Elementi di statistica non parametrica

Lucio Barabesi

Indice

Capitolo 1

L'equivalenza in distribuzione

1.1. L'equivalenza in distribuzione. L'equivalenza in distribuzione è una particolare relazione di equivalenza tra variabili casuali.

Definizione 1.1.1. Le variabili casuali X e Y, con rispettive funzioni di ripartizione F e G, sono dette equivalenti in distribuzione se

$$F(x) = G(x) \, , \forall x \in \mathbb{R} \, .$$

In modo analogo, i vettori di variabili casuali $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$, con rispettive funzioni di ripartizione congiunte F_n e G_n, sono detti equivalenti in distribuzione se

$$F_n(x_1, \ldots, x_n) = G_n(x_1, \ldots, x_n) \, , \forall (x_1, \ldots, x_n) \in \mathbb{R}^n \, . \qquad \triangle$$

Per indicare che X e Y sono equivalenti in distribuzione si adotta la notazione

$$X \stackrel{d}{=} Y \, ,$$

mentre per indicare che $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ sono equivalenti in distribuzione si adotta la notazione

$$(X_1, \ldots, X_n) \stackrel{d}{=} (Y_1, \ldots, Y_n) \, .$$

L'equivalenza in distribuzione rappresenta in effetti una relazione di equivalenza, in quanto risulta riflessiva, ovvero

$$X \stackrel{d}{=} X \, ,$$

simmetrica, ovvero

$$X \stackrel{d}{=} Y \Leftrightarrow Y \stackrel{d}{=} X \, ,$$

e transitiva, ovvero

$$X \stackrel{d}{=} Y, Y \stackrel{d}{=} Z \Rightarrow X \stackrel{d}{=} Z \, .$$

• **Esempio 1.1.1.** Si consideri le variabili casuali X e Y, con rispettive funzioni di ripartizione F e G, dove $G(x) = F(x - \lambda)$ con $\lambda \in \mathbb{R}$. La trasformata $Z = Y - \lambda$ ha per funzione di ripartizione F, per cui dalla Definizione 1.1.1 risulta $X \stackrel{d}{=} Z$. Di conseguenza, si ha $X \stackrel{d}{=} Y - \lambda$. Alternativamente, si consideri le variabili casuali X e Y, con rispettive funzioni di ripartizione F e G, dove $G(x) = F(x/\delta)$ con $\delta \in \mathbb{R}^+$. La trasformata $Z = Y/\delta$ ha per funzione di ripartizione F, per cui dalla Definizione 1.1.1 risulta $X \stackrel{d}{=} Z$ e dunque $X \stackrel{d}{=} Y/\delta$. $\triangleleft$

• **Esempio 1.1.2.** Si consideri le variabili casuali X e Y equivalenti in distribuzione e sia F la loro funzione di ripartizione comune. Inoltre, sia data una trasformata misurabile U per cui $\mathrm{E}(U(X)) < \infty$. Dalla definizione di valor medio si ha

$$\mathrm{E}(U(X)) = \int_{\mathbb{R}} U(x)\, dF(x) = \int_{\mathbb{R}} U(y)\, dF(y) = \mathrm{E}(U(Y))\,,$$

dal momento che X e Y possiedono la stessa funzione di ripartizione. Dunque, si può concludere che se X e Y sono equivalenti in distribuzione allora possiedono i medesimi momenti. Analogamente, se $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ sono vettori di variabili casuali equivalenti in distribuzione e se $(U_1, \ldots, U_k)$ è un vettore di trasformate misurabili per cui $\mathrm{E}(U_j(X_1, \ldots, X_n)) < \infty$, allora risulta

$$\mathrm{E}(U_j(X_1, \ldots, X_n)) = \mathrm{E}(U_j(Y_1, \ldots, Y_n))\,,\ j = 1, \ldots, k\,.$$

Si noti infine che la proposizione inversa non è valida, ovvero possono esistere due variabili casuali X e Y che possiedono i medesimi momenti, ma che non sono equivalenti in distribuzione. ◁

• **Esempio 1.1.3.** Si consideri le variabili casuali X e Y equivalenti in distribuzione. Tenendo presente il risultato dell'Esempio 1.1.2, per un dato insieme misurabile $A \subset \mathbb{R}$ risulta

$$\mathrm{E}(\mathbf{1}_A(X)) = \mathrm{E}(\mathbf{1}_A(Y))\,.$$

Dal momento che $\mathrm{E}(\mathbf{1}_A(X)) = \Pr(X \in A)$ e $\mathrm{E}(\mathbf{1}_A(Y)) = \Pr(Y \in A)$, allora si ha

$$\Pr(X \in A) = \Pr(Y \in A)\,.$$ ◁

• **Esempio 1.1.4.** Si consideri i vettori di variabili casuali $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ equivalenti in distribuzione. Se $A = \{(x_1, \ldots, x_n) : (x_1, \ldots, x_n) \in \mathbb{R}^n,\ x_1 < \ldots < x_n\}$, tenendo presente il risultato dell'Esempio 1.1.3, risulta

$$\mathrm{E}(\mathbf{1}_A(X_1, \ldots, X_n)) = \mathrm{E}(\mathbf{1}_A(Y_1, \ldots, Y_n))\,.$$

Si ha $\mathrm{E}(\mathbf{1}_A(X_1, \ldots, X_n)) = \Pr(X_1 < \ldots < X_n)$ e $\mathrm{E}(\mathbf{1}_A(Y_1, \ldots, Y_n)) = \Pr(Y_1 < \ldots < Y_n)$ e si deve concludere che

$$\Pr(X_1 < \ldots < X_n) = \Pr(Y_1 < \ldots < Y_n)\,.$$ ◁

• **Esempio 1.1.5.** Si consideri i vettori di variabili casuali $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$, tali che $X_i \stackrel{d}{=} Y_i$ per $i = 1, \ldots, n$. Si noti che non si può affermare che $(X_1, \ldots, X_n)$ è equivalente in distribuzione a $(Y_1, \ldots, Y_n)$, come si potrebbe ingenuamente concludere in un primo momento. Infatti, $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ possiedono in generale funzioni di ripartizione congiunte differenti, anche se con identiche funzioni di ripartizione marginali. ◁

Il seguente teorema estende l'equivalenza in distribuzione a trasformate, nel senso che l'equivalenza rimane valida applicando una trasformata misurabile ad ambo i membri della stessa.

Teorema 1.1.2. *Se X e Y sono variabili casuali equivalenti in distribuzione e se U è una trasformata misurabile, allora*

$$U(X) \stackrel{d}{=} U(Y)\,.$$

Inoltre, se $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ sono vettori di variabili casuali equivalenti in distribuzione e se $(U_1, \ldots, U_k)$ è un vettore di trasformate misurabili, allora

$$(U_1(X_1, \ldots, X_n), \ldots, U_k(X_1, \ldots, X_n)) \stackrel{d}{=} (U_1(Y_1, \ldots, Y_n), \ldots, U_k(Y_1, \ldots, Y_n))\,.$$

Dimostrazione. Se F è la funzione di ripartizione comune alle variabili casuali X e Y e se A è un insieme misurabile, allora

$$\Pr(U(X) \in A) = \int_{\mathbb{R}} \mathbf{1}_A(U(x))\, dF(x) = \int_{\mathbb{R}} \mathbf{1}_A(U(y))\, dF(y) = \Pr(U(Y) \in A)\,.$$

Dal momento che la precedente relazione è vera per ogni insieme misurabile A, dalla Definizione 1.1.1 risulta $U(X) \stackrel{d}{=} U(Y)$. Risulta analoga la dimostrazione nel caso di due vettori di variabili casuali. □

• **Esempio 1.1.6.** Se X e Y sono variabili casuali equivalenti in distribuzione, si consideri la trasformata $U(x) = (x - \lambda)/\delta$, con λ e δ costanti. Dal Teorema 1.1.2 si ha

$$\frac{X - \lambda}{\delta} \stackrel{d}{=} \frac{Y - \lambda}{\delta}\,,$$

ovvero in una equivalenza in distribuzione è possibile aggiungere o moltiplicare i membri per una costante.◁

• **Esempio 1.1.7.** Il risultato dell'Esempio 1.1.6 non rimane valido in generale se si aggiunge o si moltiplica i membri per una quantità stocastica. Si consideri infatti una variabile casuale X assolutamente continua con funzione di ripartizione F e tale che $X \stackrel{d}{=} -X$. Se si moltiplica ambo i membri dell'equivalenza per X si dovrebbe concludere che le variabili casuali $Y = X^2$ e $Z = -X^2$ sono equivalenti in distribuzione. Questa affermazione è falsa in quanto Y è ovviamente una variabile casuale il cui supporto è contenuto in $\mathbb{R}^+$, mentre Z è una variabile casuale il cui supporto è contenuto in $\mathbb{R}^-$. Infatti la funzione di ripartizione di Y risulta

$$G(y) = \Pr(Y \le y) = \Pr(X^2 \le y) = (F(\sqrt{y}) - F(-\sqrt{y}))\, \mathbf{1}_{[0,\infty)}(y)\,,$$

mentre la funzione di ripartizione di Z risulta

$$H(z) = \Pr(Z \le z) = \Pr(-X^2 \le z) = (1 - F(\sqrt{-z}) + F(-\sqrt{-z}))\, \mathbf{1}_{(-\infty,0)}(z) + \mathbf{1}_{[0,\infty)}(z)\,.$$

Questo esempio serve a sottolineare che si deve porre una certa cautela nell'applicare la nozione di equivalenza in distribuzione. In particolare, il Teorema 1.1.2 rimane valido solo se si considera trasformate misurabili. ◁

• **Esempio 1.1.8.** Se $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ sono vettori di variabili casuali equivalenti in distribuzione, allora si consideri il vettore di trasformate $(U_1, \ldots, U_n)$ tale che $U_i(x_1, \ldots, x_n) = x_i$ per $i = 1, \ldots, n$. Dal Teorema 1.1.2 risulta $X_i \stackrel{d}{=} Y_i$ per $i = 1, \ldots, n$, ovvero si deve concludere che se due vettori di variabili casuali sono equivalenti in distribuzione allora anche le singole componenti dei vettori sono ordinatamente equivalenti in distribuzione. ◁

Il seguente teorema consente di ottenere una relazione di equivalenza in distribuzione per un vettore di variabili casuali indipendenti e ugualmente distribuite.

Teorema 1.1.3. *Se $(X_1, \ldots, X_n)$ è un vettore di variabili casuali indipendenti e ugualmente distribuite e $(\alpha_1, \ldots, \alpha_n)$ è una qualsiasi permutazione di $(1, \ldots, n)$, allora*

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\alpha_1}, \ldots, X_{\alpha_n})\,.$$

Dimostrazione. Se F è la funzione di ripartizione marginale di X_i per $i = 1, \ldots, n$, allora la funzione di ripartizione congiunta di $(X_1, \ldots, X_n)$ è data da

$$F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i)\,.$$

Dal momento che la funzione di ripartizione congiunta di $(X_{\alpha_1}, \ldots, X_{\alpha_n})$ risulta

$$G_n(x_{\alpha_1},\dots,x_{\alpha_n}) = \prod_{i=1}^{n} F(x_{\alpha_i}) = \prod_{i=1}^{n} F(x_i) = F_n(x_1,\dots,x_n)\,,$$

allora si ha $G_n(x_1,\dots,x_n) = F_n(x_1,\dots,x_n)$ per ogni $(x_1,\dots,x_n) \in \mathbb{R}^n$, e dalla Definizione 1.1.1 si conclude che $(X_1,\dots,X_n) \stackrel{d}{=} (X_{\alpha_1},\dots,X_{\alpha_n})$. □

1.2. L'equivalenza in distribuzione e le variabili casuali simmetriche. In questa sezione vengono introdotte alcune relazioni di equivalenza in distribuzione per variabili casuali simmetriche. A questo fine si consideri innanzitutto la definizione di variabile casuale simmetrica.

Definizione 1.2.1. Una variabile casuale X con funzione di ripartizione F è detta simmetrica rispetto ad una costante λ se

$$F(\lambda + x) = 1 - F(\lambda - x) + \Pr(X = \lambda - x)\,, \forall x \in \mathbb{R}\,. \qquad \triangle$$

• **Esempio 1.2.1.** Se X è una variabile casuale assolutamente continua, simmetrica rispetto a λ, e con funzione di ripartizione F e funzione di densità f, allora dalla Definizione 1.2.1 risulta

$$F(\lambda + x) = 1 - F(\lambda - x)\,, \forall x \in \mathbb{R}\,,$$

da cui si ha inoltre

$$f(\lambda + x) = f(\lambda - x)\,, \forall x \in \mathbb{R}\,.$$

Se invece X è una variabile casuale discreta, simmetrica rispetto a λ, con funzione di ripartizione F e funzione di probabilità p, allora dalla Definizione 1.2.1 risulta

$$F(\lambda + x) = 1 - F(\lambda - x) + p(\lambda - x)\,, \forall x \in \mathbb{R}\,,$$

da cui si ottiene inoltre

$$p(\lambda + x) = p(\lambda - x)\,, \forall x \in \mathbb{R}\,. \qquad \triangleleft$$

Nel seguente teorema si ottiene una condizione necessaria e sufficiente per la simmetria rispetto ad una costante.

Teorema 1.2.2. *La variabile casuale X ha una distribuzione simmetrica rispetto alla costante λ se e solo se*

$$X - \lambda \stackrel{d}{=} \lambda - X\,.$$

Dimostrazione. Si dimostra innanzitutto che la condizione è necessaria. Se X è simmetrica rispetto a λ con funzione di ripartizione F, allora la funzione di ripartizione della trasformata $Y = X - \lambda$ è data da

$$G(y) = \Pr(Y \le y) = \Pr(X - \lambda \le y) = \Pr(X \le \lambda + y) = F(\lambda + y)\,,$$

mentre, tenendo presente la definizione di variabile casuale simmetrica, la funzione di ripartizione della trasformata $Z = \lambda - X$ risulta

$$\begin{aligned} H(z) &= \Pr(Z \le z) = \Pr(\lambda - X \le z) = \Pr(X \ge \lambda - z) \\ &= 1 - F(\lambda - z) + \Pr(X = \lambda - z) = F(\lambda + z)\,. \end{aligned}$$

Dal momento che Y e Z possiedono la medesima funzione di ripartizione, dalla Definizione 1.1.1 si ha $Y \stackrel{d}{=} Z$, ovvero $X - \lambda \stackrel{d}{=} \lambda - X$. Si dimostra che la condizione è sufficiente. Se $X - \lambda \stackrel{d}{=} \lambda - X$, dalla definizione di equivalenza in distribuzione risulta

$$\Pr(X - \lambda \le x) = \Pr(\lambda - X \le x)\,.$$

Tenendo presente la precedente relazione si ha

$$\begin{aligned} F(\lambda + x) &= \Pr(X \leq \lambda + x) = \Pr(X - \lambda \leq x) = \Pr(\lambda - X \leq x) \\ &= \Pr(X \geq \lambda - x) = 1 - F(\lambda - x) + \Pr(X = \lambda - x) \,, \end{aligned}$$

ovvero X è simmetrica rispetto a λ. □

• **Esempio 1.2.2.** Sia (X_1, X_2) un vettore di variabili casuali tale che $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$ e si consideri la trasformata $U(x_1, x_2) = x_1 - x_2$. Dal Teorema 1.1.2 si ottiene dunque

$$X_1 - X_2 \stackrel{d}{=} X_2 - X_1 \,,$$

ovvero

$$X_1 - X_2 \stackrel{d}{=} -(X_1 - X_2) \,.$$

Dunque, per il Teorema 1.2.2 la trasformata $(X_1 - X_2)$ è simmetrica rispetto a 0. ◁

• **Esempio 1.2.3.** Sia X una variabile casuale simmetrica rispetto a λ con $\mathrm{E}(X) < \infty$. Dal Teorema 1.2.2 si ha $X - \lambda \stackrel{d}{=} \lambda - X$. Dal momento che variabili casuali equivalenti in distribuzione hanno la stessa media (vedi Esempio 1.1.2), allora

$$\mathrm{E}(X - \lambda) = \mathrm{E}(\lambda - X) \,,$$

ovvero

$$\mathrm{E}(X) - \lambda = \lambda - \mathrm{E}(X) \,,$$

da cui infine risulta $\mathrm{E}(X) = \lambda$. ◁

Il seguente teorema considera un insieme di equivalenze in distribuzione per un vettore di variabili casuali le cui componenti sono indipendenti e simmetriche.

Teorema 1.2.3. *Se $(X_1, \ldots, X_n)$ è un vettore di variabili casuali indipendenti tali che X_i è simmetrica rispetto a λ_i per $i = 1, \ldots, n$, allora*

$$(X_1 - \lambda_1, \ldots, X_n - \lambda_n) \stackrel{d}{=} (\lambda_1 - X_1, \ldots, X_n - \lambda_n) \stackrel{d}{=} \ldots \stackrel{d}{=} (\lambda_1 - X_1, \ldots, \lambda_n - X_n) \,,$$

dove l'equivalenza in distribuzione è estesa a tutte le possibili 2^n configurazioni di vettori.

Dimostrazione. Si dimostra la prima delle equivalenze in distribuzione. Per il Teorema 1.2.2 la simmetria di X_i rispetto a λ_i implica $X_i - \lambda_i \stackrel{d}{=} \lambda_i - X_i$, ovvero dalla definizione di equivalenza in distribuzione si ottiene

$$\Pr(X_i - \lambda_i \leq x) = \Pr(\lambda_i - X_i \leq x) \,, i = 1, \ldots, n \,.$$

Data l'indipendenza delle componenti di $(X_1, \ldots, X_n)$, risulta dunque

$$\begin{aligned} \Pr(X_1 - \lambda_1 \leq x_1, \ldots, X_n - \lambda_n \leq x_n) &= \prod_{i=1}^{n} \Pr(X_i - \lambda_i \leq x_i) = \Pr(\lambda_1 - X_1 \leq x_1) \prod_{i=2}^{n} \Pr(X_i - \lambda_i \leq x_i) \\ &= \Pr(\lambda_1 - X_1 \leq x_1, \ldots, X_n - \lambda_n \leq x_n) \,, \end{aligned}$$

ovvero $(X_1 - \lambda_1, \ldots, X_n - \lambda_n) \stackrel{d}{=} (\lambda_1 - X_1, \ldots, X_n - \lambda_n)$. In modo analogo si ottiene la dimostrazione delle altre equivalenze in distribuzione. □

• **Esempio 1.2.4.** Sia $(X_1, \ldots, X_n)$ un campione casuale da una variabile casuale X simmetrica rispetto a λ. Dal Teorema 1.2.3 risulta dunque $(X_1 - \lambda, \ldots, X_n - \lambda) \stackrel{d}{=} (\lambda - X_1, \ldots, \lambda - X_n)$. Inoltre, se si considera la trasformata $U(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^{n} x_i$, dal Teorema 1.1.2 si ha

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\lambda)\stackrel{d}{=}\frac{1}{n}\sum_{i=1}^{n}(\lambda-X_i)\,,$$

ovvero

$$\bar{X}-\lambda\stackrel{d}{=}\lambda-\bar{X}\,,$$

dove $\bar{X}=n^{-1}\sum_{i=1}^{n}X_i$ è la media campionaria. Tenendo presente il Teorema 1.2.2, la precedente equivalenza in distribuzione porta a concludere che la media campionaria è simmetrica rispetto a λ. ◁

• **Esempio 1.2.5.** Sia $(X_1,\ldots,X_n)$ un campione casuale da una variabile casuale X continua e simmetrica rispetto a 0. Dal Teorema 1.2.3 risulta

$$(X_1,\ldots,X_n)\stackrel{d}{=}(-X_1,\ldots,-X_n)\,.$$

Si consideri il vettore di trasformate $(U_1,\ldots,U_n)$ tale che $U_i(x_1,\ldots,x_n)=x_{(i)}$, dove $x_{(i)}$ rappresenta l'i-esimo elemento ordinato nel vettore $(x_1,\ldots,x_n)$ per $i=1,\ldots,n$. Si ha $U_i(-x_1,\ldots,-x_n)=-x_{(n-i+1)}$ per $i=1,\ldots,n$, in quanto cambiando il segno agli elementi del vettore $(x_1,\ldots,x_n)$ se ne inverte l'ordinamento. Dal Teorema 1.1.2 si ottiene dunque

$$(X_{(1)},\ldots,X_{(n)})\stackrel{d}{=}(-X_{(n)},\ldots,-X_{(1)})\,,$$

dove $X_{(i)}$ rappresenta l'i-esima statistica ordinata per $i=1,\ldots,n$. In particolare, tenendo presente l'Esempio 1.1.8, si ha $X_{(i)}\stackrel{d}{=}-X_{(n-i+1)}$ per $i=1,\ldots,n$. Risulta analogo verificare che, se $(X_1,\ldots,X_n)$ è un campione casuale da una variabile casuale X assolutamente continua e simmetrica rispetto a λ, allora

$$(X_{(1)}-\lambda,\ldots,X_{(n)}-\lambda)\stackrel{d}{=}(\lambda-X_{(n)},\ldots,\lambda-X_{(1)})\,,$$

che implica $X_{(i)}-\lambda\stackrel{d}{=}\lambda-X_{(n-i+1)}$ per $i=1,\ldots,n$. ◁

• **Esempio 1.2.6.** Sia $(X_1,\ldots,X_n)$ un campione casuale da una variabile casuale X assolutamente continua e simmetrica rispetto a 0. Tenendo presente l'Esempio 1.2.5, si applichi all'equivalenza in distribuzione $(X_{(1)},\ldots,X_{(n)})\stackrel{d}{=}(-X_{(n)},\ldots,-X_{(1)})$ la trasformata U, tale che $U(x_1,\ldots,x_n)=x_{(l)}$ con $l=(n+1)/2$ se n è dispari e $U(x_1,\ldots,x_n)=(x_{(l)}+x_{(l+1)})/2$ con $l=n/2$ se n è pari. Dal Teorema 1.1.2 per n dispari si ha dunque

$$X_{(l)}\stackrel{d}{=}-X_{(l)}\,,$$

mentre per n pari si ha

$$\frac{1}{2}\left(X_{(l)}+X_{(l+1)}\right)\stackrel{d}{=}-\frac{1}{2}\left(X_{(l)}+X_{(l+1)}\right).$$

Dal momento che la mediana campionaria è usualmente definita come $\widetilde{X}=X_{(l)}$ con $l=(n+1)/2$ se n è dispari ed è definita come $\widetilde{X}=(X_{(l)}+X_{(l+1)})/2$ con $l=n/2$ se n è pari, allora risulta

$$\widetilde{X}\stackrel{d}{=}-\widetilde{X}\,,$$

ovvero, tenendo presente il Teorema 1.2.2, si deve concludere che in questo caso la mediana campionaria è simmetrica rispetto a 0. Risulta analogo verificare che se $(X_1,\ldots,X_n)$ è un campione casuale da una variabile casuale X assolutamente continua e simmetrica rispetto a λ, allora la mediana campionaria è simmetrica rispetto a λ. ◁

Capitolo 2

Le statistiche “distribution-free”

2.1. I modelli statistici “distribution-free”. Nella statistica inferenziale classica si assume nota la morfologia funzionale delle funzioni di ripartizione congiunte specificate dal modello statistico, e in particolare si assume che queste siano dello stesso tipo a meno di un insieme di parametri. Invece, nella statistica moderna si tende a non fare assunzioni funzionali sulla distribuzione congiunta del campione, ovvero si considera i cosiddetti modelli statistici “distribution-free” (una definizione anglosassone ormai consolidata anche nella terminologia statistica italiana). Un modello statistico “distribution-free” ha la seguente definizione formale.

Definizione 2.1.1. Si consideri il campione $(X_1, \ldots, X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}$, dove $\mathcal{F}$ è una classe di funzioni di ripartizione congiunte, ovvero un modello statistico. Se $\mathcal{F}$ contiene più di una famiglia di funzioni di ripartizione congiunte, allora è detto modello statistico “distribution-free”. $\triangle$

• **Esempio 2.1.1.** Il modello statistico

$$\mathcal{F} = \{F_n : F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} \Phi((x_i - \mu)/\sigma), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$$

non è “distribution-free”, in quanto contiene una sola famiglia di funzioni di ripartizione congiunte, ovvero le funzioni di ripartizione congiunte di un campione casuale da una distribuzione $N(\mu, \sigma^2)$. Il precedente modello statistico è tipico nell'inferenza classica. Invece, se $\mathcal{C}_n$ rappresenta la classe delle funzioni di ripartizione di un vettore di n variabili casuali assolutamente continue, allora

$$\mathcal{F} = \{F_n : F_n \in \mathcal{C}_n\}$$

costituisce un modello statistico “distribution-free”. $\triangleleft$

Frequentemente il modello statistico è del tipo $\mathcal{F}_\psi$, ovvero è indicizzato mediante un insieme di “parametri” ψ (non necessariamente reali). Importanti modelli statistici “distribution-free” possono essere indicizzati in questa maniera. Ad esempio nel seguito sarà fatto riferimento al modello statistico “distribution-free”

$$\mathcal{C}_F = \{F_n : F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i), F \in \mathcal{C}\}\ ,$$

dove $\mathcal{C}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua. Quindi, $\mathcal{C}_F$ rappresenta il modello statistico relativo ad un campione casuale proveniente da una variabile casuale assolutamente continua. Una importante sottoclasse di $\mathcal{C}_F$ è data dal modello statistico “distribution-free”

$$\mathcal{M}_{\lambda,F} = \{F_n : F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i - \lambda), F \in \mathcal{M}, \lambda \in \mathbb{R}\}\ ,$$

dove $\mathcal{M}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua con mediana pari a 0, ovvero

$$\mathcal{M} = \{F : F \in \mathcal{C}, F(0) = 1/2\} .$$

Dunque, $\mathcal{M}_{\lambda,F}$ rappresenta il modello statistico relativo ad un campione casuale proveniente da una variabile casuale assolutamente continua con funzione di ripartizione $F(x-\lambda)$ e con mediana pari a λ. Una sottoclasse di $\mathcal{M}_{\lambda,F}$ è data dal modello statistico "distribution-free"

$$\mathcal{S}_{\lambda,F} = \{F_n : F_n(x_1,\ldots,x_n) = \prod_{i=1}^{n} F(x_i-\lambda), F \in \mathcal{S}, \lambda \in \mathbb{R}\} ,$$

dove $\mathcal{S}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua e simmetrica rispetto a 0, ovvero

$$\mathcal{S} = \{F : F \in \mathcal{C}, F(x) = 1 - F(-x)\} .$$

Dunque, $\mathcal{S}_{\lambda,F}$ rappresenta il modello statistico relativo ad un campione casuale proveniente da una variabile casuale assolutamente continua con funzione di ripartizione $F(x-\lambda)$ e simmetrica rispetto a λ. Due ulteriori sottoclassi di $\mathcal{C}_F$ di uso frequente sono rappresentati dai modelli statistici "distribution-free"

$$\mathcal{L}_{\Delta,F} = \{F_n : F_n(x_1,\ldots,x_n) = \prod_{i=1}^{n_1} F(x_i) \prod_{i=n_1+1}^{n} F(x_i-\Delta), F \in \mathcal{C}, \Delta \in \mathbb{R}\} ,$$

e

$$\mathcal{V}_{\eta,F} = \{F_n : F_n(x_1,\ldots,x_n) = \prod_{i=1}^{n_1} F(x_i) \prod_{i=n_1+1}^{n} F(x_i/\eta), F \in \mathcal{C},\ \eta \in \mathbb{R}^+\} ,$$

dove $1 \leq n_1 \leq n$.

2.2. Le statistiche "distribution-free". Le statistiche "distribution-free" hanno la seguente definizione formale.

Definizione 2.2.1. Si consideri il campione $(X_1,\ldots,X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}$, dove $\mathcal{F}$ è un modello statistico. La statistica $T = T(X_1,\ldots,X_n)$ è detta "distribution-free" su $\mathcal{F}$ se la corrispondente funzione di ripartizione rimane invariata per ogni $F_n \in \mathcal{F}$. △

• **Esempio 2.2.1.** Si consideri un campione casuale $(X_1,\ldots,X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\sigma$, dove

$$\mathcal{F}_\sigma = \{F_n : F_n(x_1,\ldots,x_n) = \prod_{i=1}^{n} \Phi(x_i/\sigma),\ \sigma \in \mathbb{R}^+\} .$$

Il modello statistico $\mathcal{F}_\sigma$ è quello relativo ad un campione casuale da una distribuzione Normale $N(0,\sigma^2)$. Se $\bar{X}$ e S^2 rappresentano rispettivamente la media campionaria e la varianza campionaria corretta, allora la statistica $T = \sqrt{n}\bar{X}/S$ è distribuita come una t di Student con $(n-1)$ gradi di libertà. Dal momento che la distribuzione di T non dipende dal parametro σ, allora T è una statistica "distribution-free" su $\mathcal{F}_\sigma$. ◁

L'Esempio 2.2.1 evidenzia che una tipica statistica dell'inferenza classica può essere considerata "distribution-free" su un particolare modello statistico classico. Tuttavia, usualmente una statistica è detta "distribution-free" quando il modello statistico $\mathcal{F}$ è anch'esso "distribution-free". Quando si dispone di campioni di numerosità elevata si considerano anche statistiche "distribution-free" per grandi campioni, che sono definite formalmente di seguito.

Definizione 2.2.2. Si consideri il campione $(X_1,\ldots,X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}$, dove $\mathcal{F}$ è un modello statistico per ogni n. La statistica $T = T_n = T_n(X_1,\ldots,X_n)$ è detta "distribution-

free" per grandi campioni su $\mathcal{F}$, se per $n \to \infty$ risulta $T_n \xrightarrow{d} V$ per ogni $F_n \in \mathcal{F}$, dove V è una variabile casuale limite. △

• **Esempio 2.2.2.** Si consideri il modello statistico "distribution-free"

$$\mathcal{F}_{\sigma,F} = \{F_n : F_n \in \mathcal{C}_F, \mathrm{E}(X) = 0, \mathrm{Var}(X) = \sigma^2 < \infty\} .$$

Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\sigma,F}$, si consideri la statistica $T = T_n = \sqrt{n}\bar{X}/S$ dell'Esempio 2.2.1. La statistica T_n può essere espressa come

$$T_n = \frac{\bar{X}}{\sigma/\sqrt{n}} \frac{\sigma}{S} .$$

Dal momento che $S^2 \xrightarrow{p} \sigma^2$ per $n \to \infty$ (vedi Esempio A.3.4), per il Teorema di Sverdrup (Teorema A.3.4) si ha $S \xrightarrow{p} \sigma$ per $n \to \infty$. Inoltre, per il Teorema Fondamentale Classico del Limite (Teorema A.3.6) si ha $\sqrt{n}\bar{X}/\sigma \xrightarrow{d} N(0,1)$. Combinando questi risultati mediante il Teorema di Slutsky (Teorema A.3.5) si ottiene infine

$$T_n = \frac{\bar{X}}{S/\sqrt{n}} \xrightarrow{d} N(0,1) .$$

Si deve dunque concludere che la statistica T_n è "distribution-free" per grandi campioni su $\mathcal{F}_{\sigma,F}$. ◁

2.3. Le statistiche segno. In questa sezione viene introdotta una prima classe di statistiche "distribution-free", ovvero le cosiddette statistiche segno.

Definizione 2.3.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{0,F}$. Si definiscono statistiche segno le trasformate

$$Z_i = \mathbf{1}_{(0,\infty)}(X_i) , i = 1, \ldots, n .$$

Il vettore di statistiche $(Z_1, \ldots, Z_n)$ è detto vettore dei segni. △

Ogni statistica segno Z_i assume valore 1 se $X_i > 0$ (ovvero se X_i è maggiore della mediana) e il valore 0 altrimenti, e da questo fatto deriva la loro denominazione. Il seguente teorema fornisce la distribuzione congiunta delle statistiche segno.

Teorema 2.3.2. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{0,F}$, il relativo vettore dei segni $(Z_1, \ldots, Z_n)$ ha componenti indipendenti ed ugualmente distribuite come variabili casuali Binomiali $Bi(1, 1/2)$.*

Dimostrazione. Dal momento che dalla Definizione 2.3.1 ogni statistica segno Z_i è trasformata solo della relativa X_i, e poichè il campione casuale $(X_1, \ldots, X_n)$ ha componenti indipendenti, allora anche le componenti del vettore dei segni $(Z_1, \ldots, Z_n)$ sono indipendenti. Inoltre, il supporto della statistica segno Z_i è l'insieme $\{0, 1\}$. Dalla assunzione di continuità per X_i si ha

$$\Pr(Z_i = 1) = \Pr(X_i > 0) = 1 - F(0) = \frac{1}{2} , i = 1, \ldots, n ,$$

e di conseguenza

$$\Pr(Z_i = 0) = 1 - \Pr(Z_i = 1) = \frac{1}{2} , i = 1, \ldots, n .$$

Dunque, si deve concludere che la statistica segno Z_i ha distribuzione Binomiale $Bi(1, 1/2)$. □

Corollario 2.3.3. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{0,F}$. Se $T = T(Z_1, \ldots, Z_n)$, ovvero T è una statistica basata solo sul vettore dei segni, allora T è "distribution-free" su $\mathcal{M}_{0,F}$.*

Dimostrazione. Il risultato segue dal Teorema 2.3.2 e dalla Definizione 2.3.1, in quanto per qualsiasi distribuzione congiunta $F_n \in \mathcal{M}_{0,F}$ la statistica T è distribuita come una trasformata di un vettore di variabili casuali indipendenti ed ugualmente distribuite, ognuna con distribuzione Binomiale $Bi(1, 1/2)$. □

• **Esempio 2.3.1.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{0,F}$. Si consideri la statistica

$$B = \sum_{i=1}^{n} Z_i \, ,$$

che rappresenta il numero di osservazioni positive nel campione. Utilizzando il Teorema 2.3.2 si verifica che B ha distribuzione Binomiale $Bi(n, 1/2)$. Di conseguenza, poichè la funzione di ripartizione di B rimane invariata per ogni funzione di ripartizione congiunta di $\mathcal{M}_{0,F}$, allora B è una statistica "distribution-free" su $\mathcal{M}_{0,F}$. Questo risultato può essere ottenuto dal Corollario 2.3.3. ◁

2.4. Le statistiche rango. La classe delle statistiche rango è fondamentale per costruire statistiche "distribution-free". La seguente è la definizione formale di statistica rango.

Definizione 2.4.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Si definiscono statistiche rango le seguenti trasformate

$$R_i = \sum_{j=1}^{n} \mathbf{1}_{[0,\infty)}(X_i - X_j) \, , \, i = 1, \ldots, n \, .$$

Il vettore di statistiche $(R_1, \ldots, R_n)$ è detto vettore dei ranghi. △

La statistica R_i rappresenta la posizione di X_i all'interno del campione ordinato, ovvero

$$X_i = X_{(R_i)} \, , \, i = 1, \ldots, n \, ,$$

e da questo deriva ovviamente la denominazione di statistiche rango.

Definizione 2.4.2. Si definisce insieme delle permutazioni $\mathcal{R}_n \subset \mathbb{R}^n$ dato da

$$\mathcal{R}_n = \{(r_1, \ldots, r_n) : (r_1, \ldots, r_n) \text{ è una permutazione di } (1, \ldots, n)\} \, . \qquad \triangle$$

Nel seguente teorema viene ottenuta la distribuzione congiunta del vettore dei ranghi.

Teorema 2.4.3. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, allora la funzione di probabilità congiunta del relativo vettore dei ranghi $(R_1, \ldots, R_n)$ è data da*

$$\Pr(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{n!} \, \mathbf{1}_{\mathcal{R}_n}(r_1, \ldots, r_n) \, .$$

Dimostrazione. Il vettore di variabili casuali $(R_1, \ldots, R_n)$ è discreto con supporto $\mathcal{R}_n$. Per un prefissato $(r_1, \ldots, r_n) \in \mathcal{R}_n$ risulta

$$\Pr(R_1 = r_1, \ldots, R_n = r_n) = \Pr(X_1 = X_{(r_1)}, \ldots, X_n = X_{(r_n)}) = \Pr(X_{d_1} < \ldots < X_{d_n}) \, ,$$

dove d_i è la posizione del numero i nella permutazione $(r_1, \ldots, r_n)$. Dal Teorema 1.1.3 si ha inoltre $(X_1, \ldots, X_n) \stackrel{d}{=} (X_{d_1}, \ldots, X_{d_n})$, che implica (vedi Esempio 1.1.4)

$$\Pr(X_{d_1} < \ldots < X_{d_n}) = \Pr(X_1 < \ldots < X_n) \, .$$

Dunque, risulta

$$\Pr(R_1 = r_1, \ldots, R_n = r_n) = \Pr(X_1 < \ldots < X_n) = \Pr(R_1 = 1, \ldots, R_n = n) \, .$$

Inoltre, dal momento che $\#(\mathcal{R}_n) = n!$, si ha

$$\begin{aligned} \sum_{(r_1,\ldots,r_n)\in\mathcal{R}_n} \Pr(R_1 = r_1, \ldots, R_n = r_n) &= \sum_{(r_1,\ldots,r_n)\in\mathcal{R}_n} \Pr(R_1 = 1, \ldots, R_n = n) \\ &= n! \Pr(R_1 = 1, \ldots, R_n = n) \, . \end{aligned}$$

Infine, essendo

$$\sum_{(r_1,\ldots,r_n)\in\mathcal{R}_n} \Pr(R_1 = r_1, \ldots, R_n = r_n) = 1 \, ,$$

allora segue

$$\Pr(R_1 = 1, \ldots, R_n = n) = \frac{1}{n!} \, .$$

Si deve concludere dunque che

$$\Pr(R_1 = r_1, \ldots, R_n = r_n) = \Pr(R_1 = 1, \ldots, R_n = n) = \frac{1}{n!} \, , (r_1, \ldots, r_n) \in \mathcal{R}_n \, ,$$

ovvero il vettore dei ranghi $(R_1, \ldots, R_n)$ è distribuito uniformemente su $\mathcal{R}_n$. □

Dal Teorema 2.4.3 è possibile ricavare i seguenti quattro corollari.

Corollario 2.4.4. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. La funzione di probabilità della statistica rango R_i, per $i = 1, \ldots, n$, risulta*

$$\Pr(R_i = r) = \frac{1}{n} \, , r = 1, \ldots, n \, ,$$

la funzione di probabilità congiunta di una scelta di due statistiche rango (R_i, R_j) per $i \neq j = 1, \ldots, n$, risulta

$$\Pr(R_i = r_1, R_j = r_2) = \frac{1}{n(n-1)} \, , r_1 \neq r_2 = 1, \ldots, n \, ,$$

mentre la funzione di probabilità congiunta di una scelta di $k = 1, \ldots, n$ statistiche rango $(R_{i_1}, \ldots, R_{i_k})$ per $i_1 \neq \ldots \neq i_k = 1, \ldots, n$, risulta

$$\Pr(R_{i_1} = r_1, \ldots, R_{i_k} = r_k) = \frac{1}{n(n-1)\cdots(n-k+1)} \, , r_1 \neq \ldots \neq r_k = 1, \ldots, n \, .$$

Dimostrazione. Dal Teorema 2.4.3 si ha che ogni determinazione del vettore $(R_1, \ldots, R_n)$ in $\mathcal{R}_n$ è ugualmente probabile. Inoltre, vi sono $(n-1)!$ elementi di $\mathcal{R}_n$ per cui $R_i = r$ con $r = 1, \ldots, n$, e dunque si ha

$$\Pr(R_i = r) = (n-1)! \, \frac{1}{n!} = \frac{1}{n} \, .$$

Analogamente, vi sono $(n-2)!$ elementi di $\mathcal{R}_n$ per cui $R_i = r_1$ e $R_j = r_2$ con $r_1 \neq r_2 = 1, \ldots, n$, e quindi

$$\Pr(R_i = r_1, R_j = r_2) = (n-2)! \, \frac{1}{n!} = \frac{1}{n(n-1)} \, .$$

Infine, vi sono $(n-k)!$ elementi di $\mathcal{R}_n$ per cui $R_{i_1} = r_1, \ldots, R_{i_k} = r_k$ con $r_1 \neq \ldots \neq r_k = 1, \ldots, n$, e quindi

$$\Pr(R_{i_1} = r_1, \ldots, R_{i_k} = r_k) = (n-k)!\,\frac{1}{n!} = \frac{1}{n(n-1)\ldots(n-k+1)}\,. \qquad \square$$

Corollario 2.4.5. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Per $i = 1, \ldots, n$, si ha*

$$\mathrm{E}(R_i) = \frac{n+1}{2}\,,\ \mathrm{Var}(R_i) = \frac{n^2-1}{12}\,,$$

e per $i \neq j = 1, \ldots, n$, si ha

$$\mathrm{Cov}(R_i, R_j) = -\frac{n+1}{12}\,.$$

Dimostrazione. Dal Corollario 2.4.4 e tenendo presente il Teorema A.2.1, si ha

$$\mathrm{E}(R_i) = \frac{1}{n}\sum_{r=1}^{n} r = \frac{n+1}{2}\,.$$

Analogamente, si ha

$$\mathrm{E}(R_i^2) = \frac{1}{n}\sum_{r=1}^{n} r^2 = \frac{(n+1)(2n+1)}{6}\,,$$

da cui

$$\mathrm{Var}(R_i) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}\,.$$

Di nuovo dal Corollario 2.4.4 e tenendo presente il Teorema A.2.1, si ha

$$\begin{aligned}\mathrm{E}(R_i R_j) &= \frac{1}{n(n-1)}\sum_{r_1=1}^{n}\sum_{r_2\neq r_1=1}^{n} r_1 r_2 = \frac{1}{n(n-1)}\,\Big((\sum_{r=1}^{n} r)^2 - \sum_{r=1}^{n} r^2\Big) \\ &= \frac{1}{n(n-1)}\,\Big(\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6}\Big) = \frac{(n+1)(3n+2)}{12}\,,\end{aligned}$$

per cui

$$\mathrm{Cov}(R_i, R_j) = \frac{(n+1)(3n+2)}{12} - \frac{(n+1)^2}{4} = -\frac{n+1}{12}\,. \qquad \square$$

Corollario 2.4.6. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Se $T = T(R_1, \ldots, R_n)$, ovvero T è una statistica basata solo sul vettore dei ranghi, allora T è "distribution-free" sulla classe $\mathcal{C}_F$.*

Dimostrazione. Il risultato segue immediatamente dal Teorema 2.4.3, in quanto per qualsiasi distribuzione congiunta di $F_n \in \mathcal{C}_F$ la statistica T è distribuita come una trasformata di un vettore di variabili casuali uniformemente distribuito su $\mathcal{R}_n$. $\square$

Corollario 2.4.7. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Si consideri una scelta di $k = 1, \ldots, n$ statistiche rango $(R_{i_1}, \ldots, R_{i_k})$ e sia $(R_{(i_1)}, \ldots, R_{(i_k)})$ il relativo vettore di statistiche rango ordinato in senso crescente per $i_1 \neq \ldots \neq i_k = 1, \ldots, n$. Si ha*

$$\Pr(R_{(i_1)} = r_{(1)}, \ldots, R_{(i_k)} = r_{(k)}) = \binom{n}{k}^{-1},\ 1 \leq r_{(1)} < \ldots < r_{(k)} \leq n\,.$$

Dimostrazione. Sia $(r_{(1)}, \ldots, r_{(k)})$ una scelta di k elementi di $(1, \ldots, n)$ tale che $1 \leq r_{(1)} < \ldots < r_{(k)} \leq n$. Dal Corollario 2.4.4 si ha

$$\Pr(R_{i_1} = r_{(1)}, \ldots, R_{i_k} = r_{(k)}) = \frac{1}{n(n-1)\ldots(n-k+1)} .$$

Tuttavia, questa è solamente una possibile permutazione di $(R_{i_1}, \ldots, R_{i_k})$ per cui si ha $R_{(i_1)} = r_{(1)}, \ldots, R_{(i_k)} = r_{(k)}$. Poichè esistono $k!$ di tali permutazioni ed ognuna è ugualmente probabile, allora

$$\Pr(R_{(i_1)} = r_{(1)}, \ldots, R_{(i_k)} = r_{(k)}) = k! \, \frac{1}{n(n-1)\ldots(n-k+1)} = \binom{n}{k}^{-1} . \qquad \square$$

• **Esempio 2.4.1.** Si consideri due campioni casuali $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$ provenienti da due variabili casuali continue ed equivalenti in distribuzione. Dunque, se $n = n_1 + n_2$, il campione misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ formato dai due campioni casuali originali può essere considerato sua volta un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Di conseguenza, siano $(R_1, \ldots, R_{n_1})$ i ranghi assegnati a $(X_1, \ldots, X_{n_1})$ e $(R_{n_1+1}, \ldots, R_n)$ i ranghi assegnati a $(Y_1, \ldots, Y_{n_2})$ nel campione misto. Si consideri la statistica

$$W = \sum_{i=1}^{n_1} R_i ,$$

che fornisce la somma dei ranghi assegnati a $(X_1, \ldots, X_{n_1})$. Quando i ranghi assegnati a $(X_1, \ldots, X_{n_1})$ sono i più bassi, ovvero $1, \ldots, n_1$, si ottiene il valore minimo che W può assumere, dato da $\sum_{i=1}^{n_1} i = n_1(n_1+1)/2$. Alternativamente, quando i ranghi assegnati a $(X_1, \ldots, X_{n_1})$ sono i più elevati, ovvero $n_2 + 1, \ldots, n$, si ottiene il valore massimo che W può assumere, dato da $\sum_{i=1}^{n_1}(n_2 + i) = n_1(n + n_2 + 1)/2$. Quindi, il supporto di W risulta

$$\{n_1(n_1+1)/2, n_1(n_1+1)/2 + 1, \ldots, n_1(n+n_2+1)/2\} .$$

Se $c_{n_1,n_2}(w)$ rappresenta il numero di sottoinsiemi di n_1 interi di $(1, \ldots, n)$ la cui somma è w, dal momento che

$$W = \sum_{i=1}^{n_1} R_{(i)} ,$$

allora tenendo presente il Corollario 2.4.7 la funzione di probabilità di W è data da

$$p_{n_1,n_2}(w) = \Pr(W = w) = \binom{n}{n_1}^{-1} c_{n_1,n_2}(w) \mathbf{1}_{\{n_1(n_1+1)/2, n_1(n_1+1)/2+1, \ldots, n_1(n+n_2+1)/2\}}(w) .$$

Poichè la distribuzione di W rimane invariata per ogni funzione di ripartizione congiunta di $\mathcal{C}_F$, si deve concludere che W è "distribution-free" su $\mathcal{C}_F$. Questo risultato può essere ottenuto immediatamente dal Corollario 2.4.6. $\triangleleft$

2.5. Le statistiche rango dei valori assoluti. Le statistiche rango dei valori assoluti costituiscono un'ulteriore importante classe di statistiche "distribution-free".

Definizione 2.5.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$. Si definiscono come statistiche rango dei valori assoluti le seguenti trasformate

$$R_i^+ = \sum_{j=1}^{n} \mathbf{1}_{[0,\infty)}(|X_i| - |X_j|) , i = 1, \ldots, n .$$

Il vettore di statistiche $(R_1^+, \ldots, R_n^+)$ è detto vettore dei ranghi dei valori assoluti. $\triangle$

Si noti che $(R_1^+, \ldots, R_n^+)$ non è altro che il vettore dei ranghi assegnati ai valori assoluti degli elementi del campione casuale originale. Il seguente teorema fornisce la distribuzione congiunta delle statistiche rango dei valori assoluti.

Teorema 2.5.2. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, allora la funzione di probabilità congiunta del relativo vettore dei ranghi dei valori assoluti $(R_1^+, \ldots, R_n^+)$ è data da*

$$\Pr(R_1^+ = r_1, \ldots, R_n^+ = r_n) = \frac{1}{n!}\, \mathbf{1}_{\mathcal{R}_n}(r_1, \ldots, r_n)\,.$$

Dimostrazione. Si noti che $(|X_1|, \ldots, |X_n|)$ è ancora un campione casuale. Dunque, poichè $(R_1^+, \ldots, R_n^+)$ è un vettore dei ranghi relativo ad un campione casuale, allora dal Teorema 2.4.3 si ha che $(R_1^+, \ldots, R_n^+)$ è uniformemente distribuito su $\mathcal{R}_n$. □

Se una variabile casuale assolutamente continua è simmetrica, il punto di simmetria coincide con la mediana. Dunque, essendo $\mathcal{S}_{0,F} \subset \mathcal{M}_{0,F}$, la distribuzione congiunta del vettore dei segni è quella ottenuta nel Teorema 2.3.2. I seguenti teoremi forniscono la distribuzione congiunta del vettore dei segni e del vettore dei ranghi dei valori assoluti.

