i=1}^{m} (R_i - i) = T_X - \frac{m(m+1)}{2} \tag{19}$$

La statistica U è la *statistica di Mann e Whitney*, mentre la somma dei ranghi T_X è la *statistica di Wilcoxon*.

Usiamo ora U per calcolare media e varianza di T_X, se $F = G$. Se $F = G$, allora $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ è un campione casuale di $m+n$ osservazioni con comune f.d.r. F e

$$\mathrm{E}(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{E}(\mathbf{1}_{(Y_j,\infty)}(X_i)) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(X_i > Y_j) = mnP(X_1 > Y_1) = \frac{mn}{2}$$

in quanto $P(X_1 = Y_1) = 0$ e $P(X_1 > Y_1) = P(X_1 < Y_1) = 1/2$ perché X_1, Y_1 sono indipendenti e regolate dalla stessa f.d.r. continua F. Segue che se $F = G$, allora $\mathrm{E}(T_X) = \mathrm{E}(U) + m(m+1)/2 = m(m+n+1)/2$. Conti simili, ma più tediosi, portano al seguente valore della varianza:

$$\mathrm{Var}(T_X) = \mathrm{Var}(U) = \frac{mn(m+n+1)}{12}$$

Se $F = G$ e $m, n \le 20$, i quantili w_a della f.d.r. di T_X[12] sono tabulati. Inoltre, siccome T ha densità simmetrica rispetto alla sua media quando $F = G$, allora fra i quantili w_a e w_{1-a} sussiste la seguente relazione

$$w_a = m(m+n+1) - w_{1-a}$$

[12]Le tavole furono calcolate per la prima volta da Mann e Whitney nel 1947 per $m, n \le 8$; mentre, la prima versione del test dovuta a Wilcoxon per il caso $m = n$ è del 1945.

Infine, se $F = G$ allora la statistica test T_X è asintoticamente gaussiana con media asintotica $m(m+n+1)/2$ e varianza asintotica $mn(m+n+1)/12$. Segue che per m, n grandi, un valore approssimato del quantile di T_X di ordine a è

$$w_a \simeq \frac{m(m+n+1)}{2} + z_a \sqrt{\frac{mn(m+n+1)}{12}}$$

Alla luce di quanto detto appaiono "sensate" le seguenti regole di significatività α per decidere sull'omogeneità:

Rifiuto $H_0 : F(x) = G(x) \, \forall x$ *e accetto* H_1 : "$F(x) \geq G(x) \, \forall x \in \mathbb{R}$ *e* $F(x) > G(x)$ *per qualche* x" *se* $T_X < w_\alpha$

Rifiuto $H_0 : F(x) = G(x) \, \forall x$ *e accetto* H_1 : "$F(x) \leq G(x) \, \forall x \in \mathbb{R}$ *e* $F(x) < G(x)$ *per qualche* x" *se* $T_X > w_{1-\alpha}$

Rifiuto $H_0 : F(x) = G(x) \, \forall x$ *e accetto* $H_1 : F(x) \neq G(x)$ *se* $T_X \notin [w_{\alpha/2}, w_{1-\alpha/2}]$

Il test descritto è noto come *test della somma dei ranghi di Wilcoxon-Mann-Whitney.*

Osservazione 4.2 Se lavoriamo con le somme dei ranghi delle y_j T_Y arriviamo a costruire lo stesso test, dal momento che *essenzialmente* non ci sono ripetizioni nel campione riunito e quindi $T_X + T_Y = (m+n)(m+n+1)/2$.

Esercizio 4.3 (continuazione dell'Esempio 4.1) Verifichiamo sulla base dei dati forniti nell'Esempio 4.1 l'ipotesi $H_0 : F = G$ contro l'alternativa $H_1 : F \neq G$, a livello $\alpha = 10\%$; $T_X = 21$ e, con $m = 4$ e $n = 5$, scopro che $w_{0.10/2} = w_{0.05} = 13$ e $w_{1-0.10/2} = w_{0.95} = m(m+n+1) - w_{0.05} = 40 - 13 = 27$. Essendo $13 < 21 < 27$ accetto H_0.

Osservazione 4.4 Il test della somma dei ranghi può essere letto come la versione non parametrica del t-test per confrontare medie di popolazioni gaussiani e indipendenti. Infatti, se i campioni $X_1, \ldots, X_m$ e $Y_1, \ldots, Y_n$ sono entrambi gaussiani e hanno la stessa varianza, allora $F = G$ se e solo se le medie sono uguali e possiamo studiare l'omogeneità con il t-test opportuno, a seconda della specificazione dell'ipotesi alternativa (15) o (16) o (17). Se $F = N(\mu_X, \sigma^2)$ e $G = N(\mu_Y, \sigma^2)$, è facile dimostrare che $F > G$ equivale a $\mu_X < \mu_Y$ e $F < G$ equivale a $\mu_X > \mu_Y$.

Vi sono varie ragioni per preferire il test di Wilcoxon-Mann-Whitney al t-test.

Un primo vantaggio del test di Wilcoxon-Mann-Whitney rispetto al t-test è che essendo un test non parametrico (free-distribution) il livello di significatività calcolato coincide con quello esatto, qualunque sia la distribuzione comune ai due campioni specificata dall'ipotesi H_0. Invece, per campioni non gaussiani, anche se numerosi, la significatività α del t-test asintotico calcolata è un'approssimazione di quella esatta e le due potrebbero differire in modo non trascurabile.

Un altro vantaggio del test di Wilcoxon-Mann-Whitney è la sua robustezza rispetto a valori *outliers*; di contro, ogni t-test è affetto dai valori outliers.

Ovviamente, se sapppiamo che i dati sono gaussiani il t-test è preferibile al test di Wilcoxon-Mann-Whitney, in quanto usa più informazioni (cioè usa la normalità dei dati).