Teorema 2.5.3. *Se X è una variabile casuale con funzione di ripartizione $F \in \mathcal{S}$, allora le variabili casuali trasformate $Y = |X|$ e $Z = \mathbf{1}_{(0,\infty)}(X)$ sono indipendenti.*

Dimostrazione. Per un dato $x > 0$ si ha

$$\begin{aligned}\Pr(Y \le x \mid Z = 1) &= \frac{\Pr(Y \le x, Z = 1)}{\Pr(Z = 1)} = \frac{\Pr(|X| \le x, X > 0)}{\Pr(X > 0)} \\ &= \frac{\Pr(0 < X \le x)}{\Pr(X > 0)} = \frac{F(x) - F(0)}{1 - F(0)}\,.\end{aligned}$$

Dal momento che $F(0) = 1/2$ e che la funzione di ripartizione di Y è data da $(2F(x) - 1)\mathbf{1}_{(0,\infty)}(x)$, si ha

$$\Pr(Y \le x \mid Z = 1) = \frac{F(x) - 1/2}{1/2} = 2F(x) - 1 = \Pr(Y \le x)\,.$$

Analogamente, per un dato $x > 0$ si ha

$$\begin{aligned}\Pr(Y \le x \mid Z = 0) &= \frac{\Pr(Y \le x, Z = 0)}{\Pr(Z = 0)} = \frac{\Pr(\mid X \mid \le x, X \le 0)}{\Pr(X \le 0)} \\ &= \frac{\Pr(-x \le X \le 0)}{\Pr(X \le 0)} = \frac{F(0) - F(-x)}{F(0)}\,,\end{aligned}$$

da cui, tenendo presente che $F(-x) = 1 - F(x)$, si ha

$$\Pr(Y \le x \mid Z = 0) = \frac{1/2 - 1 + F(x)}{1/2} = 2F(x) - 1 = \Pr(Y \le x)\,.$$

Poichè il supporto di Z è l'insieme $\{0, 1\}$, allora dalle precedenti relazioni si ottiene che $\Pr(Y \le x \mid Z = z) = \Pr(Y \le x)$ per $z = 0, 1$, da cui si conclude che le variabili casuali Y e Z sono indipendenti. □

Teorema 2.5.4. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, i relativi vettori di statistiche $(Z_1, \ldots, Z_n)$ e $(R_1^+, \ldots, R_n^+)$ sono indipendenti.*

Dimostrazione. Dal Teorema 2.5.3, si ottiene che Z_i e $|X_i|$ sono indipendenti per $i = 1, \ldots, n$. Dunque ogni statistica segno Z_i è indipendente dal campione trasformato $(|X_1|, \ldots, |X_n|)$, in quanto questo vettore costituisce un campione casuale. Di conseguenza, dal momento che per la Definizione 2.5.1 ogni statistica

rango dei valori assoluti R_i^+, per $i = 1, \ldots, n$, dipende solo dal campione trasformato $(|X_1|, \ldots, |X_n|)$, allora risulta indipendente dal vettore dei segni $(Z_1, \ldots, Z_n)$. Quindi, si deve concludere $(Z_1, \ldots, Z_n)$ e $(R_1^+, \ldots, R_n^+)$ sono vettori di statistiche indipendenti. □

Corollario 2.5.5. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$. Se $T = T(Z_1, \ldots, Z_n, R_1^+, \ldots, R_n^+)$, ovvero T è una statistica basata sul vettore dei segni e sul vettore dei ranghi dei valori assoluti, allora T è "distribution-free" su $\mathcal{S}_{0,F}$.*

Dimostrazione. Segue immediatamente dai Teoremi 2.3.2 e 2.5.2, in quanto per ogni distribuzione congiunta $F_n \in \mathcal{S}_{0,F}$ la statistica T è una trasformata di un vettore di variabili casuali indipendenti ed ugualmente distribuite come variabili casuali Binomiali $Bi(1, 1/2)$ e di un vettore uniformemente distribuito su $\mathcal{R}_n$. □

Teorema 2.5.6. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$. Si consideri una scelta di $k = 0, 1, \ldots, n$ statistiche rango dei valori assoluti $(R_{i_1}^+, \ldots, R_{i_k}^+)$ e sia $(R_{(i_1)}^+, \ldots, R_{(i_k)}^+)$ il relativo vettore dei ranghi dei valori assoluti ordinato in senso crescente per $i_1 \neq \ldots \neq i_k = 1, \ldots, n$. Sia inoltre $(Z_{i_1}, \ldots, Z_{i_k})$ la scelta di k statistiche segno associata a $(R_{i_1}^+, \ldots, R_{i_k}^+)$. Se $K = \sum_{i=1}^n Z_i$, si ha*

$$\Pr(R_{(i_1)}^+ = r_{(1)}, \ldots, R_{(i_k)}^+ = r_{(k)}, K = k \mid Z_{i_1} = 1, \ldots, Z_{i_k} = 1) = 2^{-n} ,$$

con $1 \leq r_{(1)} < \ldots < r_{(k)} \leq n$.

Dimostrazione. Dal Corollario 2.4.7 si ha

$$\Pr(R_{(i_1)}^+ = r_{(1)}, \ldots, R_{(i_k)}^+ = r_{(k)} \mid K = k) = \binom{n}{k}^{-1} , 1 \leq r_{(1)} < \ldots < r_{(k)} \leq n .$$

Inoltre, dal Teorema 2.3.2 si verifica che

$$\Pr(K = k) = \binom{n}{k} 2^{-n} \mathbf{1}_{\{0,1,\ldots,n\}}(k) ,$$

da cui si ottiene

$$\begin{aligned} \Pr(R_{(i_1)}^+ = r_{(1)}, \ldots, R_{(i_k)}^+ = r_{(k)}, K = k) &= \Pr(R_{(i_1)}^+ = r_{(1)}, \ldots, R_{(i_k)}^+ = r_{(k)} \mid K = k)\Pr(K = k) \\ &= \binom{n}{k}^{-1} \binom{n}{k} 2^{-n} = 2^{-n} , k = 0, 1, \ldots, n , 1 \leq r_{(1)} < \ldots < r_{(k)} \leq n . \end{aligned}$$

Tenendo presente che $(R_{i_1}^+, \ldots, R_{i_k}^+)$ è indipendente da $(Z_{i_1}, \ldots, Z_{i_k})$ per il Teorema 2.5.4, allora si ha

$$\begin{aligned} \Pr(R_{(i_1)}^+ &= r_{(1)}, \ldots, R_{(i_k)}^+ = r_{(k)}, K = k \mid Z_{i_1} = 1, \ldots, Z_{i_k} = 1) = \\ &= \Pr(R_{(i_1)}^+ = r_{(1)}, \ldots, R_{(i_k)}^+ = r_{(k)}, K = k) \\ &= 2^{-n} , k = 0, 1, \ldots, n , 1 \leq r_{(1)} < \ldots < r_{(k)} \leq n , \end{aligned}$$

che conclude la dimostrazione. □

• **Esempio 2.5.1.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$. Si consideri dunque la statistica

$$W^+ = \sum_{i=1}^n Z_i R_i^+ .$$

Il supporto di W^+ risulta $\{0, 1, \ldots, n(n+1)/2\}$. Se $c_n(w)$ rappresenta il numero di sottoinsiemi di interi di $(1, \ldots, n)$ la cui somma è w, allora dal Teorema 2.5.6 segue che la funzione di probabilità di W^+ è data da

$$p_n(w) = \Pr(W^+ = w) = 2^{-n} c_n(w) \mathbf{1}_{\{0,1,\ldots,n(n+1)/2\}}(w) .$$

Poichè la distribuzione di W^+ rimane invariata per ogni funzione di ripartizione congiunta di $\mathcal{S}_{0,F}$, allora W^+ è “distribution-free” su $\mathcal{S}_{0,F}$. Questo risultato poteva essere ottenuto immediatamente dal Corollario 2.5.5. Infine, dal momento che ad ogni sottoinsieme di interi di $(1, \ldots, n)$ la cui somma è w corrisponde un sottoinsieme complementare la cui somma è $(n(n+1)/2 - w)$, allora si ottiene $c_n(w) = c_n(n(n+1)/2 - w)$. Quindi, si ha anche $p_n(w) = p_n(n(n+1)/2 - w)$, ovvero $p_n(n(n+1)/4 + w) = p_n(n(n+1)/4 - w)$ per $w = 0, 1, \ldots, n(n+1)/2$. Tenendo presente l'Esempio 1.2.1 si può dunque concludere che W^+ è simmetrica rispetto a $n(n+1)/4$. ◁

Capitolo 3

Il test statistico “distribution-free”

3.1. I test statistici “distribution-free”. Se si dispone di un campione $(X_1, \dots, X_n)$, si può stabilire un insieme H delle ipotesi ammissibili sulla relativa funzione di ripartizione congiunta F_n. Se gli insiemi H_0 e H_1 costituiscono una partizione di H, il problema della verifica delle ipotesi consiste nel verificare l'ipotesi di base $H_0 : F_n \in \mathcal{F}_0$ contro l'ipotesi alternativa $H_1 : F_n \in \mathcal{F} \setminus \mathcal{F}_0$, dove $\mathcal{F}$ è il modello statistico e $\mathcal{F}_0$ una sua sottoclasse. L'insieme H delle ipotesi ammissibili e la sua partizione in H_0 e H_1 è detto sistema di ipotesi. Se il modello statistico è indicizzato mediante un insieme di parametri ψ e il contesto probabilistico in cui si lavora è evidente, per semplicità di notazione il sistema di ipotesi può essere indicato con $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$, dove Ψ rappresenta lo spazio dei parametri e Ψ_0 un suo sottoinsieme.

• **Esempio 3.1.1.** Si consideri un campione casuale $(X_1, \dots, X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$. Si vuole verificare che il punto di simmetria è un particolare valore λ_0, ovvero l'ipotesi di base $H_0 : F_n \in \mathcal{S}_{\lambda_0,F}$ contro l'ipotesi alternativa $H_1 : F_n \in \mathcal{S}_{\lambda,F} \setminus \mathcal{S}_{\lambda_0,F}$. Il modello statistico è “distribution-free” ed è indicizzato da due parametri, ovvero $\psi = \{\lambda, F\}$. Il sistema di ipotesi può dunque essere espresso più semplicemente come $H_0 : \lambda = \lambda_0, F \in \mathcal{S}$, contro $H_1 : \lambda \neq \lambda_0, F \in \mathcal{S}$. In maniera analoga, in un contesto di statistica classica si potrebbe considerare un modello statistico $\mathcal{F}_\lambda \subset \mathcal{S}_{\lambda,F}$ del tipo

$$\mathcal{F}_\lambda = \{F_n : F_n(x_1, \dots, x_n) = \prod_{i=1}^{n} \Phi(x_i - \lambda), \lambda \in \mathbb{R}\} .$$

Si osservi che $\mathcal{F}_\lambda$ contiene una sola famiglia di funzioni di ripartizione congiunte ed in particolare rappresenta il modello statistico relativo ad un campione casuale da una distribuzione Normale $N(\lambda, 1)$. Il modello $\mathcal{F}_\lambda$ è indicizzato solo attraverso λ, ovvero $\psi = \lambda$, e quindi in questo caso il sistema di ipotesi si riduce a verificare $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$. Tuttavia, può risultare poco plausibile in pratica assumere un modello statistico classico, ovvero assumere una specifica morfologia funzionale per F. ◁

Lo strumento statistico che sulla base dei dati campionari consente di concludere in favore dell'una o dell'altra ipotesi è il cosiddetto test.

Definizione 3.1.1. Sia $(X_1, \dots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Se $\mathcal{X}_n$ è lo spazio campionario, ovvero il supporto di F_n, si dice test la funzione $D : \mathcal{X}_n \to \{H_0, H_1\}$. △

Il test è in effetti una regola decisionale che suddivide $\mathcal{X}_n$ negli insiemi complementari $\mathcal{X}_0$ e $\mathcal{X}_1$ in modo tale che si accetta H_0 se la realizzazione del campione è in $\mathcal{X}_0$, mentre si accetta H_1 se la realizzazione del campione è in $\mathcal{X}_1$. L'insieme $\mathcal{X}_1$ è detto regione critica del test.

Definizione 3.1.2. Sia $(X_1, \dots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Se $T = T(X_1, \dots, X_n)$ è una statistica con supporto $\mathcal{T}$, si definisce test basato sulla statistica T la funzione $D : \mathcal{T} \to \{H_0, H_1\}$. La statistica T è detta statistica test. △

Il test basato sulla statistica T è dunque una regola decisionale che suddivide $\mathcal{T}$ negli insiemi complementari $\mathcal{T}_0$ e $\mathcal{T}_1$, in modo che si accetta H_0 se la realizzazione di T è in $\mathcal{T}_0$, mentre si accetta H_1 se la realizzazione di T è in $\mathcal{T}_1$. L'insieme $\mathcal{T}_1$ è detto regione critica del test basato sulla statistica T.

Definizione 3.1.3. Sia $(X_1, \dots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Un test per questo sistema di ipotesi è detto “distribution-free” se è basato su una statistica $T = T(X_1, \dots, X_n)$ “distribution-free” su $\mathcal{F}_\psi$ per ogni $\psi \in \Psi_0$. △

Uno strumento per misurare la capacità discriminatoria del test basato su una statistica T è dato dalla funzione potenza, che possiede la seguente difinizione.

Definizione 3.1.4. Sia $(X_1, \dots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. La funzione potenza del test basato sulla statistica T è data da

$$P_T(\psi) = \Pr_\psi(T \in \mathcal{T}_1) \,,$$

dove $\Pr_\psi$ indica che la probabilità è quella indotta da F_n con il valore del parametro pari a ψ. △

Per ogni $\psi \in \Psi_0$ la funzione potenza $P_T(\psi)$ fornisce quindi la probabilità di respingere H_0 quando questa è vera, ovvero la probabilità di commettere un errore di I specie. Analogamente, per ogni $\psi \in \Psi \setminus \Psi_0$ la quantità $(1 - P_T(\psi))$ fornisce la probabilità di accettare H_0 quando è vera H_1, ovvero la probabilità di commettere un errore di II specie.

Definizione 3.1.5. Sia $(X_1, \dots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Si dice che il test basato sulla statistica T è al livello di significatività α se

$$\sup_{\psi \in \Psi_0} P_T(\psi) = \alpha \,. \qquad \triangle$$

Il livello di significatività α rappresenta dunque la massima probabilità di errore di I specie. Quando il test è basato su una statistica discreta, allora esiste solo un numero finito o al più contabile di possibili livelli di significatività, che vengono detti livelli di significatività naturali. Sebbene esistano tecniche per ottenere un qualsiasi livello di significatività per un test basato su una statistica discreta (la cosiddetta casualizzazione del test), tuttavia risultano artificiose ed è preferibile considerare in pratica solo livelli di significatività naturali.

Definizione 3.1.6. Sia $(X_1, \dots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Un test basato sulla statistica T al livello di significatività α con funzione potenza $P_T(\psi)$ è detto corretto al livello di significatività α se

$$P_T(\psi) \geq \alpha \,, \forall \psi \in \Psi \setminus \Psi_0 \,. \qquad \triangle$$

La proprietà della correttezza assicura che la probabilità di accettare H_1 quando è vera risulta maggiore della probabilità di commettere un errore di I specie.

• **Esempio 3.1.2.** Si consideri un campione casuale $(X_1, \dots, X_n)$ con funzione di ripartizione $F_n \in \mathcal{F}_\lambda$, dove il modello statistico $\mathcal{F}_\lambda$ è definito nell'Esempio 3.1.1. Si vuole verificare $H_0 : \lambda = 0$ contro $H_1 : \lambda > 0$ mediante il test basato sulla statistica $T = \sqrt{n}\bar{X}$. Se è vera H_0 la statistica T ha distribuzione Normale $N(0, 1)$. Essendo $\mathcal{T} = (-\infty, +\infty)$, si può scegliere $\mathcal{T}_0 = (-\infty, z_{1-\alpha}]$ e $\mathcal{T}_1 = (z_{1-\alpha}, \infty)$. Se è vera H_1 la statistica T ha distribuzione Normale $N(\sqrt{n}\lambda, 1)$ e perciò la funzione potenza, che in questo caso dipende dal solo parametro λ, risulta

$$P_T(\lambda) = \Pr_\lambda(T \in \mathcal{T}_1) = 1 - \Phi(z_{1-\alpha} - \sqrt{n}\lambda) \, , \, \lambda \geq 0 \, .$$

Dal momento che $P_T(0) = \alpha$, allora il test è al livello di significatività α. Inoltre, essendo $P_T(\lambda)$ una funzione crescente, il test è corretto al livello di significatività α. ◁

• **Esempio 3.1.3.** Si consideri un campione casuale $(X_1, \ldots, X_n)$ con funzione di ripartizione $F_n \in \mathcal{S}_{\lambda,F}$. Si vuole verificare $H_0 : \lambda = 0, F \in \mathcal{S}$, contro $H_1 : \lambda > 0, F \in \mathcal{S}$. Si consideri dunque il test basato sulla statistica

$$B = \sum_{i=1}^{n} Z_i \, ,$$

definita nell'Esempio 2.3.1. Se è vera H_0, la statistica B è distribuita come una Binomiale $Bi(n, 1/2)$ essendo $\mathcal{S}_{0,F} \subset \mathcal{M}_{0,F}$. Dal momento che $\mathcal{T} = \{0, 1, \ldots, n\}$, si può scegliere $\mathcal{T}_0 = \{0, 1, \ldots, b_{n,1-\alpha}\}$ e $\mathcal{T}_1 = \{b_{n,1-\alpha} + 1, \ldots, n\}$. Se è vera H_1 risulta

$$\Pr(X_i > 0) = 1 - F(-\lambda) = F(\lambda) \, ,$$

per $i = 1, \ldots, n$, per cui B ha distribuzione Binomiale $Bi(n, F(\lambda))$. La funzione potenza è dunque data da

$$P_B(\lambda, F) = \Pr_{\lambda,F}(B \in \mathcal{T}_1) = \sum_{b=b_{n,1-\alpha}+1}^{n} \binom{n}{b} F(\lambda)^b (1 - F(\lambda))^{n-b} \, , \, \lambda \geq 0 \, .$$

Essendo $F(0) = 1/2$, allora risulta $P_B(0, F) = \alpha$ per ogni $F \in \mathcal{S}$ e quindi si deve concludere che il test è al livello di significatività α. Inoltre, dal momento che si può verificare che $P_B(\lambda, F)$ è una funzione crescente di λ per ogni $F \in \mathcal{S}$, allora il test è corretto al livello di significatività α. ◁

La seguente definizione enuncia una proprietà desiderabile per grandi campioni, ovvero quando $n \to \infty$, della funzione potenza di un test.

Definizione 3.1.7. Si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Data la statistica $T = T_n$, si consideri inoltre la successione di test al livello di significatività α basati sulla successione di statistiche $(T_n)_{n\geq 1}$. La successione di test è detta coerente se

$$\lim_n P_{T_n}(\psi) = 1 \, ,$$

per ogni $\psi \in \Psi \setminus \Psi_0$. △

La proprietà della coerenza assicura dunque che la probabilità di commettere un errore di II specie tende a 0 per $n \to \infty$ per ogni valore del parametro nell'alternativa.

• **Esempio 3.1.4.** Si consideri il sistema di ipotesi dell'Esempio 3.1.2 e sia $(T_n)_{n\geq 1}$ la successione di statistiche data da $(\sqrt{n}\bar{X})_{n\geq 1}$. Dal momento che la successione di funzioni $(\Phi(z_{1-\alpha} - \sqrt{n}\lambda))_{n\geq 1}$ converge uniformemente alla funzione identicamente nulla per $\lambda > 0$, allora si ha

$$\lim_n P_{T_n}(\lambda) = 1 \, , \, \lambda > 0 \, .$$

Dunque, la successione di test basata sulla successione di statistiche $(T_n)_{n\geq 1}$ è coerente. ◁

Il seguente teorema stabilisce delle condizioni per cui una successione di test basata su una successione di statistiche $(T_n)_{n\geq 1}$ è coerente.

Teorema 3.1.8. *Si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Data la statistica $T = T_n$, si consideri inoltre la successione di test al livello di significatività α basati sulla successione di*

statistiche $(T_n)_{n\geq 1}$*, tale che la regione critica del test basato su* T_n *è data da* $\mathcal{T}_{1,n} = \{t : t \geq c_n\}$*. Se esiste una funzione* g *per cui*
i) $T_n \xrightarrow{p} g(\psi), \forall \psi \in \Psi$,
ii) $g(\psi) = g_0, \forall \psi \in \Psi_0$,
iii) $g(\psi) > g_0, \forall \psi \in \Psi \setminus \Psi_0$,
iv) $\lim_n c_n \leq g_0$,
allora la successione di test basata su $(T_n)_{n\geq 1}$ *è coerente.*

Dimostrazione. Per un dato $\psi \in \Psi \setminus \Psi_0$, sia $\epsilon = (g(\psi) - g_0)/2$. Dalla condizione *iii*) si ha $\epsilon > 0$ e dunque per n sufficientemente elevato dalla condizione *iv*) si ha $c_n \leq g_0 + \epsilon = g(\psi) - \epsilon$. Quindi, risulta

$$\begin{aligned}\lim_n P_{T_n}(\psi) &= \lim_n \Pr_\psi(T_n \geq c_n) \geq \lim_n \Pr_\psi(T_n \geq g(\psi) - \epsilon) \\ &\geq \lim_n \Pr_\psi(g(\psi) - \epsilon \leq T_n \leq g(\psi) + \epsilon) = \lim_n \Pr_\psi(|T_n - g(\psi)| \leq \epsilon) \, .\end{aligned}$$

Per la condizione *i*) si ha $\lim_n \Pr_\psi(|T_n - g(\psi)| \leq \epsilon) = 1$, da cui

$$\lim_n P_{T_n}(\psi) = 1 \, , \forall \psi \in \Psi \setminus \Psi_0 \, ,$$

ovvero dalla Definizione 3.1.7 si ha che la successione di test basata su $(T_n)_{n\geq 1}$ è coerente. □

• **Esempio 3.1.5.** Se si considera il sistema di ipotesi dell'Esempio 3.1.3, si vuole verificare che la successione di test basata sulla successione di statistiche $(B_n)_{n\geq 1}$ è coerente. Si devono dunque verificare le condizioni *i*)-*iv*) del Teorema 3.1.8. A questo fine si considera la successione di test basata sulla successione di statistiche $(B_n/n)_{n\geq 1}$, che ha le stesse proprietà di quella basata sulla successione di statistiche $(B_n)_{n\geq 1}$. Per quanto riguarda *i*), si noti che per la Legge Debole dei Grandi Numeri di Khintchine (Teorema A.3.1) si ottiene $B_n/n \xrightarrow{p} g(\lambda, F) = F(\lambda)$ per ogni $\lambda \geq 0$. Inoltre, per quanto riguarda *ii*) e *iii*), si ha $g_0 = g(0, F) = 1/2$ da cui $g(\lambda, F) = F(\lambda) > g_0$ per ogni $\lambda > 0$. Infine, per quanto riguarda *iv*), si ha $\mathcal{T}_{1,n} = \{b : b \geq b_{n,1-\alpha}/n\}$. Per il Teorema Fondamentale Classico del Limite (Teorema A.3.6), se H_0 è vera risulta

$$\frac{B_n/n - 1/2}{1/\sqrt{4n}} \xrightarrow{d} N(0,1) \, ,$$

da cui $b_{n,1-\alpha}/n \simeq 1/2 + z_{1-\alpha}/\sqrt{4n}$ per n elevato. Dunque, si ha $\lim_n b_{n,1-\alpha}/n = 1/2 = g_0$. Le condizioni *i*)-*iv*) del Teorema 3.1.8 sono pertanto soddisfatte e dunque la successione di test basata sulla successione di statistiche $(B_n/n)_{n\geq 1}$ è coerente. Di conseguenza, anche la successione di test basata sulla successione di statistiche $(B_n)_{n\geq 1}$ è coerente. ◁

3.2. L'efficienza asintotica relativa. Sia $(X_1, \ldots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Se si dispone di due statistiche test T e U sorge a questo punto il problema di determinare quale delle due è preferibile. Un modo di procedere è quello di fissare il livello di significatività dei due test ad un medesimo valore preassegnato α (ovvero di controllare l'errore di I specie) e di considerare le relative funzioni potenza $P_T(\psi)$ e $P_U(\psi)$. Se risulta $P_T(\psi) \geq P_U(\psi)$ per ogni $\psi \in \Psi \setminus \Psi_0$, viene preferita la statistica test T, mentre se risulta $P_T(\psi) \leq P_U(\psi)$ per ogni $\psi \in \Psi \setminus \Psi_0$, viene preferita invece la statistica test U. Tuttavia, quando si considera test "distribution-free" in generale non è possibile determinare un test uniformemente più potente come nella statistica classica, dato che la struttura del sistema di ipotesi non è abbastanza rigida da consentire un risultato come il Lemma di Neyman-Pearson. Per determinati sistemi di ipotesi una possibilità alternativa è quella di ottenere il test localmente più potente, ovvero il test che ha la maggiore potenza dove la discriminazione fra H_0 e H_1 è più difficoltosa, ovvero in prossimità del punto di soglia delle due ipotesi (questi test verrano discussi nei prossimi capitoli). In questa sezione viene considerato piuttosto un confronto locale dei test per grandi campioni.

Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, dove θ è un parametro reale e F rappresenta la funzione di ripartizione della variabile casuale X da cui proviene il campione. Si consideri il problema di verificare l'ipotesi $H_0 : \theta \in \Theta_0, F \in \mathcal{F}$, contro

$H_1 : \theta \in \Theta \setminus \Theta_0, F \in \mathcal{F}$, dove Θ è lo spazio parametrico relativo a θ e Θ_0 un suo sottoinsieme, mentre $\mathcal{F}$ è una classe di funzioni di ripartizione. Se per verificare questo sistema di ipotesi si considera i test basati sulle statistiche T e U, allora il problema è stabilire quale dei due test è preferibile. A questo fine, è utile la seguente definizione.

Definizione 3.2.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, e si consideri il sistema di ipotesi $H_0 : \theta \in \Theta_0, F \in \mathcal{F}$, contro $H_1 : \theta \in \Theta \setminus \Theta_0, F \in \mathcal{F}$. Data la statistica $T = T_n$, si consideri inoltre la successione di test basata sulla successione di statistiche $(T_n)_{n\geq 1}$. Se la regione critica del test basato su T_n è data da $\mathcal{T}_{1,n}$, allora per un dato $\alpha \in (0,1)$, sia

$$P_{T_n}(\theta, F) = \Pr_{\theta,F}(T_n \in \mathcal{T}_{1,n}) \leq \alpha \ , \forall \theta \in \Theta_0 \ , n = 1, 2, \ldots \ .$$

Se N è il più piccolo valore di n tale che per un dato $\beta \in (\alpha, 1)$ risulta

$$P_{T_N}(\theta, F) = \Pr_{\theta,F}(T_n \in \mathcal{T}_{1,n}) \geq \beta \ , \forall \theta \in \Theta \setminus \Theta_0 \ ,$$

allora si dice che N è la minima numerosità campionaria del test al livello di significatività α basato sulla statistica T per raggiungere la potenza β per l'alternativa $\theta \in \Theta \setminus \Theta_0$. △

La quantità N è funzione di α, β, θ e della funzione di ripartizione F, ovvero $N = N(\alpha, \beta, \theta, F)$. Per una determinata alternativa θ e per α e β fissati, un test è dunque maggiormente preferibile quanto minore risulta $N(\alpha, \beta, \theta, F)$. Questa considerazione conduce alla seguente definizione.

Definizione 3.2.2. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, e si consideri il sistema di ipotesi $H_0 : \theta \in \Theta_0, F \in \mathcal{F}$, contro $H_1 : \theta \in \Theta \setminus \Theta_0, F \in \mathcal{F}$. Siano inoltre $N(\alpha, \beta, \theta, F)$ e $M(\alpha, \beta, \theta, F)$ le minime numerosità campionarie dei test al livello di significatività α basati sulle statistiche T e U per raggiungere la potenza β per l'alternativa $\theta \in \Theta \setminus \Theta_0$. Si dice efficienza relativa del test basato sulla statistica T rispetto al test basato sulla statistica U la quantità

$$e_{T,U}(\alpha, \beta, \theta, F) = \frac{M(\alpha, \beta, \theta, F)}{N(\alpha, \beta, \theta, F)} \ , \theta \in \Theta \setminus \Theta_0 \ .$$

Se $e_{T,U}(\alpha, \beta, \theta, F) > 1$ allora il test basato su T è più efficiente di quello basato su U, mentre se $e_{T,U}(\alpha, \beta, \theta, F) < 1$ è vero l'opposto. △

L'efficienza relativa è una misura locale dell'efficienza di un test rispetto ad un altro, nel senso che dipende dalle quantità α, β, θ e F. Al fine di eliminare almeno la dipendenza da α, β, θ e quindi di ottenere una misura più globale dell'efficienza, è conveniente introdurre un indice basato su un confronto per grandi campioni in un particolare sistema di ipotesi.

Definizione 3.2.3. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, e si consideri il sistema di ipotesi $H_0 : \theta = \theta_0, F \in \mathcal{F}$, contro $H_1 : \theta = \theta_i, F \in \mathcal{F}$, dove $(\theta_i)_{i\geq 1}$ è una successione di alternative tali che $\lim_i \theta_i = \theta_0$. Date le statistiche $T = T_n$ e $U = U_n$, si consideri le due successioni di test al livello di significatività α basati sulle successioni di statistiche $(T_{n_i})_{i\geq 1}$ e $(U_{m_i})_{i\geq 1}$. Se le regioni critiche dei test basati su T_{n_i} e U_{m_i} sono date da $\mathcal{T}_{1,n_i} = \{t : t \geq c_{n_i}\}$ e $\mathcal{U}_{1,m_i} = \{u : u \geq d_{m_i}\}$, allora siano

$$P_{T_{n_i}}(\theta_i, F) = \Pr_{\theta_i,F}(T_{n_i} \geq c_{n_i})$$

e

$$P_{U_{m_i}}(\theta_i, F) = \Pr_{\theta_i,F}(U_{m_i} \geq d_{m_i})$$

le rispettive potenze per l'alternativa θ_i con $i = 1, 2, \ldots$. Se $(n_i)_{i\geq 1}$ e $(m_i)_{i\geq 1}$ sono due successioni crescenti di interi tali che

$$\alpha < \lim_i P_{T_{n_i}}(\theta_i, F) = \lim_i P_{U_{m_i}}(\theta_i, F) < 1 \ ,$$

si definisce efficienza asintotica relativa del test basato su T rispetto a quella basato su U

$$\mathrm{EAR}_{T,U} = \lim_i \frac{m_i}{n_i} \,,$$

dove il limite deve essere costante per qualsiasi successione $(n_i)_{i\geq 1}$ e $(m_i)_{i\geq 1}$ ed essere indipendente dalla successione $(\theta_i)_{i\geq 1}$. △

Si osservi che $\mathrm{EAR}_{T,U}$ è il limite del rapporto delle numerosità campionarie necessarie ad ottenere la medesima potenza dei due test per la stessa successione di alternative (che converge al valore specificato sotto ipotesi di base) a un livello di significatività costante. Si ha $\mathrm{EAR}_{T,U} = \mathrm{EAR}_{T,U}(F)$, ovvero l'efficienza asintotica relativa dipende comunque dalla struttura funzionale della funzione di ripartizione F. Il seguente teorema permette di determinare una espressione di $\mathrm{EAR}_{T,U}$ più conveniente dal punto di vista computazionale.

Teorema 3.2.4. (Teorema di Noether) *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, e si consideri il sistema di ipotesi $H_0 : \theta = \theta_0, F \in \mathcal{F}$, contro $H_1 : \theta = \theta_i, F \in \mathcal{F}$, dove $(\theta_i)_{i\geq 1}$ è una successione di alternative tali che $\lim_i \theta_i = \theta_0$. Date le statistiche $T = T_n$ e $U = U_n$, siano $(T_{n_i})_{i\geq 1}$ e $(U_{m_i})_{i\geq 1}$ due successioni di statistiche a cui sono associate le successioni $(\mu_{T_{n_i}}(\theta))_{i\geq 1}$, $(\sigma_{T_{n_i}}(\theta))_{i\geq 1}$, $(\mu_{U_{m_i}}(\theta))_{i\geq 1}$ e $(\sigma_{U_{m_i}}(\theta))_{i\geq 1}$, dove $(n_i)_{i\geq 1}$ e $(m_i)_{i\geq 1}$ sono due successioni crescenti di interi. Siano inoltre $\mathcal{T}_{1,n_i} = \{t : t \geq c_{n_i}\}$ e $\mathcal{U}_{1,m_i} = \{u : u \geq d_{m_i}\}$ le regioni critiche dei test al livello di significatività α basati sulle statistiche T_{n_i} e U_{m_i}. Se:*
i) quando θ_i è il vero valore di θ, per una variabile casuale limite V si ha

$$\frac{T_{n_i} - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_i)} \xrightarrow{d} V \,, \; \frac{U_{m_i} - \mu_{U_{m_i}}(\theta_i)}{\sigma_{U_{m_i}}(\theta_i)} \xrightarrow{d} V \,;$$

ii) stessa condizione i) con θ_0 al posto di θ_i;
iii) si ha

$$\lim_i \frac{\sigma_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_0)} = \lim_i \frac{\sigma_{U_{m_i}}(\theta_i)}{\sigma_{U_{m_i}}(\theta_0)} = 1 \,;$$

iv) le derivate

$$\frac{d}{d\theta}\,\mu_{T_n}(\theta) = \mu'_{T_n}(\theta) \,, \; \frac{d}{d\theta}\,\mu_{U_m}(\theta) = \mu'_{U_m}(\theta)$$

sono continue in intorno di $\theta = \theta_0$ e $\mu'_{T_n}(\theta_0) \neq 0$, $\mu'_{U_m}(\theta_0) \neq 0$;
v) si ha

$$\lim_i \frac{\mu'_{T_{n_i}}(\theta_i)}{\mu'_{T_{n_i}}(\theta_0)} = \lim_i \frac{\mu'_{U_{m_i}}(\theta_i)}{\mu'_{U_{m_i}}(\theta_0)} = 1 \,;$$

vi) si ha

$$K_T = \lim_n \frac{\mu'_{T_n}(\theta_0)}{\sqrt{n}\sigma_{T_n}(\theta_0)} > 0 \,, \; K_U = \lim_m \frac{\mu'_{U_m}(\theta_0)}{\sqrt{m}\sigma_{U_m}(\theta_0)} > 0 \,;$$

allora

$$\mathrm{EAR}_{T,U} = \frac{K_T^2}{K_U^2} \,.$$

Dimostrazione. In base alla condizione *iv*) si può ottenere la seguente espansione in serie di Taylor di $\mu_{T_{n_i}}(\theta)$ nel punto $\theta = \theta_0$

$$\mu_{T_{n_i}}(\theta_i) = \mu_{T_{n_i}}(\theta_0) + (\theta_i - \theta_0)\mu'_{T_{n_i}}(\theta_i^*) \, , \, \theta_0 < \theta_i^* < \theta_i \, .$$

Analogamente, per $\mu_{U_{m_i}}(\theta)$ si ha

$$\mu_{U_{m_i}}(\theta_i) = \mu_{U_{m_i}}(\theta_0) + (\theta_i - \theta_0)\mu'_{U_{m_i}}(\theta_i^{**}) \, , \, \theta_0 < \theta_i^{**} < \theta_i \, .$$

Dalle assunzioni fatte risulta

$$\lim_i P_{T_{n_i}}(\theta_0, F) = \lim_i \Pr_{\theta_0,F}(T_{n_i} \geq c_{n_i}) = \lim_i \Pr_{\theta_0,F}\left(\frac{T_{n_i} - \mu_{T_{n_i}}(\theta_0)}{\sigma_{T_{n_i}}(\theta_0)} \geq \frac{c_{n_i} - \mu_{T_{n_i}}(\theta_0)}{\sigma_{T_{n_i}}(\theta_0)}\right) = \alpha \, .$$

Tenendo presente la condizione *ii*), per il Teorema A.3.12 la precedente relazione implica che

$$\lim_i \frac{c_{n_i} - \mu_{T_{n_i}}(\theta_0)}{\sigma_{T_{n_i}}(\theta_0)} = v_{1-\alpha} \, ,$$

dove $v_{1-\alpha}$ rappresenta il quantile di ordine $(1-\alpha)$ della variabile casuale V. Mediante la precedente relazione e la condizione *iii*) si ha

$$\begin{aligned} \lim_i \frac{c_{n_i} - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_i)} &= \lim_i \frac{c_{n_i} - \mu_{T_{n_i}}(\theta_0) + \mu_{T_{n_i}}(\theta_0) - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_0)} \frac{\sigma_{T_{n_i}}(\theta_0)}{\sigma_{T_{n_i}}(\theta_i)} \\ &= v_{1-\alpha} + \lim_i \frac{\mu_{T_{n_i}}(\theta_0) - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_0)} \, . \end{aligned}$$

Si supponga che

$$\lim_i P_{T_{n_i}}(\theta_i, F) = \lim_i \Pr_{\theta_i,F}(T_{n_i} \geq c_{n_i}) = \beta \, ,$$

ovvero che il limite della potenza della successione di test basata sulla successione di statistiche $(T_{n_i})_{i\geq 1}$ per la successione di alternative $(\theta_i)_{i\geq 1}$ sia β. Con un procedimento simile a quello visto in precedenza, tenendo presente la condizione *i*), per il Teorema A.3.12 si ha che

$$\lim_i \frac{c_{n_i} - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_i)} = v_{1-\beta} \, .$$

Analogamente, si ha

$$\lim_i P_{U_{m_i}}(\theta_0, F) = \lim_i \Pr_{\theta_0,F}(U_{m_i} \geq d_{m_i}) = \alpha \, ,$$

da cui

$$\lim_i \frac{d_{m_i} - \mu_{U_{m_i}}(\theta_i)}{\sigma_{U_{m_i}}(\theta_i)} = v_{1-\alpha} + \lim_i \frac{\mu_{U_{m_i}}(\theta_0) - \mu_{U_{m_i}}(\theta_i)}{\sigma_{U_{m_i}}(\theta_0)} \, ,$$

mentre se si suppone che

$$\lim_i P_{U_{m_i}}(\theta_i, F) = \lim_i \Pr_{\theta_i,F}(U_{m_i} \geq d_{m_i}) = \beta \, ,$$

allora

$$\lim_i \frac{d_{m_i} - \mu_{U_{m_i}}(\theta_i)}{\sigma_{U_{m_i}}(\theta_i)} = v_{1-\beta} \, .$$

Combinando le relazioni ottenute si ha dunque

$$\lim_i \frac{\mu_{T_{n_i}}(\theta_0) - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_0)} = \lim_i \frac{\mu_{U_{m_i}}(\theta_0) - \mu_{U_{m_i}}(\theta_i)}{\sigma_{U_{m_i}}(\theta_0)} ,$$

ovvero

$$\lim_i \frac{\mu_{T_{n_i}}(\theta_0) - \mu_{T_{n_i}}(\theta_i)}{\sigma_{T_{n_i}}(\theta_0)} \frac{\sigma_{U_{m_i}}(\theta_0)}{\mu_{U_{m_i}}(\theta_0) - \mu_{U_{m_i}}(\theta_i)} = 1 .$$

Sostituendo in questa ultima espressione le opportune espansioni in serie di Taylor si ha

$$\lim_i \frac{(\theta_i - \theta_0)\mu'_{T_{n_i}}(\theta_i^*)}{\sigma_{T_{n_i}}(\theta_0)} \frac{\sigma_{U_{m_i}}(\theta_0)}{(\theta_i - \theta_0)\mu'_{U_{m_i}}(\theta_i^{**})} = 1 ,$$

ovvero

$$\lim_i \frac{\sqrt{n_i}}{\sqrt{m_i}} \frac{\mu'_{T_{n_i}}(\theta_i^*)}{\sqrt{n_i}\sigma_{T_{n_i}}(\theta_0)} \frac{\sqrt{m_i}\sigma_{U_{m_i}}(\theta_0)}{\mu'_{U_{m_i}}(\theta_i^{**})} = 1 .$$

Tenendo presente la condizione *iv*), *v*) *vi*) si ha dunque

$$\frac{K_T}{K_U} \lim_i \frac{\sqrt{n_i}}{\sqrt{m_i}} = 1 ,$$

da cui si ottiene infine

$$\mathrm{EAR}_{T,U} = \lim_i \frac{m_i}{n_i} = \frac{K_T^2}{K_U^2} . \qquad \square$$

Definizione 3.2.5. Se sono soddisfatte le condizioni del Teorema 3.2.4, la quantità

$$K_T = \lim_n \frac{\mu'_{T_n}(\theta_0)}{\sqrt{n}\sigma_{T_n}(\theta_0)}$$

è detta efficacia del test basato sulla statistica T ed è denotata con eff_T. $\triangle$

Nell'espressione dell'efficacia la quantità $\mu'_{T_n}(\theta_0)$ non è altro che una misura del tasso di variazione del parametro di posizione $\mu_{T_n}(\theta)$ per valori di θ prossimi a θ_0. La quantità $\sigma_{T_n}(\theta_0)$ serve invece a standardizzare la quantità al numeratore. Dunque, l'efficacia è in effetti il limite del tasso di variazione standardizzato del parametro di posizione di T_n per alternative prossime a θ_0. Più il tasso di variazione è sensibile alle alternative, più alta è l'efficacia del test. Inoltre, il Teorema 3.2.4 implica

$$\lim_i \frac{\mu_{T_{n_i}}(\theta_i) - \mu_{T_{n_i}}(\theta_0)}{\sigma_{T_{n_i}}(\theta_0)} = v_{1-\alpha} - v_{1-\beta} ,$$

ovvero

$$\lim_i \sqrt{n_i}(\theta_i - \theta_0) \frac{\mu'_{T_{n_i}}(\theta_i^*)}{\sqrt{n_i}\sigma_{T_{n_i}}(\theta_0)} = v_{1-\alpha} - v_{1-\beta} ,$$

da cui

$$\lim_i \sqrt{n_i}(\theta_i - \theta_0) = \frac{v_{1-\alpha} - v_{1-\beta}}{K_T} .$$

Dunque, la successione di alternative $(\theta_i)_{i\geq 1}$ deve essere tale che

$$\theta_i = \theta_0 + \frac{v_{1-\alpha} - v_{1-\beta}}{\sqrt{n_i} K_T} + g(n_i) \,,$$

dove $\lim_i \sqrt{n_i} g(n_i) = 0$. Quindi, il Teorema 3.2.4 assume implicitamente una struttura particolare per la successione di alternative $(\theta_i)_{i\geq 1}$, che possono essere espresse in funzione della numerosità campionaria. Infatti, la successione di alternative $(\theta_i)_{i\geq 1}$ è equivalente ad una successione di alternative del tipo $(\theta_0 + c/\sqrt{n_i})_{i\geq 1}$ con c costante, che è detto sistema di alternative di Pitman.

• **Esempio 3.2.1.** Si consideri un campione casuale $(X_1, \ldots, X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\lambda,F}$, dove

$$\mathcal{F}_{\lambda,F} = \{F_n : F_n \in \mathcal{S}_{\lambda,F}, \text{Var}(X) = \sigma^2 < \infty\} \,.$$

Dall'Esempio 1.2.3 si verifica che $\text{E}(X) = \lambda$. Si consideri il sistema di ipotesi $H_0 : \lambda = 0, F \in \mathcal{F}$, contro $H_1 : \lambda = \lambda_i, F \in \mathcal{F}$, dove $(\lambda_i)_{i\geq 1}$ è una successione di alternative tali che $\lambda_i = c/\sqrt{n_i}$ con c costante. Inoltre, $\mathcal{F}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua e simmetrica rispetto a 0 con varianza finita. Non si perde di generalità nell'assumere questo sistema di ipotesi, in quanto se risulta $H_0 : \lambda = \lambda_0, F \in \mathcal{F}$, contro $H_1 : \lambda = \lambda_i = \lambda_0 + c/\sqrt{n_i}$, $F \in \mathcal{F}$, si può ottenere il sistema di ipotesi originale considerando il campione trasformato $(X_1 - \lambda_0, \ldots, X_n - \lambda_0)$. Si vuole determinare l'efficacia del test basato sulla statistica di Student

$$T = T_n = \frac{\bar{X}}{S/\sqrt{n}} \,.$$

Si devono verificare le condizioni del Teorema 3.2.4. Innanzitutto, si sceglie $\mu_{T_{n_i}}(\lambda) = \sqrt{n_i}\lambda/\sigma$ e $\sigma_{T_{n_i}}(\lambda) = 1$. Per quanto riguarda la verifica della condizione *i*) si noti che

$$\frac{T_{n_i} - \mu_{T_{n_i}}(\lambda_i)}{\sigma_{T_{n_i}}(\lambda_i)} = \frac{\bar{X}}{S/\sqrt{n_i}} - \frac{\lambda_i}{\sigma/\sqrt{n_i}} = \frac{\bar{X} - c/\sqrt{n_i}}{\sigma/\sqrt{n_i}} \frac{\sigma}{S} + \frac{c}{\sigma} \left(\frac{\sigma}{S} - 1\right) .$$

Dal momento che $S^2 \xrightarrow{p} \sigma^2$ per $i \to \infty$ (vedi Esempio A.3.5), dal Teorema di Svedrup (Teorema A.3.4) si ha $S \xrightarrow{p} \sigma$ per $i \to \infty$. Inoltre, tenendo presente il Corollario A.3.8 al Teorema Fondamentale del Limite di Lindberg, si ottiene $\sqrt{n_i}(\bar{X} - c/\sqrt{n_i})/\sigma \xrightarrow{d} N(0,1)$ per $i \to \infty$. Combinando questi risultati mediante il Teorema di Slutsky (Teorema A.3.5) si ha

$$\frac{T_{n_i} - \mu_{T_{n_i}}(\lambda_i)}{\sigma_{T_{n_i}}(\lambda_i)} \xrightarrow{d} N(0,1) \,,$$

e quindi la condizione *i*) è verificata. Anche la condizione *ii*) è verificata, in quanto se è vera H_0 si ha $S \xrightarrow{p} \sigma$ e $\sqrt{n_i}\bar{X}/\sigma \xrightarrow{d} N(0,1)$ per $i \to \infty$ (vedi Esempio 2.2.2). Di conseguenza, applicando il Teorema di Slutski (Teorema A.3.5) si ha

$$\frac{T_{n_i} - \mu_{T_{n_i}}(0)}{\sigma_{T_{n_i}}(0)} = \frac{\bar{X}}{\sigma/\sqrt{n_i}} \frac{\sigma}{S} \xrightarrow{d} N(0,1) \,.$$

E' banale verificare la condizione *iii*). Anche la condizione *iv*) è verificata, in quanto

$$\mu'_{T_n}(\lambda) = \frac{d}{d\lambda} \mu_{T_n}(\lambda) = \frac{d}{d\lambda} \frac{\lambda}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}$$

è una funzione continua in un intorno di 0 e $\mu'_{T_n}(0) \neq 0$. Risulta banale verificare anche la condizione *v*). Infine, per quanto riguarda la condizione *vi*), si ha

$$K_T = \lim_n \frac{\mu'_{T_n}(0)}{\sqrt{n}\sigma_{T_n}(0)} = \lim_n \frac{\sqrt{n}/\sigma}{\sqrt{n}} = \frac{1}{\sigma} > 0 \, .$$

Le condizioni del Teorema 3.2.4 sono dunque soddisfatte e quindi si ha

$$\text{eff}_T = \frac{1}{\sigma} \, .$$

Questa quantità sarà utilizzata per determinare l'efficienza asintotica relativa di questo test classico rispetto ad alcuni test “distribution-free”. ◁

• **Esempio 3.2.2.** Si consideri un campione casuale $(X_1, \ldots, X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\lambda,F}$, dove

$$\mathcal{F}_{\lambda,F} = \{F_n : F_n \in \mathcal{S}_{\lambda,F}, 0 < F'(0) = f(0) < \infty\} \, .$$

Si consideri inoltre il sistema di ipotesi $H_0 : \lambda = 0, F \in \mathcal{F}$, contro $H_1 : \lambda = \lambda_i, F \in \mathcal{F}$, dove $(\lambda_i)_{i\geq 1}$ è una successione di alternative tali che $\lambda_i = c/\sqrt{n_i}$ con c costante. Inoltre, $\mathcal{F}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua e simmetrica rispetto a 0 con funzione di densità finita e non nulla a 0. Si vuole determinare l'efficacia del test basato sulla statistica considerata nell'Esempio 2.3.1

$$B = B_n = \sum_{i=1}^{n} Z_i \, .$$

Si deve dunque verificare le condizioni del Teorema 3.2.4. Se è vera H_1, allora risulta

$$\Pr(X > 0) = 1 - F(-\lambda) = F(\lambda) \, .$$

Dunque, è conveniente scegliere $\mu_{B_{n_i}}(\lambda) = n_i F(\lambda)$ e $\sigma_{B_{n_i}}(\lambda) = \sqrt{n_i F(\lambda)(1 - F(\lambda))}$. Inoltre, si ha $\text{E}(Z_i) = F(c/\sqrt{n_i})$ e $\text{Var}(Z_i) = F(c/\sqrt{n_i})(1 - F(c/\sqrt{n_i}))$. Dunque, la condizione *i*) è verificata, dal momento che per il Corollario A.3.8 al Teorema Fondamentale del Limite di Lindberg per $i \to \infty$ si ha

$$\frac{B_{n_i} - \mu_{B_{n_i}}(\lambda_i)}{\sigma_{B_{n_i}}(\lambda_i)} = \frac{B_{n_i} - n_i F(\lambda_i)}{\sqrt{n_i F(\lambda_i)(1 - F(\lambda_i))}} = \sqrt{n_i} \, \frac{B_{n_i}/n_i - F(\lambda_i)}{\sqrt{F(\lambda_i)(1 - F(\lambda_i))}} \xrightarrow{d} N(0,1) \, .$$

Anche la condizione *ii*) è verificata, in quanto dal Teorema Fondamentale del Limite Classico (Teorema A.3.6) si ha

$$\frac{B_{n_i} - \mu_{B_{n_i}}(0)}{\sigma_{B_{n_i}}(0)} = \frac{B_{n_i} - n_i/2}{\sqrt{n_i/4}} = \sqrt{n_i} \, \frac{B_{n_i}/n_i - 1/2}{\sqrt{1/4}} \xrightarrow{d} N(0,1) \, .$$

Risulta banale verificare la condizione *iii*). La condizione *iv*) è verificata, in quanto

$$\mu'_{B_n}(\lambda) = \frac{d}{d\lambda} \, \mu_{B_n}(\lambda) = \frac{d}{d\lambda} \, nF(\lambda) = nf(\lambda) \, ,$$

è una funzione continua in un intorno di 0 e $\mu'_{T_n}(0) \neq 0$. Risulta banale verificare anche la condizione *v*). Infine, per quanto riguarda la condizione *vi*), si ha

$$K_B = \lim_n \frac{\mu'_{B_n}(0)}{\sqrt{n}\sigma_{B_n}(0)} = \lim_n \frac{nf(0)}{n/2} = 2f(0) > 0 \, .$$

Le condizioni del Teorema 3.2.4 sono dunque soddisfatte e quindi si ha

$$\text{eff}_B = 2f(0) \, .$$

Utilizzando i risultati dell'Esempio 3.2.1, l'efficienza asintotica relativa del test basato sulla statistica B rispetto al test basato sulla statistica T di Student è data da

$$\mathrm{EAR}_{B,T} = \frac{K_B^2}{K_T^2} = \frac{4f(0)^2}{1/\sigma^2} = 4\sigma^2 f(0)^2 \,.$$

Se il campione casuale proviene da una distribuzione $N(\lambda, \sigma^2)$, dal momento che $f(0) = 1/\sqrt{2\pi\sigma^2}$, allora $\mathrm{EAR}_{B,T} = 2/\pi \simeq 0.6366$. Tuttavia, se il campione casuale proviene da una distribuzione $L(\lambda, \delta)$, dal momento che $\sigma^2 = 2\delta^2$ e $f(0) = 1/(2\delta)$ allora $\mathrm{EAR}_{B,T} = 2$. Quindi, se cade l'assunzione di normalità si può avere una notevole perdita di efficienza dei test classici. ◁

• **Esempio 3.2.3.** Si consideri i due campioni casuali indipendenti $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$, tali che se $n = n_1 + n_2$, allora $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ costituisce un campione misto con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\Delta,F}$, dove

$$\mathcal{F}_{\Delta,F} = \{F_n : F_n \in \mathcal{L}_{\Delta,F}, \mathrm{Var}(X) = \sigma^2 < \infty\} \,.$$

Si ha $\mathrm{Var}(Y) = \sigma^2$ e senza perdita di generalità si può supporre $\mathrm{E}(X) = 0$, da cui segue $\mathrm{E}(Y) = \Delta$. Si consideri il sistema di ipotesi $H_0 : \Delta = 0, F \in \mathcal{F}$, contro $H_1 : \Delta = \Delta_i, F \in \mathcal{F}$ dove $(\Delta_i)_{i\geq 1}$ è una successione di alternative tali che $\Delta_i = c/\sqrt{n_i}$ con c costante. In questo caso, $\mathcal{F}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua con varianza finita. Siano inoltre $(n_{1i})_{i\geq 1}$ e $(n_{2i})_{i\geq 1}$ due successioni tali che $\lim_i n_{1i}/n_i = \nu$ e $\lim_i n_{2i}/n_i = 1 - \nu$ con $\nu \in (0, 1)$. Si vuole determinare l'efficacia del test basato sulla statistica di Student per due campioni indipendenti, data da

$$T = T_n = \frac{\bar{Y} - \bar{X}}{S\sqrt{n/(n_1 n_2)}} \,,$$

dove $\bar{X}$ e $\bar{Y}$ rappresentano le medie campionarie, mentre

$$S^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2\right) = \frac{1}{n-2}\left((n_1 - 1)S_x^2 + (n_2 - 1)S_y^2\right) ,$$

dove S_x^2 e S_y^2 rappresentano le varianze campionarie corrette. Si devono dunque verificare le condizioni del Teorema 3.2.4. Innanzitutto, si sceglie $\mu_{T_{n_i}}(\Delta) = \sqrt{n_{1i}n_{2i}/n_i}(\Delta/\sigma)$ e $\sigma_{T_{n_i}}(\Delta) = 1$. Per quanto riguarda la verifica della condizione *i*), si noti che

$$\begin{aligned}\frac{T_{n_i} - \mu_{T_{n_i}}(\Delta_i)}{\sigma_{T_{n_i}}(\Delta_i)} &= \frac{\bar{Y} - \bar{X}}{S\sqrt{n_i/(n_{1i}n_{2i})}} - \frac{\Delta_i}{\sigma\sqrt{n_i/(n_{1i}n_{2i})}} \\ &= \frac{(\bar{Y} - c/\sqrt{n_i}) - \bar{X}}{\sigma\sqrt{n_i/(n_{1i}n_{2i})}}\,\frac{\sigma}{S} + \frac{c\sqrt{n_{1i}n_{2i}}}{\sigma n_i}\left(\frac{\sigma}{S} - 1\right) .\end{aligned}$$

Analogamente all'Esempio 3.2.1 si ha $S_y \xrightarrow{p} \sigma$ e $\sqrt{n_{2i}}(\bar{Y} - c/\sqrt{n_i})/\sigma \xrightarrow{d} N(0,1)$ per $i \to \infty$. Inoltre, dall'Esempio 2.2.2 si ha $S_x \xrightarrow{p} \sigma$ e $\sqrt{n_{1i}}\bar{X}/\sigma \xrightarrow{d} N(0,1)$ per $i \to \infty$. Di conseguenza, applicando il Teorema di Sverdrup (Teorema A.3.4), per $i \to \infty$ si ha

$$S^2 = \frac{(n_{1i} - 1)S_x^2}{n_i - 2} + \frac{(n_{2i} - 1)S_y^2}{n_i - 2} \xrightarrow{p} \nu\sigma^2 + (1 - \nu)\sigma^2 = \sigma^2 \,,$$

da cui $S \xrightarrow{p} \sigma$. Inoltre, applicando il Metodo Delta (Teorema A.3.10) alle variabili casuali indipendenti $\sqrt{n_{1i}}\bar{X}/\sigma$ e $\sqrt{n_{2i}}(\bar{Y} - c/\sqrt{n_i})/\sigma$ mediante la funzione $g(x, y) = y - x$, per $i \to \infty$ si ottiene

$$\frac{(\bar{Y} - c/\sqrt{n_i}) - \bar{X}}{\sqrt{\sigma^2/n_{1i} + \sigma^2/n_{2i}}} = \frac{(\bar{Y} - c/\sqrt{n_i}) - \bar{X}}{\sigma\sqrt{n_i/(n_{1i}n_{2i})}} \xrightarrow{d} N(0,1) \,.$$

Per il Teorema di Slutski (Teorema A.3.5), combinando i precedenti risultati, per $i \to \infty$ risulta infine

$$\frac{T_{n_i} - \mu_{T_{n_i}}(\Delta_i)}{\sigma_{T_{n_i}}(\Delta_i)} \xrightarrow{d} N(0,1) .$$

La condizione *ii*) è verificata, in quanto con un procedimento analogo a quello considerato per verificare la condizione *i*) per $i \to \infty$ si ha

$$\frac{T_{n_i} - \mu_{T_{n_i}}(0)}{\sigma_{T_{n_i}}(0)} \xrightarrow{d} N(0,1) .$$

Risulta banale verificare la condizione *iii*). La condizione *iv*) è verificata, dal momento che

$$\mu'_{T_n}(\Delta) = \frac{d}{d\Delta}\,\mu_{T_n}(\Delta) = \frac{d}{d\Delta}\,\frac{\Delta}{\sigma\sqrt{n/(n_1 n_2)}} = \frac{1}{\sigma\sqrt{n/(n_1 n_2)}} ,$$

è una funzione continua in un intorno di 0 e $\mu'_{T_n}(0) \neq 0$. Risulta banale verificare anche la condizione *v*). Infine, per quanto riguarda la condizione *vi*), si ha

$$K_T = \lim_n \frac{\mu'_{T_n}(0)}{\sqrt{n}\sigma_{T_n}(0)} = \lim_n \frac{1}{\sigma}\,\frac{\sqrt{n_1}}{\sqrt{n}}\,\frac{\sqrt{n_2}}{\sqrt{n}} = \frac{1}{\sigma}\,\sqrt{\nu(1-\nu)} > 0 .$$

Le condizioni del Teorema 3.2.4 sono dunque soddisfatte e quindi si ha

$$\text{eff}_T = \frac{1}{\sigma}\,\sqrt{\nu(1-\nu)} .$$

Questa quantità sarà utilizzata per determinare l'efficienza asintotica relativa di questo test classico rispetto ad alcuni test “distribution-free”. ◁

• **Esempio 3.2.4.** Si consideri i due campioni casuali indipendenti $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$, tali che se $n = n_1 + n_2$, allora $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ costituisce un campione misto con funzione di ripartizione $F_n \in \mathcal{F}_{\eta,F}$, dove

$$\mathcal{F}_{\eta,F} = \{F_n : F_n \in \mathcal{V}_{\eta,F}, \mathrm{E}(X^4) = \mu_4 < \infty\}$$

Si ha $\mathrm{E}(X^r) = \eta^r \mathrm{E}(Y^r)$. Quindi, se $\mathrm{Var}(X) = \sigma^2$ si ha $\mathrm{Var}(Y) = \eta^2\sigma^2$ e se

$$\gamma^2 = \mathrm{E}(X^4) - \mathrm{Var}(X)^2 = \mu_4 - \sigma^4$$

si ha $\mathrm{E}(Y^4) - \mathrm{Var}(Y)^2 = \eta^4\gamma^2$. Si consideri il sistema di ipotesi $H_0 : \eta = 1, F \in \mathcal{F}$, contro $H_1 : \eta = \eta_i, F \in \mathcal{F}$, dove $(\eta_i)_{i\geq 1}$ è una successione di alternative tali che $\eta_i = 1 + c/\sqrt{n_i}$ con c costante. Inoltre, $\mathcal{F}$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale assolutamente continua con quarto momento finito. Siano inoltre $(n_{1i})_{i\geq 1}$ e $(n_{2i})_{i\geq 1}$ due successioni tali che $\lim_i n_{1i}/n_i = \nu$ e $\lim_i n_{2i}/n_i = 1 - \nu$ con $\nu \in (0,1)$. Si vuole determinare l'efficacia del test basato sulla statistica di Snedecor per l'omogeneità delle varianze, data da

$$T = T_n = \frac{S_y^2}{S_x^2} ,$$

dove S_x^2 e S_y^2 rappresentano le varianze campionarie corrette. Si deve dunque verificare le condizioni del Teorema 3.2.4. Innanzitutto, si sceglie $\mu_{T_n}(\eta) = \eta^2$ e $\sigma_{T_n}(\eta) = \sqrt{n/(n_1 n_2)}(\eta^2\gamma/\sigma^2)$. Per quanto riguarda la condizione *i*), si ha

$$\frac{T_{n_i} - \mu_{T_{n_i}}(\eta_i)}{\sigma_{T_{n_i}}(\eta_i)} = \frac{S_y^2/S_x^2 - \eta_i^2}{\sqrt{n_i/(n_{1i}n_{2i})}(\eta_i^2\gamma/\sigma^2)} = \frac{(S_y^2 - \eta_i^2\sigma^2) - \eta_i^2(S_x^2 - \sigma^2)}{\eta_i^2\gamma\sqrt{n_i/(n_{1i}n_{2i})}}\,\frac{\sigma^2}{S_x^2} .$$

Applicando il Corollario A.3.8 al Teorema Fondamnetale del Limite di Lindberg, si ottiene che $\sqrt{n_{2i}}(S_y^2 - \eta_i^2\sigma^2)/(\gamma\eta_i^2) \xrightarrow{d} N(0,1)$ per $i \to \infty$. Inoltre, dall'Esempio A.3.4 si ha $S_x^2 \xrightarrow{p} \sigma^2$ per $i \to \infty$ e dall'Esempio A.3.6 si ha $\sqrt{n_{1i}}(S_x^2 - \sigma^2)/\gamma \xrightarrow{d} N(0,1)$ per $i \to \infty$. Inoltre, applicando il Metodo Delta (Teorema A.3.10) alle variabili casuali indipendenti $\sqrt{n_{1i}}(S_x^2 - \sigma^2)/\gamma$ e $\sqrt{n_{2i}}(S_y^2 - \eta_i^2\sigma^2)/(\gamma\eta_i^2)$ mediante la funzione $g(x,y) = y - x$, per $i \to \infty$ si ottiene

$$\frac{(S_y^2 - \eta_i^2\sigma^2) - \eta_i^2(S_x^2 - \sigma^2)}{\eta_i^2\sqrt{\gamma^2/n_{1i} + \gamma^2/n_{2i}}} = \frac{(S_y^2 - \eta_i^2\sigma^2) - \eta_i^2(S_x^2 - \sigma^2)}{\eta_i^2\gamma\sqrt{n_i/(n_{1i}n_{2i})}} \xrightarrow{d} N(0,1) .$$

Per il Teorema di Slutski (Teorema A.3.5), combinando i precedenti risultati, per $i \to \infty$ risulta infine

$$\frac{T_{n_i} - \mu_{T_{n_i}}(\eta_i)}{\sigma_{T_{n_i}}(\eta_i)} \xrightarrow{d} N(0,1) .$$

La condizione *ii*) è verificata, in quanto con un procedimento analogo a quello considerato per verificare la condizione *i*) per $i \to \infty$ si ha

$$\frac{T_{n_i} - \mu_{T_{n_i}}(1)}{\sigma_{T_{n_i}}(1)} \xrightarrow{d} N(0,1) .$$

La condizione *iii*) è verificata in quanto dal momento che

$$\lim_i \frac{\sigma_{T_{n_i}}(\eta_i)}{\sigma_{T_{n_i}}(1)} = \lim_i \frac{\sqrt{n_i/(n_{1i}n_{2i})}(\eta_i^2\gamma/\sigma^2)}{\sqrt{n_i/(n_{1i}n_{2i})(\gamma/\sigma^2)}} = \lim_i \eta_i^2 = 1 .$$

Inoltre, si ha

$$\mu'_{T_n}(\eta) = \frac{d}{d\eta}\,\mu_{T_n}(\eta) = \frac{d}{d\eta}\,\eta^2 = 2\eta ,$$

che è una funzione continua in un intorno di 1 e $\mu'_{T_n}(1) \neq 0$ e quindi anche la condizione *iv*) è verificata. Risulta banale verificare anche la condizione *v*). Infine, per quanto riguarda la condizione *vi*), si ha

$$K_T = \lim_n \frac{\mu'_{T_n}(1)}{\sqrt{n}\sigma_{T_n}(1)} = \lim_n \frac{2\sigma^2}{\gamma}\,\frac{\sqrt{n_1}}{\sqrt{n}}\,\frac{\sqrt{n_2}}{\sqrt{n}} = \frac{2\sigma^2}{\gamma}\sqrt{\nu(1-\nu)} > 0 .$$

Le condizioni del Teorema 3.2.4 sono dunque soddisfatte e quindi si ha

$$\text{eff}_T = \frac{2\sigma^2}{\gamma}\sqrt{\nu(1-\nu)} .$$

Questa quantità sarà utilizzata per determinare l'efficienza asintotica relativa di questo test classico rispetto ai test "distribution-free". ◁

3.3. La significatività osservata. Anche se per sviluppare la teoria è necessario fissare il livello di significatività α, quando si lavora operativamente non esiste nessuna regola ragionevole per stabilirne la scelta. Questa considerazione porta al concetto di significatività osservata o valore-P.

Definizione 3.3.1. Sia $(X_1, \ldots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \geq c\}$, per un determinato valore campionario t si definisce significatività osservata la quantità

$$\alpha_{oss} = \sup_{\psi \in \Psi_0} \text{Pr}_\psi(T \geq t) .$$

Alternativamente, se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \leq c\}$, per un determinato valore campionario t si definisce significatività osservata la quantità

$$\alpha_{oss} = \sup_{\psi \in \Psi_0} \Pr_\psi(T \leq t) \, . \qquad \triangle$$

Dalla definizione si può constatare che la significatività osservata rappresenta la probabilità di ottenere, quando H_0 è vera, un valore campionario t di T estremo (nella appropriata direzione) almeno quanto quello osservato. La significatività osservata fornisce dunque una misura su quanto l'ipotesi di base risulta compatibile con i dati campionari. Una significatività osservata esigua porta a ritenere poco compatibile con i dati campionari l'ipotesi di base, mentre con una significatività osservata elevata è vera l'affermazione contraria. Di conseguenza, in una verifica di ipotesi si può semplicemente riportare la significatività osservata, oppure si può arrivare ad una decisione sull'accettazione di H_0 fissando un livello di significatività α. Se la significatività osservata è minore o uguale ad α, allora si respinge H_0, altrimenti si accetta H_0. La significatività osservata diventa in questo caso il più elevato livello di significatività per cui si accetta H_0. Tuttavia, in questo caso la significatività osservata diventa non solo uno strumento per la decisione nella verifica di ipotesi, ma anche misura quantitativa di questa decisione. Infine, la seguente definizione di significatività osservata è utile quando la statistica test T è simmetrica.

Definizione 3.3.2. Sia $(X_1, \ldots, X_n)$ un campione con funzione di ripartizione congiunta $F_n \in \mathcal{F}_\psi$ e si consideri il sistema di ipotesi $H_0 : \psi \in \Psi_0$ contro $H_1 : \psi \in \Psi \setminus \Psi_0$. Se T è una statistica simmetrica e se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \leq c_1, t \geq c_2\}$, per un determinato valore campionario t si definisce significatività osservata la quantità

$$\alpha_{oss} = 2 \min(\sup_{\psi \in \Psi_0} \Pr_\psi(T \leq t), \sup_{\psi \in \Psi_0} \Pr_\psi(T \geq t)) \, . \qquad \triangle$$

Capitolo 4

Gli intervalli di confidenza “distribution-free”

4.1. Gli intervalli di confidenza “distribution-free”. La seguente è la definizione formale di intervallo di confidenza “distribution-free”.

Definizione 4.1.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, dove θ è un parametro reale e F rappresenta la funzione di ripartizione della variabile casuale da cui proviene il campione. Supponiamo che $\theta \in \Theta$ e $F \in \mathcal{F}$, dove Θ è lo spazio parametrico e $\mathcal{F}$ è una classe di funzioni di ripartizione. Sia inoltre $P = P(X_1, \ldots, X_n; \theta)$ una quantità pivot “distribution-free” su $\mathcal{F}_{\theta,F}$, ovvero una trasformata la cui funzione di ripartizione rimane invariata per ogni $F_n \in \mathcal{F}_{\theta,F}$. Se c_1 e c_2 sono due valori tali che

$$\Pr_{\theta,F}(c_1 < P(X_1, \ldots, X_n; \theta) < c_2) = 1 - \alpha \, , \, 0 < \alpha < 1 \, , \, \theta \in \Theta \, , \, F \in \mathcal{F} \, ,$$

e se $L = L(X_1, \ldots, X_n)$ e $U = U(X_1, \ldots, X_n)$ sono statistiche tali che per ogni θ

$$\{(x_1, \ldots, x_n) : c_1 < P(x_1, \ldots, x_n; \theta) < c_2\} \Leftrightarrow \{(x_1, \ldots, x_n) : L(x_1, \ldots, x_n) < \theta < U(x_1, \ldots, x_n)\}$$

allora l'intervallo casuale (L, U) è detto intervallo di confidenza di θ “distribution-free” su $\mathcal{F}_{\theta,F}$ al livello di confidenza $(1 - \alpha)$. $\triangle$

Quando la quantità pivot è una variabile casuale discreta, allora esiste un numero finito o contabile di livelli di confidenza ottenibili, che vengono detti livelli di confidenza naturali. Nel seguito saranno considerati solo livelli di confidenza naturali. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$ e si consideri il sistema di ipotesi $H_0 : \theta = \theta_0, F \in \mathcal{F}$, contro $H_1 : \theta \neq \theta_0, F \in \mathcal{F}$, dove θ è un parametro reale e F rappresenta la funzione di ripartizione della variabile casuale da cui proviene il campione. Si abbia inoltre un test basato sulla statistica $T = T(X_1, \ldots, X_n)$ “distribution-free” su $\mathcal{F}_{\theta,F}$ al livello di significatività α e sia

$$\mathcal{X}_0 = \{(x_1, \ldots, x_n) : c_1 < T(x_1, \ldots, x_n) < c_2\}$$

la relativa regione di accettazione. Si ha

$$\Pr_{\theta_0,F}(c_1 < T < c_2) = 1 - \alpha \, ,$$

che evidenzia come la regione di accettazione $\mathcal{X}_0$ dipenda dal valore prefissato θ_0 di θ. Dunque, esiste una quantità pivot $P = P(X_1, \ldots, X_n; \theta)$ per cui si ha

$$\Pr_{\theta,F}(c_1 < P(X_1, \ldots, X_n; \theta) < c_2) = \Pr_{\theta_0,F}(c_1 < T < c_2) = 1 - \alpha \, .$$

Di conseguenza, se è possibile costruire l'intervallo di confidenza (L, U) a partire dalla quantità pivot P, allora si deve concludere che esiste una equivalenza tra la regione critica del test nel sistema di ipotesi considerato e l'intervallo di confidenza di θ. Questa considerazione consente di costruire un intervallo di confidenza per un dato parametro partendo da un opportuno sistema di ipotesi.

• **Esempio 4.1.1.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{\lambda,F}$ e si consideri il sistema di ipotesi $H_0 : \lambda = 0, F \in \mathcal{M}$, contro $H_1 : \lambda \neq 0, F \in \mathcal{M}$. Dal momento che λ rappresenta la mediana della variabile casuale da cui proviene il campione, a partire da questo sistema di ipotesi si può costruire un intervallo di confidenza per la mediana. In questo caso, un test opportuno è basato

sulla statistica $B = B(X_1, \ldots, X_n)$ dell'Esempio 2.3.1. La statistica test B è "distribution-free" su $\mathcal{M}_{0,F}$ e se è vera l'ipotesi di base si distribuisce come una variabile casuale Binomiale $Bi(n, 1/2)$. Dunque, scelto un livello di significatività α, si ha

$$\Pr_{0,F}(b_{n,\alpha/2} < B(X_1, \ldots, X_n) < n - b_{n,\alpha/2}) = 1 - \alpha \, .$$

In questo caso, la quantità pivot associata a B è data da $P = B(X_1 - \lambda, \ldots, X_n - \lambda)$, per cui tenendo presente l'Esempio 1.1.1 e la definizione della classe $\mathcal{M}_{\lambda,F}$, si ha

$$\Pr_{\lambda,F}(b_{n,\alpha/2} < B(X_1 - \lambda, \ldots, X_n - \lambda) < n - b_{n,\alpha/2}) = 1 - \alpha \, .$$

La quantità pivot $B(X_1 - \lambda, \ldots, X_n - \lambda)$ rappresenta il numero di X_i maggiori di λ per $i = 1, \ldots, n$. Se $(X_{(1)}, \ldots, X_{(n)})$ è la statistica ordinata, si ottiene che

$$\{(x_1, \ldots, x_n) : B(x_1 - \lambda, \ldots, x_n - \lambda) < n - b_{n,\alpha/2}\} \Leftrightarrow \{(x_1, \ldots, x_n) : x_{(b_{n,\alpha/2}+1)} < \lambda\} \, .$$

Analogamente

$$\{(x_1, \ldots, x_n) : b_{n,\alpha/2} < B(x_1 - \lambda, \ldots, x_n - \lambda)\} \Leftrightarrow \{(x_1, \ldots, x_n) : \lambda < x_{(n-b_{n,\alpha/2})}\} \, .$$

Dunque, $L = X_{(b_{n,\alpha/2}+1)}$ e $U = X_{(n-b_{n,\alpha/2})}$, ovvero $(X_{(b_{n,\alpha/2}+1)}, X_{(n-b_{n,\alpha/2})})$, è un intervallo di confidenza della mediana λ "distribution-free" su $\mathcal{M}_{\lambda,F}$ al livello di confidenza $(1 - \alpha)$. ◁

Il seguente teorema permette di costruire intervalli di confidenza di un parametro θ "distribution-free" su una classe $\mathcal{F}_{\theta,F}$ se la statistica test su cui si basa il sistema di ipotesi associato ha una particolare struttura.

Teorema 4.1.2. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, dove $\theta \in \Theta$ è un parametro reale e $F \in \mathcal{F}$ rappresenta la funzione di ripartizione della variabile casuale da cui proviene il campione. Si consideri inoltre l'insieme di statistiche $V_i = V_i(X_1, \ldots, X_n)$ per $i = 1, \ldots, k$, tali che*

$$T = T(V_1, \ldots, V_k) = \sum_{i=1}^{k} \mathbf{1}_{(0,\infty)}(V_i)$$

è una statistica "distribution-free" su $\mathcal{F}_{0,F}$. Sia inoltre

$$\Pr_{0,F}(c_1 < T(V_1, \ldots, V_k) < c_2) = \Pr_{\theta,F}(c_1 < T(V_1 - \theta, \ldots, V_k - \theta) < c_2) = 1 - \alpha \, .$$

Se $(V_{(1)}, \ldots, V_{(k)})$ è la statistica ordinata relativa al vettore di statistiche $(V_1, \ldots, V_k)$, allora $(V_{(k+1-c_2)}, V_{(k-c_1)})$ è un intervallo di confidenza di θ "distribution-free" su $\mathcal{F}_{\theta,F}$ al livello di confidenza $(1 - \alpha)$.

Dimostrazione. Si noti che $T(V_1 - \theta, \ldots, V_k - \theta)$ è una quantità pivot. Inoltre, dal momento che $T(V_1 - \theta, \ldots, V_k - \theta)$ rappresenta il numero di V_i maggiori di θ per $i = 1, \ldots, k$, si ha che

$$\{(v_1, \ldots, v_k) : T(v_1 - \theta, \ldots, v_k - \theta) < c_2\} \Leftrightarrow \{(v_1, \ldots, v_k) : v_{(k+1-c_2)} < \theta\} \, .$$

Analogamente

$$\{(v_1, \ldots, v_k) : c_1 < T(v_1 - \theta, \ldots, v_k - \theta)\} \Leftrightarrow \{(v_1, \ldots, v_k) : \theta < v_{(k-c_1)}\} \, .$$

Di conseguenza, in base alla Definizione 4.1.1, si ha che $(V_{(k+1-c_2)}, V_{(k-c_1)})$ è un intervallo di confidenza di θ "distribution-free" su $\mathcal{F}_{\theta,F}$ al livello di confidenza $(1 - \alpha)$. □

• **Esempio 4.1.2.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$. In questo caso, λ rappresenta la mediana della variabile casuale da cui proviene il campione. Si consideri inoltre la statistica W^+ "distribution-free" su $\mathcal{S}_{0,F}$ dell'Esempio 2.5.1. Se $(X_{(1)}, \ldots, X_{(n)})$ è la statistica ordinata, allora si ha $\mathbf{1}_{(0,\infty)}((X_{(i)} + X_{(j)})/2) = 1$ se e solo se $X_{(i)} > 0$ e $|X_{(i)}| > |X_{(j)}|$ per $j < i$, mentre si

ha $\mathbf{1}_{(0,\infty)}((X_{(i)} + X_{(i)})/2) = 1$ se e solo se $X_{(i)} > 0$. Di conseguenza, le variabili casuali $Z_i R_i^+$ possono essere espresse come

$$Z_i R_i^+ = \sum_{j=1}^{i} \mathbf{1}_{(0,\infty)}((X_{(i)} + X_{(j)})/2) \, , \, i = 1, \ldots, n \, ,$$

per cui una rappresentazione alternativa di W^+ è data da

$$W^+ = \sum_{i=1}^{n} \sum_{j=1}^{i} \mathbf{1}_{(0,\infty)}((X_{(i)} + X_{(j)})/2) = \sum_{i=1}^{n} \sum_{j=1}^{i} \mathbf{1}_{(0,\infty)}((X_i + X_j)/2) \, .$$

Dunque, W^+ è basata sulle $k = n(n+1)/2$ statistiche $V_{ij} = (X_i + X_j)/2$ per $i \geq j = 1, \ldots, n$, dette anche medie di Walsh. Se si indicizzano di nuovo le statistiche V_{ij} denotandole con W_i per $i = 1, \ldots, k$, si ottiene

$$W^+(W_1, \ldots, W_k) = \sum_{i=1}^{k} \mathbf{1}_{(0,\infty)}(W_i) \, ,$$

ovvero la statistica W^+ può essere espressa nella forma richiesta dal Teorema 4.1.2. Inoltre, dal momento che

$$V_{ij} - \lambda = \frac{1}{2}(X_i + X_j) - \lambda = \frac{1}{2}(X_i - \lambda + X_j - \lambda) \, , \, i \geq j = 1, \ldots, n \, ,$$

tenendo presente l'Esempio 1.1.1 e la definizione della classe $\mathcal{S}_{\lambda,F}$, si ha

$$\begin{aligned} \mathrm{Pr}_{0,F}(w_{\alpha/2} < W^+(W_1, \ldots, W_k) < k - w_{\alpha/2}) = \\ = \mathrm{Pr}_{\lambda,F}(w_{\alpha/2} < T(W_1 - \lambda, \ldots, W_k - \lambda) < k - w_{\alpha/2}) = 1 - \alpha \, , \end{aligned}$$

dove w_α rappresenta il quantile di ordine α della distribuzione di W^+ e dove si è tenuto presente che W^+ è simmetrica rispetto a $k/2$ (vedi Esempio 2.5.1). Dunque, se $(W_{(1)}, \ldots, W_{(k)})$ rappresenta la statistica ordinata relativo al vettore di statistiche $(W_1, \ldots, W_k)$, risulta che $(W_{(w_{\alpha/2}+1)}, W_{(k-w_{\alpha/2})})$ è un intervallo di confidenza della mediana λ "distribution-free" su $\mathcal{S}_{\lambda,F}$ al livello di confidenza $(1 - \alpha)$. Questo intervallo di confidenza è "distribution-free" su una classe più ristretta di quello ottenuto nell'Esempio 4.1.1, dal momento che $\mathcal{S}_{\lambda,F} \subset \mathcal{M}_{\lambda,F}$. ◁

• **Esempio 4.1.3.** Si consideri i due campioni casuali indipendenti $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$, tali che se $n = n_1 + n_2$, allora $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ costituisce un campione misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$. Si consideri inoltre la statistica W "distribution-free" su $\mathcal{L}_{0,F}$ dell'Esempio 2.4.1. Si noti che le classi $\mathcal{L}_{0,F}$ e $\mathcal{C}_F$ coincidono. Se $(X_{(1)}, \ldots, X_{(n_1)})$ e $(Y_{(1)}, \ldots, Y_{(n_2)})$ sono le statistiche ordinate relative ai due campioni, si ha (vedi Esempio 2.4.1)

$$R_{(i)} = \sum_{j=1}^{n_2} \mathbf{1}_{(0,\infty)}(X_{(i)} - Y_{(j)}) + i \, , \, i = 1, \ldots, n_1 \, ,$$

per cui una rappresentazione alternativa di W risulta

$$W = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{(0,\infty)}(X_{(i)} - Y_{(j)}) + \frac{n_1(n_1+1)}{2} \, .$$

Dalla precedente espressione si ha che la statistica W è equivalente alla statistica U data da

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{(0,\infty)}(X_{(i)} - Y_{(j)}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{(0,\infty)}(X_i - Y_j) \, .$$

La statistica U ha supporto $\{0, 1, \ldots, n_1 n_2\}$ ed è simmetrica rispetto a $n_1 n_2/2$. Inoltre, U è basata sulle $k = n_1 n_2$ statistiche $V_{ij} = X_i - Y_j$ per $i = 1, \ldots, n_1, j = 1, \ldots, n_2$. Se si indicizzano di nuovo le statistiche V_{ij} denotandole con D_i per $i = 1, \ldots, k$, si ha

$$U(D_1, \ldots, D_k) = \sum_{i=1}^{k} \mathbf{1}_{(0,\infty)}(D_i) \,,$$

ovvero la statistica U può essere espressa nella forma richiesta dal Teorema 4.1.2. Inoltre, dal momento che

$$V_{ij} - \Delta = X_i - (Y_j + \Delta) \,, i = 1, \ldots, n_1 \,, j = 1, \ldots, n_2 \,,$$

allora, tenendo presente l'Esempio 1.1.1 e la definizione della classe $\mathcal{L}_{\Delta,F}$, si ha

$$\begin{aligned} \Pr_{0,F}(u_{\alpha/2} < U(D_1, \ldots, D_k) < k - u_{\alpha/2}) = \\ = \Pr_{\Delta,F}(u_{\alpha/2} < U(D_1 - \Delta, \ldots, D_k - \Delta) < k - u_{\alpha/2}) = 1 - \alpha \,, \end{aligned}$$

dove u_α rappresenta il quantile di ordine α della distribuzione di U e dove si è tenuto presente che U è simmetrica rispetto a $k/2$. Dunque, se $(D_{(1)}, \ldots, D_{(k)})$ rappresenta la statistica ordinata relativa al vettore di statistiche $(D_1, \ldots, D_k)$, allora $(D_{(u_{\alpha/2}+1)}, D_{(k-u_{\alpha/2})})$ è un intervallo di confidenza per Δ “distribution-free” su $\mathcal{L}_{\Delta,F}$ al livello di confidenza $(1 - \alpha)$. ◁

4.2. Gli intervalli di confidenza “distribution-free” per grandi campioni. Di seguito viene definito il concetto di intervallo di confidenza “distribution-free” per grandi campioni.

Definizione 4.2.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\theta,F}$, dove θ è un parametro reale e F rappresenta la funzione di ripartizione della variabile casuale da cui proviene il campione. Supponiamo inoltre che $\theta \in \Theta$ e $F \in \mathcal{F}$, dove Θ è lo spazio parametrico, mentre $\mathcal{F}$ è una classe di funzioni di ripartizione. Sia $P_n = P_n(X_1, \ldots, X_n; \theta)$ una quantità pivot “distribution-free” per grandi campioni su $\mathcal{F}_{\theta,F}$, ovvero una trasformata tale che $P_n \xrightarrow{d} V$ con $n \to \infty$ per ogni $F_n \in \mathcal{F}_{\theta,F}$. Se c_1 e c_2 sono due valori tali che

$$\lim_n \Pr_{\theta,F}(c_1 < P_n(X_1, \ldots, X_n; \theta) < c_2) = 1 - \alpha \,, 0 < \alpha < 1 \,, \theta \in \Theta \,, F \in \mathcal{F} \,,$$

e se $L_n = L_n(X_1, \ldots, X_n)$ e $U_n = U_n(X_1, \ldots, X_n)$ sono statistiche tali che per ogni θ

$$\{(x_1, \ldots, x_n) : c_1 < P_n(x_1, \ldots, x_n; \theta) < c_2\} \Leftrightarrow \{(x_1, \ldots, x_n) : L_n(x_1, \ldots, x_n) < \theta < U_n(x_1, \ldots, x_n)\}$$

allora l'intervallo casuale (L_n, U_n) è detto intervallo di confidenza di θ “distribution-free” per grandi campioni su $\mathcal{F}_{\theta,F}$ al livello di confidenza $(1 - \alpha)$. △

• **Esempio 4.2.1.** Si consideri un campione casuale $(X_1, \ldots, X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\lambda,F}$, dove la classe $\mathcal{F}_{\lambda,F}$ è definita nell'Esempio 3.2.1. Sia inoltre $P_n = \sqrt{n}(\overline{X} - \lambda)/S$ la quantità pivot. Dal momento che per ogni $F_n \in \mathcal{F}_{\lambda,F}$ si ha $P_n \xrightarrow{d} N(0, 1)$, allora P_n è una quantità pivot “distribution-free” per grandi campioni su $\mathcal{F}_{\lambda,F}$. Scelto dunque un livello di confidenza pari a $(1 - \alpha)$, si ha

$$\lim_n \Pr_{\lambda,F}(-z_{1-\alpha/2} < \frac{\sqrt{n}(\overline{X} - \lambda)}{S} < z_{1-\alpha/2}) = 1 - \alpha \,.$$

Dal momento che l'insieme

$$\{(x_1, \ldots, x_n) : -z_{1-\alpha/2} < \frac{\sqrt{n}(\overline{x} - \lambda)}{s} < z_{1-\alpha/2}\}$$

è equivalente a

$$\{(x_1,\ldots,x_n) : \overline{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} < \lambda < \overline{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\}\,,$$

allora $L_n = \bar{X} - z_{1-\alpha/2}S/\sqrt{n}$ e $U_n = \bar{X} + z_{1-\alpha/2}S/\sqrt{n}$, e $(\bar{X} - z_{1-\alpha/2}S/\sqrt{n},\ \bar{X} + z_{1-\alpha/2}S/\sqrt{n})$ è un intervallo di confidenza di λ "distribution-free" per grandi campioni su $\mathcal{F}_{\lambda,F}$ al livello di confidenza $(1-\alpha)$. ◁

• **Esempio 4.2.2.** Si consideri un campione casuale $(X_1,\ldots,X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{F}_{\sigma,F}$, dove

$$\mathcal{F}_{\sigma,F} = \{F_n : F_n \in \mathcal{C}_F, \mathrm{Var}(X) = \sigma^2 < \infty, \mathrm{E}(X^4) = \mu_4 < \infty\}\,.$$

Dall'Esempio A.3.6 si ha $\sqrt{n}(S^2-\sigma^2)/\gamma \xrightarrow{d} N(0,1)$, dove $\gamma^2 = \mu_4 - \sigma^4$. Inoltre, se M_4 rappresenta il quarto momento centrale campionario, allora applicando la Legge Debole dei Grandi Numeri di Khintchine (Teorema A.3.1) e il Teorema di Sverdrup (Teorema A.3.4), si può dimostrare che $G^2 = M_4 - S^4 \xrightarrow{p} \gamma^2$. Applicando il Teorema di Slutsky risulta $P_n = \sqrt{n}(S^2-\sigma^2)/G \xrightarrow{d} N(0,1)$ per ogni $F_n \in \mathcal{F}_{\sigma,F}$, ovvero la quantità pivot P_n è "distribution-free" per grandi campioni su $\mathcal{F}_{\sigma,F}$. Scelto dunque un livello di confidenza pari a $(1-\alpha)$, si ha

$$\lim_n \mathrm{Pr}_{\sigma,F}(-z_{1-\alpha/2} < \frac{\sqrt{n}(S^2-\sigma^2)}{G} < z_{1-\alpha/2}) = 1-\alpha\,.$$

Dal momento che l'insieme

$$\{(x_1,\ldots,x_n) : -z_{1-\alpha/2} < \frac{\sqrt{n}(s^2-\sigma^2)}{g} < z_{1-\alpha/2}\}$$

è equivalente a

$$\{(x_1,\ldots,x_n) : s^2 - z_{1-\alpha/2}\,\frac{g}{\sqrt{n}} < \sigma^2 < s^2 + z_{1-\alpha/2}\,\frac{g}{\sqrt{n}}\}\,,$$

si ha $L_n = S^2 - z_{1-\alpha/2}G/\sqrt{n}$, $U_n = S^2 + z_{1-\alpha/2}G/\sqrt{n}$ e $(S^2 - z_{1-\alpha/2}G/\sqrt{n}, S^2 + z_{1-\alpha/2}G/\sqrt{n})$ è un intervallo di confidenza di σ^2 "distribution-free" per grandi campioni su $\mathcal{F}_{\sigma,F}$ al livello di confidenza $(1-\alpha)$.

◁

• **Esempio 4.2.3.** Si consideri un campione casuale $(X_1,\ldots,X_n)$ con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{\lambda,F}$. Sia inoltre

$$P_n = B_n(X_1-\lambda,\ldots,X_n-\lambda) = \sum_{i=1}^{n} \mathbf{1}_{(0,\infty)}(X_i-\lambda)$$

la quantità pivot dell'Esempio 4.1.1. Dal Teorema Fondamentale Classico del Limite (Teorema A.3.6) si ha

$$\frac{P_n - n/2}{\sqrt{n/4}} \xrightarrow{d} N(0,1)$$

per ogni $F_n \in \mathcal{M}_{\lambda,F}$, ovvero la quantità pivot P_n è "distribution-free" per grandi campioni su $\mathcal{M}_{\lambda,F}$. Scelto dunque un livello di confidenza pari a $(1-\alpha)$, si ha

$$\lim_n \mathrm{Pr}_{\lambda,F}(-z_{1-\alpha/2} < \frac{P_n - n/2}{\sqrt{n}/2} < z_{1-\alpha/2}) = 1-\alpha\,.$$

Tenendo presente che $B_n(X_1-\lambda,\ldots,X_n-\lambda)$ può assumere solo valori interi, posto $l = \lfloor n/2 - z_{1-\alpha/2}\sqrt{n/4}\rfloor$, dove $\lfloor\cdot\rfloor$ rappresenta la funzione troncamento, analogamente all'Esempio 4.1.1 si ha che l'insieme

$$\{(x_1,\ldots,x_n):\ -z_{1-\alpha/2} < \frac{B_n(x_1-\lambda,\ldots,x_n-\lambda)-n/2}{\sqrt{n}/2} < z_{1-\alpha/2}\}$$

è equivalente a

$$\{(x_1,\ldots,x_n): \frac{n}{2} - z_{1-\alpha/2}\frac{\sqrt{n}}{2} < B_n(x_1-\lambda,\ldots,x_n-\lambda) < \frac{n}{2} + z_{1-\alpha/2}\frac{\sqrt{n}}{2}\}\,,$$

da cui

$$\{(x_1,\ldots,x_n): l < B_n(x_1-\lambda,\ldots,x_n-\lambda) < n-l\} \Leftrightarrow \{x_{(l+1)} < \lambda < x_{(n-l)}\}\,.$$

Di conseguenza, $(X_{(l+1)}, X_{(n-l)})$ è un intervallo di confidenza di λ “distribution-free” per grandi campioni su $\mathcal{M}_{\lambda,F}$ al livello di confidenza $(1-\alpha)$. ◁

Capitolo 5

I test basati su statistiche lineari dei ranghi con segno

5.1. I test basati su statistiche lineari dei ranghi con segno. Sia $(X_1, \dots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$ e si consideri il sistema di ipotesi $H_0 : \lambda = 0, F \in \mathcal{S}$ contro una alternativa bilaterale $H_1 : \lambda \neq 0, F \in \mathcal{S}$, o direzionale $H_1 : \lambda > 0\,(\lambda < 0), F \in \mathcal{S}$. Questo sistema di ipotesi è del tutto generale. Infatti, se l'ipotesi di base è data da $H_0 : \lambda = \lambda_0, F \in \mathcal{S}$, allora si ritorna al sistema di ipotesi precedente semplicemente considerando il campione trasformato $(X_1 - \lambda_0, \dots, X_n - \lambda_0)$. Una classe di statistiche test "distribution-free" opportuna in questo sistema di ipotesi è quella delle statistiche lineari dei ranghi con segno.

Definizione 5.1.1. Sia $(X_1, \dots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$ e siano $(Z_1, \dots, Z_n)$ e $(R_1^+, \dots, R_n^+)$ i relativi vettori dei segni e dei ranghi dei valori assoluti. Siano inoltre i punteggi un insieme di valori $a(i)$ per $i = 1, \dots, n$, tali che

$$0 \leq a(1) \leq \dots \leq a(n)\,,\, a(n) > 0\,.$$

Una statistica del tipo

$$T^+ = \sum_{i=1}^{n} Z_i a(R_i^+)\,,$$

è detta statistica lineare dei ranghi con segno. △

In base al Corollario 2.5.5 la statistica T^+ è "distribution-free" sulla classe $\mathcal{S}_{0,F}$, ovvero fornisce un test "distribution-free" per il sistema di ipotesi considerato. Inoltre, la statistica test T^+ è sensibile a variazioni nel parametro di posizione, in quanto se non è vera l'ipotesi di base T^+ tende ad assumere o valori piccoli o valori elevati. La scelta dei punteggi influisce sul peso che si vuole assegnare ad ogni singolo rango dei valori assoluti.

• **Esempio 5.1.1.** Se i punteggi sono scelti come $a(i) = i$ per $i = 1, \dots, n$, si ottiene la cosiddetta statistica di Wilcoxon (introdotta nell'Esempio 2.5.1), ovvero

$$W^+ = \sum_{i=1}^{n} Z_i R_i^+\,.$$

Se i punteggi sono scelti come $a(i) = 1$ per $i = 1, \dots, n$, si ottiene la cosiddetta statistica dei segni (introdotta nell'Esempio 2.3.1), ovvero

$$B = \sum_{i=1}^{n} Z_i\,.$$

Per la statistica W^+ i punteggi vengono scelti come funzione lineare crescente dei ranghi dei valori assoluti, ovvero in modo da assegnare un peso maggiore ai ranghi assoluti più elevati. Al contrario, per la statistica B i punteggi sono costanti per tutti i ranghi dei valori assoluti. ◁

Il seguente teorema permette di ottenere una importante equivalenza in distribuzione per le statistiche lineari dei ranghi con segno quando è vera l'ipotesi di base.

Teorema 5.1.2. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, per una statistica lineare dei ranghi con segno T^+ si ha*

$$T^+ = \sum_{i=1}^{n} Z_i a(R_i^+) \stackrel{d}{=} \sum_{i=1}^{n} Z_i a(i) \, .$$

Dimostrazione. Dal momento che per il Teorema 2.5.4 i vettori di statistiche $(Z_1, \ldots, Z_n)$ e $(R_1^+, \ldots, R_n^+)$ sono indipendenti, allora per ogni $(r_1, \ldots, r_n) \in \mathcal{R}_n$ si ha

$$(T^+ \mid R_1^+ = r_1, \ldots, R_n^+ = r_n) = (\sum_{i=1}^{n} Z_i a(R_i^+) \mid R_1^+ = r_1, \ldots, R_n^+ = r_n) = \sum_{i=1}^{n} Z_i a(r_i) \, .$$

Se d_i è la posizione dell'intero i nel vettore $(r_1, \ldots, r_n)$, allora si ha

$$\sum_{i=1}^{n} Z_i a(r_i) = \sum_{j=1}^{n} Z_{d_j} a(j) \, .$$

Tuttavia, dal momento che il vettore dei segni $(Z_1, \ldots, Z_n)$ ha componenti indipendenti per il Teorema 2.3.2, allora dal Teorema 1.1.3 risulta $(Z_1, \ldots, Z_n) \stackrel{d}{=} (Z_{d_1}, \ldots, Z_{d_n})$ essendo $(d_1, \ldots, d_n)$ una permutazione di $(1, \ldots, n)$. Dunque, si ha

$$(T^+ \mid R_1^+ = r_1, \ldots, R_n^+ = r_n) = \sum_{j=1}^{n} Z_{d_j} a(j) \stackrel{d}{=} \sum_{i=1}^{n} Z_i a(i) \, .$$

Questa equivalenza in distribuzione vale per ogni $(r_1, \ldots, r_n) \in \mathcal{R}_n$ e il teorema è dimostrato. □

• **Esempio 5.1.2.** Si consideri la statistica W^+ dell'Esempio 5.1.1. Dal Teorema 5.1.2 si verifica che

$$W^+ = \sum_{i=1}^{n} Z_i R_i^+ \stackrel{d}{=} \sum_{i=1}^{n} i Z_i \, ,$$

ovvero, tenendo presente il Teorema 2.3.2, la statistica W^+ è distribuita come una combinazione lineare di variabili casuali distribuite come Binomiali $Bi(1, 1/2)$, i cui i pesi sono dati da $(1, \ldots, n)$. ◁

Il seguente teorema fornisce la media e la varianza di una statistica lineare dei ranghi con segno quando è vera l'ipotesi di base.

Teorema 5.1.3. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, allora per una statistica lineare dei ranghi con segno T^+ si ha*

$$\mathrm{E}(T^+) = \frac{1}{2} \sum_{i=1}^{n} a(i) \, , \, \mathrm{Var}(T^+) = \frac{1}{4} \sum_{i=1}^{n} a(i)^2 \, .$$

Dimostrazione. Dal momento che il Teorema 2.3.2 implica che $\mathrm{E}(Z_i) = 1/2$, dal Teorema 5.1.2 si ha

$$\mathrm{E}(T^+) = \mathrm{E}(\sum_{i=1}^{n} Z_i a(R_i^+)) = \mathrm{E}(\sum_{i=1}^{n} Z_i a(i)) = \sum_{i=1}^{n} \mathrm{E}(Z_i) a(i) = \frac{1}{2} \sum_{i=1}^{n} a(i) \, .$$

Inoltre, il Teorema 2.3.2 implica che $\mathrm{Var}(Z_i) = 1/4$ per $i = 1, \ldots, n$, e tenendo presente che $(Z_1, \ldots, Z_n)$ ha componenti indipendenti, dal Teorema 5.1.2 si ha

$$\mathrm{Var}(T^+) = \mathrm{Var}(\sum_{i=1}^{n} Z_i a(R_i^+)) = \mathrm{Var}(\sum_{i=1}^{n} Z_i a(i)) = \sum_{i=1}^{n} \mathrm{Var}(Z_i) a(i)^2 = \frac{1}{4}\sum_{i=1}^{n} a(i)^2 \,. \qquad \square$$

• **Esempio 5.1.3.** Si consideri la statistica W^+ dell'Esempio 5.1.1. Tenendo presente il Teorema A.2.1, dal Teorema 5.1.3 si verifica che

$$\mathrm{E}(W^+) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}$$

e

$$\mathrm{Var}(W^+) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24} \,. \qquad \triangleleft$$

Nel seguente teorema si dimostra la simmetria rispetto alla media di ogni statistica lineare dei ranghi con segno quando è vera l'ipotesi di base.

Teorema 5.1.4. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, allora una statistica lineare dei ranghi con segno T^+ è simmetrica rispetto a* $\mathrm{E}(T^+)$.