Osservazione 4.5 (Test non parametrico sulla varianza) Esiste una variante del test della somma dei ranghi dovuta a Siegel e Tukey (1960) (per esempio in Conover 1999) che permette di confrontare le varianze di due popolazioni diverse e indipendenti, quando le medie sono eguali, in un contesto non parametrico. Siegel e Tukey cambiano la regola di assegnazione delle etichette nel seguente modo: un volta ordinate tutte le osservazioni dalla più piccola

alla più grande, assegnano etichetta 1 alla più piccola e 2 alla più grande, 3 alla penultima e 4 alla seconda, 5 alla terza e 6 alla terzultima e così via:

$$1,\ 4,\ 5,\ 8,\ 9,\ ,\ldots,(m+n-1),(m+n),\ldots\ 7,\ 6,\ 3,\ 2$$

Quindi contano la somma delle etichette di X. Se questa somma è piccola, vuol dire che X è molto più dispersa di Y e rifiutano $H_0 : \mathrm{Var}(X) = \mathrm{Var}(Y)$ a favore di $H_1 : \mathrm{Var}(X) > \mathrm{Var}(Y)$. Opportune tavole sono costruite. Il test di Siegel e Tukey è una sorta di variante non parametrica del test F di Fisher per il confronto di varianze di popolazioni gaussiane indipendenti.

4.2 Test dei segni di Wilcoxon per dati accoppiati

Estraiamo un campione casuale di dati accoppiati $(X_1, Y_1), \ldots, (X_n, Y_n)$ da una f.d.r. congiunta H con f.d.r. marginali F, G e tale che $P(X = Y) = 0$. Per esempio, se H è f.d.r. congiunta continua effettivamente $P(X = Y) = 0$. Contiamo poi il numero T^+ di coppie in cui la coordinata X è più grande di Y. Ci aspettiamo $T^+ \approx n/2$ se $F = G$, T^+ grande se $F < G$ e T^+ piccola se $F > G$. Poiché tutte le coppie di osservazioni sono i.i.d. *la statistica T^+ ha densità binomiale di parametri n e $p = P(X > Y)$*. Inoltre, sotto l'ipotesi nulla $H_0 : F = G$, $T^+ \sim Binom(n, 1/2)$ perché $P(X = Y) = 0$ per la continuità di F e $P(X > Y) = P(Y > X) = 1/2$.

Alla luce di quanto detto usiamo le seguenti regole di livello α per decidere su $F = G$ con dati accoppiati. Sia q_a^+ è il quantile di ordine a della f.d.r. Binomiale$(n, 1/2)$.

Rifiuto $H_0 : F(x) = G(x)\,\forall x$ e accetto H_1 : "$F(x) \geq G(x)\,\forall x \in \mathbb{R}$ e $F(x) > G(x)$ per qualche x" se $T^+ < q_\alpha^+$. Se $T^+ = k$ il p-value del test è dato da $\displaystyle\sum_{j=0}^{k-1} \binom{n}{j} \frac{1}{2^n}$

Rifiuto $H_0 : F(x) = G(x)\,\forall x$ e accetto H_1 : "$F(x) \leq G(x)\,\forall x \in \mathbb{R}$ e $F(x) < G(x)$ per qualche x" se $T^+ > q_{1-\alpha}^+$. Se $T^+ = k$ il p-value del test è dato da $1 - \displaystyle\sum_{j=0}^{k} \binom{n}{j} \frac{1}{2^n}$

Rifiuto $H_0 : F(x) = G(x)\,\forall x$ e accetto $H_1 : F(x) \neq G(x)$ se $T^+ \notin [q_{\alpha/2}^+, q_{1-\alpha/2}^+]$. Se $T^+ = k$ il p-value del test è dato da $2\min\{p_1, p_2\}$ *dove* $p_1 = \displaystyle\sum_{j=0}^{k-1} \binom{n}{j} \frac{1}{2^n}$ *e* $p_2 = 1 - p_1$.

Il test appena descritto è noto come *test dei segni di Wilcoxon*. Per n grande, in virtù del teorema centrale del limite, abbiamo $q_+(a) \simeq n/2 + z_a\sqrt{n}/2$.

Osservazione 4.6 Il test dei segni di Wilcoxon può essere letto come la versione non parametrica del t-test per confrontare medie di dati accoppiati gaussiani.

Riferimenti bibliografici

[1] Cifarelli, D.M. Conti, L., Regazzini, E. (1996) On the asymptotic distribution of a general measure of monotone dependence. *Ann. Statist.* **24**, 1386–1399

[2] Conover, W.J. (1999) *Practical Nonparametric Statistics 3ª Ed*, Wiley, New York

[3] Del Barrio, E. Cuesta-Albertos, J. A.; Matrán, C. (2000) Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. (With comments) *Test* **9**, 1–96

[4] Fisher, R. (1922) On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, A*, **222**, 309–368

[5] Fisher, R.A. (1924) The conditions under which χ^2 measures the discepancy between observation and hypohesis. *J. Roy. Statist. Soc.*, **87**, 442–450

[6] Mann, H.B. and Wald, A. (1942) On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math. Stat.*,**13**, 306–317

[7] Karl Pearson (1894) Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society A*, **185**, 71–110

[8] Pestman, Wiebe R. (1998) *Mathematical Statistics An Introduction* De Gruyter

[9] *R: A language and environment for statistical computing* R Development Core Team (2003) `http://www.R-project.org` , R Foundation for Statistical Computing Vienna, Austria

[10] Rohatgi, V.K e Saleh, A.K. Md. E. (1999) *An Introduction to Probability and Statistics* Wiley, New York

[11] Silvey, S.D (1975) *Statistical Inference* Chapman & Hall London