Dimostrazione. Il Teorema 2.3.2 implica la simmetria di ogni statistica segno Z_i per $i = 1, \ldots, n$ rispetto alla media, ovvero tenendo presente il Teorema 1.2.2 risulta

$$(Z_1 - 1/2, \ldots, Z_n - 1/2) \stackrel{d}{=} (1/2 - Z_1, \ldots, 1/2 - Z_n) \,.$$

Tenendo presente l'Esempio 1.1.6, dalla precedente relazione si ha

$$(Z_1, \ldots, Z_n) \stackrel{d}{=} (1 - Z_1, \ldots, 1 - Z_n) \,.$$

Di conseguenza, dal Teorema 5.1.2 segue che

$$\begin{aligned} T^+ - \mathrm{E}(T^+) &\stackrel{d}{=} \sum_{i=1}^{n} Z_i a(i) - \mathrm{E}(T^+) \stackrel{d}{=} \sum_{i=1}^{n} (1 - Z_i) a(i) - \mathrm{E}(T^+) \\ &= \sum_{i=1}^{n} a(i) - \sum_{i=1}^{n} Z_i a(i) - \mathrm{E}(T^+) \stackrel{d}{=} 2\mathrm{E}(T^+) - T^+ - \mathrm{E}(T^+) = \mathrm{E}(T^+) - T^+ \,. \end{aligned}$$

Il Teorema 1.2.2 permette infine di concludere che T^+ è simmetrica rispetto a $\mathrm{E}(T^+)$. $\square$

Il seguente teorema fornisce la funzione generatrice di probabilità di una statistica lineare dei ranghi con segno nel caso che i punteggi siano valori interi.

Teorema 5.1.5. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, allora la funzione generatrice di probabilità di una statistica lineare dei ranghi con segno T^+ risulta*

$$L_{T^+}(t) = 2^{-n} \prod_{i=1}^{n} (1 + t^{a(i)}) \,, \; |t| < 1 \,.$$

Dimostrazione. Tenendo presente il Teorema 5.1.2 e il Teorema 2.3.2, dalla definizione di funzione generatrice di probabilità si ha

$$L_{T^+}(t) = \mathrm{E}(t^{T^+}) = \mathrm{E}(\prod_{i=1}^{n} t^{a(i)Z_i}) = \prod_{i=1}^{n} \mathrm{E}(t^{a(i)Z_i}) = \prod_{i=1}^{n} 2^{-1}(1 + t^{a(i)}) = 2^{-n} \prod_{i=1}^{n} (1 + t^{a(i)}) \,. \qquad \square$$

Quando i punteggi non sono interi allora per impiegare il precedente teorema è sufficiente determinare una trasformata biunivoca che discretizzi i punteggi originali. Dal momento che la distribuzione di una statistica lineare dei ranghi con segno sotto l'ipotesi di base $H_0 : \lambda = 0, F \in \mathcal{S}$, si può ottenere mediante il Teorema 5.1.5, allora si può determinare le appropriate regioni critiche del test. Se l'alternativa è bilaterale, ovvero $H_1 : \lambda \neq 0, F \in \mathcal{S}$, allora si ha $\Pr(Z_i = 1) \neq 1/2$ e quindi si respingere l'ipotesi di base per determinazioni sia troppo elevate che troppo piccole di T^+. Fissato quindi un livello di significatività α, poichè la distribuzione della statistica T^+ è simmetrica se è vera l'ipotesi di base, allora la regione critica è data dall'insieme

$$\mathcal{T}_1 = \{t^+ : t^+ \leq t^+_{n,\alpha/2}, t^+ \geq t^+_{n,1-\alpha/2}\} ,$$

dove $t^+_{n,\alpha}$ rappresenta il quantile di ordine α della distribuzione di T^+ per una numerosità campionaria pari a n. Data la simmetria di T^+, si ha $t^+_{n,1-\alpha/2} + 1 = \sum_{i=1}^n a(i) - t^+_{n,\alpha/2}$. Se l'alternativa è direzionale del tipo $H_1 : \lambda > 0, F \in \mathcal{S}$, allora si ha $\Pr(Z_i = 1) > 1/2$ e quindi si respinge l'ipotesi di base per determinazioni troppo elevate di T^+. Fissato quindi un livello di significatività α, si ha la seguente regione critica

$$\mathcal{T}_1 = \{t^+ : t^+ \geq t^+_{n,1-\alpha}\} .$$

Al contrario, se l'alternativa è direzionale del tipo $H_1 : \lambda < 0, F \in \mathcal{S}$, allora $\Pr(Z_i = 1) < 1/2$ e quindi si respinge l'ipotesi di base per determinazioni troppo piccole di T^+. Fissato quindi un livello di significatività α, si ha la seguente regione critica

$$\mathcal{T}_1 = \{t^+ : t^+ \leq t^+_{n,\alpha}\} .$$

Il test basato sulla T^+ per i precedenti sistemi di ipotesi è corretto al livello di significatività α. Infatti, dal momento che si può dimostrare che $P_{T^+}(\lambda, F) = \Pr_{\lambda,F}(T^+ \in \mathcal{T}_1)$ è una funzione monotona crescente per $\lambda > 0$ e monotona decrescente per $\lambda < 0$ per ogni $F \in \mathcal{S}$, allora risulta $P_{T^+}(\lambda, F) > \alpha$, ovvero il test è corretto.

5.2. I test basati su statistiche lineari dei ranghi con segno localmente più potenti. Si desidera determinare la scelta ottima dei punteggi della statistica test per verificare il ipotesi del tipo $H_0 : \lambda = 0, F = F_0$, contro un'alternativa direzionale $H_1 : \lambda > 0\,(\lambda < 0), F = F_0$, ovvero quando si verifica ipotesi sul parametro di posizione fissando la struttura della funzione di ripartizione F. Dal momento che non è possibile determinare una scelta ottima dei punteggi per ottenere un test uniformemente più potente, allora è utile introdurre un concetto di test localmente più potente.

Definizione 5.2.1. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$ e si consideri il sistema di ipotesi $H_0 : \lambda = 0, F = F_0$, contro $H_1 : \lambda > 0\,(\lambda < 0), F = F_0$, dove $F_0 \in \mathcal{S}$. Il test basato sulla statistica lineare dei ranghi con segno T^+_* è detto localmente più potente se esiste un $\epsilon > 0$ tale che per ogni livello di significatività naturale si ha

$$P_{T^+_*}(\lambda) \geq P_{T^+}(\lambda) , 0 < \lambda < \epsilon ,$$

per ogni statistica lineare dei ranghi con segno T^+. △

Il prossimo teorema fornisce la scelta ottima dei punteggi per la costruzione del test localmente più potente per verificare ipotesi sul parametro di posizione. L'utilità di questo teorema consiste solamente nell'evidenziare la struttura ottima dei punteggi al variare della funzione di ripartizione, dal momento che questa non è mai nota in pratica.

Teorema 5.2.2. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$ e sia f la funzione di densità relativa a F. Si assuma inoltre che f' esista, sia assolutamente continua e che*

$$\int_{\mathbb{R}} |f'(x)|\, dx < \infty .$$

Il test localmente più potente per verificare il sistema di ipotesi $H_0 : \lambda = 0, F = F_0$, *contro* $H_1 : \lambda > 0$ *(*$\lambda < 0$*),* $F = F_0$, *è basato sulla statistica lineare dei ranghi con segno*

$$T_*^+ = \sum_{i=1}^{n} Z_i a_*(R_i^+) \, ,$$

dove

$$a_*(i) = \mathrm{E}(- \frac{f'(Y_{(i)})}{f(Y_{(i)})}) \, , \, i = 1, \ldots, n \, ,$$

e $(Y_{(1)}, \ldots, Y_{(n)})$ *è la statistica ordinata relativa al campione casuale* $(|X_1|, \ldots, |X_n|)$.

Dimostrazione. Si veda Hettmansperger e McKean (1998). □

Supponiamo che $F(x) = G(x/\delta)$, ovvero $f(x) = (1/\delta)g(x/\delta)$, dove G e g sono rispettivamente la funzione di ripartizione e la funzione di densità di una variabile casuale standard. Se $f_{(i)}$ rappresenta la funzione di densità di $Y_{(i)}$, allora risulta $f_{(i)}(x) = (1/\delta)g_{(i)}(x/\delta)$, dove $g_{(i)}$ è la funzione di densità della variabile casuale standard ordinata $V_{(i)} = Y_{(i)}/\delta$ per $i = 1, \ldots, n$. Dunque, si ha

$$a_*(i) = \mathrm{E}(- \frac{f'(Y_{(i)})}{f(Y_{(i)})}) = \int_{\mathbb{R}} - \frac{f'(x)}{f(x)} \, f_{(i)}(x) \, dx = \frac{1}{\delta} \int_{\mathbb{R}} - \frac{g'(x)}{g(x)} \, g_{(i)}(x) \, dx = \frac{1}{\delta} \, \mathrm{E}(- \frac{g'(V_{(i)})}{g(V_{(i)})}) \, .$$

Di conseguenza, se i punteggi ottimi sono ottenuti sulla base della distribuzione standardizzata, allora la relativa statistica test ottima è data da δT_*^+, che è una statistica test equivalente a T_*^+. Dal momento che la statistica test ottima non dipende dal parametro di scala della distribuzione, allora i punteggi ottimi possono essere calcolati semplicemente a partire dalla distribuzione standard.

• **Esempio 5.2.1.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$, dove F è la funzione di ripartizione di una distribuzione Normale $N(0, 1)$. Si può verificare che le condizioni del Teorema 5.2.2 sono soddisfatte. Inoltre, dal momento che risulta $f'(x) = \, - xf(x)$, allora si ha

$$- \frac{f'(x)}{f(x)} = x \, .$$

Di conseguenza, la scelta ottimale dei punteggi è data da

$$a_*(i) = \mathrm{E}(Y_{(i)}) \, , \, i = 1, \ldots, n \, ,$$

dove $(Y_{(1)}, \ldots, Y_{(n)})$ è la statistica ordinata relativa ai valori assoluti di un campione casuale da una distribuzione Normale $N(0, 1)$. ◁

• **Esempio 5.2.2.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$, dove F è la funzione di ripartizione di una distribuzione Logistica $Lo(0, 1)$. Si può verificare che le condizioni del Teorema 5.2.2 sono soddisfatte. Inoltre, si ha $f(x) = F(x) - F(x)^2$, ovvero risulta $f'(x) = \, - f(x)(2F(x) - 1)$, da cui

$$- \frac{f'(x)}{f(x)} = 2F(x) - 1 \, .$$

Di conseguenza, la scelta ottimale dei punteggi è data da

$$a_*(i) = \mathrm{E}(2F(Y_{(i)}) - 1) \, , \, i = 1, \ldots, n \, ,$$

dove $(Y_{(1)}, \ldots, Y_{(n)})$ è la statistica ordinata relativa ai valori assoluti di un campione casuale da una distribuzione $Lo(0, 1)$. Dal momento che dal Teorema dell'Integrale di Probabilità (vedi Feller, 1971) si ha

$2F(Y_{(i)}) - 1 \stackrel{d}{=} U_{(i)}$ per $i = 1, \ldots, n$, dove $(U_{(1)}, \ldots, U_{(n)})$ rappresenta la statistica ordinata relativa ad un campione casuale da una distribuzione Uniforme $U(0,1)$, allora risulta

$$a_*(i) = \mathrm{E}(U_{(i)}) = \frac{i}{n+1}\,,\ i = 1, \ldots, n\,.$$

La statistica lineare dei ranghi con segno costruita su questi punteggi è data da

$$T_*^+ = \frac{1}{n+1}\sum_{i=1}^{n} Z_i R_i^+ = \frac{W^+}{n+1}\,,$$

dove W^+ è la statistica di Wilcoxon. Quindi T_*^+ e W^+ forniscono test equivalenti, ovvero la scelta dei punteggi fatta per la statistica di Wilcoxon risulta ottima per una variabile casuale Logistica. ◁

• **Esempio 5.2.3.** Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$, dove F è la funzione di ripartizione di una distribuzione di Laplace $L(0,1)$. Si può verificare che le condizioni del Teorema 5.2.2 sono soddisfatte. Inoltre, dal momento che risulta $f'(x) = -\,\mathrm{segn}(x) f(x)$, dove $\mathrm{segn}(x) = 2\mathbf{1}_{(0,\infty)}(x) - 1$, allora

$$-\,\frac{f'(x)}{f(x)} = \mathrm{segn}(x)\,.$$

Dunque, la scelta ottimale dei punteggi è data da

$$a_*(i) = \mathrm{E}(\mathrm{segn}(Y_{(i)}))\,,\ i = 1, \ldots, n\,,$$

dove $(Y_{(1)}, \ldots, Y_{(n)})$ è la statistica ordinata relativa ai valori assoluti di un campione casuale da una distribuzione di Laplace $L(0,1)$. Dal momento che $\mathrm{E}(\mathrm{segn}(Y_{(i)})) = 1$, allora risulta $a_*(i) = 1$ per $i = 1, \ldots, n$. Questa scelta dei punteggi fornisce dunque la statistica dei segni B. ◁

5.3. La distribuzione per grandi campioni delle statistiche lineari dei ranghi con segno. In questa sezione vengono considerate le proprietà per grandi campioni delle statistiche lineari dei ranghi con segno. Di seguito vengono considerati due risultati preliminari.

Lemma 5.3.1. *Se i punteggi sono tali che*

$$a_n(i) = \phi(i/(n+1))\,,\ i = 1, \ldots, n\,,$$

dove ϕ è una funzione punteggio non negativa e non decrescente che non dipende da n per cui

$$0 < \int_0^1 \phi(u)^2\, du < \infty\,,$$

allora

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} a_n(i)^2 = \int_0^1 \phi(u)^2\, du\,.$$

Dimostrazione. Sia

$$q_n(u) = \sum_{i=1}^{n} \phi(i/(n+1))\, \mathbf{1}_{[(i-1)/n, i/n]}(u)$$

una discretizzazione della funzione ϕ. Dal momento che si ha $\mathbf{1}_{[(i-1)/n,i/n]}(u)^2 = \mathbf{1}_{[(i-1)/n,i/n]}(u)$ per $i = 1, \ldots, n$, e che $\mathbf{1}_{[(i-1)/n,i/n]}(u)\mathbf{1}_{[(j-1)/n,j/n]}(u) = 0$ per $i \neq j = 1, \ldots, n$, allora

$$\int_0^1 q_n(u)^2\,du = \int_0^1 (\sum_{i=1}^n \phi(i/(n+1))\mathbf{1}_{[(i-1)/n,i/n]}(u))^2 du = \int_0^1 \sum_{i=1}^n \phi(i/(n+1))^2\,\mathbf{1}_{[(i-1)/n,i/n]}(u)du$$
$$= \sum_{i=1}^n \phi(i/(n+1))^2 \int_0^1 \mathbf{1}_{[(i-1)/n,i/n]}(u)du = \frac{1}{n}\sum_{i=1}^n \phi[i/(n+1)]^2 = \frac{1}{n}\sum_{i=1}^n a_n(i)^2\,.$$

Inoltre, poichè ϕ è una funzione non negativa e non decrescente, si ha

$$\frac{1}{n}\,\phi(i/(n+1))^2 \le \frac{1}{n}\,\phi(i/n)^2 \le \int_{i/n}^{(i+1)/n} \phi(u)^2 du\,,\ 1 \le i \le n-1\,,$$

e

$$\frac{1}{n+1}\,\phi(n/(n+1))^2 < \frac{1}{n}\,\phi(n/(n+1))^2 \le \frac{n+1}{n}\int_{n/(n+1)}^1 \phi(u)^2 du\,.$$

Di conseguenza, si ha

$$\int_0^1 q_n(u)^2\,du = \frac{1}{n}\sum_{i=1}^n \phi(i/(n+1))^2 \le \sum_{i=1}^{n-1}\int_{i/n}^{(i+1)/n} \phi(u)^2 du + \frac{n+1}{n}\int_{n/(n+1)}^1 \phi(u)^2 du$$
$$= \int_{1/n}^1 \phi(u)^2 du + \frac{n+1}{n}\int_{n/(n+1)}^1 \phi(u)^2 du\,,$$

da cui

$$\limsup_n \int_0^1 q_n(u)^2\,du \le \int_0^1 \phi(u)^2\,du\,.$$

Inoltre, per il Lemma di Fatou si ha

$$\int_0^1 \phi(u)^2\,du \le \liminf_n \int_0^1 q_n(u)^2\,du\,.$$

Si deve concludere dunque che

$$\lim_n \int_0^1 q_n(u)^2\,du = \int_0^1 \phi(u)^2\,du\,,$$

ovvero

$$\lim_n \frac{1}{n}\sum_{i=1}^n a_n(i)^2 = \int_0^1 \phi(u)^2\,du\,. \qquad \square$$

Lemma 5.3.2. *Se sono valide le condizioni del Lemma 5.3.1 si ha*

$$\lim_n \frac{\sum_{i=1}^n a_n(i)^2}{\max\limits_{1\le i\le n} a_n(i)^2} = \infty\,.$$

Dimostrazione. Si ha

$$\lim_n \frac{\sum_{i=1}^n a_n(i)^2}{\max\limits_{1\le i\le n} a_n(i)^2} = \lim_n \frac{(1/n)\sum_{i=1}^n a_n(i)^2}{(1/n)\max\limits_{1\le i\le n} a_n(i)^2} = \frac{\lim\limits_n (1/n)\sum_{i=1}^n a_n(i)^2}{\lim\limits_n (1/n)\max\limits_{1\le i\le n} a_n(i)^2}\,.$$

Dal Lemma 5.3.1 si ottiene che il numeratore della precedente espressione è finito e positivo. Inoltre, poichè ϕ è non decrescente, allora

$$\frac{1}{n}\max_{1\le i\le n} a_n(i)^2 = \frac{1}{n}\,\phi(n/(n+1))^2 \le \frac{n+1}{n}\int_{n/(n+1)}^{1}\phi(u)^2\,du\,,$$

per cui

$$\lim_n \frac{1}{n}\max_{1\le i\le n} a_n(i)^2 = 0\,,$$

da cui segue la tesi. □

Nel prossimo teorema si ottiene la distribuzione per grandi campioni delle statistiche lineari dei ranghi con segno sotto ipotesi di base.

Teorema 5.3.3. *Se $(X_1,\dots,X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{0,F}$, allora per una statistica lineare dei ranghi con segno*

$$T^+ = T_n^+ = \sum_{i=1}^{n} Z_i a_n(R_i^+)\,,$$

i cui punteggi $a_n(i)$ per $i = 1,\dots,n$, soddisfano le condizioni del Lemma 5.3.1, risulta

$$\frac{T_n^+ - \mathrm{E}(T_n^+)}{\sqrt{\mathrm{Var}(T_n^+)}} \xrightarrow{d} N(0,1)\,,$$

dove $\mathrm{E}(T_n^+)$ *e* $\mathrm{Var}(T_n^+)$ *sono definite nel Teorema 5.1.3.*

Dimostrazione. Dal Teorema 5.1.2 si ottiene che

$$T_n^+ = \sum_{i=1}^{n} Z_i a_n(R_i^+) \stackrel{d}{=} \sum_{i=1}^{n} Z_i a_n(i)\,.$$

Tenendo presente il Teorema 2.3.2 si ha $\mathrm{E}[|Z_i - 1/2|^3] = 1/8$ e $\mathrm{Var}(Z_i) = 1/4$, da cui

$$\frac{\sum_{i=1}^n \mathrm{E}(|Z_i a_n(i) - a_n(i)/2|^3)}{(\sum_{i=1}^n \mathrm{Var}(Z_i a_n(i)))^{3/2}} = \frac{\sum_{i=1}^n a_n(i)^3 \mathrm{E}(|Z_i - 1/2|^3)}{(\sum_{i=1}^n a_n(i)^2 \mathrm{Var}(Z_i))^{3/2}} = \frac{\sum_{i=1}^n a_n(i)^3}{(\sum_{i=1}^n a_n(i)^2)^{3/2}}$$
$$\le \frac{\sum_{i=1}^n a_n(i)^2 \max_{1\le j\le n}|a_n(j)|}{(\sum_{i=1}^n a_n(i)^2)^{3/2}} = \frac{\max_{1\le i\le n}|a_n(i)| \sum_{i=1}^n a_n(i)^2}{(\sum_{i=1}^n a_n(i)^2)^{3/2}} = \sqrt{\frac{\max_{1\le i\le n} a_n(i)^2}{\sum_{i=1}^n a_n(i)^2}}\,.$$

Di conseguenza, dal Lemma 5.3.2 si ha

$$\lim_n \frac{\sum_{i=1}^n \mathrm{E}(|Z_i a_n(i) - a_n(i)/2|^3)}{(\sum_{i=1}^n \mathrm{Var}[Z_i a_n(i)])^{3/2}} = \lim_n \sqrt{\frac{\max_{1\le i\le n} a_n(i)^2}{\sum_{i=1}^n a_n(i)^2}} = 0\,.$$

Applicando il Teorema Fondamentale del Limite di Lyapunov (Teorema A.3.9) con $\delta = 1$, si ottiene

$$\frac{\sum_{i=1}^n Z_i a_n(i) - \mathrm{E}(T_n^+)}{\sqrt{\mathrm{Var}(T_n^+)}} \xrightarrow{d} N(0,1)\,,$$

ovvero, dato che

$$\frac{T_n^+ - \mathrm{E}(T_n^+)}{\sqrt{\mathrm{Var}(T_n^+)}} \stackrel{d}{=} \frac{\sum_{i=1}^n Z_i a_n(i) - \mathrm{E}(T_n^+)}{\sqrt{\mathrm{Var}(T_n^+)}}\,,$$

si ha la tesi. □

Se si considera la statistica $T_n^{+\prime}$ i cui punteggi sono dati da

$$a'_n(i) = b_n\phi(i/(n+1)) = b_n a_n(i)\ ,\ i = 1,\dots,n\ ,$$

allora risulta $T_n^{+\prime} = b_n T_n^+$ con

$$\mathrm{E}(T_n^{+\prime}) = b_n\mathrm{E}(T_n^+)\ ,\ \mathrm{Var}(T_n^{+\prime}) = b_n^2\mathrm{Var}(T_n^+)\ .$$

Di conseguenza, si ha

$$\frac{T_n^{+\prime} - \mathrm{E}(T_n^{+\prime})}{\sqrt{\mathrm{Var}(T_n^{+\prime})}} = \frac{b_n T_n^+ - b_n\mathrm{E}(T_n^+)}{\sqrt{b_n^2\mathrm{Var}(T_n^+)}} = \frac{T_n^+ - \mathrm{E}(T_n^+)}{\sqrt{\mathrm{Var}(T_n^+)}}\ ,$$

ovvero T_n^+ e $T_n^{+\prime}$ hanno le medesime proprietà per grandi campioni. Dunque, la costante b_n non influenza il comportamento per grandi campioni della statistica dei ranghi con segno, mentre la scelta della funzione punteggio ϕ è determinante sotto questo punto di vista.

• **Esempio 5.3.1.** Si consideri la statistica $W^+ = W_n^+$ di Wilcoxon. Al fine di determinare la distribuzione per grandi campioni di W_n^+ è conveniente considerare la statistica $T_n^+ = b_n W_n^+$ con $b_n = 1/(n+1)$ che ha le medesime proprietà per grandi campioni della statistica W_n^+. La funzione punteggio relativa alla statistica T_n^+, data da $\phi(u) = u\mathbf{1}_{[0,1]}(u)$, è una funzione positiva e crescente. Inoltre, si ha

$$\int_0^1 \phi(u)^2\,du = \int_0^1 u^2 du = \frac{1}{3} < \infty\ .$$

Le condizioni del Teorema 5.3.3 sono dunque soddisfatte. Tenendo presente l'Esempio 5.1.3 si deve concludere che

$$\frac{W_n^+ - \mathrm{E}(W_n^+)}{\sqrt{\mathrm{Var}(W_n^+)}} = \frac{W_n^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{d} N(0,1)\ . \qquad \triangleleft$$

Tenendo presente i precedenti risultati, fissato un livello di significatività α, la regione critica per verificare $H_0 : \lambda = 0, F \in \mathcal{S}$, contro l'alternativa $H_1 : \lambda \neq 0, F \in \mathcal{S}$, può essere dunque approssimata dall'insieme

$$\{t^+ : t^+ \leq \mathrm{E}(T_n^+) + z_{\alpha/2}\sqrt{\mathrm{Var}(T_n^+)}, t^+ \geq \mathrm{E}(T_n^+) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(T_n^+)}\}\ .$$

Analogamente, la regione critica per verificare l'alternativa $H_1 : \lambda > 0, F \in \mathcal{S}$, può essere approssimata dall'insieme

$$\{t^+ : t^+ \geq \mathrm{E}(T_n^+) + z_{1-\alpha}\sqrt{\mathrm{Var}(T_n^+)}\}\ ,$$

mentre la regione critica per verificare l'alternativa $H_1 : \lambda < 0, F \in \mathcal{S}$, può essere approssimata dall'insieme

$$\{t^+ : t^+ \leq \mathrm{E}(T_n^+) + z_{\alpha}\sqrt{\mathrm{Var}(T_n^+)}\}\ .$$

Inoltre, mediante il Teorema 3.1.8 si può dimostrare che la successione di test basata su $(T_n^+)_{n\geq 1}$ è coerente. Il seguente teorema permette di ottenere l'efficacia di una statistica lineare dei ranghi con segno.

Teorema 5.3.4. *Sia $(X_1,\dots,X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$ e si consideri il sistema di ipotesi $H_0 : \lambda = 0, F \in \mathcal{S}$, contro $H_1 : \lambda = \lambda_i, F \in \mathcal{S}$, dove $(\lambda_i)_{i\geq 1}$ è una successione di alternative tali che $\lambda_i = c/\sqrt{n_i}$ con c costante. Data una statistica lineare dei ranghi con segno*

$$T^+ = T_n^+ = \sum_{i=1}^n Z_i\, a_n(R_i^+)\ ,$$

i cui punteggi $a_n(i)$ per $i = 1,\dots,n$ soddisfano le condizioni del Lemma 5.3.1, allora l'efficacia del test basato su T_n^+ risulta

$$\text{eff}_{T^+} = \frac{\int_0^1 \phi(u)\phi_f(u)\,du}{(\int_0^1 \phi(u)^2 du)^{1/2}} ,$$

dove

$$\phi_f(u) = -\frac{f'(F^{-1}(1/2+u/2))}{f(F^{-1}(1/2+u/2))} .$$

Dimostrazione. Si veda Háyek e Šidák (1967). □

• **Esempio 5.3.2.** Si consideri la statistica dei segni $B = B_n$. La relativa funzione punteggio $\phi(u) = \mathbf{1}_{[0,1]}(u)$, è una funzione positiva e non decrescente, per cui si ha

$$\int_0^1 \phi(u)^2\,du = \int_0^1 du = 1 < \infty .$$

Le condizioni del Teorema 5.3.4 sono dunque soddisfatte. Inoltre, risulta

$$\int_0^1 \phi(u)\phi_f(u)\,du = \int_0^1 -\frac{f'(F^{-1}(1/2+u/2))}{f(F^{-1}(1/2+u/2))}\,du ,$$

da cui, mediante la trasformazione di variabile $x = F^{-1}(1/2+u/2)$ con $u = 2F(x)-1$, tenendo presente la simmetria di f, si ha

$$\int_0^1 \phi(u)\phi_f(u)\,du = -2\int_0^\infty \frac{f'(x)}{f(x)}\,f(x)\,dx = 2f(0) .$$

Quindi, l'efficacia del test basato sulla statistica dei segni risulta

$$\text{eff}_B = 2f(0) ,$$

come era già noto dall'Esempio 3.2.2. □

• **Esempio 5.3.3.** Si consideri la statistica $W^+ = W_n^+$ del test dei ranghi con segno di Wilcoxon. Dall'Esempio 5.3.1 è noto che le condizioni del Lemma 5.3.1 sono soddisfatte. Inoltre, risulta

$$\int_0^1 \phi(u)\phi_f(u)\,du = \int_0^1 -\frac{f'(F^{-1}(1/2+u/2))}{f(F^{-1}(1/2+u/2))}\,u\,du ,$$

da cui, mediante la trasformazione di variabile $x = F^{-1}(1/2+u/2)$ con $u = 2F(x)-1$, tenendo presente la simmetria di f, si ha

$$\int_0^1 \phi(u)\phi_f(u)\,du = -2\int_0^\infty f'(x)(2F(x)-1)\,dx = 2\int_{-\infty}^\infty f(x)^2\,dx .$$

Quindi, l'efficacia del test basato sulla statistica di Wilcoxon risulta

$$\text{eff}_{W^+} = 2\sqrt{3}\int_{-\infty}^\infty f(x)^2\,dx .$$

Utilizzando i risultati dell'Esempio 3.2.1, l'efficienza asintotica relativa del test basato sulla statistica W^+ rispetto al test basato sulla statistica T è data dunque da

$$\text{EAR}_{W^+,T} = \frac{K_{W^+}^2}{K_T^2} = 12\,\sigma^2\,(\int_{-\infty}^\infty f(x)^2\,dx)^2 .$$

Se X ha distribuzione Normale $N(\lambda,\sigma^2)$, allora dal momento che

$$\int_{-\infty}^{\infty} f(x)^2 \, dx = \frac{1}{2\sqrt{\pi}\sigma} \, ,$$

si ha $\text{EAR}_{W^+,T} = 3/\pi \simeq 0.9549$. Dunque, il test basato su W^+ è quasi equivalente al test basato su T dal punto di vista dell'efficienza asintotica relativa anche sotto assunzione di normalità. ◁

Capitolo 6

I test per un parametro di posizione: un campione e due campioni appaiati

6.1. Il test dei segni. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{\lambda,F}$. Il test dei segni è basato sulla statistica dei segni

$$B = \sum_{i=1}^{n} Z_i \,,$$

descritta nell'Esempio 2.3.1. Mediante il test dei segni si può dunque verificare l'ipotesi di base H_0: $\lambda = 0, F \in \mathcal{M}$. La statistica B è una statistica lineare dei ranghi con segno, con la scelta dei punteggi $a(i) = 1$ per $i = 1, \ldots, n$, come già evidenziato nell'Esempio 5.1.1. Tuttavia, il test dei segni può essere costruito anche senza ipotizzare la simmetria della variabile casuale dalla quale proviene il campione e quindi questa assunzione sarà evitata. Se l'ipotesi di base è vera, dall'Esempio 2.3.1 è noto che B ha distribuzione Binomiale $Bi(n, 1/2)$. Dunque, se $p_n(b) = \Pr(B = b)$, allora

$$p_n(b) = \binom{n}{b} 2^{-n} \mathbf{1}_{\{0,1,\ldots,n\}}(b) \,.$$

Risulta

$$\mathrm{E}(B) = \frac{n}{2} \,, \mathrm{Var}(B) = \frac{n}{4} \,.$$

Inoltre, dal Corollario 2.3.3 la statistica B è "distribution-free" su $\mathcal{M}_{0,F}$ e dunque il test dei segni è "distribution-free". Dal momento che la distribuzione della statistica B sotto ipotesi di base è specificata, le appropriate regioni critiche del test per alternative bilaterali o direzionali si possono facilmente ottenere tenendo presente la discussione fatta nella §5.1.

• **Esempio 6.1.1.** Una fabbrica produce un tipo specifico di sfere di acciaio del diametro di 1 micron. Alla fine di una giornata di produzione è stato estratto un campione casuale di 10 sfere ed è stato misurato il loro diametro, ottenendo i dati della Tavola 6.1.1.

Tavola 6.1.1. Diametro delle sfere (in micron).

sfera	x_i	$x_i - 1$	z_i
1	1.18	0.18	1
2	1.42	0.42	1
3	0.69	− 0.31	0
4	0.88	− 0.12	0
5	1.62	0.62	1
6	1.09	0.09	1
7	1.53	0.53	1
8	1.02	0.02	1
9	1.19	0.19	1
10	1.32	0.32	1

Fonte: Romano (1977)

Per vedere se il livello medio della produzione si discosta dallo standard, si vuole verificare dunque il sistema di ipotesi $H_0: \lambda = 1, F \in \mathcal{M}$, contro $H_1: \lambda \neq 1, F \in \mathcal{M}$. Dal momento che per questi dati si ha $b = 8$, allora per $n = 10$ risulta $\Pr(B \geq 8) = 0.0547$ e quindi la significatività osservata è data da

$$\alpha_{oss} = 2 \times 0.0547 = 0.1094 \, .$$

Di conseguenza, si può accettare H_0 ad ogni livello di significatività $\alpha < 0.1094$. ◁

Per quanto riguarda la distribuzione per grandi campioni della statistica $B = B_n$, dal Teorema Fondamentale Classico del Limite (Teorema A.3.6), se è vera l'ipotesi di base si ha

$$\frac{B_n - n/2}{\sqrt{n/4}} \xrightarrow{d} N(0,1) \, .$$

La convergenza della statistica B_n alla normalità è particolarmente rapida anche per campioni moderati, ovvero $n \geq 15$. Le approssimazioni per grandi campioni delle regioni critiche del test per alternative bilaterali o direzionali si possono ottenere tenendo presente la discussione fatta nella §5.3.

• **Esempio 6.1.2.** Gli antichi greci indicavano come rettangolo aureo un rettangolo con un rapporto fra i lati dato da $\rho = 1 \div (\sqrt{5} + 1)/2 \simeq 0.618$. Questo tipo di rettangolo era spesso adoperato nella loro architettura, come ad esempio nella struttura del Partenone. I dati della Tavola 6.1.2 riguardano il rapporto fra i lati di rettangoli usati dai nativi americani Shoshoni per decorare le loro tende.

Tavola 6.1.2. Rapporto dei lati dei rettangoli.

rettangolo	x_i	$x_i - \rho$	z_i
1	0.693	0.075	1
2	0.662	0.044	1
3	0.690	0.072	1
4	0.606	− 0.012	0
5	0.570	− 0.048	0
6	0.749	0.131	1
7	0.672	0.054	1
8	0.628	0.010	1
9	0.609	− 0.009	0
10	0.844	0.226	1
11	0.654	0.036	1
12	0.615	− 0.003	0
13	0.668	0.050	1
14	0.601	− 0.017	0
15	0.576	− 0.042	0
16	0.670	0.052	1
17	0.606	− 0.012	0
18	0.611	− 0.007	0
19	0.553	− 0.065	0
20	0.933	0.315	1

Fonte: Dubois (1970)

Si vuole verificare se gli Shoshoni avessero conoscenza del rettangolo aureo, ovvero si vuole verificare il sistema di ipotesi $H_0: \lambda = \rho, F \in \mathcal{M}$, contro $H_1: \lambda \neq \rho, F \in \mathcal{M}$. Dal momento che per questi dati si verifica che $b = 11$, allora si ha

$$\Pr(B \geq 11) \simeq 1 - \Phi((11 - 10)/\sqrt{5}) = 1 - \Phi(0.4472) = 0.3272 \, ,$$

per cui la significatività osservata risulta $\alpha_{oss} \simeq 2 \times 0.3272 = 0.6544$. Dato che la significatività osservata è piuttosto elevata l'evidenza empirica porta a concludere che gli Shoshoni avevano conoscenza del rettangolo aureo. In effetti, si può accettare H_0 ad ogni livello di significatività $\alpha < 0.6544$. ◁

6.2. Le prestazioni del test dei segni. Risulta interessante confrontare il test dei segni con la sua controparte classica, ovvero il test di Student. Sebbene sia possibile determinare analiticamente la funzione

potenza del test dei segni (vedi Esempio 3.1.3), questa operazione risulta proibitiva nel caso del test di Student quando il campione proviene da una variabile casuale che non ha distribuzione Normale. Di conseguenza, le potenze di questi test per varie distribuzioni simmetriche state calcolate mediante simulazione per le numerosità campionarie $n = 5, 10, 15$. Le alternative scelte sono state $\lambda = 0.0\,\sigma, 0.2\,\sigma, 0.4\,\sigma, 0.6\,\sigma, 0.8\,\sigma$, dove σ^2 rappresenta la varianza della distribuzione ipotizzata. Nel caso della distribuzione di Cauchy si noti che σ denota il valore per cui $\Pr(X \leq \sigma) = \Phi(1)$. Inoltre, il test dei segni è stato casualizzato al fine di ottenere un livello di significatività esattamente pari a $\alpha = 0.05$ per entrambi i test. I risultati della simulazione sono riportati nella Tavola 6.2.1.

Tavola 6.2.1. Potenza del test dei segni (potenza del test di Student).

distribuzione	0.0σ	0.2σ	0.4σ	0.6σ	0.8σ
			$n = 5$		
$U(\lambda - 1/2, 1)$	0.05(0.05)	0.08(0.10)	0.12(0.16)	0.18(0.25)	0.26(0.37)
$N(\lambda, 1)$	0.05(0.05)	0.09(0.11)	0.15(0.19)	0.26(0.30)	0.35(0.43)
$Lo(\lambda, 1)$	0.05(0.05)	0.10(0.11)	0.18(0.20)	0.28(0.33)	0.40(0.48)
$L(\lambda, 1)$	0.05(0.04)	0.13(0.12)	0.24(0.23)	0.35(0.38)	0.46(0.52)
$C(\lambda, 1)$	0.05(0.02)	0.12(0.07)	0.21(0.15)	0.31(0.24)	0.39(0.33)
			$n = 10$		
$U(\lambda - 1/2, 1)$	0.05(0.05)	0.10(0.13)	0.18(0.28)	0.29(0.52)	0.43(0.74)
$N(\lambda, 1)$	0.05(0.05)	0.13(0.15)	0.26(0.32)	0.42(0.55)	0.60(0.76)
$Lo(\lambda, 1)$	0.05(0.05)	0.14(0.15)	0.30(0.33)	0.49(0.56)	0.68(0.77)
$L(\lambda, 1)$	0.05(0.05)	0.19(0.16)	0.41(0.36)	0.60(0.59)	0.76(0.78)
$C(\lambda, 1)$	0.05(0.02)	0.17(0.09)	0.37(0.18)	0.53(0.29)	0.67(0.39)
			$n = 15$		
$U(\lambda - 1/2, 1)$	0.05(0.05)	0.12(0.17)	0.22(0.41)	0.38(0.70)	0.58(0.91)
$N(\lambda, 1)$	0.05(0.05)	0.15(0.18)	0.32(0.44)	0.55(0.71)	0.76(0.90)
$Lo(\lambda, 1)$	0.05(0.05)	0.17(0.18)	0.37(0.45)	0.63(0.72)	0.82(0.90)
$L(\lambda, 1)$	0.05(0.05)	0.25(0.20)	0.51(0.47)	0.75(0.73)	0.89(0.89)
$C(\lambda, 1)$	0.05(0.02)	0.22(0.09)	0.47(0.20)	0.69(0.30)	0.82(0.41)

Anche se il test dei segni dimostra scarse prestazioni per distribuzioni a code leggere come l'Uniforme, si osservi come diventi invece particolarmente efficiente per distribuzioni a code pesanti come la Laplace. Questo comportamento risulta in generale ancora più evidente con distribuzioni a code particolamente pesanti quali per esempio quelle che non posseggono varianza finita come la distribuzione di Cauchy. Per quest'ultima distribuzione, il test di Student non mantiene neppure il livello di significatività. Si può verificare inoltre che le prestazioni del test dei segni risultano generalmente superiori (talvolta in modo molto marcato) a quelle del test di Student quando si considerano distribuzioni asimmetriche. Per quanto riguarda le prestazioni per grandi campioni del test dei segni, dall'Esempio 3.2.2 si ha

$$\text{eff}_B = 2f(0)\ .$$

Inoltre, ancora dall'Esempio 3.2.2, è noto che l'efficienza asintotica relativa del test dei segni rispetto al test di Student risulta

$$\text{EAR}_{B,T} = 4\sigma^2 f(0)^2\ .$$

La Tavola 6.2.2 fornisce i valori dell'efficienza asintotica relativa $\text{EAR}_{B,T}$ per alcune distribuzioni. Anche per grandi campioni il test dei segni dimostra scarse prestazioni per distribuzioni a code leggere, e ottime prestazioni per distribuzioni a code pesanti. In generale, Noether (1967) ha provato che $\text{EAR}_{B,T} \geq 1/3$ e il limite inferiore è raggiunto nel caso di una distribuzione Uniforme.

Tavola 6.2.2. EAR del test dei segni rispetto al test di Student.

distribuzione	$\text{EAR}_{B,T}$
$U(\lambda, \delta)$	$1/3 \simeq 0.3333$
$N(\mu, \sigma^2)$	$2/\pi \simeq 0.6366$
$Lo(\lambda, \delta)$	$\pi^2/12 \simeq 0.8224$
$L(\mu, \delta)$	2
$C(\lambda, \delta)$	∞

6.3. Il test dei segni e gli intervalli di confidenza per la mediana. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{M}_{\lambda,F}$. Dall'Esempio 4.1.1 è noto che se $(X_{(1)}, \ldots, X_{(n)})$ è la statistica ordinata allora $(X_{(b_{n,\alpha/2}+1)}, X_{(n-b_{n,\alpha/2})})$ è un intervallo di confidenza della mediana "distribution-free" su $\mathcal{M}_{\lambda,F}$ al livello di confidenza $(1-\alpha)$.

• **Esempio 6.3.1.** Si considerino ancora i dati dell'Esempio 6.1.1. Dal momento che $\Pr(B \leq 1) = 0.0107$ per $n = 10$, si può scegliere un livello di confidenza naturale pari a $1 - \alpha = 1 - 2 \times 0.0107 = 0.9786$. Dunque, una volta ordinato il campione si ottiene che $(0.88, 1.53)$ è un intervallo di confidenza della mediana "distribution-free" su $\mathcal{M}_{\lambda,F}$ al livello di confidenza del 97.86%. ◁

Alternativamente, quando la numerosità campionaria è elevata, dall'Esempio 4.2.3 si ha che $(X_{(l+1)}, X_{(n-l)})$, dove $l = \lfloor n/2 - z_{1-\alpha/2}\sqrt{n/4} \rfloor$, è un intervallo di confidenza della mediana "distribution-free" per grandi campioni su $\mathcal{M}_{\lambda,F}$ al livello di confidenza $(1-\alpha)$.

• **Esempio 6.3.2.** I dati della Tavola 6.3.1 riguardano i tempi di sopravvivenza dal momento della diagnosi di 43 pazienti malati di leucemia.

Tavola 6.3.1. Tempi di sopravvivenza dei malati di leucemia (in giorni).

paziente	x_i	paziente	x_i
1	7	23	715
2	47	24	779
3	58	25	881
4	74	26	900
5	177	27	930
6	232	28	968
7	273	29	1077
8	285	30	1109
9	317	31	1314
10	429	32	1334
11	440	33	1367
12	445	34	1534
13	455	35	1712
14	468	36	1784
15	495	37	1877
16	497	38	1886
17	532	39	2045
18	571	40	2056
19	579	41	2260
20	581	42	2429
21	650	43	2509
22	702		

Fonte: Bryson e Siddiqui (1969)

Scelto un livello di confidenza pari a $1 - \alpha = 0.95$, si ha

$$l = \lfloor 43/2 - 1.9577\sqrt{43/4} \rfloor = \lfloor 15.08 \rfloor = 15 \,,$$

per cui $(497, 968)$ è un intervallo di confidenza della mediana "distribution-free" per grandi campioni su $\mathcal{M}_{\lambda,F}$ al livello di confidenza del 95%. ◁

6.4. Il test dei segni per due campioni appaiati. I campioni appaiati si ottengono quando sulle unità campionarie si effettuano misure ripetute. Ad esempio, le misure sulle unità campionarie potrebbero essere fatte prima e dopo un certo trattamento e l'obiettivo potrebbe essere quello di verificare se il trattamento stesso è efficace. Più formalmente, supponiamo che $(X_1, Y_1), \ldots, (X_n, Y_n)$ sia un campione casuale proveniente dal vettore di variabili casuali bivariato (X, Y), dove X è riferita al pre-trattamento ed Y è riferita al post-trattamento. Supponiamo inoltre che il campione casuale trasformato $(D_1, \ldots, D_n)$, dove $D_i = Y_i - X_i$ per $i = 1, \ldots, n$, abbia funzione di ripartizione congiunta $F_n \in \mathcal{M}_{\lambda,F}$. Basandosi sul

campione trasformato, la verifica dell'efficacia del trattamento si riduce di conseguenza alla verifica dell'ipotesi $H_0: \lambda = 0, F \in \mathcal{M}$, contro una opportuna alternativa, e dunque il test dei segni può essere applicato in maniera analoga a quanto visto nella §6.1.

• **Esempio 6.4.1.** Su 8 pazienti con anemia cronica grave è stato misurato l'indice di infarto prima e dopo un trattamento medico e i relativi dati sono stati riportati nella Tavola 6.4.1.

Tavola 6.4.1. Indice di infarto dei pazienti con anemia (in ml/battito/m^2).

paziente	prima	dopo	d_i	z_i
1	109	56	− 53	0
2	57	44	− 13	0
3	53	55	2	1
4	57	40	− 17	0
5	68	62	− 6	0
6	72	46	− 26	0
7	51	49	− 2	0
8	65	41	− 24	0

Fonte: Bhatia, Manchanda e Roy (1969)

Per vedere se il trattamento è stato effettivo e quindi per determinare se l'indice di infarto è diminuito, si vuole verificare dunque il sistema di ipotesi $H_0: \lambda = 0, F \in \mathcal{M}$, contro $H_1: \lambda < 0, F \in \mathcal{M}$. Dal momento che per questi dati si ha $b = 1$, per $n = 8$ risulta $\Pr(B \leq 1) = 0.0352$ e quindi la significatività osservata risulta $\alpha_{oss} = 0.0352$. Sulla base dell'evidenza empirica si deve dunque concludere che il trattamento è efficace, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha > 0.0352$. ◁

6.5. Il test di Wilcoxon. Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$. Il test di Wilcoxon è basato sulla statistica

$$W^+ = \sum_{i=1}^{n} Z_i R_i^+ ,$$

descritta nell'Esempio 2.5.1, con cui si può verificare l'ipotesi di base $H_0: \lambda = 0, F \in \mathcal{S}$. La statistica W^+ è una statistica lineare dei ranghi con segno con i punteggi scelti come $a(i) = i$ per $i = 1, \ldots, n$. Inoltre, dal Teorema 5.1.2 si verifica che

$$W^+ \stackrel{d}{=} \sum_{i=1}^{n} i Z_i ,$$

ovvero la statistica W^+ è equivalente in distribuzione ad una combinazione lineare di variabili casuali indipendenti con distribuzione Binomiale $Bi(1, 1/2)$, i cui pesi sono dati da $\{1, \ldots, n\}$. Se l'ipotesi di base è vera, denotando la funzione di probabilità di W^+ con $p_n(w) = \Pr(W^+ = w)$, dall'Esempio 2.5.1 si ha

$$p_n(w) = 2^{-n} c_n(w) \mathbf{1}_{\{0,1,\ldots,n(n+1)/2\}}(w) ,$$

dove $c_n(w)$ rappresenta il numero di sottoinsiemi di interi di $\{1, \ldots, n\}$ la cui somma è w. Sebbene esistano delle relazioni ricorrenti per il calcolo diretto della funzione di probabilità di W^+, la distribuzione di questa statistica si ottiene più facilmente attraverso la funzione generatrice delle probabilità. Infatti, dal Teorema 5.1.5 risulta

$$L_{W^+}(t) = 2^{-n} \prod_{i=1}^{n} (1 + t^i) , |t| < 1 .$$

• **Esempio 6.5.1.** Per $n = 3$ si ha

$$L_{W^+}(t) = 2^{-3}(1 + t)(1 + t^2)(1 + t^3) = \frac{1}{8}(1 + t + t^2 + 2t^3 + t^4 + t^5 + t^6) .$$

Poichè le probabilità $p_3(w)$ corrispondono ai coefficienti del polinomio $L_{W^+}(t)$, si ha la Tavola 6.5.1.

Tavola 6.5.1. Funzione di probabilità di W^+ per $n = 3$.

w	0	1	2	3	4	5	6
$p_3(w)$	1/8	1/8	1/8	2/8	1/8	1/8	1/8

◁

Se l'ipotesi di base è vera, dall'Esempio 5.1.3 si ha

$$\mathrm{E}(W^+) = \frac{n(n+1)}{4}\,,\ \mathrm{Var}(W^+) = \frac{n(n+1)(2n+1)}{24}\,.$$

Inoltre, dal Teorema 5.1.4 W^+ è simmetrica rispetto a $\mathrm{E}(W^+)$. Infine, dal Corollario 2.5.5 la statistica W^+ è "distribution-free" su $\mathcal{S}_{0,F}$ e di conseguenza anche il test di Wilcoxon è "distribution-free". Dal momento che la distribuzione della statistica W^+ sotto ipotesi di base è specificata, le appropriate regioni critiche del test per alternative bilaterali o direzionali si possono ottenere tenendo presente la discussione fatta nella §5.1.

• **Esempio 6.5.2.** Il numero trascendente π può essere calcolato mediante simulazione alla seguente maniera. Si consideri un cerchio di raggio unitario ed il relativo quadrato circoscritto e si generi uniformemente un certo numero di punti casuali all'interno del quadrato. La probabilità che un punto cada all'interno del cerchio è data da $\pi/4$, per cui se p rappresenta la proporzione di punti casuali caduti all'interno del cerchio, allora $4p$ è una stima di π. Con questa procedura si sono ottenute 10 stime di π, ognuna basata sulla generazione di 10 000 punti pseudo-casuali, che sono state riportate nella Tavola 6.5.2.

Tavola 6.5.2. Stime di π.

stima	x_i	$x_i - \pi$	r_i^+	$z_i r_i^+$
1	3.1348	− 0.0068	3	0
2	3.1520	0.0104	6	6
3	3.1332	− 0.0084	5	0
4	3.1540	0.0124	8	8
5	3.1298	− 0.0118	7	0
6	3.1404	− 0.0012	1	0
7	3.1400	− 0.0016	2	0
8	3.1240	− 0.0176	9	0
9	3.1336	− 0.0080	4	0
10	3.1744	0.0328	10	10

Al fine di verificare se si sono ottenute delle stime credibili di π, si deve dunque verificare il sistema di ipotesi $H_0\colon \lambda = \pi \simeq 3.1416, F \in \mathcal{S}$, contro $H_1\colon \lambda \neq \pi, F \in \mathcal{S}$. Dal momento che per questi dati si ha $w = 24$, per $n = 10$ si ottiene $\Pr(W^+ \leq 24) = 0.3848$ e quindi la significatività osservata risulta $\alpha_{oss} = 2 \times 0.3848 = 0.7696$, un valore elevato che porta ad accettare l'ipotesi di base. In particolare si può accettare H_0 ad ogni livello di significatività $\alpha < 0.7696$. ◁

Per quanto riguarda la distribuzione per grandi campioni della statistica $W^+ = W_n^+$, se è vera l'ipotesi di base dall'Esempio 5.3.1 risulta

$$\frac{W_n^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{d} N(0,1)\,.$$

La convergenza della statistica di Wilcoxon è abbastanza rapida anche per campioni moderati, ovvero $n \geq 25$. Inoltre, le approssimazioni per grandi campioni delle regioni critiche del test per alternative bilaterali o direzionali si possono ottenere tenendo presente la discussione fatta nella §5.3.

• **Esempio 6.5.3.** Su un campione di 20 soggetti in sovrappeso (con un peso corporeo superiore a 100 chilogrammi) è stato misurato il livello di colesterolo ottenendo i dati della Tavola 6.5.3.

Tavola 6.5.3. Livello di colesterolo (in mg per 100 ml).

soggetto	x_i	$x_i - 190$	r_i^+	$z_i r_i^+$
1	334	144	20	20
2	185	− 5	3	0
3	263	73	19	19
4	246	56	16	16
5	224	34	10	10
6	212	22	8	8
7	188	− 2	1	0
8	250	60	17	17
9	148	− 42	13	0
10	169	− 21	7	0
11	226	36	11	11
12	175	− 15	6	0
13	242	52	14	14
14	252	62	18	18
15	153	− 37	12	0
16	183	− 7	4	0
17	137	− 53	15	0
18	202	12	5	5
19	194	4	2	2
20	213	23	9	9

Fonte: Selvin (1991)

Da numerose esperienze cliniche risulta che il livello di colesterolo in una persona sana è di circa 190 mg per 100 ml. Si sospetta che i soggetti in forte sovrappeso abbiano un livello del colesterolo più alto della norma e dunque si vuole verificare il sistema di ipotesi $H_0: \lambda = 190, F \in \mathcal{S}$, contro $H_1: \lambda > 190, F \in \mathcal{S}$. Dal momento che per questi dati si ha $w = 149$, allora

$$\Pr(W^+ \geq 149) \simeq \Phi((149 - 20 \times 21/4)/\sqrt{20 \times 21 \times 41/24}) = 1 - \Phi(1.6426) = 0.0501\,,$$

per cui la significatività osservata risulta $\alpha_{oss} \simeq 0.0501$. L'evidenza empirica sembra confermare che i soggetti in sovrappeso hanno un livello del colesterolo più alto della norma, dal momento che si può respingere H_0 ad ogni livello di significatività $\alpha > 0.0501$. L'approssimazione normale è molto buona, in quanto il valore esatto risulta $\Pr(W^+ \geq 149) = 0.0527$. ◁

Tavola 6.6.1. Potenza del test di Wilcoxon (potenza del test di Student).

distribuzione	0.0σ	0.2σ	0.4σ	0.6σ	0.8σ
			$n = 5$		
$U(\lambda - 1/2, 1)$	0.05(0.05)	0.09(0.10)	0.14(0.16)	0.22(0.25)	0.32(0.37)
$N(\lambda, 1)$	0.05(0.05)	0.10(0.11)	0.17(0.19)	0.29(0.30)	0.42(0.43)
$Lo(\lambda, 1)$	0.05(0.05)	0.11(0.11)	0.20(0.20)	0.32(0.33)	0.46(0.48)
$L(\lambda, 1)$	0.05(0.04)	0.13(0.12)	0.26(0.23)	0.39(0.38)	0.51(0.52)
$C(\lambda, 1)$	0.05(0.02)	0.12(0.07)	0.22(0.15)	0.33(0.24)	0.42(0.33)
			$n = 10$		
$U(\lambda - 1/2, 1)$	0.05(0.05)	0.13(0.13)	0.28(0.28)	0.46(0.52)	0.68(0.74)
$N(\lambda, 1)$	0.05(0.05)	0.14(0.15)	0.31(0.32)	0.52(0.55)	0.73(0.76)
$Lo(\lambda, 1)$	0.05(0.05)	0.16(0.15)	0.35(0.33)	0.56(0.56)	0.77(0.77)
$L(\lambda, 1)$	0.05(0.05)	0.19(0.16)	0.41(0.36)	0.62(0.59)	0.80(0.78)
$C(\lambda, 1)$	0.05(0.02)	0.16(0.09)	0.32(0.18)	0.46(0.29)	0.58(0.39)
			$n = 15$		
$U(\lambda - 1/2, 1)$	0.05(0.05)	0.17(0.17)	0.37(0.41)	0.64(0.70)	0.85(0.91)
$N(\lambda, 1)$	0.05(0.05)	0.18(0.18)	0.42(0.44)	0.70(0.71)	0.89(0.90)
$Lo(\lambda, 1)$	0.05(0.05)	0.20(0.18)	0.45(0.45)	0.73(0.72)	0.90(0.90)
$L(\lambda, 1)$	0.05(0.05)	0.24(0.20)	0.52(0.47)	0.78(0.73)	0.93(0.89)
$C(\lambda, 1)$	0.05(0.02)	0.19(0.09)	0.40(0.20)	0.60(0.30)	0.73(0.41)

6.6. Le prestazioni del test di Wilcoxon. Analogamente a quanto fatto nella §6.2 per il test dei segni, è stato confrontato il test di Wilcoxon con il test di Student. Le potenze dei due test sono state calcolate mediante simulazione per le medesime distribuzioni e numerosità considerate nella §6.2. Inoltre, il test di Wilcoxon è stato casualizzato al fine di ottenere un livello di significatività esattamente pari a $\alpha = 0.05$ per

entrambi i test. I risultati della simulazione sono riportati nella Tavola 6.6.1. Da questa tavola è evidente che il test di Wilcoxon dimostra ottime prestazioni rispetto al test di Student per tutte le distribuzioni, anche sotto ipotesi di normalità. Inoltre, confrontando la Tavola 6.6.1 con la Tavola 6.2.1, si nota che il test di Wilcoxon risulta generalmente superiore al test dei segni eccetto che per la distribuzione di Cauchy.

Per quanto riguarda le prestazioni per grandi campioni del test di Wilcoxon, dall'Esempio 5.3.3 si ha che

$$\text{eff}_{W^+} = 2\sqrt{3}\int_{-\infty}^{\infty} f(x)^2 dx\ ,$$

per cui l'efficienza asintotica relativa del test di Wilcoxon rispetto al test di Student risulta

$$\text{EAR}_{W^+,T} = 12\sigma^2(\int_{-\infty}^{\infty} f(x)^2 dx)^2\ .$$

La Tavola 6.6.2 fornisce i valori dell'efficienza asintotica relativa $\text{EAR}_{W^+,T}$ per alcune distribuzioni.

Tavola 6.6.2. EAR del test di Wilcoxon rispetto al test di Student.

distribuzione	$\text{EAR}_{W^+,T}$
$U(\lambda,\delta)$	1
$N(\mu,\sigma^2)$	$3/\pi \simeq 0.9549$
$Lo(\lambda,\delta)$	$\pi^2/9 \simeq 1.0966$
$L(\mu,\delta)$	$3/2 = 1.5$
$C(\lambda,\delta)$	∞

Anche per grandi campioni il test di Wilcoxon dimostra ottime prestazioni per tutte le distribuzioni considerate. In generale Noether (1967) ha provato che $\text{EAR}_{W^+,T} \geq 105/125 \simeq 0.864$ quando la variabile casuale in oggetto di studio è distribuita in modo simmetrico. Infine, l'efficienza asintotica relativa del test di Wilcoxon rispetto al test dei segni risulta

$$\text{EAR}_{W^+,B} = \frac{3}{f(0)^2}\,(\int_{-\infty}^{\infty} f(x)^2 dx)^2\ .$$

La Tavola 6.6.3 fornisce i valori dell'efficienza asintotica relativa $\text{EAR}_{W^+,B}$ per alcune distribuzioni. Anche per grandi campioni il test dei segni fornisce prestazioni buone per distribuzioni a code pesanti, mentre risulta molto meno efficiente per distribuzionia code leggere.

Tavola 6.6.3. EAR del test di Wilcoxon rispetto al test dei segni.

distribuzione	$\text{EAR}_{W^+,B}$
$U(\lambda,\delta)$	3
$N(\mu,\sigma^2)$	$3/2 = 1.5$
$Lo(\lambda,\delta)$	$4/3 \simeq 1.3333$
$Ed(\mu,\delta)$	$3/4 = 0.75$
$Ch(\lambda,\delta)$	$3/4 = 0.75$

6.7. Il test di Wilcoxon e gli intervalli di confidenza per la mediana. Se $(X_1,\ldots,X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$, è possibile costruire mediante il test dei segni un intervallo di confidenza per la mediana λ "distribution-free" su $\mathcal{S}_{\lambda,F}$. Analogamente all'Esempio 4.1.2, si denoti le medie di Walsh relative al campione casuale con $(W_1,\ldots,W_k)$, dove $k = n(n+1)/2$. Se $(W_{(1)},\ldots,W_{(k)})$ è la statistica ordinata relativa alle medie di Walsh, allora $(W_{(w_{n,\alpha/2}+1)}, W_{(k-w_{n,\alpha/2})})$ è un intervallo di confidenza della mediana "distribution-free" su $\mathcal{S}_{\lambda,F}$ al livello di confidenza $(1-\alpha)$.

• **Esempio 6.7.1.** Su un campione di cinque regioni italiane è stata considerata l'età media delle donne alla nascita del primogenito e si sono ottenuti i dati della Tavola 6.7.1. Vi sono $5 \times 6/2 = 15$ medie di Walsh e il relativo vettore ordinato risulta

$$(25.1, 25.4, 25.7, 26.3, 26.4, 26.6, 26.7, 26.7, 27.0, 27.5, 27.6, 27.7, 27.9, 28.0, 28.3)\ .$$

Dal momento che si ha $\Pr(W^+ \leq 1) = 0.0625$ per $n = 5$, allora si può scegliere un livello di confidenza naturale pari a $1 - \alpha = 1 - 2 \times 0.0625 = 0.875$. Dunque, si deve concludere che $(25.4, 28.0)$ è un intervallo di confidenza "distribution-free" su $\mathcal{S}_{\lambda,F}$ per la mediana al livello di confidenza del 87.5%. ◁

Tavola 6.7.1. Età media della madre alla nascita del primogenito (in anni).

regione	x_i
Sicilia	25.1
Toscana	27.7
Liguria	28.3
Lombardia	27.5
Puglia	25.7

Fonte: "Repubblica" del 4 novembre 1995, inserto Salute

Quando la numerosità campionaria è elevata, con un procedimento analogo a quello dell'Esempio 4.2.3, si ha che $(W_{(l+1)}, W_{(k-l)})$, dove $l = \lfloor n(n+1)/4 - z_{1-\alpha/2}\sqrt{n(n+1)(2n+1)/24} \rfloor$, è un intervallo di confidenza della mediana "distribution-free" per grandi campioni su $\mathcal{S}_{\lambda,F}$ al livello di confidenza $(1 - \alpha)$.

6.8. Il test di Wilcoxon per due campioni appaiati. Analogamente a quanto fatto per il test dei segni, quando si dispone di un campione casuale di osservazioni appaiate $(X_1, Y_1), \ldots, (X_n, Y_n)$, si può considerare il campione casuale trasformato $(D_1, \ldots, D_n)$, dove $D_i = Y_i - X_i$ per $i = 1, \ldots, n$, con funzione di ripartizione congiunta $F_n \in \mathcal{S}_{\lambda,F}$. Basandosi sul campione trasformato, si vuole verificare dunque l'ipotesi $H_0 \colon \lambda = 0, F \in \mathcal{S}$, contro una opportuna alternativa e di conseguenza il test di Wilcoxon può essere applicato in maniera analoga a quanto visto nella §6.4.

• **Esempio 6.8.1.** Sono stati considerati 15 coppie di semi della medesima età, di cui uno prodotto con fecondazione incrociata e l'altro con auto-fecondazione, e le piante relative ad ogni coppia sono state cresciute vicine nelle medesime condizioni ambientali. La Tavola 6.8.1 riporta le altezze finale delle coppie di piante ottenute in questo modo dopo un periodo fissato di tempo.

Tavola 6.8.1. Altezze delle piante (pollici).

coppia	incrociata	auto	d_i	r_i^+	$z_i r_i^+$
1	23.5	17.4	6.1	11	11
2	12.0	20.4	− 8.4	14	0
3	21.0	20.0	1.0	2	2
4	22.0	20.0	2.0	4	4
5	19.1	18.4	0.7	1	1
6	21.5	18.6	2.9	5	5
7	22.1	18.6	3.5	7	7
8	20.4	15.3	5.1	9	9
9	18.3	16.5	1.8	3	3
10	21.6	18.0	3.6	8	8
11	23.3	16.3	7.0	12	12
12	21.0	18.0	3.0	6	6
13	22.1	12.8	9.3	15	15
14	23.0	15.5	7.5	13	13
15	12.0	18.0	− 6.0	10	0

Fonte: Darwin (1876)

Al fine di determinare se la fecondazione incrociata risulta più efficace dell'auto-fecondazione, si vuole verificare dunque il sistema di ipotesi $H_0 \colon \lambda = 0, F \in \mathcal{S}$, contro $H_1 \colon \lambda > 0, F \in \mathcal{S}$. Dal momento che per questi dati si ottiene $w = 96$, allora per $n = 15$ si ha $\Pr(W^+ \geq 96) = 0.0206$ e quindi la significatività osservata risulta $\alpha_{oss} = 0.0206$. Si deve dunque concludere che la fecondazione incrociata è più efficace dell'auto-fecondazione, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha > 0.0206$. ◁

Capitolo 7

I test basati su statistiche lineari dei ranghi

7.1. Le statistiche lineari dei ranghi. Una classe molto generale di statistiche "distribution-free" è definita come segue.

Definizione 7.1.1. Sia $(X_1, \dots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$ e sia $(R_1, \dots, R_n)$ il relativo vettore dei ranghi. Una statistica del tipo

$$T = \sum_{i=1}^{n} c(i)a(R_i)$$

è detta statistica lineare dei ranghi, mentre le costanti $a(1), \dots, a(n)$ e $c(1), \dots, c(n)$ sono dette rispettivamente punteggi e costanti di regressione. $\triangle$

In base al Corollario 2.4.6 la statistica T è "distribution-free" sulla classe $\mathcal{C}_F$. Differenti scelte dei punteggi e delle costanti di regressione consentono di costruire statistiche test per una vasta gamma di sistemi di ipotesi, come sarà evidenziato nel seguito.

• **Esempio 7.1.1.** Se $(X_1, \dots, X_{n_1})$ e $(Y_1, \dots, Y_{n_2})$ sono campioni casuali indipendenti provenienti dalla stessa variabile casuale assolutamente continua e dove $n = n_1 + n_2$, il campione misto $(X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2})$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Se $(R_1, \dots, R_{n_1})$ sono i ranghi assegnati a $(X_1, \dots, X_{n_1})$ e se $(R_{n_1+1}, \dots, R_n)$ sono i ranghi assegnati a $(Y_1, \dots, Y_{n_2})$ nel campione misto, con la scelta delle costanti di regressione

$$c(i) = \begin{cases} 1 & i = 1, \dots, n_1 \\ 0 & i = n_1 + 1, \dots, n \end{cases}$$

si ottiene la statistica

$$T = \sum_{i=1}^{n} c(i)a(R_i) = \sum_{i=1}^{n_1} a(R_i)\,,$$

che rappresenta la somma dei punteggi assegnati a $(X_1, \dots, X_{n_1})$. In particolare, se i punteggi sono scelti come $a(i) = i$ per $i = 1, \dots, n$, allora si ottiene la cosiddetta statistica di Mann-Whitney-Wilcoxon, ovvero

$$W = \sum_{i=1}^{n_1} R_i\,,$$

che fornisce la somma dei ranghi assegnati a $(X_1, \dots, X_{n_1})$. Se i punteggi sono scelti invece come

$$a(i) = \begin{cases} 1 & i > \lfloor n/2 \rfloor \\ 0 & i \leq \lfloor n/2 \rfloor \end{cases}, i = 1, \dots, n\,,$$

si ottiene la statistica

$$L = \sum_{i=1}^{n_1} a(R_i)\,,$$

che fornisce il numero di $(X_1, \ldots, X_{n_1})$ maggiori della mediana del campione misto. Per questo motivo L è detta statistica della mediana. ◁

Il seguente risultato preliminare è utile nella determinazione della media e della varianza di una statistica lineare dei ranghi.

Lemma 7.1.2. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, allora*

$$\mathrm{E}(a(R_i)) = \overline{a} = \frac{1}{n}\sum_{r=1}^{n} a(r)\,,\ i = 1, \ldots, n\,,$$

$$\mathrm{Var}(a(R_i)) = s_a^2 = \frac{1}{n}\sum_{r=1}^{n} (a(r) - \overline{a})^2\,,\ i = 1, \ldots, n\,,$$

$$\mathrm{Cov}(a(R_i), a(R_j)) = -\frac{s_a^2}{n-1}\,,\ i \neq j = 1, \ldots, n\,.$$

Dimostrazione. Dal Corollario 2.4.4 si ottiene

$$\mathrm{E}(a(R_i)) = \sum_{r=1}^{n} a(r)\mathrm{Pr}(R_i = r) = \frac{1}{n}\sum_{r=1}^{n} a(r) = \overline{a}\,,\ i = 1, \ldots, n\,,$$

e

$$\mathrm{Var}(a(R_i)) = \sum_{r=1}^{n} (a(r) - \overline{a})^2\mathrm{Pr}(R_i = r) = \frac{1}{n}\sum_{r=1}^{n} (a(r) - \overline{a})^2 = s_a^2\,,\ i = 1, \ldots, n\,.$$

Dal Corollario 2.4.4 si ottiene anche

$$\begin{aligned}
\mathrm{Cov}(a(R_i), a(R_j)) &= \sum_{r=1}^{n}\sum_{s\neq r=1}^{n} (a(r) - \overline{a})(a(s) - \overline{a})\mathrm{Pr}(R_i = r, R_j = s) \\
&= \frac{1}{n(n-1)}\sum_{r=1}^{n}\sum_{s\neq r=1}^{n} (a(r) - \overline{a})(a(s) - \overline{a}) \\
&= \frac{1}{n(n-1)}\,((\sum_{r=1}^{n} (a(r) - \overline{a}))^2 - \sum_{r=1}^{n} (a(r) - \overline{a})^2) \\
&= -\frac{1}{n(n-1)}\sum_{r=1}^{n} (a(r) - \overline{a})^2 = -\frac{s_a^2}{n-1}\,,\ i \neq j = 1, \ldots, n\,,
\end{aligned}$$

che completa la dimostrazione. □

Il seguente teorema fornisce la media e la varianza di una statistica lineare dei ranghi.

Teorema 7.1.3. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, per una statistica lineare dei ranghi T risulta*

$$\mathrm{E}(T) = n\overline{a}\,\overline{c}\,,\ \mathrm{Var}(T) = \frac{n^2}{n-1}\,s_a^2 s_c^2\,,$$

dove

$$\overline{c} = \frac{1}{n}\sum_{i=1}^{n} c(i)\ ,\ s_c^2 = \frac{1}{n}\sum_{i=1}^{n}(c(i)-\overline{c})^2\ .$$

Dimostrazione. Tenendo presente il Lemma 7.1.2 si ha

$$\mathrm{E}(T) = \sum_{i=1}^{n} c(i)\mathrm{E}(a(R_i)) = \overline{a}\sum_{i=1}^{n} c(i) = n\overline{a}\,\overline{c}$$

e

$$\begin{aligned}\mathrm{Var}(T) &= \sum_{i=1}^{n} c(i)^2\mathrm{Var}(a(R_i)) + \sum_{i=1}^{n}\sum_{j\neq i=1}^{n} c(i)c(j)\mathrm{Cov}(a(R_i),a(R_j))\\ &= s_a^2\sum_{i=1}^{n} c(i)^2 - \frac{s_a^2}{n-1}\sum_{i=1}^{n}\sum_{j\neq i=1}^{n} c(i)c(j) = \frac{s_a^2}{n-1}\left((n-1)\sum_{i=1}^{n} c(i)^2 - \sum_{i=1}^{n}\sum_{j\neq i=1}^{n} c(i)c(j)\right)\\ &= \frac{s_a^2}{n-1}\left(n\sum_{i=1}^{n} c(i)^2 - (\sum_{i=1}^{n} c(i))^2\right) = \frac{n^2 s_a^2}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(c(i)-\overline{c})^2\right) = \frac{n^2}{n-1}\,s_a^2 s_c^2\ .\end{aligned}$$

che completa la dimostrazione. □

• **Esempio 7.1.2.** Si consideri la statistica W dell'Esempio 7.1.1. Dal momento che

$$\overline{a} = \frac{1}{n}\sum_{r=1}^{n} r = \frac{n+1}{2}\ ,\ \overline{c} = \frac{1}{n}\sum_{i=1}^{n_1} 1 = \frac{n_1}{n}\ ,$$

dal Teorema 7.1.3 si ottiene

$$\mathrm{E}(W) = n\overline{a}\,\overline{c} = \frac{n_1(n+1)}{2}\ .$$

Inoltre, risulta

$$s_a^2 = \frac{1}{n}\sum_{r=1}^{n} r^2 - \frac{(n+1)^2}{4} = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{(n^2-1)}{12}$$

e

$$s_c^2 = \frac{1}{n}\sum_{i=1}^{n_1} 1 - \frac{n_1^2}{n^2} = \frac{n_1(n-n_1)}{n^2} = \frac{n_1 n_2}{n^2}\ ,$$

da cui

$$\mathrm{Var}(W) = \frac{n^2}{n-1}\,s_a^2 s_c^2 = \frac{n_1 n_2(n+1)}{12}\ .$$ ◁

• **Esempio 7.1.3.** Si consideri la statistica L dell'Esempio 7.1.1. Dal momento che

$$\overline{a} = \frac{1}{n}\sum_{r=1}^{\lfloor n/2\rfloor} 1 = \frac{\lfloor n/2\rfloor}{n}$$

e

$$s_a^2 = \frac{1}{n}\sum_{r=1}^{\lfloor n/2\rfloor} 1 - \frac{\lfloor n/2\rfloor^2}{n^2} = \frac{\lfloor n/2\rfloor(n-\lfloor n/2\rfloor)}{n^2}\ ,$$

mentre $\overline{c}$ e s_c^2 sono stati ottenuti nell'Esempio 7.1.2, allora dal Teorema 7.1.3 si ha

$$\mathrm{E}(L) = n\overline{a}\,\overline{c} = \frac{n_1\lfloor n/2\rfloor}{n}$$

e

$$\mathrm{Var}(L) = \frac{n^2}{n-1}\, s_a^2 s_c^2 = \frac{n_1 n_2 \lfloor n/2\rfloor (n - \lfloor n/2\rfloor)}{n^2(n-1)} \,.$$

Nel caso particolare che n sia pari, allora risulta

$$\mathrm{E}(L) = \frac{n_1}{2} \,,\, \mathrm{Var}(L) = \frac{n_1 n_2}{4(n-1)} \,. \qquad \triangleleft$$

Il seguente risultato preliminare viene impiegato per ottenere importanti equivalenze in distribuzione per le statistiche lineari dei ranghi.

Lemma 7.1.4. *Sia $(X_1, \ldots, X_n)$ un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Se $(R_1, \ldots, R_n)$ è il vettore dei ranghi e se $(U_1, \ldots, U_n) : \mathcal{R}_n \to \mathcal{R}_n$ è un vettore di trasformate biunivoche tali che $(U_1(R_1, \ldots, R_n), \ldots, U_n(R_1, \ldots, R_n))$, allora si ha*

$$\Pr(U_1 = u_1, \ldots, U_n = u_n) = \frac{1}{n!} \,,\, (u_1, \ldots, u_n) \in \mathcal{R}_n \,.$$

Dimostrazione. Dal momento che ad ogni $(u_1, \ldots, u_n)$ corrisponde un $(r_1, \ldots, r_n) \in \mathcal{R}_n$, dove $r_i = T_i^{-1}(u_1, \ldots, u_n)$ per $i = 1, \ldots, n$, allora

$$\Pr(U_1 = u_1, \ldots, U_n = u_n) = \Pr(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{n!} \,.$$

Questa relazione è valida per ogni $(u_1, \ldots, u_n) \in \mathcal{R}_n$ e il teorema è dimostrato. □

Teorema 7.1.5. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, per una statistica lineare dei ranghi T si ha*

$$T = \sum_{i=1}^{n} c(i)a(R_i) \stackrel{d}{=} \sum_{i=1}^{n} a(i)c(R_i) \,.$$

Dimostrazione. Si consideri il vettore di trasformate $(U_1, \ldots, U_n)$, tale che U_i rappresenta la posizione dell'intero i nel vettore $(R_1, \ldots, R_n)$. Per $(U_1, \ldots, U_n)$ valgono le condizioni del Lemma 7.1.4 e quindi si ha $(U_1, \ldots, U_n) \stackrel{d}{=} (R_1, \ldots, R_n)$, da cui

$$T = \sum_{i=1}^{n} c(i)a(R_i) = \sum_{i=1}^{n} a(i)c(U_i) \stackrel{d}{=} \sum_{i=1}^{n} a(i)c(R_i) \,. \qquad \square$$

Teorema 7.1.6. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$ e se inoltre $\{c'(1), \ldots, c'(n)\}$ e $\{a'(1), \ldots, a'(n)\}$ sono rispettivamente permutazioni di $\{c(1), \ldots, c(n)\}$ e $\{a(1), \ldots, a(n)\}$, per una statistica lineare dei ranghi T si ha*

$$T = \sum_{i=1}^{n} c(i)a(R_i) \stackrel{d}{=} \sum_{i=1}^{n} c'(i)a'(R_i) \,.$$

Dimostrazione. Sia α_i la posizione di $c'(i)$ nella permutazione $\{c(1), \ldots, c(n)\}$ e sia β_i la posizione di $a'(i)$ nella permutazione $\{a(1), \ldots, a(n)\}$. Si ha $\{c(\alpha_1), \ldots, c(\alpha_n)\} = \{c'(1), \ldots, c'(n)\}$ e $\{a(\beta_1), \ldots, a(\beta_n)\} = \{a'(1), \ldots, a'(n)\}$, per cui

$$\sum_{i=1}^{n} c'(i)a'(R_i) = \sum_{i=1}^{n} c(\alpha_i)a(\beta_{R_i}) .$$

Dal momento che per il vettore casuale $(\beta_{R_1}, \ldots, \beta_{R_n})$ valgono le condizioni del Lemma 7.1.4, allora si ottiene $(\beta_{R_1}, \ldots, \beta_{R_n}) \stackrel{d}{=} (R_1, \ldots, R_n)$, da cui

$$\sum_{i=1}^{n} c(\alpha_i)a(\beta_{R_i}) \stackrel{d}{=} \sum_{i=1}^{n} c(\alpha_i)a(R_i) .$$

Se γ_i rappresenta la posizione dell'intero i nella permutazione $(\alpha_1, \ldots, \alpha_n)$, si ottiene inoltre

$$\sum_{i=1}^{n} c(\alpha_i)a(R_i) = \sum_{i=1}^{n} c(i)a(R_{\gamma_i}) .$$

Dal momento che per $(R_{\gamma_1}, \ldots, R_{\gamma_n})$ valgono le condizioni del Teorema 7.1.4, allora si ha $(R_{\gamma_1}, \ldots, R_{\gamma_n}) \stackrel{d}{=} (R_1, \ldots, R_n)$, da cui

$$\sum_{i=1}^{n} c(i)a(R_{\gamma_i}) \stackrel{d}{=} \sum_{i=1}^{n} c(i)a(R_i) .$$

Si deve dunque concludere che

$$\sum_{i=1}^{n} c'(i)a'(R_i) \stackrel{d}{=} \sum_{i=1}^{n} c(\alpha_i)a(R_i) \stackrel{d}{=} \sum_{i=1}^{n} c(i)a(R_i) = T . \qquad \square$$

Il seguente teorema fornisce le condizioni per cui una statistica lineare dei ranghi possiede una distribuzione simmetrica rispetto alla media.

Teorema 7.1.7. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$ e se $c^*(1) \leq \ldots \leq c^*(n)$ e $a^*(1) \leq \ldots \leq a^*(n)$ rappresentano rispettivamente i valori ordinati di $c(1), \ldots, c(n)$ e $a(1), \ldots, a(n)$, allora una statistica lineare dei ranghi T è simmetrica rispetto a* $\mathrm{E}(T) = n\overline{a}\,\overline{c}$ *se*

$$a^*(i) + a^*(n+1-i) = k \, , \, i = 1, \ldots, n \, ,$$

o

$$c^*(i) + c^*(n+1-i) = k \, , \, i = 1, \ldots, n \, ,$$

con k costante.

Dimostrazione. Si dimostra il teorema per la prima condizione. Se si assumono vere le n relazioni, sommando e dividendo per n si ha

$$\begin{aligned} k &= \frac{1}{n}\sum_{i=1}^{n} a^*(i) + \frac{1}{n}\sum_{i=1}^{n} a^*(n+1-i) = \frac{1}{n}\sum_{i=1}^{n} a(i) + \frac{1}{n}\sum_{i=1}^{n} a(n+1-i) \\ &= \frac{1}{n}\sum_{i=1}^{n} a(i) + \frac{1}{n}\sum_{i=1}^{n} a(i) = 2\overline{a} . \end{aligned}$$

Sostituendo il valore ottenuto per k nelle relazioni originali si ha quindi

$$2\overline{a} = a^*(i) + a^*(n+1-i) \, , \, i = 1, \ldots, n \, ,$$

ovvero

$$a^*(i) - \overline{a} = \overline{a} - a^*(n+1-i) \, , \, i = 1, \ldots, n .$$

Data ora la statistica lineare dei ranghi

$$T^* = \sum_{i=1}^{n} c(i)a^*(R_i) \, ,$$

per il Teorema 7.1.6 si ha $T^* \stackrel{d}{=} T$. Tenendo presente questa equivalenza in distribuzione, allora è sufficiente dimostrare che T^* è simmetrica rispetto a $\mathrm{E}(T) = \mathrm{E}(T^*) = n\overline{a}\,\overline{c}$. Dalla relazione ottenuta in precedenza si ottiene

$$T^* - n\overline{a}\,\overline{c} = \sum_{i=1}^{n} c(i)(a^*(R_i) - \overline{a}) = \sum_{i=1}^{n} c(i)(\overline{a} - a^*(n+1-R_i)) \, .$$

Dal momento che per il Lemma 7.1.4 si ha

$$(n+1-R_1, \ldots, n+1-R_n) \stackrel{d}{=} (R_1, \ldots, R_n) \, ,$$

allora

$$T^* - n\overline{a}\,\overline{c} = \sum_{i=1}^{n} c(i)(\overline{a} - a^*(n+1-R_i)) \stackrel{d}{=} \sum_{i=1}^{n} c(i)(\overline{a} - a^*(R_i)) = n\overline{a}\,\overline{c} - T^* \, ,$$

ovvero per il Teorema 1.2.2 si ha che T^* è simmetrica rispetto a $n\overline{a}\,\overline{c}$. La dimostrazione della seconda condizione segue immediatamente dal Teorema 7.1.5 mediante la stessa dimostrazione della prima condizione. □

• **Esempio 7.1.4.** Si consideri la scelta di costanti di regressione dell'Esempio 7.1.1 e si supponga $n_1 = n_2$. Dal momento si ha $c^*(i) = c(n+1-i)$ per $i = 1, \ldots, n$, allora risulta

$$c^*(i) + c^*(n+1-i) = 1 \, , i = 1, \ldots, n \, ,$$

ovvero dal Teorema 7.1.7 si ottiene che la relativa statistica lineare dei ranghi T è simmetrica. ◁

• **Esempio 7.1.5.** Si consideri la statistica W dell'Esempio 7.1.1. Dal momento che $a^*(i) = a(i) = i$ per $i = 1, \ldots, n$, risulta

$$a^*(i) + a^*(n+1-i) = i + n + 1 - i = n + 1 \, , i = 1, \ldots, n \, ,$$

per cui dal Teorema 7.1.7 e tenendo presente l'Esempio 7.1.2, si ha che W è simmetrica rispetto a $\mathrm{E}(W) = n_1(n+1)/2$. ◁

• **Esempio 7.1.6.** Si consideri la statistica L dell'Esempio 7.1.1 e si supponga che n sia pari. Dal momento che $a^*(i) = a(i)$ per $i = 1, \ldots, n$, risulta

$$a^*(i) + a^*(n+1-i) = 1 \, , i = 1, \ldots, n \, ,$$

per cui dal Teorema 7.1.7 e tenendo presente l'Esempio 7.1.3, si ha che L è simmetrica rispetto a $\mathrm{E}(L) = n_1/2$. ◁

7.2. La distribuzione per grandi campioni delle statistiche lineari dei ranghi. In questa sezione vengono discusse le proprietà delle statistiche lineari dei ranghi per grandi campioni.

Definizione 7.2.1. Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, data una statistica lineare dei ranghi

$$T = T_n = \sum_{i=1}^{n} c_n(i)a_n(R_i) \, ,$$

si dice che le costanti di regressione $c_n(1), \ldots, c_n(n)$ soddisfano alla condizione di Noether se

$$\lim_n \frac{\sum_{i=1}^n (c_n(i) - \overline{c}_n)^2}{\max\limits_{1\le i\le n} (c_n(i) - \overline{c}_n)^2} = \infty \,,$$

dove

$$\overline{c}_n = \frac{1}{n}\sum_{i=1}^n c_n(i) \,. \qquad \triangle$$

• **Esempio 7.2.1.** Si consideri le costanti di regressione definite nell'Esempio 7.1.1. Dal momento che $\overline{c}_n = n_1/n$, si ha

$$\sum_{i=1}^n (c_n(i) - \overline{c}_n)^2 = \sum_{i=1}^n c_n(i)^2 - n\overline{c}_n^2 = \frac{n_1 n_2}{n} \,.$$

Inoltre si ha

$$\max_{1\le i\le n} (c_n(i) - \overline{c}_n)^2 = \max(\overline{c}_n^2, (1-\overline{c}_n)^2) = \frac{1}{n^2}\max(n_1^2, n_2^2) = \frac{1}{n^2}\left(\max(n_1, n_2)\right)^2 \,,$$

per cui

$$\frac{\sum_{i=1}^n (c_n(i) - \overline{c}_n)^2}{\max\limits_{1\le i\le n} (c_n(i) - \overline{c}_n)^2} = \frac{n_1 n_2 n}{(\max(n_1, n_2))^2} = n\,\frac{\min(n_1, n_2)}{\max(n_1, n_2)} \,,$$

ovvero la condizione di Noether è soddisfatta quando $n_1, n_2 \to \infty$ contemporaneamente. $\triangleleft$

Nel seguente teorema si ottiene la distribuzione per grandi campioni delle statistiche lineari dei ranghi.

Teorema 7.2.2. *Si consideri punteggi tali che*

$$a_n(i) = \phi(i/(n+1)) \,,\, i = 1, \ldots, n \,,$$

dove ϕ è una funzione punteggio esprimibile come differenza di due funzioni non decrescenti che non dipende da n per cui

$$0 < \int_0^1 (\phi(u) - \overline{\phi})^2 \, du < \infty \,,$$

dove $\overline{\phi} = \int_0^1 \phi(u) du$. Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, allora per una statistica lineare dei ranghi

$$T_n = \sum_{i=1}^n c_n(i) a_n(R_i) \,,$$

le cui costanti di regressione soddisfano la condizione di Noether, si ha

$$\frac{T_n - \mathrm{E}(T_n)}{\sqrt{\mathrm{Var}(T_n)}} \xrightarrow{d} N(0,1) \,,$$

dove $\mathrm{E}(T_n)$ *e* $\mathrm{Var}(T_n)$ *sono definite nel Teorema 7.1.3.*

Dimostrazione. Vedi Hájek e Šidák (1967). $\square$

Se si considera la statistica T_n' i cui punteggi sono dati da

$$a_n'(i) = b_n\phi(i/(n+1)) + d_n = b_n a_n(i) + d_n \,,\, i = 1, \ldots, n \,,$$

allora risulta $T'_n = b_n T_n + n\overline{c}_n d_n$ con

$$\mathrm{E}(T'_n) = b_n \mathrm{E}(T_n) + n\overline{c}_n d_n$$

e

$$\mathrm{Var}(T'_n) = b_n^2 \mathrm{Var}(T_n) \, .$$

Di conseguenza, si ha

$$\frac{T'_n - \mathrm{E}(T'_n)}{\sqrt{\mathrm{Var}(T'_n)}} = \frac{b_n T_n + n\overline{c}_n d_n - b_n \mathrm{E}(T_n) - n\overline{c}_n d_n}{\sqrt{b_n^2 \mathrm{Var}(T_n)}} = \frac{T_n - \mathrm{E}(T_n)}{\sqrt{\mathrm{Var}(T_n)}} \, ,$$

da cui segue che T_n e T'_n hanno le medesime proprietà per grandi campioni. Dunque, le costanti b_n e d_n non influenzano il comportamento per grandi campioni della statistica dei ranghi, mentre la scelta della funzione punteggio ϕ è determinante sotto questo punto di vista.

• **Esempio 7.2.2.** Si consideri la statistica $W = W_n$ di Mann-Whitney-Wilcoxon. Dall'Esempio 7.2.1 è noto che le relative costanti di regressione soddisfano alla condizione di Noether. Inoltre, al fine di determinare la distribuzione per grandi campioni di W_n è conveniente considerare la statistica $T_n = b_n W_n$ con $b_n = 1/(n+1)$ che ha le medesime proprietà per grandi campioni della statistica W_n. La funzione punteggio relativa alla statistica T_n, data da $\phi(u) = u\mathbf{1}_{[0,1]}(u)$, può essere espressa come differenza di due funzioni non decrescenti. Inoltre, risulta

$$\int_0^1 (\phi(u) - \overline{\phi})^2 \, du = \int_0^1 (u - 1/2)^2 \, du = \frac{1}{12} < \infty \, .$$

Dunque, tutte le condizioni del Teorema 7.2.2 sono soddisfatte. Quindi, tenendo presente anche l'Esempio 7.1.2, si ha

$$\frac{W_n - \mathrm{E}(W_n)}{\sqrt{\mathrm{Var}(W_n)}} = \frac{W_n - n_1(n+1)/2}{\sqrt{n_1 n_2 (n+1)/12}} \xrightarrow{d} N(0,1) \, . \qquad \triangleleft$$

• **Esempio 7.2.3.** Si consideri la statistica $L = L_n$ della mediana. Dall'Esempio 7.2.1 è noto che le relative costanti di regressione soddisfano alla condizione di Noether. La funzione punteggio della statistica L_n, data da $\phi(u) = \mathbf{1}_{[1/2,1]}(u)$, può essere espressa come differenza di due funzioni non decrescenti. Inoltre, risulta

$$\int_0^1 (\phi(u) - \overline{\phi})^2 \, du = \int_0^{1/2} (1/2)^2 \, du + \int_{1/2}^1 (1 - 1/2)^2 du = \frac{1}{4} < \infty \, .$$

Dunque, tutte le condizioni del Teorema 7.2.2 sono soddisfatte. Quindi, tenendo presente anche l'Esempio 7.1.3, dal Teorema 7.2.2 si ottiene

$$\frac{L_n - \mathrm{E}(L_n)}{\sqrt{\mathrm{Var}(L_n)}} = \frac{L_n - n_1 \lfloor n/2 \rfloor / n}{\sqrt{n_1 n_2 \lfloor n/2 \rfloor (n - \lfloor n/2 \rfloor)/(n^2(n-1))}} \xrightarrow{d} N(0,1) \, . \qquad \triangleleft$$

Capitolo 8

I test per i parametri di posizione: due campioni indipendenti

8.1. Le statistiche lineari dei ranghi per i parametri di posizione. Consideriamo due campioni casuali indipendenti $(X_1, \dots, X_{n_1})$ e $(Y_1, \dots, Y_{n_2})$ e supponiamo che il campione misto $(X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2})$ abbia funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$, dove $n = n_1 + n_2$. Tenedo presente la definizione della classe $\mathcal{L}_{\Delta,F}$, il parametro Δ rappresenta in effetti la differenza fra i parametri di posizione relativi alle variabili casuali da cui provengono i due campioni. Si vuole verificare il sistema di ipotesi $H_0 : \Delta = 0, F \in \mathcal{C}$, ovvero l'omogeneità dei parametri di posizione, contro un'alternativa bilaterale $H_1 : \Delta \neq 0, F \in \mathcal{C}$, o direzionale $H_1 : \Delta > 0\,(\Delta < 0), F \in \mathcal{C}$. Una classe di statistiche test "distribution-free" opportuna in questo sistema di ipotesi è definita di seguito.

Definizione 8.1.1. Siano $(X_1, \dots, X_{n_1})$ e $(Y_1, \dots, Y_{n_2})$ due campioni casuali indipendenti, tali che il campione misto abbia funzione di ripartizione congiunta $F_n \in \mathcal{L}_{0,F}$, dove $n = n_1 + n_2$. Siano inoltre $(R_1, \dots, R_{n_1})$ i ranghi assegnati a $(X_1, \dots, X_{n_1})$ e siano $(R_{n_1+1}, \dots, R_n)$ i ranghi assegnati a $(Y_1, \dots, Y_{n_2})$ nel campione misto. Se le costanti di regressione nella Definizione 7.1.1 sono date da

$$c(i) = \begin{cases} 1 & i = 1, \dots, n_1 \\ 0 & i = n_1 + 1, \dots, n \end{cases}$$

e i punteggi $a(i)$, per $i = 1, \dots, n$, sono tali che

$$0 \leq a(1) \leq \dots \leq a(n)\,,\, a(1) \neq a(n)\,,$$

una statistica del tipo

$$T = \sum_{i=1}^{n} c(i)a(R_i) = \sum_{i=1}^{n_1} a(R_i)$$

è detta statistica lineare dei ranghi per i parametri di posizione. △

Per il Corollario 2.4.6, sotto ipotesi di base la statistica T è "distribution-free" sulla classe $\mathcal{L}_{0,F} = \mathcal{C}_F$. La statistica T è sensibile a variazioni nel parametro Δ, in quanto se non è vera l'ipotesi di base T tende ad assumere valori piccoli o elevati.

• **Esempio 8.1.1.** Dall'Esempio 7.1.1 si verifica che la statistica W di Mann-Whitney-Wilcoxon e la statistica L della mediana sono statistiche lineari dei ranghi per i parametri di posizione. ◁

Il seguente teorema fornisce la media e la varianza di una statistica lineare dei ranghi per i parametri di posizione quando è vera l'ipotesi di base.

Teorema 8.1.2. *Se il campione casuale misto $(X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2})$ ha funzione di ripartizione congiunta $F_n \in \mathcal{L}_{0,F}$, allora per una statistica lineare dei ranghi per i parametri di posizione T risulta*

$$\mathrm{E}(T) = n_1\overline{a}\,,\, \mathrm{Var}(T) = \frac{n_1 n_2}{n-1}\, s_a^2\,,$$

dove $\overline{a}$ e s_a^2 sono definite nel Lemma 7.1.2.

Dimostrazione. Con la stessa simbologia del Teorema 7.1.3 si ha

$$\overline{c} = \frac{1}{n}\sum_{i=1}^{n_1} 1 = \frac{n_1}{n}$$

e

$$s_c^2 = \frac{1}{n}\sum_{i=1}^{n_1} 1 - \frac{n_1^2}{n^2} = \frac{n_1}{n} - \frac{n_1^2}{n^2} = \frac{n_1(n-n_1)}{n^2} = \frac{n_1 n_2}{n^2}\,,$$

e quindi ancora dal Teorema 7.1.3 risulta

$$\mathrm{E}(T) = n\overline{a}\,\frac{n_1}{n} = n_1\overline{a}$$

e

$$\mathrm{Var}(T) = \frac{n^2}{n-1}\,s_a^2\,\frac{n_1 n_2}{n^2} = \frac{n_1 n_2}{n-1}\,s_a^2\,. \qquad \square$$

Il seguente teorema fornisce le condizioni per cui una statistica lineare dei ranghi per i parametri di posizione risulta simmetrica quando è vera l'ipotesi di base.

Teorema 8.1.3. *Se il campione casuale misto $(X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2})$ ha funzione di ripartizione congiunta $F_n \in \mathcal{L}_{0,F}$, allora una statistica lineare dei ranghi per i parametri di posizione T è simmetrica rispetto a* $\mathrm{E}(T)$ *se si ha*

$$n_1 = n_2\,,$$

o

$$a(i) + a(n+1-i) = k\,,\ i = 1,\ldots,n\,,$$

dove k è una costante.

Dimostrazione. Risulta immediata considerando il Teorema 7.1.7 e l'Esempio 7.1.4. $\square$

Il seguente teorema consente di ottenere la funzione generatrice di probabilità di una statistica lineare dei ranghi per i parametri di posizione nel caso che i punteggi siano valori interi.

Teorema 8.1.4. *Se il campione casuale misto $(X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2})$ ha funzione di ripartizione congiunta $F_n \in \mathcal{L}_{0,F}$, allora la funzione generatrice di probabilità di una statistica lineare dei ranghi per i parametri di posizione T è data da $\binom{n}{n_1}^{-1}$ volte il coefficiente di ordine n_1 del polinomio in u*

$$\prod_{i=1}^{n}(1+uv^{a(i)})\,.$$

Dimostrazione. In base al Teorema 7.1.5 risulta

$$T \stackrel{d}{=} \sum_{i=1}^{n} Z_i a(i)\,,$$

dove $Z_i = c(R_i)$ per $i = 1,\ldots,n$. La variabile casuale Z_i vale 1 se $R_i \le n_1$ e 0 altrimenti, per $i = 1,\ldots,n$, ed inoltre risulta $\sum_{i=1}^{n} Z_i = n_1$. Supponiamo inoltre che n_1 sia la realizzazione di una variabile casuale N_1 con distribuzione Binomiale $Bi(n,1/2)$. Per un dato n_1, dal momento che vi sono $\binom{n}{n_1}$ modi di assegnare gli n_1 ranghi più bassi al campione casuale, si ha

$$\Pr(Z_1 = z_1, \ldots, Z_n = z_n \mid N_1 = n_1) = \binom{n}{n_1}^{-1} \mathbf{1}_A(z_1, \ldots, z_n) \,,$$

dove $A = \{(z_1, \ldots, z_n) : z_i = 0, 1, i = 1, \ldots, n, \sum_{i=1}^n z_i = n_1\}$, da cui

$$\begin{aligned} \Pr(Z_1 = z_1, \ldots, Z_n = z_n, N_1 = n_1) &= \Pr(Z_1 = z_1, \ldots, Z_n = z_n \mid N_1 = n_1)\Pr(N_1 = n_1) \\ &= 2^{-n}\mathbf{1}_A(z_1, \ldots, z_n)\mathbf{1}_{\{0,1,\ldots,n\}}(n_1) \,. \end{aligned}$$

Dunque, risulta anche

$$\Pr(Z_1 = z_1, \ldots, Z_n = z_n) = \sum_{n_1=0}^{n} \Pr(Z_1 = z_1, \ldots, Z_n = z_n, N_1 = n_1) = 2^{-n} \prod_{i=1}^{n} \mathbf{1}_{\{0,1\}}(z_i) \,.$$

Le Z_i risultano quindi indipendenti ed ugualmente distribuite con distribuzione Binomiale $Bi(1, 1/2)$, per $i = 1, \ldots, n$. Inoltre, dal momento che la funzione generatrice delle probabilità congiunta di Z_i e di $T_i = Z_i a(i)$ risulta

$$L_{Z_i,T_i}(u_i, v_i) = \frac{1}{2} \sum_{z_i=0}^{1} u_i^{z_i} v_i^{z_i a(i)} = \frac{1}{2} \left(1 + u_i v_i^{a(i)}\right) , i = 1, \ldots, n \,,$$

e dal momento che $N_1 = \sum_{i=1}^n Z_i$ e $T = \sum_{i=1}^n T_i$ sono somme di n variabili casuali indipendenti, allora la funzione generatrice delle probabilità congiunta di N_1 e T è data da

$$L_{N_1,T}(u, v) = \prod_{i=1}^{n} L_{Z_i,T_i}(u, v) = 2^{-n} \prod_{i=1}^{n} \left(1 + uv^{a(i)}\right) .$$

Inoltre, denotando con $L_T(v \mid n_1)$ la funzione generatrice delle probabilità di T condizionata al valore n_1 di N_1, risulta anche

$$L_{N_1,T}(u, v) = \sum_{n_1=0}^{n} L_T(v \mid n_1)\Pr(N_1 = n_1)u^{n_1} = 2^{-n} \sum_{n_1=0}^{n} \binom{n}{n_1} L_T(v \mid n_1)u^{n_1} \,,$$

ovvero $\binom{n}{n_1} L_T(v \mid n_1)$ è il coefficiente di ordine n_1 del polinomio in u dato da $\prod_{i=1}^n (1 + uv^{a(i)})$. La distribuzione di N_1 è stata scelta semplicemente per convenienza. □

Quando i punteggi non sono interi il precedente teorema può essere ancora impiegato determinando una trasformata biunivoca che discretizzi i punteggi originali. Dal momento che la distribuzione di una statistica lineare dei ranghi per i parametri di posizione sotto l'ipotesi di base $H_0 : \Delta = 0, F \in \mathcal{C}$, si può ottenere mediante il Teorema 8.1.4, allora si può determinare le appropriate regioni critiche del test. Se l'alternativa è bilaterale, ovvero $H_1 : \Delta \neq 0, F \in \mathcal{C}$, allora il primo campione tende ad assumere i ranghi più bassi o elevati e quindi si respingere l'ipotesi di base per determinazioni sia troppo elevate che troppo piccole di T. Fissato quindi un livello di significatività α, allora si sceglie come regione critica l'insieme

$$\mathcal{T}_1 = \{t : t \leq t_{n_1,n_2,\alpha/2}, t \geq t_{n_1,n_2,1-\alpha/2}\} \,,$$

dove $t_{n_1,n_2,\alpha}$ rappresenta il quantile di ordine α della distribuzione di T per numerosità campionarie pari a n_1 e n_2. Se l'alternativa è direzionale del tipo $H_1 : \Delta > 0, F \in \mathcal{C}$, allora il secondo campione tende ad assumere i ranghi più elevati e quindi si respinge l'ipotesi di base per determinazioni basse di T. Fissato quindi un livello di significatività α, si ha la seguente regione critica

$$\mathcal{T}_1 = \{t : t \leq t_{n_1,n_2,\alpha}\} \,.$$

Al contrario se l'alternativa è direzionale del tipo $H_1 : \Delta < 0, F \in \mathcal{C}$, allora si il primo campione tende ad assumere i ranghi più elevati e quindi si respinge l'ipotesi di base per determinazioni troppo elevate di T. Fissato quindi un livello di significatività α, si ha la seguente regione critica

$$\mathcal{T}_1 = \{t : t \geq t_{n_1,n_2,1-\alpha}\} .$$

Infine, il test basato sulla T per i precedenti sistemi di ipotesi è corretto al livello di significatività α. Infatti, dal momento che si può dimostrare che $P_T(\Delta, F) = \Pr_{\Delta,F}(T \in \mathcal{T}_1)$ è una funzione monotona crescente per $\Delta > 0$ e monotona decrescente per $\Delta < 0$ per ogni $F \in \mathcal{C}$, allora si ha $P_T(\Delta, F) > \alpha$, ovvero il test è corretto. Risulta interessante determinare la scelta dei punteggi che fornisce il test localmente più potente per verificare il sistema di ipotesi $H_0 : \Delta = 0, F = F_0$, contro $H_1 : \Delta > 0, F = F_0$. In questo caso, si ha la seguente definizione.

Definizione 8.1.5. Se il campione casuale misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ ha funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$, dove $n = n_1 + n_2$, si consideri il sistema di ipotesi $H_0 : \Delta = 0, F = F_0$, contro $H_1 : \Delta > 0, F = F_0$, dove $F_0 \in \mathcal{C}$. Il test basato sulla statistica lineare dei ranghi per i parametri di posizione T_* è detto localmente più potente se esiste un $\epsilon > 0$ tale che per ogni livello di significatività naturale si ha

$$P_{T_*}(\Delta) \geq P_T(\Delta) \, , \, 0 < \Delta < \epsilon \, ,$$

per ogni statistica lineare dei ranghi per i parametri di posizione T. △

Il prossimo teorema fornisce la scelta ottima dei punteggi per la costruzione del test localmente più potente. L'utilità di questo teorema consiste solamente nell'evidenziare la struttura ottima dei punteggi al variare della funzione di ripartizione, dal momento che questa non è mai nota in pratica.

Teorema 8.1.6. *Si consideri il campione casuale misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$, e sia f la funzione di densità corrispondente alla funzione di ripartizione F. Si assuma inoltre che f' esista, sia assolutamente continua e*

$$\int_{\mathbb{R}} |f'(x)| \, dx < \infty .$$

Il test localmente più potente per verificare il sistema di ipotesi $H_0 : \Delta = 0, F = F_0$, contro $H_1 : \Delta > 0, F = F_0$, è basato sulla statistica lineare dei ranghi per i parametri di posizione

$$T_* = \sum_{i=1}^{n_1} a_*(R_i) \, ,$$

dove

$$a_*(i) = \mathrm{E}(- \frac{f'(V_{(i)})}{f(V_{(i)})}) \, , \, i = 1, \ldots, n \, ,$$

e $(V_{(1)}, \ldots, V_{(n)})$ è la statistica ordinata relativa a $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$.

Dimostrazione. Si veda Hettmansperger e McKean (1998). □

Analogamente a quanto visto per la scelta ottima dei punteggi nel caso di una statistica lineare dei ranghi con segno, anche in questo caso si può dimostrare che la scelta ottima dei punteggi non dipende dal parametro di posizione e di scala della distribuzione.

• **Esempio 8.1.2.** Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$, dove F è la funzione di ripartizione di una variabile casuale Normale $N(0, 1)$ e f rappresenta la relativa funzione di densità. Le condizioni del Teorema 8.1.6 sono soddisfatte. Inoltre, analogamente all'Esempio 5.2.1, si ha $-f'(x)/f(x) = x$, per cui la scelta ottimale dei punteggi è data da

$$a^*(i) = \mathrm{E}(V_{(i)}) \, , \, i = 1, \ldots, n \, ,$$

dove $(V_{(1)}, \ldots, V_{(n)})$ è la statistica ordinata relativa ad un campione casuale proveniente da una distribuzione Normale $N(0,1)$. ◁

• **Esempio 8.1.3.** Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$, dove F è la funzione di ripartizione di una variabile casuale Logistica $Lo(0,1)$ e f rappresenta la relativa funzione di densità. Le condizioni del Teorema 8.1.6 sono soddisfatte. Inoltre, analogamente all'Esempio 5.2.2, si ha $-f'(x)/f(x) = 2F(x) - 1$, per cui la scelta ottimale dei punteggi è data da

$$a^*(i) = \mathrm{E}(2F(V_{(i)}) - 1) = 2\mathrm{E}(F(V_{(i)})) - 1 \, , \, i = 1, \ldots, n \, ,$$

dove $(V_{(1)}, \ldots, V_{(n)})$ è la statistica ordinata relativa ad un campione casuale proveniente da una distribuzione Logistica $Lo(0,1)$. Tuttavia, dal momento che $F(V_{(i)}) \stackrel{d}{=} U_{(i)}$, per $i = 1, \ldots, n$, dove $(U_{(1)}, \ldots, U_{(n)})$ rappresenta la statistica ordinata relativa ad una campione casuale proveniente da una distribuzione Uniforme $U(0,1)$, si ottiene anche

$$a^*(i) = 2\mathrm{E}(U_{(i)}) - 1 = \frac{2i}{n+1} - 1 \, , \, i = 1, \ldots, n \, .$$

La statistica lineare dei ranghi per i parametri di posizione costruita su questi punteggi è data da

$$T_* = \sum_{i=1}^{n_1} \frac{2R_i}{n+1} - n_1 = \frac{2}{n+1} W - n_1 \, ,$$

dove W è la statistica lineare dei ranghi di Mann-Whitney-Wilcoxon. Quindi T_* e W forniscono test equivalenti, ovvero la scelta fatta per la statistica di Mann-Whitney-Wilcoxon risulta ottima per una variabile casuale Logistica. ◁

8.2. La distribuzione per grandi campioni delle statistiche lineari dei ranghi per i parametri di posizione. In questa sezione vengono considerate le proprietà per grandi campioni delle statistiche lineari dei ranghi per i parametri di posizione.

Teorema 8.2.1. *Se $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ è un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{0,F}$, allora per una statistica lineare dei ranghi per i parametri di posizione*

$$T = T_n = \sum_{i=1}^{n_1} a_n(R_i) \, ,$$

i cui punteggi $a_n(i)$, per $i = 1, \ldots, n$, soddisfano le condizioni del Teorema 7.2.2, risulta

$$\frac{T_n - \mathrm{E}(T_n)}{\sqrt{\mathrm{Var}(T_n)}} \stackrel{d}{\to} N(0,1) \, ,$$

dove $\mathrm{E}(T_n)$ e $\mathrm{Var}(T_n)$ sono definite nel Teorema 8.1.2.

Dimostrazione. Segue dall'Esempio 7.2.1 e dal Teorema 7.2.2. □

Fissato un livello di significatività α, per $n \to \infty$ la regione critica per verificare il sistema di ipotesi $H_0 : \Delta = 0, F \in \mathcal{C}$, contro l'alternativa $H_1 : \Delta \neq 0, F \in \mathcal{C}$, può essere approssimata dall'insieme

$$\{t : t \leq \mathrm{E}(T_n) + z_{\alpha/2}\sqrt{\mathrm{Var}(T_n)}, t \geq \mathrm{E}(T_n) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(T_n)}\} \, .$$

Analogamente, per $n \to \infty$ la regione critica per verificare l'alternativa $H_1 : \Delta > 0, F \in \mathcal{C}$, può essere approssimata dall'insieme

$$\{t : t \leq \mathrm{E}(T_n) + z_{\alpha}\sqrt{\mathrm{Var}(T_n)}\} \, ,$$

mentre la regione critica per verificare l'alternativa $H_1 : \Delta < 0, F \in \mathcal{C}$, può essere approssimata dall'insieme

$$\{t : t \geq \mathrm{E}(T_n) + z_{1-\alpha}\sqrt{\mathrm{Var}(T_n)}\} .$$

Mediante il Teorema 3.1.8 si può dimostrare inoltre che la successione di test basata su $(T_n)_{n\geq 1}$ è coerente. Per quanto riguarda l'efficacia delle statistiche lineari dei ranghi per i parametri di posizione si ha il seguente teorema.

Teorema 8.2.2. *Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$, e si consideri il sistema di ipotesi $H_0 : \Delta = 0, F \in \mathcal{C}$, contro $H_1 : \Delta = \Delta_i, F \in \mathcal{C}$, dove $(\Delta_i)_{i\geq 1}$ è una successione di alternative tali che $\Delta_i = c/\sqrt{n_i}$ con c costante. Data una statistica lineare dei ranghi per i parametri di posizione*

$$T = T_n = \sum_{i=1}^{n_1} a_n(R_i) ,$$

i cui punteggi $a_n(i)$, per $i = 1, \ldots, n$, soddisfano le condizioni del Teorema 8.2.1, allora l'efficacia del test basato su T_n risulta

$$\mathrm{eff}_T = \sqrt{\nu(1-\nu)}\, \frac{\int_0^1 \phi(u)\phi_f(u)du}{(\int_0^1 (\phi(u) - \overline{\phi})^2 du)^{1/2}} ,$$

dove $\nu = \lim_n n_1/n$, per $0 < \nu < 1$, e

$$\phi_f(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))} .$$

Dimostrazione. Vedi Háyek e Šidák (1967). □

• **Esempio 8.2.1.** Si consideri la statistica $W = W_n$ del test di Mann-Whitney-Wilcoxon dell'Esempio 8.1.1. Dall'Esempio 7.2.2 risulta che le condizioni del Teorema 8.2.2 sono soddisfatte. Inoltre, si ha

$$\int_0^1 \phi(u)\phi_f(u)\, du = \int_0^1 -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\, u\, du ,$$

da cui, mediante la trasformazione di variabile $x = F^{-1}(u)$ con $u = F(x)$, si ha

$$\int_0^1 \phi(u)\phi_f(u)du = \int_{-\infty}^{\infty} -\frac{f'(x)}{f(x)}\, F(x)f(x)dx = -\int_{-\infty}^{\infty} f'(x)F(x)dx = \int_{-\infty}^{\infty} f(x)^2 dx .$$

Quindi, tenendo presente l'Esempio 7.2.2, l'efficacia del test di Mann-Whitney-Wilcoxon risulta

$$\mathrm{eff}_W = \sqrt{12\nu(1-\nu)} \int_{-\infty}^{\infty} f(x)^2\, dx . \qquad \triangleleft$$

• **Esempio 8.2.2.** Si consideri la statistica $L = L_n$ del test della mediana dell'Esempio 8.1.1. Dall'Esempio 7.2.3 risulta che le condizioni del Teorema 8.2.2 sono soddisfatte. Inoltre, si ha

$$\int_0^1 \phi(u)\phi_f(u)\, du = \int_{1/2}^1 -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\, du ,$$

da cui, mediante la trasformazione di variabile $x = F^{-1}(u)$ con $u = F(x)$, se $F(x_{0.5}) = 1/2$, si ha

$$\int_0^1 \phi(u)\phi_f(u)du = \int_{x_{0.5}}^{\infty} -\frac{f'(x)}{f(x)}\, f(x)dx = -\int_{x_{0.5}}^{\infty} f'(x)dx = f(x_{0.5}) .$$

Quindi, tenendo presente l'Esempio 7.2.3, l'efficacia del test della mediana risulta

$$\text{eff}_L = \sqrt{\nu(1-\nu)}\,\frac{f(x_{0.5})}{\sqrt{1/4}} = \sqrt{4\nu(1-\nu)}\,f(x_{0.5})\,. \qquad \triangleleft$$

8.3. Il test di Mann-Whitney-Wilcoxon. Sia $(X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$. Il test di Mann-Whitney-Wilcoxon è basato sulla statistica

$$W = \sum_{i=1}^{n_1} R_i\,,$$

descritta nell'Esempio 2.4.1. Con questo test si può verificare l'ipotesi di base $H_0 : \Delta = 0, F \in \mathcal{C}$. La statistica W è una statistica lineare dei ranghi per i parametri di posizione con i punteggi scelti come $a(i) = i$ per $i = 1,\ldots,n$. Se l'ipotesi di base è vera, dall'Esempio 2.4.1 la funzione di probabilità della statistica W risulta

$$p_{n_1,n_2}(w) = \binom{n}{n_1}^{-1} c_{n_1,n_2}(w)\mathbf{1}_{\{n_1(n_1+1)/2,\ldots,n_1(n+n_2+1)/2\}}(w)\,,$$

dove $c_{n_1,n_2}(w)$ è il numero di sottoinsiemi di n_1 interi dell'insieme $\{1,\ldots,n\}$ la cui somma è w. Sebbene esistano delle relazioni ricorrenti per il calcolo della funzione di probabilità di W, la distribuzione di questa statistica si ottiene più facilmente attraverso la funzione generatrice di probabilità. Infatti, dal Teorema 8.1.4 la funzione generatrice di probabilità $L_W(v)$ di W è $\binom{n}{n_1}^{-1}$ volte il coefficiente di ordine n_1 del polinomio in u dato da $\prod_{i=1}^{n}(1+uv^i)$.

• **Esempio 8.3.1.** Per $n_1 = 2$ e $n_2 = 2$ si ha

$$\prod_{i=1}^{4}(1+uv^i) = 1 + (v+v^2+v^3+v^4)u + \\ + (v^3+v^4+2v^5+v^6+v^7)u^2 + (v^6+v^7+v^8+v^9)u^3 + v^{10}u^4$$

da cui

$$L_W(v) = \frac{1}{6}\,(v^3+v^4+2v^5+v^6+v^7)\,.$$

Dal momento che le probabilità $p_{2,2}(w)$ corrispondono ai coefficienti del polinomio $L_W(v)$, si ha la Tavola 8.3.1.

Tavola 8.3.1. Funzione di probabilità di W per $n_1 = 2$ e $n_2 = 2$.

w	3	4	5	6	7
$p_{2,2}(w)$	1/6	1/6	2/6	1/6	1/6

$\triangleleft$

Se l'ipotesi di base è vera, dall'Esempio 7.1.2 si ha

$$\text{E}(W) = \frac{n_1(n+1)}{2}\,,\ \text{Var}(W) = \frac{n_1 n_2(n+1)}{12}\,.$$

Inoltre, dall'Esempio 7.1.5 si ha che W è simmetrica rispetto a $\text{E}(W)$. Infine, dal Corollario 2.4.6 la statistica W è "distribution-free" su $\mathcal{L}_{0,F} = \mathcal{C}_F$ e di conseguenza anche il test di Mann-Whitney-Wilcoxon è "distribution-free". Dal momento che la distribuzione della statistica W sotto ipotesi di base è specificata, le appropriate regioni critiche del test per alternative bilaterali o direzionali si possono ottenere tenendo presente la discussione fatta nella §8.1.

• **Esempio 8.3.2.** In un esperimento due gruppi di piccioni sono stati presi dai loro nidi nella campagna a Siena e sono stati portati in una località vicino Roma. Il gruppo di controllo (composto da 8 piccioni) è stato trasportato in un container non coperto, dove passava aria naturale durante il trasporto. Un secondo gruppo (composto da 10 piccioni) invece è stato trasportato in un container coperto e riceveva solamente aria condizionata. I piccioni sono stati lasciati liberi dopo 172 km, in una località vicino Roma (Siena rimane rispetto a questa località in una direzione di 325°). Le direzioni (in gradi, dove il nord è 0) verso le quali i piccioni sono volati sono contenute nella Tavola 8.3.2.

Tavola 8.3.2. Direzioni di fuga (in gradi).

	guppo controllo		gruppo esperimento	
piccione	x_i	r_i	y_i	r_i
1	247	16	107	4
2	153	10	109	5
3	202	14	186	12
4	264	17	121	8
5	24	2	171	11
6	228	15	4	1
7	333	18	110	6
8	192	13	82	3
9			131	9
10			117	7

Fonte: Batschelet (1981)

Si vuole verificare se le condizioni di viaggio hanno alterato la capacità di orientamento dei piccioni, ovvero si vuole verificare il sistema di ipotesi $H_0 : \Delta = 0, F \in \mathcal{C}$, contro $H_1 : \Delta \neq 0, F \in \mathcal{C}$. Dal momemto che per questi dati risulta $w = 105$, per $n_1 = 8$ e $n_2 = 10$ si ha $\Pr(W \geq 105) = 0.0043$ e quindi la significatività osservata risulta $\alpha_{oss} = 2 \times 0.0043 = 0.0086$. Questa è una significatività osservata piuttosto bassa che porta a respingere l'ipotesi di base. Si può concludere che le condizioni di trasporto hanno alterato le capacità di orientamento dei piccioni, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha > 0.0086$. ◁

Per quanto riguarda la distribuzione per grandi campioni della statistica $W = W_n$ sotto ipotesi di base, dall'Esempio 7.2.2 si ha

$$\frac{W_n - n_1(n+1)/2}{\sqrt{n_1 n_2 (n+1)/12}} \xrightarrow{d} N(0,1) \, .$$

La convergenza della statistica di Mann-Whitney-Wilcoxon è abbastanza veloce anche per campioni moderati, posto che entrambe le numerosità siano abbastanza elevate, ovvero $n_1 \geq 10$ e $n_2 \geq 10$. Le approssimazioni per grandi campioni delle regioni critiche del test per alternative bilaterali o direzionali si possono ottenere tenendo presente la discussione nella §8.2.

8.4. Il test della mediana. Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\Delta,F}$. Il test della mediana è basato sulla statistica

$$L = \sum_{i=1}^{n_1} a(R_i) \, ,$$

descritta nell'Esempio 7.1.1, con cui si può verificare l'ipotesi di base $H_0 : \Delta = 0, F \in \mathcal{C}$. La statistica L è una statistica lineare dei ranghi per i parametri di posizione i cui punteggi sono definiti nell'Esempio 7.1.1. Come evidenziato nell'Esempio 7.1.1, la statistica L rappresenta il numero di $(X_1, \ldots, X_{n_1})$ maggiori della mediana del campione misto. Il seguente teorema fornisce la funzione di probabilità della statistica L quando è vera l'ipotesi di base.

Teorema 8.4.1. *Se $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ è un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{0,F}$, la funzione di probabilità di L è data da*

$$\Pr(L = l) = \frac{\binom{n_1}{l}\binom{n_2}{\lfloor n/2\rfloor - l}}{\binom{n}{\lfloor n/2\rfloor}} \mathbf{1}_{\{\max(0,\lfloor n/2\rfloor + n_1 - n),\ldots,\min(\lfloor n/2\rfloor, n_1)\}} .$$

Dimostrazione. Sotto ipotesi di base ogni scelta di ranghi è ugualmente probabile. Quindi l'assegnazione di l degli $(n - \lfloor n/2\rfloor)$ ranghi più elevati ad $(X_1, \ldots, X_{n_1})$ è equivalente ad uno schema probabilistico Ipergeometrico, dove si estrae in blocco $\lfloor n/2\rfloor$ elementi da una popolazione di numerosità $n_1 + n_2 = n$ e gli elementi di interesse sono quelli provenienti dalla sottopopolazione di numerosità n_1. □

Se l'ipotesi di base è vera, dall'Esempio 7.1.3 si ha

$$\mathrm{E}(L) = \frac{n_1\lfloor n/2\rfloor}{n} \text{ , } \mathrm{Var}(L) = \frac{n_1 n_2 \lfloor n/2\rfloor (n - \lfloor n/2\rfloor)}{n^2(n-1)} .$$

Inoltre, dall'Esempio 7.1.4 e dall'Esempio 7.1.6 si ha che L è simmetrica rispetto a $\mathrm{E}(L)$ quando $n_1 = n_2$ o quando n è pari. Infine, dal Corollario 2.4.6 risulta che la statistica L è "distribution-free" su $\mathcal{L}_{0,F} = \mathcal{C}_F$ e di conseguenza anche il test della mediana è "distribution-free". Dal momento che la distribuzione della statistica L sotto ipotesi di base è specificata, le appropriate regioni critiche del test per alternative bilaterali o direzionali si possono facilmente ottenere tenendo presente la discussione fatta nella §8.1. Per quanto riguarda la distribuzione per grandi campioni della statistica $L = L_n$ sotto ipotesi di base, dall'Esempio 7.2.3 risulta

$$\frac{L_n - n_1\lfloor n/2\rfloor/n}{\sqrt{n_1 n_2 \lfloor n/2\rfloor (n - \lfloor n/2\rfloor)/(n^2(n-1))}} \xrightarrow{d} N(0,1) .$$

La convergenza della statistica della mediana è abbastanza veloce anche per campioni moderati, posto che entrambe le numerosità siano abbastanza elevate, ovvero $n_1 \geq 10$ e $n_2 \geq 10$. Le approssimazioni per grandi campioni delle regioni critiche del test per alternative bilaterali o direzionali si possono ottenere tenendo presente la discussione fatta nella §8.2.

• **Esempio 8.4.1.** In un esperimento è stata misurata la secrezione di tromboglobulina urinaria in 12 pazienti sani e in 12 pazienti diabetici, ottenendo in questa maniera i dati della Tavola 8.4.1.

Tavola 8.4.1. Secrezione di tromboglobulina.

	sani		diabetici	
paziente	x_i	r_i	y_i	r_i
1	4.1	1	11.6	8
2	6.3	2	12.1	10
3	7.8	3	16.1	12
4	8.5	4	17.8	14
5	8.9	5	24.0	15
6	10.4	6	28.8	17
7	11.5	7	33.9	18
8	12.0	9	40.7	20
9	13.8	11	51.3	21
10	17.6	13	56.2	22
11	24.3	16	61.7	23
12	37.2	19	69.2	24

Fonte: van Oost, Veldhayzen, Timmermans e Sixma (1983)

Si sospetta che i pazienti diabetici abbiano una secrezione di tromboglobulina più elevata dei pazienti sani, ovvero si vuole verificare il sistema di ipotesi $H_0 : \Delta = 0, F \in \mathcal{C}$, contro $H_1 : \Delta > 0, F \in \mathcal{C}$. La mediana del campione misto è compresa fra 16.1 e 17.6, per cui si ha $l = 3$. Dal momento che

$$\Pr(L \leq 3) \simeq \Phi((3 - 12 \times 12/24)/\sqrt{12^4/(24^2 \times 23)}) = \Phi(-2.3979) = 0.0080 ,$$

allora la significatività osservata risulta $\alpha_{oss} \simeq 0.0080$, un valore piuttosto basso che porta a respingere l'ipotesi di base. Sulla base dell'evidenza empirica si può concludere che i pazienti diabetici hanno una

secrezione di tromboglobulina più elevata dei pazienti sani, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha > 0.0080$. L'approssimazione normale è ragionevole, in quanto il valore esatto risulta $\Pr(L \leq 3) = 0.0196$. ◁

8.5. Le prestazioni del test di Mann-Whitney-Wilcoxon e del test della mediana. La potenza del test di Mann-Whitney-Wilcoxon, del test della mediana e del test di Student per due campioni sono state calcolate mediante simulazione per alcune distribuzioni e per numerosità $n_1 = n_2 = 5, 10, 15$. Le alternative scelte sono state $\Delta = 0.0\,\sigma, 0.3\,\sigma, 0.6\,\sigma, 0.9\,\sigma$, dove σ^2 rappresenta la varianza della distribuzione ipotizzata. Per la distribuzione di Cauchy σ denota il valore per cui $\Pr(X \leq \sigma) = \Phi(1)$. I test di Mann-Whitney-Wilcoxon e della mediana sono stati casualizzati al fine di ottenere un livello di significatività pari a $\alpha = 0.05$ per tutti i test. I risultati della simulazione sono riportati nella Tavola 8.5.1. Dalla Tavola 8.5.1 è evidente che il test di Mann-Whitney-Wilcoxon dimostra ottime prestazioni rispetto agli altri due test per tutte le distribuzioni. Il test della mediana dimostra buone prestazioni solo per una distribuzione a code particolamente pesante quale la distribuzione di Cauchy. Il test di Student per due campioni dimostra prestazioni quasi sempre inferiori al test di Mann-Whitney-Wilcoxon. Inoltre, per distribuzioni quali la Cauchy e l'Esponenziale, il test di Student per due campioni non mantiene neppure il livello di significatività.

Tavola 8.5.1. Potenza del test di Mann-Whitney-Wilcoxon, del test della mediana e del test di Student per due campioni.

distribuzione	0.0σ	0.3σ	0.6σ	0.9σ
		$n_1 = 5\, n_2 = 5$		
$U(\lambda{-}1/2, 1)$	0.05(0.05)0.05	0.11(0.09)0.10	0.20(0.13)0.19	0.33(0.19)0.33
$N(\lambda, 1)$	0.05(0.05)0.05	0.11(0.10)0.11	0.21(0.16)0.22	0.36(0.24)0.35
$Lo(\lambda, 1)$	0.05(0.05)0.05	0.12(0.11)0.11	0.23(0.17)0.22	0.38(0.27)0.38
$L(\lambda, 1)$	0.05(0.05)0.05	0.13(0.12)0.12	0.27(0.22)0.25	0.43(0.33)0.42
$C(\lambda, 1)$	0.05(0.05)0.03	0.12(0.11)0.06	0.21(0.19)0.13	0.30(0.29)0.20
$E(\Delta{-}1, 1)$	0.05(0.05)0.04	0.17(0.11)0.14	0.33(0.22)0.29	0.50(0.34)0.46
		$n_1 = 10\, n_2 = 10$		
$U(\lambda{-}1/2, 1)$	0.05(0.05)0.05	0.15(0.10)0.16	0.32(0.17)0.35	0.55(0.29)0.61
$N(\lambda, 1)$	0.05(0.05)0.05	0.15(0.14)0.16	0.32(0.25)0.36	0.60(0.43)0.62
$Lo(\lambda, 1)$	0.05(0.05)0.05	0.15(0.13)0.16	0.37(0.27)0.37	0.62(0.47)0.63
$L(\lambda, 1)$	0.05(0.05)0.05	0.19(0.18)0.17	0.44(0.40)0.38	0.70(0.60)0.65
$C(\lambda, 1)$	0.05(0.05)0.03	0.15(0.16)0.07	0.32(0.33)0.14	0.53(0.54)0.22
$E(\Delta{-}1, 1)$	0.05(0.05)0.04	0.24(0.16)0.19	0.53(0.35)0.41	0.78(0.58)0.65
		$n_1 = 15\, n_2 = 15$		
$U(\lambda{-}1/2, 1)$	0.05(0.05)0.05	0.19(0.11)0.19	0.47(0.26)0.48	0.72(0.40)0.78
$N(\lambda, 1)$	0.05(0.05)0.05	0.19(0.15)0.20	0.48(0.36)0.47	0.78(0.60)0.77
$Lo(\lambda, 1)$	0.05(0.05)0.05	0.20(0.17)0.20	0.49(0.39)0.50	0.81(0.67)0.78
$L(\lambda, 1)$	0.05(0.05)0.05	0.29(0.27)0.21	0.58(0.54)0.51	0.85(0.78)0.78
$C(\lambda, 1)$	0.05(0.05)0.03	0.21(0.22)0.08	0.44(0.48)0.15	0.69(0.74)0.24
$E(\Delta{-}1, 1)$	0.05(0.05)0.05	0.34(0.20)0.23	0.71(0.49)0.53	0.92(0.72)0.80

Per quanto riguarda le prestazioni asintotiche del test di Mann-Whitney-Wilcoxon, dall'Esempio 8.2.1 si ha

$$\text{eff}_W = \sqrt{12\nu(1-\nu)} \int_{-\infty}^{\infty} f(x)^2 dx \,,$$

per cui l'efficienza asintotica relativa del test di Mann-Whitney-Wilcoxon rispetto al test di Student per due campioni risulta

$$\text{EAR}_{W,T} = 12\sigma^2 (\int_{-\infty}^{\infty} f(x)^2 dx)^2 \,.$$

Si è ottenuto la stessa efficienza asintotica del test di Wilcoxon rispetto al test di Student per un campione, per cui si può considerare la Tavola 6.6.2 dove sono tabulati i valori di $\text{EAR}_{W,T} = \text{EAR}_{W^+,T}$ per alcune distribuzioni simmetriche. Dunque, anche in questo caso, dal punto di vista asintotico il test di Mann-Whitney-Wilcoxon dimostra ottime prestazioni. Per una distribuzione non simmetrica quale l'Esponenziale

si verifica inoltre che $\text{EAR}_{W,T} = 3$. Per quanto riguarda l'efficacia del test della mediana, dall'Esempio 8.2.2 si ha

$$\text{eff}_L = \sqrt{4\nu(1-\nu)}\, f(x_{0.5}) \,,$$

per cui l'efficienza asintotica relativa del test della mediana rispetto al test di Student per due campioni risulta

$$\text{EAR}_{L,T} = 4\sigma^2 f(x_{0.5})^2 \,.$$

Si è ottenuto la stessa efficienza asintotica del test dei segni rispetto al test di Student per un campione, per cui si può considerare la Tavola 6.2.2, dove sono tabulati i valori di $\text{EAR}_{L,T} = \text{EAR}_{B,T}$ per alcune distribuzioni simmetriche. Per una distribuzione non simmetrica quale l'Esponeziale risulta $\text{EAR}_{L,T} = 1$. Infine, l'efficienza asintotica relativa del test di Mann-Whitney-Wilcoxon rispetto al test della mediana risulta

$$\text{EAR}_{W,L} = \frac{3}{f(0)^2} \left(\int_{-\infty}^{\infty} f(x)^2 dx\right)^2 \,.$$

Anche in questo caso si è ottenuto la stessa efficienza asintotica del test di Wilcoxon rispetto al test dei segni, per cui si può considerare la Tavola 6.6.3, dove sono tabulati i valori di $\text{EAR}_{W,L} = \text{EAR}_{W^+,B}$ per alcune distribuzioni simmetriche. Per una distribuzione non simmetrica quale l'Esponeziale si verifica inoltre che $\text{EAR}_{W,L} = 3$.

Capitolo 9

I test per i parametri di scala: due campioni indipendenti

9.1. Le statistiche lineari dei ranghi per i parametri di scala. Si considerino due campioni casuali indipendenti $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$, tali che il campione misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ abbia funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$, dove $n = n_1 + n_2$. Dalla definizione della classe $\mathcal{V}_{\eta,F}$, il parametro η rappresenta il rapporto fra i parametri di scala relativi alle variabili casuali da cui provengono i campioni. Inoltre, dalla definizione di $\mathcal{V}_{\eta,F}$ si suppone implicitamente che le due distribuzioni abbiano uguali parametri di posizione. Si verifica il sistema di ipotesi $H_0 : \eta = 1, F \in \mathcal{C}$, contro l'alternativa bilaterale $H_1 : \eta \neq 1, F \in \mathcal{C}$, o direzionale $H_1 : \eta > 1\,(\eta < 1), F \in \mathcal{C}$. Una classe di statistiche test "distribution-free" opportuna in questo sistema di ipotesi è definita di seguito.

Definizione 9.1.1. Si considerino due campioni casuali indipendenti $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$, tali che il campione misto abbia funzione di ripartizione congiunta $F_n \in \mathcal{V}_{1,F}$, dove $n = n_1 + n_2$ e siano $(R_1, \ldots, R_{n_1})$ i ranghi assegnati a $(X_1, \ldots, X_{n_1})$ e $(R_{n_1+1}, \ldots, R_n)$ i ranghi assegnati a $(Y_1, \ldots, Y_{n_2})$ nel campione misto. Se le costanti di regressione nella Definizione 7.2.1 sono date da

$$c(i) = \begin{cases} 1 & i = 1, \ldots, n_1 \\ 0 & i = n_1 + 1, \ldots, n \end{cases}$$

e i punteggi $a(i)$, per $i = 1, \ldots, n$, sono tali che (tipo 1)

$$0 \leq a(1) \leq \ldots \leq a(\lfloor (n+1)/2 \rfloor)\,,\; a(\lfloor (n+1)/2 \rfloor) \geq \ldots \geq a(n) \geq 0\,,$$

o (tipo 2)

$$a(1) \geq \ldots \geq a(\lfloor (n+1)/2 \rfloor) \geq 0\,,\, 0 \leq a(\lfloor (n+1)/2 \rfloor) \leq \ldots \leq a(n)\,,$$

allora una statistica del tipo

$$T = \sum_{i=1}^{n} c(i) a(R_i) = \sum_{i=1}^{n_1} a(R_i)$$

è detta statistica lineare dei ranghi per i parametri di scala. △

Per il Corollario 2.4.6, sotto ipotesi di base la statistica T è "distribution-free" sulla classe $\mathcal{V}_{1,F} = \mathcal{C}_F$. La statistica T è sensibile a variazioni del parametro η, in quanto se non è vera l'ipotesi di base T tende ad assumere valori piccoli o elevati.

• **Esempio 9.1.1.** I punteggi

$$a(i) = (i - (n+1)/2)^2\,,\, i = 1, \ldots, n\,,$$

soddisfano le condizioni della Definizione 9.1.1 (punteggi tipo 2). In questo caso, si ottiene la cosiddetta statistica di Mood data da

$$M = \sum_{i=1}^{n_1} (R_i - (n+1)/2)^2 \, .$$

Anche i punteggi

$$a(i) = (n+1)/2 - |i - (n+1)/2| \, , i = 1, \ldots, n \, ,$$

soddisfano le condizioni della Definizione 9.1.1 (punteggi tipo 1). In questo caso, si ha la cosiddetta statistica di Ansari-Bradley, data da

$$A = \sum_{i=1}^{n_1} ((n+1)/2 - |R_i - (n+1)/2|) \, . \qquad \triangleleft$$

Il seguente teorema fornisce la media e la varianza di una statistica lineare dei ranghi per i parametri di scala quando è vera l'ipotesi di base.

Teorema 9.1.2. *Se il campione casuale misto* $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ *ha funzione di ripartizione congiunta* $F_n \in \mathcal{V}_{1,F}$*, per una statistica lineare dei ranghi per i parametri di scala* T *risulta*

$$\mathrm{E}(T) = n_1 \overline{a} \, , \, \mathrm{Var}(T) = \frac{n_1 n_2}{n-1} \, s_a^2 \, ,$$

dove $\overline{a}$ *e* s_a^2 *sono definite nel Lemma 7.1.2.*

Dimostrazione. E' analoga a quella del Teorema 8.1.2. □

Il seguente teorema fornisce le condizioni per cui una statistica lineare dei ranghi per i parametri di scala è simmetrica quando è vera l'ipotesi di base.

Teorema 9.1.3. *Se il campione casuale misto* $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ *possiede funzione di ripartizione congiunta* $F_n \in \mathcal{V}_{1,F}$*, allora una statistica lineare dei ranghi per i parametri di scala* T *è simmetrica rispetto a* $\mathrm{E}(T)$ *se*

$$n_1 = n_2 \, ,$$

o

$$a^*(i) + a^*(n+1-i) = k \, , i = 1, \ldots, n \, ,$$

dove k *è una costante.*

Dimostrazione. E' analoga a quella del Teorema 8.1.3. □

• **Esempio 9.1.2.** Si consideri la statistica M dell'Esempio 9.1.1. Tenendo presente il Teorema A.2.1, è immediato verificare che

$$\overline{a} = \frac{1}{n} \sum_{r=1}^{n} (r - \frac{n+1}{2})^2 = \frac{1}{n} \sum_{r=1}^{n} r^2 - \frac{(n+1)^2}{4} = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12} \, ,$$

da cui

$$\mathrm{E}(M) = n_1 \overline{a} = \frac{n_1 (n^2-1)}{12} \, .$$

Inoltre, sempre tenendo presente il Teorema A.2.1,

$$
\begin{aligned}
s_a^2 &= \frac{1}{n}\sum_{r=1}^{n}(r-\frac{n+1}{2})^4 - \frac{(n^2-1)^2}{144} \\
&= \frac{1}{n}\sum_{r=1}^{n} r^4 - \frac{2(n+1)}{n}\sum_{r=1}^{n} r^3 + \frac{3(n+1)^2}{2n}\sum_{r=1}^{n} r^2 - \frac{(n+1)^3}{2n}\sum_{r=1}^{n} r + \frac{(n+1)^4}{16} - \frac{(n^2-1)^2}{144} \\
&= \frac{(n^2-1)(n^2-4)}{180}\,,
\end{aligned}
$$

da cui

$$
\mathrm{Var}(M) = \frac{n_1 n_2}{n-1}\, s_a^2 = \frac{n_1 n_2 (n+1)(n^2-4)}{180}\,.
$$

Dal momento che

$$
a(i) + a(n+1-i) = (i-(n+1)/2)^2 + (n+1-i-(n+1)/2)^2 = 2(i-(n+1)/2)^2\,,
$$

allora dal Teorema 9.1.3 risulta che M è simmetrica solo se $n_1 = n_2$. ◁

• **Esempio 9.1.3.** Si consideri la statistica A dell'Esempio 9.1.1. Tenendo presente il Teorema A.2.1, per n pari si ha

$$
\begin{aligned}
\overline{a} &= \frac{1}{n}\sum_{r=1}^{n}(\frac{n+1}{2} - |r-\frac{n+1}{2}|) = \frac{n+1}{2} - \frac{2}{n}\sum_{r=1}^{n/2}(\frac{n+1}{2} - r) \\
&= \frac{n+1}{2} - \frac{1}{n}\sum_{r=1}^{n/2}(n+1) + \frac{2}{n}\sum_{r=1}^{n/2} r = \frac{n+2}{4}\,,
\end{aligned}
$$

e inoltre, tenendo presente l'espressione di $\overline{a}$ nell'Esempio 9.1.2, si ha

$$
\begin{aligned}
s_a^2 &= \frac{1}{n}\sum_{r=1}^{n}(\frac{n+1}{2} - |r-\frac{n+1}{2}| - \frac{n+2}{4})^2 = \frac{1}{n}\sum_{r=1}^{n}(r-\frac{n+1}{2})^2 - \frac{n^2}{16} \\
&= \frac{n^2-1}{12} - \frac{n^2}{16} = \frac{n^2-4}{48}\,.
\end{aligned}
$$

Dunque, per n pari si ha

$$
\mathrm{E}(A) = n_1\overline{a} = \frac{n_1(n+2)}{4}\,,\ \mathrm{Var}(A) = \frac{n_1 n_2}{n-1}\, s_a^2 = \frac{n_1 n_2 (n^2-4)}{48(n-1)}\,.
$$

Inoltre, per n dispari si ha

$$
\begin{aligned}
\overline{a} &= \frac{1}{n}\sum_{r=1}^{n}(\frac{n+1}{2} - |r-\frac{n+1}{2}|) = \frac{n+1}{2} - \frac{2}{n}\sum_{r=1}^{(n-1)/2}(\frac{n+1}{2} - r) \\
&= \frac{n+1}{2} - \frac{1}{n}\sum_{r=1}^{(n-1)/2}(n+1) + \frac{2}{n}\sum_{r=1}^{(n-1)/2} r \\
&= \frac{n+1}{2} - \frac{n^2-1}{2n} + \frac{n^2-12}{4n} = \frac{(n+1)^2}{4n}
\end{aligned}
$$

e inoltre, sempre tenendo presente l'espressione di $\overline{a}$ nell'Esempio 9.1.2, si ha

$$s_a^2 = \frac{1}{n}\sum_{r=1}^{n}(\frac{n+1}{2} - |r - \frac{n+1}{2}| - \frac{(n+1)^2}{4n})^2 = \frac{1}{n}\sum_{r=1}^{n}(r - \frac{n+1}{2})^2 - \frac{(n^2-1)^2}{16n^2}$$
$$= \frac{n^2-1}{12} - \frac{n^2-1}{16n^2} = \frac{(n^2-1)(n^2+3)}{48n^2} .$$

Dunque, per n dispari si ha

$$\mathrm{E}(A) = n_1\overline{a} = \frac{n_1(n+1)^2}{4n} , \mathrm{Var}(A) = \frac{n_1 n_2}{n-1}\, s_a^2 = \frac{n_1 n_2 (n+1)(n^2+3)}{48n^2} .$$

Inoltre, dal momento che

$$a(i) + a(n+1-i) = (n+1)/2 - |i - (n+1)/2| + (n+1)/2 - |n+1-i-(n+1)/2|$$
$$= n + 1 - 2\,|i - (n+1)/2| , i = 1, \ldots, n ,$$

allora dal Teorema 9.1.3 si ha che A è simmetrica solo se $n_1 = n_2$. ◁

Il seguente teorema consente di ottenere la funzione generatrice di probabilità di una statistica lineare dei ranghi per i parametri di scala nel caso che i punteggi siano valori interi.

Teorema 9.1.4. *Se il campione casuale misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ ha funzione di ripartizione congiunta $F_n \in \mathcal{V}_{1,F}$, allora la funzione generatrice di probabilità di una statistica lineare dei ranghi per i parametri di scala T è data da $\binom{n}{n_1}^{-1}$ volte il coefficiente di ordine n_1 del polinomio in u*

$$\prod_{i=1}^{n}(1 + uv^{a(i)}) .$$

Dimostrazione. E' identica a quella del Teorema 8.1.4. □

Quando i punteggi non sono interi il precedente teorema può essere ancora impiegato determinando una trasformata biunivoca che discretizzi i punteggi originali. Dal momento che la distribuzione di una statistica lineare dei ranghi per i parametri di scala sotto l'ipotesi di base $H_0 : \eta = 1, F \in \mathcal{C}$, si può ottenere mediante il Teorema 9.1.4, allora si può determinare le appropriate regioni critiche del test. Se l'alternativa è bilaterale, ovvero $H_1 : \eta \neq 1, F \in \mathcal{C}$, allora il primo campione tende ad assumere i ranghi più bassi o elevati e quindi si respinge l'ipotesi di base per determinazioni sia troppo elevate che troppo piccole di T. Fissato quindi un livello di significatività α, la regione critica è data dall'insieme

$$\mathcal{T}_1 = \{t : t \leq t_{n_1,n_2,\alpha/2}, t \geq t_{n_1,n_2,1-\alpha/2}\} ,$$

dove $t_{n_1,n_2,\alpha}$ rappresenta il quantile di ordine α della distribuzione di T per numerosità campionarie pari a n_1 e n_2. Se l'alternativa è direzionale del tipo $H_1 : \eta > 1, F \in \mathcal{C}$, il primo campione tende ad assumere i ranghi più bassi se il test è basato sui punteggi tipo 2 e quindi si respinge l'ipotesi di base per determinazioni troppo basse di T. Analogamente, il primo campione tende ad assumere i ranghi più elevati se il test è basato sui punteggi tipo 1 e quindi si respinge l'ipotesi di base per determinazioni troppo elevate di T. Fissato quindi un livello di significatività α, per i punteggi tipo 2 si ha la regione critica

$$\mathcal{T}_1 = \{t : t \leq t_{n_1,n_2,\alpha}\} ,$$

mentre per i punteggi tipo 1 si ha la regione critica

$$\mathcal{T}_1 = \{t : t \geq t_{n_1,n_2,1-\alpha}\} .$$

Al contrario, se l'alternativa è direzionale del tipo $H_1 : \eta < 1, F \in \mathcal{C}$, allora il primo campione tende ad assumere i ranghi più elevati se il test è basato sui punteggi tipo 2 e quindi si respinge l'ipotesi di base per determinazioni troppo elevate di T. Analogamente, il primo campione tende ad assumere i ranghi più bassi

se il test è basato sui punteggi tipo 1 e quindi si respinge l'ipotesi di base per determinazioni troppo basse di T. Fissato quindi un livello di significatività α, per i punteggi tipo 2 si ha la regione critica

$$\mathcal{T}_1 = \{t : t \geq t_{n_1,n_2,1-\alpha}\} ,$$

mentre per i punteggi tipo 1 si ha la regione critica

$$\mathcal{T}_1 = \{t : t \leq t_{n_1,n_2,\alpha}\} .$$

Infine, il test basato sulla T per i precedenti sistemi di ipotesi è corretto al livello di significatività α. Infatti, dal momento che si può dimostrare che $P_T(\eta, F) = \Pr_{\eta,F}(T \in \mathcal{T}_1)$ è una funzione monotona crescente per $\eta > 1$ e monotona decrescente per $\eta < 1$ per ogni $F \in \mathcal{C}$, allora risulta $P_T(\eta, F) > \alpha$, ovvero il test è corretto.

Risulta interessante determinare la scelta dei punteggi che fornisce il test localmente più potente per verificare il sistema di ipotesi $H_0 : \eta = 1, F = F_0$, contro $H_1 : \eta > 1, F = F_0$. In questo caso, si ha la seguente definizione di test localmente più potente.

Definizione 9.1.5. Se il campione casuale misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ ha funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$, dove $n = n_1 + n_2$, si consideri il sistema di ipotesi $H_0 : \eta = 1, F = F_0$, contro $H_1 : \eta > 1, F = F_0$, dove $F_0 \in \mathcal{C}$. Il test basato sulla statistica lineare dei ranghi per i parametri di scala T_* è detto localmente più potente se esiste un $\epsilon > 0$ tale che per ogni livello di significatività naturale si ha

$$P_{T_*}(\eta) \geq P_T(\eta) , 1 < \eta < 1 + \epsilon ,$$

per ogni statistica lineare dei ranghi per i parametri di scala T. △

Il prossimo teorema fornisce la scelta ottima dei punteggi per la costruzione del test localmente più potente. Al solito, l'utilità di questo teorema consiste solamente nell'evidenziare la struttura ottima dei punteggi al variare della funzione di ripartizione, dal momento che questa non è mai nota in pratica.

Teorema 9.1.6. *Si consideri il campione casuale misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ con funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$, e sia f la funzione di densità corrispondente a F. Si assuma inoltre che f' esista, sia assolutamente continua e che*

$$\int_{\mathbb{R}} |f'(x)| \, dx < \infty .$$

Il test localmente più potente per verificare il sistema di ipotesi $H_0 : \eta = 1, F = F_0$, contro $H_1 : \eta > 1, F = F_0$ è basato sulla statistica lineare dei ranghi per i parametri di scala

$$T_* = \sum_{i=1}^{n_1} a_*(R_i) ,$$

dove

$$a_*(i) = E(- V_{(i)} \frac{f'(V_{(i)})}{f(V_{(i)})}) , i = 1, \ldots, n ,$$

e $(V_{(1)}, \ldots, V_{(n)})$ è la statistica ordinata relativa a $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$.

Dimostrazione. Si veda Hettmansperger e McKean (1998). □

Analogamente a quanto visto per la scelta ottima dei punteggi nel caso di una statistica lineare dei ranghi con segno, anche in questo caso si può dimostrare che la scelta ottima dei punteggi non dipende dal parametro di posizione e di scala della distribuzione.

• **Esempio 9.1.4.** Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$, dove F è la funzione di ripartizione di una variabile casuale con distribuzione Normale $N(0, 1)$ e f rappresenta la relativa funzione di densità. Le condizioni del Teorema 9.1.6 sono

soddisfatte. Inoltre, dall'Esempio 5.2.1 si ha $-f'(x)/f(x) = x$, per cui la scelta ottimale dei punteggi è data da

$$a^*(i) = \mathrm{E}(V_{(i)}^2)\,,\, i = 1, \ldots, n\,,$$

dove $(V_{(1)}, \ldots, V_{(n)})$ è la statistica ordinata relativa ad un campione casuale proveniente dalla distribuzione Normale $N(0,1)$. ◁

9.2. La distribuzione per grandi campioni delle statistiche lineari dei ranghi per i parametri di scala. In questa sezione vengono considerate le proprietà per grandi campioni delle statistiche lineari dei ranghi per i parametri di scala.

Teorema 9.2.1. *Se $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ è un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{V}_{1,F}$, allora per una statistica lineare dei ranghi per i parametri di scala*

$$T = T_n = \sum_{i=1}^{n_1} a_n(R_i)\,,$$

i cui punteggi $a_n(i)$, per $i = 1, \ldots, n$, soddisfano le condizioni del Teorema 7.2.2, risulta

$$\frac{T_n - \mathrm{E}(T_n)}{\sqrt{\mathrm{Var}(T_n)}} \xrightarrow{d} N(0,1)\,,$$

dove $\mathrm{E}(T_n)$ e $\mathrm{Var}(T_n)$ sono definite nel Teorema 9.1.2.

Dimostrazione. Segue dall'Esempio 7.2.1 e dal Teorema 7.2.2. □

Fissato un livello di significatività α, per $n \to \infty$ la regione critica per verificare $H_0 : \eta = 1, F \in \mathcal{C}$, contro l'alternativa $H_1 : \eta \neq 1, F \in \mathcal{C}$, può essere approssimata dall'insieme

$$\{t : t \leq \mathrm{E}(T_n) + z_{\alpha/2}\sqrt{\mathrm{Var}(T_n)}, t \geq \mathrm{E}(T_n) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(T_n)}\}\,.$$

Analogamente, per $n \to \infty$ la regione critica per verificare l'alternativa $H_1 : \eta > 1, F \in \mathcal{C}$, per i punteggi tipo 2 può essere approssimata dall'insieme

$$\{t : t \leq \mathrm{E}(T_n) + z_{\alpha}\sqrt{\mathrm{Var}(T_n)}\}\,,$$

mentre per i punteggi tipo 1 può essere approssimata dall'insieme

$$\{t : t \geq \mathrm{E}(T_n) + z_{1-\alpha}\sqrt{\mathrm{Var}(T_n)}\}\,.$$

Al contrario, la regione critica per verificare l'alternativa $H_1 : \eta < 1, F \in \mathcal{C}$, per i punteggi tipo 2 può essere approssimata dall'insieme

$$\{t : t \geq \mathrm{E}(T_n) + z_{1-\alpha}\sqrt{\mathrm{Var}(T_n)}\}\,,$$

mentre per i punteggi tipo 1 può essere approssimata dall'insieme

$$\{t : t \leq \mathrm{E}(T_n) + z_{\alpha}\sqrt{\mathrm{Var}(T_n)}\}\,.$$

Mediante il Teorema 3.1.8 si può dimostrare inoltre che la successione di test basata su $\{T_n\}$ è coerente. Per quanto riguarda l'efficacia delle statistiche lineari dei ranghi per i parametri di scala si ha il seguente teorema.

Teorema 9.2.2. *Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$, e si consideri il sistema di ipotesi $H_0 : \eta = 1, F \in \mathcal{C}$, contro $H_1 : \eta = \eta_i, F \in \mathcal{C}$, dove $(\eta_i)_{i\geq 1}$ è una successione di alternative tali che $\eta_i = 1 + c/\sqrt{n_i}$ con c costante. Data una statistica lineare dei ranghi per i parametri di scala*

$$T = T_n = \sum_{i=1}^{n_1} a_n(R_i) \,,$$

i cui punteggi $a_n(i)$ *per* $i = 1, \ldots, n$*, soddisfano le condizioni del Teorema 9.2.1, allora l'efficacia del test basato su* T_n *risulta*

$$\textit{eff}_T = \sqrt{\nu(1-\nu)}\, \frac{\int_0^1 \phi(u)\phi_f(u)\, du}{(\int_0^1 (\phi(u) - \overline{\phi})^2 du)^{1/2}} \,,$$

dove $\nu = \lim_n n_1/n$ *con* $0 < \nu < 1$*, e*

$$\phi_f(u) = -1 - F^{-1}(u)\, \frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \,.$$

Dimostrazione. Vedi Háyek e Šidák (1967). □

• **Esempio 9.2.1.** Si consideri la statistica $M = M_n$ del test di Mood introdotta nell'Esempio 9.1.1. Analogamente all'Esempio 7.2.2, al fine di determinare la distribuzione asintotica di M_n, è conveniente considerare la statistica $T_n = b_n M_n$ con $b_n = (n+1)^2$. La funzione punteggio relativa alla statistica T_n, data da $\phi(u) = (u - 1/2)^2 \mathbf{1}_{[0,1]}(u)$, è tale che

$$\int_0^1 (\phi(u) - \overline{\phi})^2\, du = \int_0^1 ((u-1/2)^2 - 1/12)^2\, du = \frac{1}{180} < \infty$$

e dunque le condizioni del Teorema 7.2.2 sono soddisfatte. Inoltre, risulta

$$\int_0^1 \phi(u)\phi_f(u) du = -\frac{1}{12} - \int_0^1 F^{-1}(u)\, \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\, (u-1/2)^2\, du \,,$$

da cui, mediante la trasformazione di variabile $x = F^{-1}(u)$ con $u = F(x)$, e integrando successivamente per parti, si ha

$$\int_0^1 \phi(u)\phi_f(u) du = -\frac{1}{12} - \int_{-\infty}^{\infty} x\, \frac{f'(x)}{f(x)}\, (F(x) - 1/2)^2 f(x)\, dx = 2\int_{-\infty}^{\infty} x(F(x) - 1/2) f(x)^2\, dx \,.$$

Quindi l'efficacia del test di Mood risulta

$$\text{eff}_M = \sqrt{720\nu(1-\nu)} \int_{-\infty}^{\infty} x(F(x) - 1/2) f(x)^2\, dx \,. \qquad \triangleleft$$

• **Esempio 9.2.2.** Si consideri la statistica $A = A_n$ del test di Ansari-Bradley dell'Esempio 9.1.1. Analogamente all'Esempio 7.2.2 al fine di determinare la distribuzione asintotica di A_n è conveniente considerare la statistica $T_n = b_n A_n$ con $b_n = (n+1)$. La funzione punteggio relativa alla statistica T_n, data da $\phi(u) = |u - 1/2| \mathbf{1}_{[0,1]}(u)$, è tale che

$$\int_0^1 (\phi(u) - \overline{\phi})^2\, du = \int_0^1 (|u-1/2| - 1/4)^2\, du = \frac{1}{48} < \infty$$

e dunque le condizioni del Teorema 7.2.2 sono soddisfatte. Inoltre, risulta

$$\int_0^1 \phi(u)\phi_f(u) du = -\frac{1}{4} - \int_0^1 F^{-1}(u)\, \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\, |u-1/2|\, du \,,$$

da cui, mediante la trasformazione di variabile $x = F^{-1}(u)$ con $u = F(x)$, se $F(x_{0.5}) = 1/2$, e integrando successivamente per parti, si ha

$$\int_0^1 \phi(u)\phi_f(u)du = -\frac{1}{4} - \int_{-\infty}^{x_{0.5}} x(1/2 - F(x))f'(x)\,dx - \int_{x_{0.5}}^{\infty} x(F(x) - 1/2)f'(x)\,dx$$
$$= \int_{x_{0.5}}^{\infty} xf(x)^2\,dx - \int_{-\infty}^{x_{0.5}} xf(x)^2\,dx\,.$$

Quindi l'efficacia del test di Ansari-Bradley risulta

$$\text{eff}_A = \sqrt{48\nu(1-\nu)}\,(\int_{x_{0.5}}^{\infty} xf(x)^2\,dx - \int_{-\infty}^{x_{0.5}} xf(x)^2\,dx)\,. \qquad \triangleleft$$

9.3. Il test di Mood. Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$. Il test di Mood è basato sulla statistica

$$M = \sum_{i=1}^{n_1} (R_i - \frac{1}{2}(n+1))^2\,,$$

descritta nell'Esempio 9.1.1. Mediante questo test si può verificare l'ipotesi di base $H_0 : \eta = 1, F \in \mathcal{C}$. La statistica M è una statistica lineare dei ranghi per i parametri di scala con i punteggi di tipo 2 scelti come $a(i) = (i - (n+1)/2)^2$ per $i = 1, \ldots, n$. Al fine di calcolare la funzione di probabilità $p_{n_1,n_2}(m)$ di M, è conveniente considerare i punteggi $a'(i) = 4a(i)$ al posto dei punteggi originali. In questa maniera la statistica M' basata sui nuovi punteggi prende valori solo sugli interi ed è equivalente alla statistica originale essendo $M' = 4M$. Dunque, dal Teorema 9.1.4, la funzione generatrice di probabilità $L_{M'}(v)$ di M' è data da $\binom{n}{n_1}^{-1}$ volte il coefficiente di ordine n_1 del polinomio in u

$$\prod_{i=1}^{n} (1 + uv^{(2i-(n+1))^2})\,.$$

• **Esempio 9.3.1.** Per $n_1 = 2$ e $n_2 = 2$ si ha

$$\prod_{i=1}^{4} (1 + uv^{(2i-5)^2}) = 1 + (2v + 2v^9)u + (v^2 + 4v^{10} + v^{18})u^2 + (2v^{11} + 2v^{19})u^3 + v^{20}u^4\,,$$

da cui

$$L_{M'}(v) = \frac{1}{6}(v^2 + 4v^{10} + v^{18})\,.$$

Poichè le probabilità $p_{2,2}(m')$ corrispondono ai coefficienti del polinomio $L_{M'}(v)$, si ottiene la Tavola 9.3.1.

Tavola 9.3.1. Funzione di probabilità di M' per $n_1 = 2$ e $n_2 = 2$.

m'	2	10	18
$p_{2,2}(m')$	1/6	4/6	1/6

Inoltre, tenendo presente la relazione fra M e M', dalla Tavola 9.3.1 si ottiene anche la Tavola 9.3.2.

Tavola 9.3.2. Funzione di probabilità di M per $n_1 = 2$ e $n_2 = 2$.

m	1/2	5/2	9/2
$p_{2,2}(m)$	1/6	4/6	1/6

$\triangleleft$

Se l'ipotesi di base è vera, dall'Esempio 9.1.2 si ha

$$\text{E}(M) = \frac{n_1(n^2-1)}{12}\,,\ \text{Var}(M) = \frac{n_1 n_2(n+1)(n^2-4)}{180}\,.$$

Inoltre, dall'Esempio 9.1.2, risulta che la distribuzione di M è simmetrica solo se $n_1 = n_2$. Infine, dal Corollario 2.4.6 si ha che M è "distribution-free" su $\mathcal{V}_{1,F} = \mathcal{C}_F$ e di conseguenza anche il test di Mood è "distribution-free". Dal momento che la distribuzione della statistica M sotto ipotesi di base è specificata, le appropriate regioni critiche del test per alternative bilaterali o direzionali si possono facilmente ottenere tenendo presente la discussione fatta nella §9.1.

• **Esempio 9.3.2.** Nella Tavola 9.3.3 sono riportati i tempi in anni dall'elezione alla morte per gli ultimi 8 presidenti degli Stati Uniti e degli ultimi 8 papi (dati aggiornati al 1991).

Tavola 9.3.3. Tempi di sopravvivenza dall'elezione (in anni).

	presidente	x_i	r_i	$4a(r_i)$	papa	y_i	r_i	$4a(r_i)$
1	Harding	2	2	169	Leone XIII	25	14	121
2	Coolidge	10	7	9	Pio X	11	8	1
3	Hoover	36	16	225	Benedetto XV	8	5	49
4	Roosvelt	12	9	1	Pio XI	17	12	49
5	Truman	28	15	169	Pio XII	19	13	81
6	Kennedy	3	3	121	Giovanni XXIII	5	4	81
7	Eisenhower	16	11	25	Paolo VI	15	10	9
8	Johnson	9	6	25	Giovanni Paolo	0	1	225

Fonte: Lunn e McNeil (1991)

Si vuole verificare se esiste una differente dispersione fra i due gruppi, ovvero si vuole verificare il sistema di ipotesi $H_0 : \eta = 1, F \in \mathcal{C}$, contro $H_1 : \eta \neq 1, F \in \mathcal{C}$. Dal momento che per questi dati si verifica che $m = 186$, per $n_1 = 8$ e $n_2 = 8$ si ha $\Pr(M \geq 186) = 0.3525$ e quindi la significatività osservata risulta $\alpha_{oss} = 2 \times 0.3525 = 0.7050$. Questo è una significatività osservata piuttosto elevata che porta ad accettare l'ipotesi di base. ◁

Per quanto riguarda la distribuzione per grandi campioni della statistica $M = M_n$ sotto ipotesi di base, dal Teorema 9.2.1 si ha

$$\frac{M_n - n_1(n^2-1)/12}{\sqrt{n_1 n_2 (n+1)(n^2-4)/180}} \xrightarrow{d} N(0,1) \, .$$

La convergenza della statistica di Mood è abbastanza rapida anche per campioni moderati, posto che entrambe le numerosità siano abbastanza elevate, ovvero $n_1 \geq 15$ e $n_2 \geq 15$. Le approssimazioni per grandi campioni delle regioni critiche del test per alternative bilaterali o direzionali si possono facilmente ottenere tenendo presente la discussione fatta nella §9.2.

9.4. Il test di Ansari-Bradley. Sia $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ un campione casuale misto con funzione di ripartizione congiunta $F_n \in \mathcal{V}_{\eta,F}$. Il test di Ansari-Bradley è basato sulla statistica

$$A = \sum_{i=1}^{n_1} \left(\frac{n+1}{2} - \left|R_i - \frac{n+1}{2}\right|\right)$$

descritta nell'Esempio 9.1.1. Mediante questo test si può verificare l'ipotesi di base $H_0 : \eta = 1, F \in \mathcal{C}$. La statistica A è una statistica lineare dei ranghi per i parametri di scala con i punteggi di tipo 1 scelti come $a(i) = (n+1)/2 - |i - (n+1)/2|$ per $i = 1, \ldots, n$. Al fine di calcolare la funzione di probabilità $p_{n_1,n_2}(a)$ di A, è conveniente considerare i punteggi $a'(i) = 2a(i)$ al posto dei punteggi originali. In questa maniera la statistica A' basata sui nuovi punteggi prende valori solo sugli interi ed è equivalente alla statistica originale essendo $A' = 2A$. Dunque, dal Teorema 9.1.4, la funzione generatrice di probabilità $L_{A'}(v)$ di A' è data da $\binom{n}{n_1}^{-1}$ volte il coefficiente di ordine n_1 del polinomio in u

$$\prod_{i=1}^{n} \left(1 + uv^{(n+1)-|2i-(n+1)|}\right) \, .$$

• **Esempio 9.4.1.** Per $n_1 = 2$ e $n_2 = 2$ si ha

$$\prod_{i=1}^{4}(1 + uv^{5-|2i-5|}) = 1 + (2v^2 + 2v^4)u + (v^4 + 4v^6 + v^8)u^2 + (2v^8 + 2v^{10})u^3 + v^{12}u^4 ,$$

da cui

$$L_{A'}(v) = \frac{1}{6}\,(v^4 + 4v^6 + v^8) .$$

Poichè le probabilità $p_{2,2}(a')$ corrispondono ai coefficienti del polinomio $L_{A'}(v)$, si ottiene la Tavola 9.4.1.

Tavola 9.4.1. Funzione di probabilità di A' per $n_1 = 2$ e $n_2 = 2$.

a'	4	6	8
$p_{2,2}(a')$	1/6	4/6	1/6

Inoltre, tenendo presente la relazione fra A e A' dalla Tavola 9.4.1 si ottiene anche la Tavola 9.4.2.

Tavola 9.4.2. Funzione di probabilità di A per $n_1 = 2$ e $n_2 = 2$.

a	2	3	4
$p_{2,2}(a)$	1/6	4/6	1/6

◁

Se l'ipotesi di base è vera, dall'Esempio 9.1.3 per n pari si ha

$$\mathrm{E}(A) = \frac{n_1(n+2)}{4} , \mathrm{Var}(A) = \frac{n_1 n_2(n^2-4)}{48(n-1)} ,$$

mentre per n dispari si ha

$$\mathrm{E}(A) = \frac{n_1(n+1)^2}{4n} , \mathrm{Var}(A) = \frac{n_1 n_2(n+1)(n^2+3)}{48n^2} .$$

Inoltre, dall'Esempio 9.1.3, risulta che la distribuzione di A è simmetrica solo se $n_1 = n_2$. Infine, dal Corollario 2.4.6 la statistica A è "distribution-free" su $\mathcal{V}_{1,F} = \mathcal{C}_F$ e di conseguenza anche il test di Ansari-Bradley è "distribution-free". Dal momento che la distribuzione della statistica A sotto ipotesi di base è specificata, le appropriate regioni critiche del test per alternative bilaterali o direzionali si possono facilmente ottenere tenendo presente la discussione fatta nella §9.1.

• **Esempio 9.4.2.** Sebbene i dati dell'Esempio 6.1.1 siano relativi ad un solo campione casuale di sfere di acciaio prodotte da un macchinario, sono tuttavia disponibili anche i dati relativi ad un secondo campione casuale proveniente da un altro macchianario che produce ancora sfere di acciaio del diametro di 1 micron. Si dispone quindi di due campioni casuali indipendenti e si hanno quindi i dati della Tavola 9.4.3.

Tavola 9.4.3. Diametro delle sfere (in micron).

	macchianario 1			macchinario 2		
sfera	x_i	r_i	$2a(r_i)$	y_i	r_i	$2a(r_i)$
1	1.18	8	16	1.72	18	6
2	1.42	12	18	1.63	16	10
3	0.69	1	2	1.69	17	8
4	0.88	4	8	0.79	3	6
5	1.62	15	12	1.79	19	4
6	1.09	7	14	0.77	2	4
7	1.53	14	14	1.44	13	16
8	1.02	6	12	1.29	10	20
9	1.19	9	18	1.96	20	2
10	1.32	11	20	0.99	5	10

Fonte: Romano (1977)

Si sospetta che il secondo macchinario produca sfere con minore precisione del primo, ovvero si vuole verificare il sistema di ipotesi $H_0 : \eta = 1, F \in \mathcal{C}$, contro $H_1 : \eta > 1, F \in \mathcal{C}$. Dal momento che per questi dati è immediato verificare che $a = 67$, per $n_1 = 10$ e $n_2 = 10$ si ha $\Pr(A \geq 67) = 0.0403$ e quindi la significatività osservata risulta $\alpha_{oss} = 0.0403$. Questa è una significatività osservata bassa che porta a respingere l'ipotesi di base, ovvero sulla base dell'evidenza empirica si deve ritenere che il secondo macchinario sia meno preciso del primo. ◁

Per quanto riguarda la distribuzione per grandi campioni della statistica $A = A_n$ sotto ipotesi di base, dal Teorema 9.2.1, si ha

$$\frac{A_n - \mathrm{E}(A_n)}{\sqrt{\mathrm{Var}(A_n)}} \xrightarrow{d} N(0,1) \, .$$

La convergenza della statistica di Ansari-Bradley è abbastanza rapida anche per campioni moderati, posto che entrambe le numerosità siano abbastanza elevate, ovvero $n_1 \geq 15$ e $n_2 \geq 15$. Le approssimazioni per grandi campioni delle regioni critiche del test per alternative bilaterali o direzionali si possono facilmente ottenere tenendo presente la discussione fatta nella §9.2.

9.5. Le prestazioni del test di Mood e del test di Ansari-Bradley. Per quanto riguarda le prestazioni asintotiche del test di Mood, dall'Esempio 9.2.1 si ha

$$\mathrm{eff}_M = \sqrt{720\nu(1-\nu)} \int_{-\infty}^{\infty} x(F(x) - 1/2) f(x)^2 \, dx \, ,$$

per cui, utilizzando i risultati dell'Esempio 3.2.4, l'efficienza asintotica relativa del test di Mood rispetto al test di Snedecor risulta

$$\mathrm{EAR}_{M,F} = \frac{180\gamma^2 (\int_{-\infty}^{\infty} x(F(x) - 1/2) f(x)^2 \, dx)^2}{\sigma^4} \, .$$

La Tavola 9.5.2 fornisce i valori dell'efficienza asintotica relativa $\mathrm{EAR}_{M,F}$ per alcune distribuzioni. Risulta evidente che, anche dal punto di vista asintotico, il test di Mood dimostra buone prestazioni.

Tavola 9.5.2. EAR del test di Mood rispetto al test di Snedecor.

distribuzione	$\mathrm{EAR}_{M,F}$
$U(\lambda, \delta)$	1
$N(\mu, \sigma^2)$	$15/(2\pi^2) = 0.7599$
$Lo(\lambda, \delta)$	1
$L(\mu, \delta)$	$625/576 \simeq 1.0851$
$C(\lambda, \delta)$	∞
$E(\lambda, \sigma)$	$115/144 \simeq 0.7986$

Per quanto riguarda le prestazioni asintotiche del test di Ansari-Bradley, dall'Esempio 9.2.2 si ha

$$\mathrm{eff}_A = \sqrt{48\nu(1-\nu)} \, (\int_{x_{0.5}}^{\infty} x f(x)^2 \, dx - \int_{-\infty}^{x_{0.5}} x f(x)^2 \, dx) \, ,$$

per cui utilizzando i risultati dell'Esempio 3.2.4, l'efficienza asintotica relativa del test di Ansari-Bradley rispetto al test di Snedecor risulta

$$\mathrm{EAR}_{A,F} = \frac{12\gamma^2 (\int_{x_{0.5}}^{\infty} x f(x)^2 \, dx - \int_{-\infty}^{x_{0.5}} x f(x)^2 \, dx)^2}{\sigma^4} \, .$$

La Tavola 9.5.3 fornisce i valori dell'efficienza asintotica relativa $\mathrm{EAR}_{A,F}$ per alcune distribuzioni. Le prestazioni del test di Ansari-Bradley sono relativamente scarse anche in questo caso, eccetto che per una distribuzione a code pesanti quale la Cauchy.

Tavola 9.5.3. EAR del test di Ansari-Bradley rispetto al test di Snedecor.

distribuzione	$\mathrm{EAR}_{A,F}$
$U(\lambda,\delta)$	$3/5 = 0.6$
$N(\mu,\sigma^2)$	$6/\pi^2 \simeq 0.6079$
$Lo(\lambda,\delta)$	$4(4\log 2-1)^2/15 \simeq 0.8379$
$L(\mu,\delta)$	$15/16 \simeq 0.9375$
$C(\lambda,\delta)$	∞
$E(\lambda,\sigma)$	$69(2\log 2-1)^2/16 \simeq 0.6435$

Infine, l'efficienza asintotica relativa del test di Mood rispetto al test di Ansari-Bradley risulta

$$\mathrm{EAR}_{M,A} = \frac{15\,(\int_{-\infty}^{\infty} x(F(x)-1/2)f(x)^2\,dx)^2}{(\int_{x_{0.5}}^{\infty} xf(x)^2\,dx - \int_{-\infty}^{x_{0.5}} xf(x)^2\,dx)^2}\ .$$

La Tavola 9.5.4 fornisce i valori dell'efficienza asintotica relativa $\mathrm{EAR}_{M,A}$ per alcune distribuzioni. Dunque, dal punto di vista asintotico, il test di Mood risulta superiore al test di Ansari-Bradley.

Tavola 9.5.4. EAR del test di Mood rispetto al test di Ansari-Bradley.

distribuzione	$\mathrm{EAR}_{M,A}$
$U(\lambda,\delta)$	$5/3 = 1.6667$
$N(\mu,\sigma^2)$	$5/4 = 1.25$
$Lo(\lambda,\delta)$	$15/(4(4\log 2-1)^2) \simeq 1.1935$
$L(\mu,\delta)$	$125/108 \simeq 1.1574$
$C(\lambda,\delta)$	$15/16 = 0.9375$
$E(\lambda,\sigma)$	$5/(27(2\log 2-1)^2) \simeq 1.2410$

Capitolo 10

I test per l'associazione

10.1. Verifica di ipotesi sull'associazione. Si consideri il seguente modello statistico "distribution-free"

$$\mathcal{C}_F^2 = \{F_n : F_n(x_1, \dots, x_n, y_1, \dots, y_n) = \prod_{i=1}^{n} F(x_i, y_i), F \in \mathcal{C}^2\} ,$$

dove $\mathcal{C}^2$ rappresenta la classe delle funzioni di ripartizione di un vettore bivariato di variabili casuali assolutamente continue. Dunque, $\mathcal{C}_F^2$ rappresenta il modello statistico relativo ad un campione casuale proveniente da un vettore bivariato di variabili casuali assolutamente continue. Una sottoclasse di $\mathcal{C}_F^2$ è data dal modello statistico "distribution-free"

$$\mathcal{I}_{F_1,F_2}^2 = \{F_n : F_n(x_1, \dots, x_n, y_1, \dots, y_n) = \prod_{i=1}^{n} F_1(x_i)F_2(y_i), F_1, F_2 \in \mathcal{C}\} .$$

Quindi, $\mathcal{I}_{F_1,F_2}^2$ rappresenta il modello statistico relativo ad un campione casuale proveniente da un vettore bivariato di variabili casuali assolutamente continue a componenti indipendenti. Il sistema di ipotesi da verificare risulta $H_0 : F_n \in \mathcal{I}_{F_1,F_2}^2$ contro l'ipotesi alternativa $H_1 : F_n \in \mathcal{C}_F^2 - \mathcal{I}_{F_1,F_2}^2$, ovvero si vuole verificare l'indipendenza delle componenti del vettore bivariato di variabili casuali da cui proviene il campione casuale.

10.2. Il test di correlazione di Spearman. Se $(X_1, Y_1), \dots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F^2$, si vuole verificare il sistema di ipotesi considerato nella §10.1. Una statistica test conveniente in questo caso è il coefficiente di correlazione campionario di Spearman, che non è altro che l'ordinario coefficiente di correlazione calcolato sul vettore dei ranghi relativo a $(X_1, \dots, X_n)$ e sul vettore dei ranghi relativo a $(Y_1, \dots, Y_n)$. Si indichi dunque con $(R_1, \dots, R_n)$ il vettore dei ranghi relativo a $(X_1, \dots, X_n)$ e con $(U_1, \dots, U_n)$ il vettore dei ranghi relativo a $(Y_1, \dots, Y_n)$. Tenendo presente che

$$\frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} U_i = \frac{n+1}{2}$$

e che

$$\frac{1}{n}\sum_{i=1}^{n} (R_i - \frac{n+1}{2})^2 = \frac{1}{n}\sum_{i=1}^{n} (U_i - \frac{n+1}{2})^2 = \frac{n^2-1}{12} ,$$

il coefficiente di correlazione campionario di Spearman è dunque dato da

$$R_S = \frac{\frac{1}{n}\sum_{i=1}^{n} U_i R_i - (n+1)^2/4}{(n^2-1)/12} = \frac{12}{n(n^2-1)}\sum_{i=1}^{n} U_i R_i - \frac{3(n+1)}{n-1} .$$

Il seguente teorema permette di ottenenre un'espressione più semplice per il coefficiente di correlazione di Spearman.

Teorema 10.2.1. *Se $(X_1, Y_1), \dots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}_{F_1,F_2}^2$, per il coefficiente di correlazione di Spearman si ha*

$$R_S = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} U_i R_i - \frac{3(n+1)}{n-1} \stackrel{d}{=} \frac{12}{n(n^2-1)} \sum_{i=1}^{n} i R_i - \frac{3(n+1)}{n-1} .$$

Dimostrazione. Dal momento che i vettori dei ranghi $(R_1, \ldots, R_n)$ e $(U_1, \ldots, U_n)$ sono indipendenti in quanto trasformate di variabili casuali indipendenti, allora per ogni $(u_1, \ldots, u_n) \in \mathcal{R}_n$ si ha

$$(R_S \mid U_1 = u_1, \ldots, U_n = u_n) \stackrel{d}{=} \frac{12}{n(n^2-1)} \sum_{i=1}^{n} u_i R_i - \frac{3(n+1)}{n-1} .$$

Se d_i è la posizione del numero i nel vettore $(u_1, \ldots, u_n)$, allora si ha

$$\frac{12}{n(n^2-1)} \sum_{i=1}^{n} u_i R_i - \frac{3(n+1)}{n-1} = \frac{12}{n(n^2-1)} \sum_{j=1}^{n} j R_{d_j} - \frac{3(n+1)}{n-1} .$$

Dal momento che per il Teorema 2.4.3 $(R_1, \ldots, R_n)$ è distribuito uniformente su $\mathcal{R}_n$, essendo il vettore dei ranghi relativo ad un campione casuale, allora si ha $(R_1, \ldots, R_n) \stackrel{d}{=} (R_{d_1}, \ldots, R_{d_n})$ per qualsiasi permutazione $(d_1, \ldots, d_n)$ di $(1, \ldots, n)$. Dunque, risulta

$$(R_S \mid U_1 = u_1, \ldots, U_n = u_n) \stackrel{d}{=} \frac{12}{n(n^2-1)} \sum_{j=1}^{n} j R_{d_j} - \frac{3(n+1)}{n-1} \stackrel{d}{=} \frac{12}{n(n^2-1)} \sum_{i=1}^{n} i R_i - \frac{3(n+1)}{n-1} .$$

Questa equivalenza in distribuzione vale per ogni $(u_1, \ldots, u_n) \in \mathcal{R}_n$ e il teorema è dimostrato. □

In base al precedente teorema, il coefficiente di correlazione di Spearman può essere calcolato semplicemente ordinando rispetto alle realizzazioni di $(Y_1, \ldots, Y_n)$ e successivamente assegnando il vettore dei ranghi alle realizzazioni di $(X_1, \ldots, X_n)$. Se risulta $R_i = i$ per $i = 1, \ldots, n$, allora esiste associazione positiva perfetta fra i ranghi ed infatti si ha

$$R_S = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} i^2 - \frac{3(n+1)}{n-1} = \frac{12}{n(n^2-1)} \frac{n(n+1)(2n+1)}{6} - \frac{3(n+1)}{n-1} = 1 .$$

Al contrario, se $R_i = n + 1 - i$ per $i = 1, \ldots, n$, allora esiste associazione negativa perfetta fra i ranghi ed infatti

$$\begin{aligned} R_S &= \frac{12}{n(n^2-1)} \sum_{i=1}^{n} i(n+1-i) - \frac{3(n+1)}{n-1} = \frac{12}{n(n-1)} \sum_{i=1}^{n} i - \frac{12}{n(n^2-1)} \sum_{i=1}^{n} i^2 - \frac{3(n+1)}{n-1} \\ &= \frac{12}{n(n-1)} \frac{n(n+1)}{2} - \frac{12}{n(n^2-1)} \frac{n(n+1)(2n+1)}{6} - \frac{3(n+1)}{n-1} = -1 . \end{aligned}$$

Posto $S = \sum_{i=1}^{n} i R_i$, R_S è una trasformata lineare di S. Dunque, R_S e S forniscono test equivalenti. Tenendo presente la Definizione 7.1.1, S è dunque una statistica lineare dei ranghi con le costanti di regressione pari a $c(i) = i$ e i punteggi pari a $a(i) = i$. Il supporto di S è dato da $\{n(n+1)(n+2)/6, \ldots, n(n+1)(2n+1)/6\}$. Tenendo presente il Teorema 2.4.3 e denotando la funzione di probabilità di S con $p_n(s) = \Pr(S = s)$, si ha

$$p_n(s) = \frac{1}{n!} c_n(s) \mathbf{1}_{\{n(n+1)(n+2)/6, \ldots, n(n+1)(2n+1)/6\}}(s) ,$$

dove $c_n(s)$ rappresenta il numero di permutazioni di $(1, \ldots, n)$ per cui S assume valore s. La funzione di probabilità di S (e di conseguenza quella di R_S) è solitamente ottenuta mediante enumerazione, dal momento che non esistono metodi che ne facilitino il calcolo. Se l'ipotesi di base è vera, dal momento che in questo caso si ha $\overline{a} = \overline{c} = (n+1)/2$ e $s_a^2 = s_c^2 = (n^2-1)/12$, dal Teorema 7.1.3 risulta

$$\mathrm{E}(S) = \frac{n(n+1)^2}{4} , \mathrm{Var}(S) = \frac{n^2(n-1)(n+1)^2}{144} ,$$

da cui

$$\mathrm{E}(R_S) = \frac{12}{n(n^2-1)}\,\mathrm{E}(S) - \frac{3(n+1)}{n-1} = 0$$

e

$$\mathrm{Var}(R_S) = \frac{144}{n^2(n^2-1)^2}\,\mathrm{Var}(S) = \frac{1}{n-1}\ .$$

Dal momento che $a(i) = i$, in modo analogo all'Esempio 7.1.5, si ottiene che S è simmetrica rispetto a $\mathrm{E}(S) = n(n^2+1)/4$. Di conseguenza, anche R_S è simmetrica rispetto a $\mathrm{E}(R_S) = 0$. Inoltre, dal Corollario 2.4.4, la statistica S è "distribution-free" su $\mathcal{C}_{F_1}$ e dunque anche su $\mathcal{I}^2_{F_1,F_2}$, ovvero anche il test di Spearman è "distribution-free". Se l'ipotesi di base non è vera, S tende ad assumere valori bassi o elevati. Fissato quindi un livello di significatività α, allora si sceglie come regione critica l'insieme

$$\mathcal{T}_1 = \{s : s \le s_{n,\alpha/2}, s \ge s_{n,1-\alpha/2}\}\ ,$$

dove $s_{n,\alpha}$ rappresenta il quantile di ordine α della distribuzione di S per una numerosità campionaria pari a n.

• **Esempio 10.2.1.** I dati della Tavola 10.2.1 si riferiscono alla mortalità per cirrosi e al consumo di maiale nell'anno 1978 nelle 10 provincie del Canada.

Tavola 10.2.1. Mortalità per cirrosi (per 100 000 abitanti) e consumo di maiale (in kg per persona).

provincia	y_i	x_i	i	r_i
Prince Edward Island	6.5	5.8	1	5
Newfoundland	10.2	6.8	2	6
Nova Scotia	10.6	3.6	3	1
Saskatchewan	13.4	4.3	4	2
New Brunswick	14.5	4.4	5	3
Alberta	16.4	5.7	6	4
Manitoba	16.6	6.9	7	7
Ontario	18.2	7.2	8	8
Quebec	19.0	14.9	9	10
British Columbia	27.5	8.4	10	9

Fonte: Nanji e French (1985)

Si vuole verificare se esiste associazione fra la mortalità per cirrosi e il consumo di maiale. Dal momento che per questi dati si ha $s = 360$, allora per $n = 10$ risulta $\Pr(S \ge 360) = 0.0153$ e quindi la significatività osservata è data da $\alpha_{oss} = 2 \times 0.0153 = 0.0306$. Dunque, esiste associazione fra la mortalità per cirrosi e il consumo di maiale, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha > 0.0306$. In effetti, esiste una associazione positiva, dato che risulta $r_S = 23/33 \simeq 0.6969$. ◁

Per quanto riguarda la distribuzione per grandi campioni della statistica $S = S_n$ sotto ipotesi di base, i relativi coefficienti di regressione verificano la condizione di Noether dal momento che

$$\lim_n \frac{\sum_{i=1}^n (c_n(i) - \overline{c}_n)^2}{\max\limits_{1\le i\le n} (c_n(i) - \overline{c}_n)^2} = \lim_n \frac{\sum_{i=1}^n (i - (n+1)/2)^2}{\max\limits_{1\le i\le n} (i - (n+1)/2)^2} = \lim_n \frac{n(n^2-1)/12}{(n-1)^2/4} = \lim_n \frac{n(n+1)}{3(n-1)} = \infty\ .$$

In questo caso, dal momento che alla statistica S è associata la funzione punteggio $\phi(u) = u\mathbf{1}_{[0,1]}$ e tenendo presente l'Esempio 5.3.1, le condizioni del Teorema 7.2.2 sono soddifatte e quindi si ha

$$\frac{S_n - n(n+1)^2/4}{\sqrt{n^2(n-1)(n+1)^2/144}} \xrightarrow{d} N(0,1)\ .$$

Tenendo presente la discussione che segue il Teorema 7.2.2 si ha inoltre

$$\frac{R_{S,n}}{\sqrt{1/(n-1)}} = R_{S,n}\sqrt{n-1} \xrightarrow{d} N(0,1)\ .$$

Fissato un livello di significatività α, per $n \to \infty$ la regione critica può essere dunque approssimata dall'insieme

$$\{s : s \leq \mathrm{E}(S_n) + z_{\alpha/2}\sqrt{\mathrm{Var}(S_n)}, s \geq \mathrm{E}(S_n) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(S_n)}\}\ .$$

• **Esempio 10.2.2.** I dati della Tavola 10.2.2 si riferiscono alle misure (in m) fatte nel lancio del peso e del giavellotto dalle 25 atlete partecipanti alla gara di eptathlon femminile alle Olimpiadi del 1988.

Tavola 10.2.2. Misure nel lancio del peso e del giavellotto (in m).

atleta	y_i	x_i	i	r_i
Joyner-Kersee - USA	15.80	45.66	24	22
John - GDR	16.23	42.56	25	16
Behmer - GDR	14.20	44.54	19	19
Sablovskaite - URS	15.23	42.78	23	17
Choubenkova - URS	14.76	47.64	22	24
Schultz - GDR	13.50	42.82	17	15
Fleming - AUS	12.88	40.28	12	13
Greiner - USA	14.13	38.00	18	5
Lajbnerova - CZE	14.28	42.20	21	14
Bouraga - URS	12.62	39.06	8	7
Wijnsma - HOL	13.01	37.86	15	4
Dimitrova - BUL	12.89	40.27	13	12
Schneider - SWI	11.58	47.50	3	25
Braun - FRG	13.16	44.58	16	20
Ruotsalainen - FIN	12.32	45.44	7	21
Yuping - CHN	14.21	38.60	20	6
Hagger - GB	12.75	35.76	11	2
Brown - USA	12.69	44.34	10	18
Mulliner - GB	12.68	37.76	9	3
Hautenauve - BEL	11.81	35.68	6	1
Kytola - FIN	11.66	39.48	4	10
Geremias - BRA	12.95	39.64	14	11
Hui-Ing - TAI	10.00	39.14	1	8
Jeong-Me - KOR	10.83	39.26	2	9
Launa - PNG	11.78	46.38	5	23

Fonte: Lunn e McNeil (1991)

I dati sono ordinati in base alla posizione finale dell'atleta nella gara. Si è interessati a conoscere se vi è associazione fra il risultato ottenuto dalle atlete nella gara del lancio del peso e quello ottenuto dalle atlete nella gara del lancio del giavellotto. Dal momento che per questi dati si ha $s = 4493$, allora

$$\Pr(S \geq 4493) \simeq 1 - \Phi((4493 - 25 \times 676/4)/\sqrt{625 \times 24 \times 676/144}) = 1 - \Phi(1.0080) = 0.1566\ ,$$

e la significatività osservata risulta $\alpha_{oss} = 2 \times 0.1566 = 0.3132$. Dunque, non esiste associazione fra i risultati ottenuti nel lancio del peso e del giavellotto nella gara di epthatlon, dal momento che si può accettare H_0 ad ogni livello di significatività $\alpha < 0.3132$. Il coefficiente di correlazione di Spearman risulta $r_S = 67/325 \simeq 0.2061$, un valore relativamente basso che conferma la mancanza di associazione. ◁

10.3. Il test di correlazione di Kendall. Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F^2$, si vuole verificare il sistema di ipotesi considerato nella §10.1. Una statistica test opportuna in questo sistema di ipotesi è data dalla statistica di Kendall, data da

$$\tau = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{segn}(X_j - X_i)\mathrm{segn}(Y_j - Y_i)\ ,$$

dove $\text{segn}(x) = 2\mathbf{1}_{(0,\infty)}(x) - 1$. La statistica τ può essere scritta come

$$\tau = \frac{2}{n(n-1)} (C - D) ,$$

dove

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(X_j - X_i)\text{segn}(Y_j - Y_i))$$

rappresenta il numero di coppie (X_i, Y_i) e (X_j, Y_j) concordanti, mentre

$$D = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(- \text{segn}(X_j - X_i)\text{segn}(Y_j - Y_i))$$

rappresenta il numero di coppie (X_i, Y_i) e (X_j, Y_j) discordanti. Dal momento che si ha $C + D = n(n-1)/2$, allora

$$\tau = \frac{4C}{n(n-1)} - 1 .$$

In caso di perfetta associazione positiva tutte le coppie sono concordanti, per cui risulta $C = n(n-1)/2$ e di conseguenza $\tau = 1$. Al contrario, in caso di perfetta associazione negativa tutte le coppie sono discordanti, per cui risulta $C = 0$ e di conseguenza $\tau = -1$. Infine, se il numero delle coppie concordanti è uguale al numero delle coppie discordanti, ovvero se non vi è associazione, allora risulta $C = n(n-1)/4$ per cui $\tau = 0$. Dunque, τ è a tutti gli effetti una valida misura di associazione. La statistica τ è una trasformata lineare della statistica C. Dunque, τ e C forniscono test equivalenti. Il seguente teorema fornisce una utile equivalenza in distribuzione per la statistica C.

Teorema 10.3.1. *Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}^2_{F_1,F_2}$, per la statistica C risulta*

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(X_j - X_i)\text{segn}(Y_j - Y_i)) \stackrel{d}{=} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(R_j - R_i) ,$$

dove $(R_1, \ldots, R_n)$ è il vettore dei ranghi relativo a $(X_1, \ldots, X_n)$.
Dimostrazione. Si ha

$$\text{segn}(X_j - X_i) \stackrel{d}{=} \text{segn}(R_j - R_i) , i = 1, \ldots, n-1 , j = i+1, \ldots, n ,$$

e

$$\text{segn}(Y_j - Y_i) \stackrel{d}{=} \text{segn}(U_j - U_i) , i = 1, \ldots, n-1 , j = i+1, \ldots, n ,$$

dove $(U_1, \ldots, U_n)$ è il vettore dei ranghi relativo a $(Y_1, \ldots, Y_n)$. Dunque, risulta

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(X_j - X_i)\text{segn}(Y_j - Y_i)) \stackrel{d}{=} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_j - R_i)\text{segn}(U_j - U_i)) .$$

Dal momento che $(R_1, \ldots, R_n)$ e $(U_1, \ldots, U_n)$ sono indipendenti in quanto trasformate di variabili casuali indipendenti, allora per ogni $(u_1, \ldots, u_n) \in \mathcal{R}_n$ si ha

$$(C \mid U_1 = u_1, \ldots, U_n = u_n) \stackrel{d}{=} (\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_j - R_i)\text{segn}(U_j - U_i)) \mid U_1 = u_1, \ldots, U_n = u_n)$$
$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_j - R_i)\text{segn}(u_j - u_i)) \, .$$

Se d_i è la posizione del numero i nel vettore $(u_1, \ldots, u_n)$, allora si ha

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_j - R_i)\text{segn}(u_j - u_i)) = \sum_{l=1}^{n-1}\sum_{m=l+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_{d_m} - R_{d_l})\text{segn}(m - l))$$
$$= \sum_{l=1}^{n-1}\sum_{m=l+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_{d_m} - R_{d_l})) \, .$$

Dal momento che per il Teorema 2.4.3 $(R_1, \ldots, R_n)$ è distribuito uniformente su $\mathcal{R}_n$, essendo il vettore dei ranghi relativo ad un campione casuale, si ha $(R_1, \ldots, R_n) \stackrel{d}{=} (R_{d_1}, \ldots, R_{d_n})$ per qualsiasi permutazione $(d_1, \ldots, d_n)$ di $(1, \ldots, n)$. Dunque, risulta

$$(C \mid U_1 = u_1, \ldots, U_n = u_n) \stackrel{d}{=} \sum_{l=1}^{n-1}\sum_{m=l+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_{d_l} - R_{d_m})) \stackrel{d}{=} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(\text{segn}(R_j - R_i)) \, .$$

Dal momento che questa equivalenza in distribuzione vale per ogni $(u_1, \ldots, u_n) \in \mathcal{R}_n$, allora il teorema è dimostrato. □

Nel prossimo teorema si ottiene una importante equivalenza in distribuzione per la statistica C.

Teorema 10.3.2. *Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}^2_{F_1,F_2}$, allora*

$$C \stackrel{d}{=} \sum_{i=2}^{n} V_i \, ,$$

dove V_i, per $i = 2, 3, \ldots, n$, sono $(n-1)$ variabili casuali indipendenti con funzione di probabilità data da

$$p_i(v) = \frac{1}{i} \, \mathbf{1}_{\{0,1,\ldots,i-1\}}(v) \, , \, i = 2, 3, \ldots, n \, .$$

Dimostrazione. Dal Teorema 10.3.1, si ha

$$C \stackrel{d}{=} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(R_j - R_i) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n-1} \mathbf{1}_{[0,\infty)}(R_j - R_i) + \sum_{i=1}^{n-1} \mathbf{1}_{[0,\infty)}(R_n - R_i)$$
$$\stackrel{d}{=} \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} \mathbf{1}_{[0,\infty)}(X_j - X_i) + V_n \, ,$$

dove $V_n = \sum_{j=1}^{n-1} \mathbf{1}_{[0,\infty)}(X_n - X_j)$. Dalla Definizione 2.4.1 si ha dunque che $V_n \stackrel{d}{=} R_n - 1$, ovvero V_n ha funzione di probabilità

$$p_n(v) = \frac{1}{n} \, \mathbf{1}_{\{0,1,\ldots,n-1\}}(v) \, .$$

Inoltre, si noti che $\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} \mathbf{1}_{[0,\infty)}(X_j - X_i)$ è indipendente da V_n in quanto eliminando l'ultima osservazione l'ordinamento rimane comunque invariato. Questa variabile casuale è in effetti la statistica C calcolata sul campione casuale $(X_1, \ldots, X_{n-1})$. Iterando, si ha dunque

$$C \stackrel{d}{=} \sum_{i=2}^{n} V_i \,,$$

dove $V_i = \sum_{j=1}^{i-1} \mathrm{I}_{[0,\infty)}(R_i - R_j)$, per $i = 2, 3, \ldots, n$, sono $(n-1)$ variabili casuali indipendenti con funzione di probabilità data da

$$p_i(v) = \frac{1}{i}\, \mathbf{1}_{\{0,1,\ldots,i-1\}}(v) \,,\, i = 2, 3, \ldots, n \,. \qquad \square$$

Il supporto di C è dunque dato da $\{0, 1, \ldots, n(n-1)/2\}$. Tenendo presente il Teorema 2.4.3 e denotando la funzione di probabilità di C con $p_n(c) = \Pr(C = c)$, se l'ipotesi di base è vera si ha

$$p_n(c) = \frac{1}{n!}\, c_n(c)\, \mathbf{1}_{\{0,1,\ldots,n(n-1)/2\}}(c) \,,$$

dove $c_n(c)$ rappresenta il numero di permutazioni di $(1, \ldots, n)$ per cui C assume valore c. Sebbene esistano delle relazioni ricorrenti per il calcolo diretto della funzione di probabilità di C, la distribuzione della statistica si ottiene più facilmente mediante la funzione generatrice delle probabilità.

Teorema 10.3.3. *Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}^2_{F_1,F_2}$, la funzione generatrice delle probabilità di C risulta*

$$L_C(t) = \prod_{i=2}^{n} \frac{1 - t^i}{i(1-t)} \,,\, |t| < 1 \,.$$

Dimostrazione. Con la notazione del Teorema 10.3.2, la funzione generatrice delle probabilità di V_i risulta

$$L_{V_i}(t) = \frac{1}{i} \sum_{j=0}^{i-1} t^j \,,\, |t| < 1 \,,\, i = 2, 3, \ldots, n \,.$$

Dall'indipendenza delle V_i si ottiene dunque

$$L_C(t) = \prod_{i=2}^{n} \frac{1}{i} \sum_{j=0}^{i-1} t^i = \prod_{i=2}^{n} \frac{1 - t^i}{i(1-t)} \,,\, |t| < 1 \,. \qquad \square$$

• **Esempio 10.3.1.** Per $n = 4$ si ha

$$L_C(t) = \frac{1}{24}\,(1+t)(1+t+t^2)(1+t+t^2+t^4) = \frac{1}{24}\,(1 + 3t + 5t^2 + 6t^3 + 5t^4 + 3t^5 + t^6) \,.$$

Poichè le probabilità $p_4(c)$ corrispondono ai coefficienti del polinomio $L_C(t)$, si otteniene la Tavola 10.3.1.

Tavola 10.3.1. Funzione di probabilità di C per $n = 4$.

c	0	1	2	3	4	5	6
τ	-1	$-2/3$	$-1/3$	0	$1/3$	$2/3$	1
$p_4(c)$	$1/24$	$3/24$	$5/24$	$6/24$	$5/24$	$3/24$	$1/24$

$\triangleleft$

Nei seguenti teoremi si ottiene la media e la varianza di C e la sua simmetria rispetto alla media.

Teorema 10.3.4. *Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}^2_{F_1,F_2}$, allora*

$$\mathrm{E}(C) = \frac{n(n-1)}{4} \, , \, \mathrm{Var}(C) = \frac{n(n-1)(2n+5)}{72} \, .$$

Dimostrazione. Per le variabili casuali V_i, per $i = 2, 3, \ldots, n$, definite nel Teorema 10.3.2, si ha

$$\mathrm{E}(V_i) = \sum_{v=0}^{i-1} v p_i(v) = \frac{1}{i} \sum_{v=0}^{i-1} v = \frac{i-1}{2}$$

e

$$\mathrm{Var}(V_i) = \sum_{v=0}^{i-1} v^2 p_i(v) - \mathrm{E}(V_i)^2 = \frac{1}{i} \sum_{v=0}^{i-1} v^2 - \frac{(i-1)^2}{4} = \frac{i^2-1}{12} \, .$$

Di conseguenza, tenendo presente il Teorema 10.3.2, si ha

$$\mathrm{E}(C) = \sum_{i=2}^{n} \mathrm{E}(V_i) = \frac{1}{2} \sum_{i=2}^{n} (i-1) = \frac{1}{2} \sum_{i=1}^{n-1} i = \frac{n(n-1)}{4}$$

e

$$\mathrm{Var}(C) = \sum_{i=2}^{n} \mathrm{Var}(V_i) = \frac{1}{12} \sum_{i=2}^{n} (i^2-1) = \frac{1}{12} \sum_{i=1}^{n} (i^2-1) = \frac{n(n-1)(2n+5)}{72} \, . \qquad \square$$

Dal precedente teorema si ha inoltre

$$\mathrm{E}(\tau) = \frac{4}{n(n-1)} \mathrm{E}(C) - 1 = 0$$

e

$$\mathrm{Var}(\tau) = \frac{16}{n^2(n-1)^2} \mathrm{Var}(C) = \frac{2(2n+5)}{9n(n-1)} \, .$$

Teorema 10.3.5. *Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}^2_{F_1,F_2}$, allora C è simmetrica rispetto a* $\mathrm{E}(C) = n(n-1)/4$.

Dimostrazione. Dal momento che ogni V_i definita nel Teorema 10.3.2 è simmetrica rispetto a $\mathrm{E}(V_i) = (i-1)/2$ per $i = 2, 3, \ldots, n$, tenendo presente il Teorema 1.2.3, si ha

$$C - \mathrm{E}(C) \stackrel{d}{=} \sum_{i=2}^{n} (V_i - \mathrm{E}(V_i)) \stackrel{d}{=} \sum_{i=2}^{n} (\mathrm{E}(V_i) - V_i) \stackrel{d}{=} \mathrm{E}(C) - C \, ,$$

per cui dal Teorema 1.2.2 si deve concludere che C è simmetrica rispetto a $\mathrm{E}(C)$. $\square$

Dal Teorema 10.3.5 si ha che anche τ è simmetrica rispetto a $\mathrm{E}(\tau) = 0$. Dal Teorema 10.3.1, tenendo presente il Corollario 2.4.4, risulta che la statistica C è "distribution-free" su $\mathcal{C}_{F_1}$ e dunque anche su $\mathcal{I}^2_{F_1,F_2}$, ovvero anche il test di Kendall è "distribution-free". Se l'ipotesi di base non è vera, C tende ad assumere valori bassi o elevati. Fissato quindi un livello di significatività α, allora si sceglie come regione critica l'insieme

$$\mathcal{T}_1 = \{c : c \leq c_{n,\alpha/2}, c \geq c_{n,1-\alpha/2}\} \, ,$$

dove $c_{n,\alpha}$ rappresenta il quantile di ordine α della distribuzione di C per una numerosità campionaria pari a n.

• **Esempio 10.3.2.** I dati della Tavola 10.3.2 si riferiscono ai tempi fatti nei $100m$ piani dai primi 10 atleti classificati nella gara di decathlon alle Olimpiadi del 1988.

Tavola 10.3.2. Tempi nei 100*m* piani (in *sec*).

atleta	y_i	x_i	r_i	$\sum_{j=i+1}^{n} \mathrm{I}_{[0,\infty)}(R_j - R_i)$
Schenk - GDR	1	11.25	10	0
Voss - GDR	2	10.87	3	6
Steen - CAN	3	11.19	8	1
Thompson - GB	4	10.62	1	6
Blondel - FRA	5	11.02	4	4
Plaziat - FRA	6	10.83	2	4
Bright - USA	7	11.18	7	1
DeWit - HOL	8	11.05	5	2
Johnson - USA	9	11.15	6	1
Tarnovetsky - URS	10	11.23	9	0

Fonte: Lunn e McNeil (1991)

Si è interessati a conoscere se vi è associazione fra la posizione finale dell'atleta e il risultato fatto nella gara dei $100m$. Dal momento che è immediato ottenere che $c = 25$, per $n = 10$ si ha $\Pr(C \geq 25) = 0.364$ e quindi la significatività osservata risulta $\alpha_{oss} = 2 \times 0.364 = 0.718$. Dunque, la sola gara dei $100m$ non condiziona la posizione finale dell'atleta, dal momento che si può accettare H_0 ad ogni livello di significatività $\alpha < 0.718$. Si noti che in questo caso si ha $\tau = 1/9 \simeq 0.1111$, un valore basso del coefficiente di Kendall che conferma la mancanza di associazione. ◁

Per quanto riguarda la distribuzione per grandi campioni di C_n, si ha il seguente teorema.

Teorema 10.3.6. *Se $(X_1, Y_1), \ldots, (X_n, Y_n)$ è un campione casuale bivariato con funzione di ripartizione congiunta $F_n \in \mathcal{I}^2_{F_1,F_2}$, allora per $n \to \infty$*

$$\frac{C_n - n(n-1)/4}{\sqrt{n(n-1)(2n+5)/72}} \xrightarrow{d} N(0,1)\,.$$

Dimostrazione. Per il Teorema 10.3.2, C_n è data dalla somma di variabili casuali indipendenti, ma non ugualmente distribuite, ed è opportuno applicare il Teorema Fondamentale del Limite di Lindberg (Teorema A.3.7). Si ha

$$\sigma^2 = \sum_{i=2}^{n} \mathrm{Var}(V_i) = \frac{1}{12} \sum_{i=2}^{n} (i^2 - 1) \sim n^3\,.$$

Inoltre, tenendo presente che il supporto di V_i è dato da$\{0, 1, \ldots, i-1\}$ per $i = 2, 3, \ldots, n$, dalla disugualianza di Chebicheff si ha

$$\begin{aligned}\mathrm{E}((V_i - \frac{i-1}{2})^2 \mathbf{1}_{(-\infty,(i-1)/2-\epsilon\sigma)\cup((i-1)/2+\epsilon\sigma,\infty)}(V_i)) &\leq \frac{(i-1)^2}{4}\,\mathrm{E}(\mathbf{1}_{(-\infty,(i-1)/2-\epsilon\sigma)\cup((i-1)/2+\epsilon\sigma,\infty)}(V_i)) \\ = \frac{(i-1)^2}{4}\,\Pr(|V_i - \frac{i-1}{2}| > \epsilon\sigma)) &\leq \frac{(i-1)^2}{4}\,\frac{i^2-1}{12\epsilon^2\sigma^2} = \frac{(i-1)^2(i^2-1)}{48\epsilon^2\sigma^2}\,.\end{aligned}$$

Dunque, si ha

$$\frac{1}{\sigma^2}\sum_{i=2}^{n} \mathrm{E}((V_i - \frac{i-1}{2})^2 \mathbf{1}_{(-\infty,(i-1)/2-\epsilon\sigma)\cup((i-1)/2+\epsilon\sigma,\infty)}(V_i)) \leq \frac{1}{48\epsilon^2\sigma^4}\sum_{i=2}^{n}(i-1)^2(i^2-1) \sim \frac{n^5}{\epsilon^2 n^6} = \frac{1}{\epsilon^2 n}$$

da cui

$$\lim_n \frac{1}{\sigma^2}\sum_{i=2}^{n} \mathrm{E}((V_i - \frac{i-1}{2})^2 \mathbf{1}_{(-\infty,(i-1)/2-\epsilon\sigma)\cup((i-1)/2+\epsilon\sigma,\infty)}(V_i)) = 0\,.$$

Dal momento che la precedente relazione è soddisfatta per ogni $\epsilon > 0$, allora le condizioni del Teorema Fondamentale del Limite di Lindberg sono soddisfatte e il teorema è dimostrato. □

Dal Teorema 10.3.6, segue inoltre che

$$\frac{\tau_n}{\sqrt{2(2n+5)/(9n(n-1))}} \xrightarrow{d} N(0,1) .$$

Fissato un livello di significatività α, per $n \to \infty$ la regione critica può essere dunque approssimata dall'insieme

$$\{c : c \leq \mathrm{E}(C_n) + z_{\alpha/2}\sqrt{\mathrm{Var}(C_n)}, c \geq \mathrm{E}(C_n) + z_{1-\alpha/2}\sqrt{\mathrm{Var}(C_n)}\} .$$

• **Esempio 10.3.3.** I dati della Tavola 10.3.3 si riferiscono alle misurazioni relative al contenuto totale di acqua corporea e alla stime della massa magra corporea fornite dallo spessore della pliche cutanea misurato mediante il plicometro in 23 bambini.

Tavola 10.3.3. Acqua (in *l*) e massa magra (in *kg*).

soggetto	y_i	x_i	r_i	$\sum_{j=i+1}^{n} \mathbf{1}_{[0,\infty)}(R_j - R_i)$
1	7.35	11.0	2	21
2	7.56	11.5	3	20
3	7.65	10.7	1	20
4	9.03	12.9	5	18
5	9.91	12.3	4	18
6	10.12	14.4	7	16
7	10.22	14.0	6	16
8	10.52	15.1	8	15
9	10.59	17.4	11	12
10	11.73	18.0	12	11
11	11.86	15.3	9	12
12	12.33	16.6	10	11
13	12.62	19.7	14	9
14	13.52	19.2	13	9
15	14.43	21.6	15	8
16	15.83	22.9	17	6
17	15.97	21.7	16	6
18	17.49	25.7	18	5
19	18.96	31.4	20	3
20	22.75	29.0	19	3
21	27.18	40.0	21	2
22	27.20	43.4	22	1
23	30.24	44.3	23	0

Fonte: Brook (1971)

Si vuole verificare se esiste associazione fra il contenuto di acqua corporea e la pliche cutanea, ovvero si vuole verificare che le stime della massa magra corporea ottenute mediante il plicometro sono credibili. Dal momento che $c = 242$, allora

$$\Pr(C \geq 242) \simeq 1 - \Phi[(242 - 23 \times 22/4)/\sqrt{23 \times 22 \times 51/72}] = 1 - \Phi(6.1008) < 0.0005 ,$$

e quindi la significatività osservata risulta $\alpha_{oss} < 2 \times 0.0005 = 0.001$. In questo caso, si può respingere H_0 ad ogni livello di significatività $\alpha > 0.001$. Il coefficiente di Kendall risulta $\tau = 21/23 \simeq 0.9130$, un valore elevato che conferma la presenza di una forte associazione positiva. ◁

Il test classico per la verifica di ipotesi sull'indipendenza è basato sulla statistica test

$$F = \sqrt{n-2}\,\frac{R}{\sqrt{1-R^2}} ,$$

dove R è il coefficiente di correlazione campionario di Pearson. La Tavola 10.3.4 fornisce i valori dell'efficienza asintotica relativa $\text{EAR}_{C,F}$ per alcune distribuzioni. Il test di Kendall presenta buone prestazioni per grandi campioni anche sotto ipotesi di normalità bivariata.

Tavola 10.3.4. EAR del test di Kendall rispetto al test di Pearson.

distribuzione	$\text{EAR}_{C,F}$
$U(\lambda, \delta)$	1
$N(\mu, \sigma^2)$	$9/\pi^2 = 0.9119$
$L(\mu, \delta)$	$81/64 = 1.2666$

Capitolo 11

L'analisi della varianza

11.1. Ulteriori risultati per le statistiche rango. Si consideri il modello statistico "distribution-free"

$$\mathcal{L}_{\lambda_1,\ldots,\lambda_k,F} = \{F_n : F_n(x_{11},\ldots,x_{1n_1},\ldots,x_{k1},\ldots,x_{kn_k}) = \prod_{j=1}^{k}\prod_{i=1}^{n_j} F(x_{ji}-\lambda_j), F \in \mathcal{C}, \lambda_1,\ldots,\lambda_k \in \mathbb{R}\} ,$$

dove $n = \sum_{j=1}^{k} n_j$. Si noti che $\mathcal{L}_{\lambda_1,\ldots,\lambda_k,F}$ rappresenta il modello statistico relativo a k campioni casuali indipendenti provenienti da k variabili casuali assolutamente continue, che sono equivalenti in distribuzione a meno di un parametro di posizione. Inoltre, risulta $\mathcal{L}_{0,\ldots,0,F} = \mathcal{C}_F$. In questa struttura, è conveniente ampliare la definizione di variabile casuale rango.

Definizione 11.1.1. Siano $(X_{j1},\ldots,X_{jn_j})$, per $j = 1,\ldots,k$, k campioni casuali indipendenti con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Si definiscono statistiche rango le seguenti trasformate

$$R_{ji} = \sum_{l=1}^{n_j}\sum_{h=1}^{k} \mathbf{1}_{[0,\infty)}(X_{ji} - X_{hl}) , j = 1,\ldots,k , i = 1,\ldots,n_j .$$

Inoltre, si dice somma dei ranghi del j-esimo campione la trasformata

$$R_{j+} = \sum_{i=1}^{n_j} R_{ji} , j = 1,\ldots,k . \qquad \triangle$$

La statistica R_{ji} rappresenta la posizione di X_{ji} all'interno del campione misto ordinato. Dunque, le statistiche R_{ji}, per $i = 1,\ldots,n_j, j = 1,\ldots,k$, sono a tutti gli effetti statistiche rango, anche se con una differente indicizzazione. Quindi, tutte le proprietà discusse nella §2.4 sono valide anche in questo caso. Per quanto riguarda le statistiche somma dei ranghi si ha il seguente teorema.

Teorema 11.1.1. *Siano $(X_{j1},\ldots,X_{jn_j})$, per $j = 1,\ldots,k$, k campioni casuali indipendenti con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Si ha*

$$\mathrm{E}(R_{j+}) = \frac{n_j(n+1)}{2} , \mathrm{Var}(R_{j+}) = \frac{n_j(n-n_j)(n+1)}{12} , j = 1,\ldots,k ,$$

e

$$\mathrm{Cov}(R_{j+}, R_{h+}) = -\frac{n_j n_h (n+1)}{12} , j \neq h = 1,\ldots,k .$$

Dimostrazione. Tenendo presente il Corollario 2.4.5, si ha

$$\mathrm{E}(R_{j+}) = \sum_{i=1}^{n_j} \mathrm{E}(R_{ji}) = \sum_{i=1}^{n_j} \frac{n+1}{2} = \frac{n_j(n+1)}{2} , j = 1,\ldots,k ,$$

e

$$\mathrm{Var}(R_{j+}) = \sum_{i=1}^{n_j} \mathrm{Var}(R_{ji}) + \sum_{i=1}^{n_j}\sum_{l\neq i=1}^{n_j} \mathrm{Cov}(R_{ji}, R_{jl}) = \sum_{i=1}^{n_j} \frac{n^2-1}{12} - \sum_{i=1}^{n_j}\sum_{l\neq i=1}^{n_j} \frac{n+1}{12}$$
$$= \frac{n_j(n^2-1)}{12} - \frac{n_j(n_j-1)(n+1)}{12} = \frac{n_j(n-n_j)(n+1)}{12}\,,\ j=1,\ldots,k\,.$$

Inoltre, ancora dal Corollario 2.4.5 risulta

$$\mathrm{Cov}(R_{j+}, R_{h+}) = \sum_{i=1}^{n_j}\sum_{l=1}^{n_h} \mathrm{Cov}(R_{ji}, R_{hl}) = -\frac{n_j n_h(n+1)}{12}\,,\ j\neq h=1,\ldots,k\,. \qquad \square$$

Il seguente Teorema è un'estenione del Corollario 2.4.7 e sarà utile nel seguito.

Teorema 11.1.2. *Siano $(X_{j1},\ldots,X_{jn_j})$, per $j=1,\ldots,k$, k campioni casuali indipendenti con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Se $(R_{j(1)},\ldots,R_{j(n_j)})$ denota il vettore ordinato dei ranghi relativo a $(R_{j1},\ldots,R_{jn_j})$, per $j=1,\ldots,k$, allora*

$$\Pr(R_{1(1)}=r_{1(1)},\ldots,R_{1(n_1)}=r_{1(n_1)},\ldots,R_{k(1)}=r_{k(1)},\ldots,R_{k(n_k)}=r_{k(n_k)}) =$$
$$= \binom{n}{n_1\ldots n_k}^{-1}\,,\ 1\le r_{j(1)}<\ldots<r_{j(n_j)}\le n\,,\ j=1,\ldots,k\,.$$

Dimostrazione. Dal Corollario 2.4.7 si ha

$$\Pr(R_{1(1)}=r_{1(1)},\ldots,R_{1(n_1)}=r_{1(n_1)}) = \binom{n}{n_1}^{-1}\,.$$

Dal momento che si ha

$$\Pr(R_{2(1)}=r_{2(1)},\ldots,R_{2(n_2)}=r_{2(n_2)} \mid R_{1(1)}=r_{1(1)},\ldots,R_{1(n_1)}=r_{1(n_1)}) = \binom{n-n_1}{n_2}^{-1}$$

allora

$$\Pr(R_{1(1)}=r_{1(1)},\ldots,R_{1(n_1)}=r_{1(n_1)},R_{2(1)}=r_{2(1)},\ldots,R_{2(n_2)}=r_{2(n_2)}) =$$
$$= \binom{n-n_1}{n_2}^{-1}\binom{n}{n_1}^{-1} = \binom{n}{n_1 n_2}^{-1}\,.$$

Procedendo iterativamente si ha la dimostrazione. $\square$

11.2. Il test di Kruskal-Wallis. Siano $(X_{j1},\ldots,X_{jn_j})$, per $j=1,\ldots,k$, k campioni casuali indipendenti con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\lambda_1,\ldots,\lambda_k,F}$. Il test di Kruskal-Wallis è basato sulla statistica

$$H = \frac{12}{n(n+1)}\sum_{j=1}^{k} n_j(R_{j+}/n_j - (n+1)/2)^2\,.$$

La statistica H è in effetti una misura della differenza fra le somme dei ranghi attesi e le somme dei ranghi osservati. Con il test di Kruskal-Wallis si può verificare l'ipotesi $H_0: \lambda_1=\ldots=\lambda_k, F\in\mathcal{C}$, contro $H_1: \lambda_j\neq\lambda_h, \exists j\neq h=1,\ldots,k, F\in\mathcal{C}$. Questa struttura di ipotesi caratterizza la cosiddetta analisi della varianza ad un criterio. La statistica H è opportuna in questo sistema di ipotesi, in quanto se non è vera l'ipotesi di base tende ad assumere valori elevati. La statistica H può essere espressa alternativamente come

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j (R_{j+}^2/n_j^2 - (n+1)R_{j+}/n_j + (n+1)^2/4)$$
$$= \frac{12}{n(n+1)} \sum_{j=1}^{k} R_{j+}^2/n_j - \frac{12}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} R_{ji} + 3(n+1) = \frac{12}{n(n+1)} \sum_{j=1}^{k} R_{j+}^2/n_j - 3(n+1) .$$

Tenendo presente il Teorema 11.1.2, la funzione di probabilità della statistica H risulta

$$p_{n_1,\ldots,n_k}(h) = \binom{n}{n_1 \ldots n_k}^{-1} c_{n_1,\ldots,n_k}(h) ,$$

dove $c_{n_1,\ldots,n_k}(h)$ è il numero di k sottoinsiemi di $(n_1, \ldots, n_k)$ interi dell'insieme $(1, \ldots, n)$ per cui la statistica vale h. Dal momento che è molto laborioso tabulare la distribuzione esatta di H, dal momento che questa operazione dovrebbe essere fatto per ogni numerosità campionaria e per ogni k, è conveniente impiegare il seguente risultato per grandi campioni.

Teorema 11.2.1. *Siano $(X_{j1}, \ldots, X_{jn_j})$, per $j = 1, \ldots, k$, k campioni casuali indipendenti con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Se $\nu_j = \lim_n n_j/n$, dove $0 < \nu_j < 1$, $j = 1, \ldots, k$, allora per $n \to \infty$ si ha*

$$H_n \xrightarrow{d} \chi_{k-1}^2 .$$

Dimostrazione. Si veda Hettmansperger (1984). □

Un'approssimazione per grandi campioni della regione critica del test è quindi data dall'insieme

$$\{h : h \geq \chi_{k-1,1-\alpha}^2\} .$$

• **Esempio 11.2.1.** I dati della Tavola 11.2.1 forniscono i coefficienti di salinità per alcuni saggi di acque prelevati in tre zone nei pressi di Bimini Lagoon nelle Bahamas.

Tavola 11.2.1. Coefficienti di salinità (numero di parti per mille).

	sito 1		sito 2		sito 3	
saggio	x_{1i}	r_{1i}	x_{2i}	r_{2i}	x_{3i}	r_{3i}
1	37.54	10	40.17	27	39.04	19
2	37.02	4	40.80	30	39.21	21
3	36.71	1	39.76	24	39.05	20
4	37.04	6	39.70	23	38.24	12
5	37.32	7	40.79	29	38.53	13
6	37.01	3	40.44	28	38.71	16
7	37.03	5	39.75	25	38.89	18
8	37.70	11	39.38	22	38.66	15
9	37.36	8			38.51	14
10	36.75	2			40.08	26
11	37.45	9				
12	38.85	17				

Fonte: Till (1974)

Si è interessati a determinare se la salinità è identica nelle tre zone considerate, ovvero si vuole verificare il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2 = \lambda_3, F \in \mathcal{C}$, contro $H_1 : \lambda_j \neq \lambda_h, \exists j \neq h = 1, 2, 3, F \in \mathcal{C}$. Dal momento che $r_{1+} = 83$, $r_{2+} = 208$ e $r_{3+} = 174$, da cui $h \simeq 23.2539$, allora

$$\Pr(H \geq 23.2539) \simeq \Pr(\chi_2^2 \geq 23.2539) < 0.001 ,$$

per cui la significatività osservata risulta $\alpha_{oss} < 0.001$. Dato che si può respingere H_0 ad ogni livello di significatività $\alpha \geq 0.001$, l'evidenza empirica porta a concludere che vi è una differente salinità nei tre siti considerati. ◁

• **Esempio 11.2.2.** I dati della Tavola 11.2.2 forniscono i tempi di sopravvivenza di una serie di pazienti con cancro avanzato (allo stomaco, ai bronchi, al colon, alle ovarie e al seno) che sono stati trattati con ascorbato.

Tavola 11.2.2. Tempi di sopravvivenza (in giorni).

	stomaco		bronchi		colon		ovarie		seno	
paziente	x_{1i}	r_{1i}	x_{2i}	r_{2i}	x_{3i}	r_{3i}	x_{4i}	r_{4i}	x_{5i}	r_{5i}
1	124	18	81	14	248	32	1234	57	1235	58
2	42	7	461	45	377	38	89	15	24	3
3	25	4	21	2	189	27	201	28	1581	59
4	45	8	450	42	1843	61	356	35	1166	56
5	412	41	246	31	180	26	2970	62	40	6
6	51	10	167	25	537	47	456	44	727	49
7	1112	55	63	11	519	46			3808	64
8	46	9	64	12	455	43			791	51
9	103	17	155	22	406	40			1804	60
10	876	53	859	52	365	36			3460	63
11	146	20	151	21	942	54				
12	340	34	166	24	776	50				
13	396	39	37	5	372	37				
14			223	29	163	23				
15			138	19	101	16				
16			72	13	20	1				
17			245	30	283	33				

Fonte: Cameron e Pauling (1974)

Si è interessati a determinare se i tempi di sopravvivenza sono differenti a secondo degli organi malati, ovvero si vuole verificare il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5, F \in \mathcal{C}$, contro $H_1 : \lambda_j \neq \lambda_h, \exists j \neq h = 1, 2, 3, 4, 5, F \in \mathcal{C}$. Dal momento che $r_{1+} = 315$, $r_{2+} = 397$, $r_{3+} = 610$, $r_{4+} = 241$ e $r_{5+} = 517$, da cui $h \simeq 14.9169$, allora si ha

$$0.001 < \Pr(H \geq 14.9169) \simeq \Pr(\chi^2_4 \geq 14.9169) < 0.005\,,$$

ovveero la significatività osservata risulta $0.001 < \alpha_{oss} < 0.005$. Dunque, l'evidenza empirica porta a concludere che i tempi di sopravvivenza differiscono a secondo dell'organo colpito dal cancro, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha \geq 0.005$. ◁

Infine, per quanto riguarda le prestazioni per grandi campioni del test di Kruskal-Wallis, si può dimostrare che l'efficienza asintotica relativa del test basato sulla statistica H rispetto al test basato sulla classica statistica F dell'analisi della varianza a un criterio risulta (Hájek e Šidák, 1967)

$$\mathrm{EAR}_{H,F} = 12\sigma^2\Big(\int_{-\infty}^{\infty} f(x)^2\,dx\Big)^2\,.$$

Questa efficienza asintotica relativa è identica a quella ottenuta per il test W^+ di Wilcoxon rispetto al test T di Student per un campione e per il test W di Mann-Whitney-Wilcoxon rispetto al test T di Student per due campioni, per cui si può considerare ancora la Tavola 6.6.2, dove si hanno i valori di $\mathrm{EAR}_{H,F} = \mathrm{EAR}_{W^+,T}$ per alcune distribuzioni.

11.3. Il test di Friedman. Questo test viene utilizzato per la verifica di ipotesi nell'analisi della varianza a due criteri nel cosiddetto disegno campionario con blocchi casualizzati con una osservazione per cella. Più esattamente, in questo disegno si hanno nk soggetti che vengono divisi in n blocchi, in maniera tale che in ogni blocco i soggetti sono assegnati casualmente ai k trattamenti. Il modello statistico "distribution-free" opportuno in questo caso è quindi dato da

$$\mathcal{L}_{\lambda_1,\ldots,\lambda_k,F_1,\ldots,F_n} = \{F_n : F_n(x_{11},\ldots,x_{k1},\ldots,x_{1n},\ldots,x_{kn}) = \\ = \prod_{j=1}^{k}\prod_{i=1}^{n} F_i(x_{ji}-\lambda_j), F_1,\ldots,F_n \in \mathcal{C}, \lambda_1,\ldots,\lambda_k \in \mathbb{R}\} .$$

Si noti che $\mathcal{L}_{\lambda_1,\ldots,\lambda_k,F_1,\ldots,F_n}$ rappresenta il modello statistico relativo a nk campioni casuali indipendenti di una osservazione provenienti da nk variabili casuali assolutamente continue, che in ogni blocco sono equivalenti in distribuzione a meno di un parametro di posizione. Inoltre, $\mathcal{L}_{0,\ldots,0,F_1,\ldots,F_n}$ rappresenta il modello statistico relativo a n campioni casuali indipendenti di k osservazioni provenienti da n variabili casuali assolutamente continue. Siano dunque X_{ji}, per $j = 1,\ldots,k$, $i = 1,\ldots,n$, nk variabili casuali da cui si dispone di una sola osservazione, con funzione di ripartizione congiunta $F_n \in \mathcal{L}_{\lambda_1,\ldots,\lambda_k,F_1,\ldots,F_n}$. Si vuole verificare l'ipotesi di base del tipo $H_0 : \lambda_1 = \ldots = \lambda_k, F_1,\ldots,F_n \in \mathcal{C}$, contro l'alternativa $H_1 : \lambda_j \neq \lambda_h, \exists j \neq h = 1,\ldots,k, F_1,\ldots,F_n \in \mathcal{C}$, ovvero si vuole verificare se i trattamenti hanno effetto differente. Se è vera l'ipotesi di base, allora $(X_{1i},\ldots,X_{ki})$, per $i = 1,\ldots,n$, è un campione casuale e sia $(R_{1(i)},\ldots,R_{k(i)})$, per $i = 1,\ldots,n$, il relativo vettore dei ranghi. La statistica del test di Friedman è quindi data da

$$G = \frac{12}{nk(k+1)}\sum_{j=1}^{k}\left(R_{j(+)} - n(k+1)/2\right)^2 ,$$

dove in questo caso $R_{j(+)} = \sum_{i=1}^{n} R_{j(i)}$ rappresenta la somma dei ranghi associati al j-esimo trattamento in ogni blocco. Se è vera l'ipotesi di base, il rango associato ad un dato trattamento per un blocco è uniformente distribuito (vedi Corollario 2.4.4), per cui in conseguenza del Corollario 2.4.5 si ha $\mathrm{E}(R_{j(+)}) = \sum_{i=1}^{n}\mathrm{E}(R_{j(i)}) = n(k+1)/2$. Dunque, la statistica G, che è una misura della differenza fra le somme attese dei ranghi e le somme osservate dei ranghi per un trattamento, è una statistica opportuna per verificare l'ipotesi di base. La statistica G può essere espressa alternativamente come

$$G = \frac{12}{nk(k+1)}\sum_{j=1}^{k}\left(R_{j(+)}^2 - n(k+1)R_{j(+)} + n^2(k+1)^2/4\right) \\ = \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_{j(+)}^2 - \frac{12}{k}\sum_{j=1}^{k}\sum_{i=1}^{n} R_{j(i)} + 3n(k+1)\frac{12}{nk(k+1)}\sum_{j=1}^{k} R_{j(+)}^2 - 3n(k+1) .$$

Dal momento che è molto laborioso tabulare la distribuzione esatta di G, è conveniente adoperare il seguente risultato per grandi campioni.

Teorema 11.3.1. *Siano X_{ji}, per $j = 1,\ldots,k$, $i = 1,\ldots,n$, nk variabili casuali da cui si ha a disposizione una osservazione, con funzione di ripartizione congiunta data da $F_n \in \mathcal{L}_{0,\ldots,0,F_1,\ldots,F_n}$. Per $n \to \infty$ si ha*

$$G_n \xrightarrow{d} \chi^2_{k-1} .$$

Dimostrazione. Si veda Hettmansperger (1984). □

Un'approssimazione per grandi campioni della regione critica del test è quindi data dall'insieme

$$\{g : g \geq \chi^2_{k-1,1-\alpha}\} .$$

• **Esempio 11.3.1.** In un esperimento si vuole verificare l'efficacia delle scorie degli altoforni come materiale fertilizzante in agricoltura su tre tipi di suolo, ovvero fertile sabbioso, fertile argilloso e sabbia fertile. Vari tipi di fertilizzante per acro sono stati applicati e i dati della Tavola 11.3.1 si riferiscono al granturco prodotto. In questo caso, i vari tipi di fertilizzanti rappresentano i trattamenti, mentre i tipi di suolo sono i blocchi. Si è quindi interessati a verificare se i trattamenti sono equivalenti, ovvero l'ipotesi di base $H_0 : \lambda_1 = \ldots = \lambda_7, F_1, F_2, F_3 \in \mathcal{C}$, contro l'alternativa $H_1 : \lambda_j \neq \lambda_h, \exists j \neq h = 1,\ldots,7, F_1, F_2, F_3 \in \mathcal{C}$. Si

ha $r_{1(+)} = 7$, $r_{2(+)} = 10$, $r_{3(+)} = 8$, $r_{4(+)} = 12$, $r_{5(+)} = 13$, $r_{6(+)} = 20$ e $r_{7(+)} = 14$, da cui $g \simeq 8.1428$, e risulta

$$0.10 < \Pr(G \geq 8.1428) \simeq \Pr(\chi^2_6 \geq 8.1428) < 0.25 \, .$$

Quindi, la significatività osservata risulta $0.10 < \alpha_{oss} < 0.25$. L'evidenza empirica porta dunque a concludere che i fertilizzanti sono equivalenti, dal momento che si può accettare H_0 ad ogni livello di significatività $\alpha \leq 0.10$. ◁

Tavola 11.3.1. Granturco prodotto (in staio per acro).

Fertilizzante	fertile sabbioso		fertile argilloso		sabbia fertile	
	x_{j1}	$r_{j(1)}$	x_{j2}	$r_{j(2)}$	x_{j2}	$r_{j(2)}$
Nessuno	11.1	1	32.6	1	63.3	5
Scorie grezze	15.3	2	40.8	2	65.0	6
Scorie medie	22.7	3	52.1	3	58.8	2
Scorie per agricoltura	23.8	4	52.8	4	61.4	4
Calcare per agricoltura	25.6	5	63.1	7	41.1	1
Scorie per agricoltura+additivi	31.2	7	59.5	6	78.1	7
Calcare per agricoltura+additivi	25.8	6	55.3	5	60.2	3

Fonte: Johnson e Graybill (1972)

• **Esempio 11.3.2.** In un esperimento si vuole verificare l'aumento di peso di alcuni topi sottoposti a 4 tipi di dieta. Ogni dieta è caratterizzata da differenti apporti proteici (1 - moderato apporto proteico animale, 2 - elevato apporto proteico animale, 3 - moderato apporto proteico vegetale, 4 - elevato apporto proteico vegetale). I dati della Tavola 11.3.2 forniscono gli aumenti di peso in 10 gruppi di 4 topi ciascuno che sono stati assegnati casualmente alle diete.

Tavola 11.3.2. Aumenti di peso dei topi (in grammi).

j	x_{j1}	$r_{j(1)}$	x_{j2}	$r_{j(2)}$	x_{j3}	$r_{j(3)}$	x_{j4}	$r_{j(4)}$	x_{j5}	$r_{j(5)}$	x_{j6}	$r_{j(6)}$	x_{j7}	$r_{j(7)}$	x_{j8}	$r_{j(8)}$	x_{j9}	$r_{j(9)}$	x_{j10}	$r_{j(10)}$
1	90	2	76	2	90	2	64	1	86	2	51	1	72	1	90	4	95	3	78	2
2	73	1	102	4	118	4	104	3	81	1	107	4	100	4	87	3	117	4	111	4
3	107	4	95	3	97	3	80	2	98	4	74	2	74	2	67	1	89	2	58	1
4	98	3	74	1	56	1	111	4	95	3	88	3	82	3	77	2	86	1	92	3

Fonte: Snedecor e Cochran (1967)

In questo caso, i vari tipi di dieta rappresentano i trattamenti, mentre i gruppi di topi assegnati casualmete ai 4 trattamenti sono i blocchi. Si è quindi interessati a verificare se le diete sono equivalenti, ovvero l'ipotesi di base $H_0 : \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4, F_1, \ldots, F_{10} \in \mathcal{C}$, contro $H_1 : \lambda_j \neq \lambda_h, \exists j \neq h = 1, 2, 3, 4, F_1, \ldots, F_{10} \in \mathcal{C}$. Dal momento che $r_{1(+)} = 20$, $r_{2(+)} = 32$, $r_{3(+)} = 24$ e $r_{4(+)} = 24$, da cui $g \simeq 4.56$, allora si ha

$$0.10 < \Pr(G \geq 4.56) \simeq \Pr(\chi^2_3 \geq 4.56) < 0.25 \, ,$$

e quindi la significatività osservata risulta $0.10 < \alpha_{oss} < 0.25$. L'evidenza empirica porta dunque a concludere che le diete sono equivalenti, in quanto si può accettare H_0 ad ogni livello di significatività $\alpha \leq 0.10$. ◁

Per quanto riguarda l'efficacia del test di Friedman, si può dimostrare che l'efficienza asintotica relativa del test basato sulla statistica G rispetto al test basato sulla classica statistica F dell'analisi della varianza a due criteri risulta (Hájek e Šidák, 1967),

$$\mathrm{EAR}_{G,F} = \frac{12\sigma^2 k}{k+1} \left(\int_{-\infty}^{\infty} f(x)^2 \, dx \right)^2 .$$

Questa efficienza asintotica relativa è identica a quella ottenuta per il test H di Kruskal-Wallis rispetto al test per l'analisi della varianza a un criterio solo per $k \to \infty$. In particolare, se è vera l'assunzione di normalità, allora per $k = 2$ si ottiene $\mathrm{EAR}_{G,F} = 2/\pi$, ovvero l'efficienza asintotica relativa del test dei segni rispetto al test T di Student per un campione. Questa perdita di efficienza è dovuta all'ordinamento all'interno dei blocchi, specie per un piccolo numero di trattamenti. Tenendo presente la relazione

$\mathrm{EAR}_{G,F} = (k/(k+1))\mathrm{EAR}_{W^+,T}$, si può ottenere i valori di $\mathrm{EAR}_{G,F}$ per alcune distribuzioni mediante la Tavola 6.6.2.

11.4. Il test di concordanza di Kendall. La statistica del test di Friedman può essere adoperata anche nel problema dell'associazione quando si hanno a disposizione campioni da variabili casuali k dimensionate. Ad esempio, questo problema sorge quando si hanno k oggetti su cui viene misurato un certo attributo da n persone in maniera indipendente e si vuole avere una conferma della credibilità delle n misurazioni fatte, ovvero si vuole verificare l'associazione tra le misurazioni. Questa struttura dei dati è dunque analoga a quella relativa a k trattamenti fatti su n blocchi. Si consideri il modello statistico "distribution-free"

$$\mathcal{C}_F^k = \{F_n : F_n(x_{11}, \ldots, x_{k1}, \ldots, x_{1n}, \ldots, x_{kn} = \prod_{i=1}^{n} F(x_{1i}, \ldots, x_{ki}), F \in \mathcal{C}^k\} ,$$

dove $\mathcal{C}^k$ rappresenta la classe delle funzioni di ripartizione di un vettore k-variato di variabili casuali assolutamente continue. Si noti che $\mathcal{C}_F^k$ rappresenta il modello statistico relativo ad un campione casuale proveniente da un vettore k-variato di variabili casuali assolutamente continue. Una sottoclasse di $\mathcal{C}_F^k$ è data dal modello statistico "distribution-free"

$$\mathcal{I}_{F_1,\ldots,F_k}^k = \{F_n : F_n(x_{11}, \ldots, x_{k1}, \ldots, x_{1n}, \ldots, x_{kn}) = \prod_{j=1}^{k}\prod_{i=1}^{n} F_j(x_{ji}), F_1, \ldots, F_k \in \mathcal{C}\} .$$

Dunque, $\mathcal{I}_{F_1,\ldots,F_k}^k$ rappresenta il modello statistico relativo ad un campione casuale proveniente da un vettore k-variato di variabili casuali assolutamente continue a componenti indipendenti. Il sistema di ipotesi d'interesse è quindi $H_0 : F_n \in \mathcal{I}_{F_1,\ldots,F_k}^k$ contro $H_1 : F_n \in \mathcal{C}_F^k - \mathcal{I}_{F_1,\ldots,F_k}^k$, ovvero si vuole verificare l'indipendenza delle componenti del vettore k-variato di variabili casuali da cui proviene il campione casuale. Se è vera l'ipotesi di base allora $\mathcal{L}_{0,\ldots,0,F_1,\ldots,F_n} = \mathcal{I}_{F_1,\ldots,F_k}^k$ e quindi una statistica test opportuna è data dalla statistica di Kendall

$$K = \frac{G}{n(k-1)} .$$

Questa statistica è equivalente alla statistica di Friedman ed è stata standardizzata in modo che $0 \leq K \leq 1$, ovvero in modo che possa essere interpretata come coefficiente di associazione. Infatti, se non esiste associazione, allora $R_{j(+)} = n(k+1)/2$, per $j = 1, \ldots, k$, e quindi $K = 0$. In caso contrario, si ha $R_{j(+)} = nj$, per $j = 1, \ldots, k$, e quindi $K = 1$. Sotto ipotesi di base per $n \to \infty$ si ha

$$n(k-1)K \xrightarrow{d} \chi_{k-1}^2 .$$

Dunque, un'approssimazione per grandi campioni della regione critica del test è data dall'insieme

$$\{k : k \geq \chi_{k-1,1-\alpha}^2\} .$$

Capitolo 12

I test funzionali

13.1. Il test Chi-quadrato per la bontà di adattamento. Sia $(X_1, \ldots, X_n)$ un campione casuale da una variabile casuale discreta X. Supponiamo che il supporto di X sia finito e che la relativa funzione di probabilità sia data da

$$p(x_i) = \Pr(X = x_i) = \pi_i \, , i = 1, \ldots, r \, ,$$

Sia inoltre f_i, per $i = 1, \ldots, r$, la frequenza osservata di determinazioni del valore x_i nel campione. Si ha $\sum_{i=1}^{r} f_i = n$. Le quantità $(f_1, \ldots, f_r)$ sono dette frequenze osservate, mentre $(n\pi_1, \ldots, n\pi_r)$ sono dette frequenze attese. In questo caso, il sistema di ipotesi è dato da $H_0 : \pi_i = \pi_{i0}(\boldsymbol{\theta}), i = 1, \ldots, r$, contro $H_1 : \pi_i \neq \pi_{i0}(\boldsymbol{\theta}), \exists i = 1, \ldots, r$, dove $\boldsymbol{\theta}$ è un vettore di parametri tale che $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$. Con questo sistema di ipotesi si vuole dunque verificare che la distribuzione di X appartiene ad una famiglia di distribuzioni specificata (eventualmente) a meno di un vettore di parametri. Dal momento che le probabilità π_i, per $i = 1, \ldots, r$, specificano completamente la funzione di ripartizione di X, la precedente ipotesi è a tutti gli effetti una ipotesi funzionale. Una statistica test conveniente in questo caso è la statistica Chi-quadrato per la bontà d'adattamento data da

$$Q = \sum_{i=1}^{r} \frac{(f_i - n\pi_{i0}(\widehat{\boldsymbol{\Theta}}))^2}{n\pi_{i0}(\widehat{\boldsymbol{\Theta}})} \, ,$$

dove $\widehat{\boldsymbol{\Theta}}$ è uno stimatore di $\boldsymbol{\theta}$ coerente, efficiente per grandi campioni e che converge in distribuzione alla distribuzione Normale. La statistica Q può essere alternativamente espressa come

$$Q = \frac{1}{n} \sum_{i=1}^{r} \frac{f_i^2}{\pi_{i0}(\widehat{\boldsymbol{\Theta}})} - n \, .$$

Se le frequenze osservate si discostano molto da quelle attese stimate, si ottengono determinazioni elevate di Q che conseguentemente portano a respingere l'ipotesi di base. La distribuzione esatta di Q è proibitiva da calcolare e quindi è conveniente impiegare il seguente risultato per grandi campioni.

Teorema 12.1.1. *Sia $(X_1, \ldots, X_n)$ un campione casuale da una variabile casuale discreta X con funzione di probabilità data da $p(x_i) = \pi_{i0}(\boldsymbol{\theta})$, per $i = 1, \ldots, r$, dove $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$, e siano $(f_1, \ldots, f_r)$ le relative frequenze osservate. Se $\pi_{i0}(\boldsymbol{\theta})$ ammette derivate del primo e secondo ordine rispetto ad ogni $\boldsymbol{\theta} \in \Theta$ e se $\widehat{\boldsymbol{\Theta}}$ è uno stimatore di $\boldsymbol{\theta}$ coerente, efficiente per grandi campioni e che converge in distribuzione alla distribuzione Normale, allora per $n \to \infty$*

$$Q = \sum_{i=1}^{r} \frac{(f_i - n\pi_{i0}(\widehat{\boldsymbol{\Theta}}))^2}{n\pi_{i0}(\widehat{\boldsymbol{\Theta}})} \xrightarrow{d} \chi^2_{r-1-q} \, .$$

Dimostrazione. Si veda Serfling (1980). □

Un'approssimazione per grandi campioni della regione critica del test è quindi data dall'insieme

$$\{q : q \geq \chi^2_{r-1-q,1-\alpha}\} \, .$$

• **Esempio 12.1.1.** I dati della Tavola 12.1.1 si riferiscono al numero di maschi nei primi sette figli di 1334 pastori prostestanti svedesi. Si vuole verificare che questi dati provengono da una distribuzione Binomiale $Bi(7,\theta)$, ovvero si vuole verificare l'ipotesi di base $H_0 : \pi_i = \binom{7}{i-1}\theta^{i-1}(1-\theta)^{7-i+1}, i = 1,\ldots,8$, contro l'alternativa $H_1 : \pi_i \neq \binom{7}{i-1}\theta^{i-1}(1-\theta)^{7-i+1}, \exists i = 1,\ldots,8$, dove $\theta \in (0,1)$.

Tavola 12.1.1. Numero dei figli maschi.

Numero figli maschi	Frequenze osservate
0	6
1	57
2	206
3	362
4	365
5	256
6	69
7	13

Fonte: Edwards e Fraccaro (1960)

Il parametro θ può essere stimato col metodo della massima verosimiglianza, che fornisce stimatori coerenti, efficienti per grandi campioni e che convergono in distribuzione alla distribuzione Normale. La funzione di log-verosimiglianza risulta

$$\begin{aligned}\log L(\theta) &= c + \sum_{i=1}^{8} f_i \log \pi_i(\theta) = c + \sum_{i=1}^{8} f_i \log(\theta^{i-1}(1-\theta)^{7-i+1}) \\ &= c + \log\theta \sum_{i=1}^{8}(i-1)f_i + \ln(1-\theta)\sum_{i=1}^{8}(7-i+1)f_i \,, \theta \in (0,1) \,.\end{aligned}$$

La funzione di log-verosimiglianza è massimizzata per

$$\widehat{\theta} = \frac{1}{7n}\sum_{i=1}^{8}(i-1)f_i \,,$$

ovvero per $\widehat{\theta} \simeq 0.5140$. Dunque, si ha $\pi_1(\widehat{\theta}) \simeq 0.0064$, $\pi_2(\widehat{\theta}) \simeq 0.0474$, $\pi_3(\widehat{\theta}) \simeq 0.1504$, $\pi_4(\widehat{\theta}) \simeq 0.2652$, $\pi_5(\widehat{\theta}) \simeq 0.2804$, $\pi_6(\widehat{\theta}) \simeq 0.1780$, $\pi_7(\widehat{\theta}) \simeq 0.0627$, $\pi_8(\widehat{\theta}) \simeq 0.0095$, da cui $q \simeq 5.9426$. Di conseguenza, si ha

$$0.25 < \Pr(Q \geq 5.9426) \simeq \Pr(\chi^2_6 \geq 5.9426) < 0.50 \,,$$

e quindi la significatività osservata risulta $0.25 < \alpha_{oss} < 0.50$. In questo caso, si può accettare H_0, ovvero che il campione proviene da una distribuzione Binomiale $Bi(7,\theta)$, ad ogni livello di significatività $\alpha \leq 0.25$. ◁

• **Esempio 12.1.2** I dati della Tavola 12.1.2 forniscono la distribuzione della prima cifra dei numeri contenuti in un volume della rivista Reader's Digest scelto casualmente. Un modello teorico per questi dati è la cosiddetta distribuzione anomala (vedi Feller, 1971). La funzione di probabilità di una variabile casuale anomala è data da

$$p(x) = \log_{10}(1 + 1/x)\mathbf{1}_{\{1,\ldots,9\}}(x) \,.$$

Si vuole verificare che i dati provengono da una variabile casuale anomala, ovvero si vuole verificare l'ipotesi $H_0 : \pi_i = \log_{10}(1+1/i), i = 1,\ldots,9$, contro $H_1 : \pi_i \neq \log_{10}(1+1/i), \exists i = 1,\ldots,9$. Non vi sono parametri da stimare, e dunque si ha $q \simeq 3.2776$, da cui

$$0.90 < \Pr(Q \geq 3.2776) \simeq \Pr(\chi^2_8 \geq 3.2776) < 0.95 \,,$$

e quindi la significatività osservata risulta $0.90 < \alpha_{oss} < 0.95$. In questo caso, vi è una forte evidenza empirica ad accettare H_0, ovvero che il campione proviene da una distribuzione anomala, in quanto si potrebbe accettare questa ipotesi ad ogni livello di significatività $\alpha \leq 0.90$. ◁

Tavola 12.1.2. Prime cifre dei numeri contenuti in un volume.

Prima cifra	Frequenze osservate
1	103
2	57
3	38
4	23
5	22
6	20
7	17
8	15
9	13

Benford, F. (1938)

Sia $(X_1, \ldots, X_n)$ un campione casuale proveniente da una variabile casuale X discreta con supporto numerabile, con relativa funzione di probabilità $p(x_i) = \Pr(X = x_i) = \pi_i$, per $i = 1, 2, \ldots$. In questo caso, per rendere applicabile il test Chi-quadrato si deve considerare un raggruppamento di valori. Più esattamente, si considera solamente i primi $(r-1)$ valori x_i con relative probabilità $\pi_i = \Pr(X = x_i)$, per $i = 1, \ldots, r-1$, e l'insieme di valori $x_r, x_{r+1}, \ldots$ con relativa probabilità $\pi_r = \sum_{i=r}^{\infty} \Pr(X = x_i)$. Dunque, in questo caso f_r rappresenta la frequenza osservata dei valori $x_r, x_{r+1}, \ldots$.

• **Esempio 12.1.3.** I dati della Tavola 12.1.3 si riferiscono al numero di taxi arrivati in intervalli di un minuto alla stazione di Euston a Londra fra le 9.00 e le 10.00 in una mattina del 1950. Se gli arrivi sono casuali allora è noto dalla teoria dei processi stocastici di punto che i dati provengono da una distribuzione di Poisson $Po(\theta)$. Raggruppando opportunamente i valori, si vuole verificare l'ipotesi di base $H_0 : \pi_i = e^{-\theta}\theta^{i-1}/(i-1)!, i = 1, \ldots, 5, \pi_6 = 1 - \sum_{i=1}^{5} e^{-\theta}\theta^{i-1}/(i-1)!$, dove $\theta \in \mathbb{R}^+$.

Tavola 12.1.3. Numero dei taxi arrivati alla stazione in un ora.

numero di taxi per minuto	frequenza
0	18
1	18
2	14
3	7
4	3
più di 5	0

Fonte: Kendall (1951)

Il parametro θ può essere stimato col metodo della massima verosimiglianza. Tenendo presente che $f_6 = 0$, la funzione di log-verosimiglianza risulta

$$\log L(\theta) = c + \sum_{i=1}^{6} f_i \log \pi_i(\theta) = c + \sum_{i=1}^{5} f_i \log(e^{-\theta}\theta^{i-1}/(i-1)!) + f_6 \log(1 - \sum_{i=1}^{5} e^{-\theta}\theta^{i-1}/(i-1)!)$$

$$= c + \sum_{i=1}^{5} f_i(-\theta + (i-1)\log\theta) = c + \log\theta \sum_{i=1}^{5} (i-1) f_i - n\theta \,, \theta > 0 \,.$$

La funzione di log-verosimiglianza viene massimizzata per

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{5} (i-1) f_i \,,$$

ovvero $\widehat{\theta} \simeq 1.3167$. Dunque, si ha $\pi_1(\widehat{\theta}) = 0.2680$, $\pi_2(\widehat{\theta}) = 0.3529$, $\pi_3(\widehat{\theta}) = 0.2323$, $\pi_4(\widehat{\theta}) = 0.1020$, $\pi_5(\widehat{\theta}) = 0.0336$, $\pi_6(\widehat{\theta}) = 0.0112$, da cui $q \simeq 1.9841$. Di conseguenza

$$0.50 < \Pr(Q \geq 1.9841) \simeq \Pr(\chi^2_4 \geq 1.9841) < 0.90 \,,$$

quindi la significatività osservata risulta $0.50 < \alpha_{oss} < 0.90$. In questo caso, si può accettare H_0, ovvero che il campione proviene da una variabile casuale di Poisson, ad ogni livello di significatività $\alpha \leq 0.50$. ◁

• **Esempio 12.1.4** I dati della Tavola 12.1.4 si riferiscono al numero di scintillazioni in intervalli di 72 secondi causate dal decadimento radioattivo del polonio. Se le scintillazioni avvengono casualmente, allora è noto dalla teoria dei processi stocastici di punto che i dati provengono da una distribuzione di Poisson $Po(\theta)$. Raggruppando i valori, si vuole dunque verificare l'ipotesi $H_0 : \pi_i = e^{-\theta}\theta^{i-1}/(i-1)!$, $i = 1, \ldots, 12$, $\pi_{13} = 1 - \sum_{i=1}^{12} e^{-\theta}\theta^{i-1}/(i-1)!$, dove $\theta \in \mathbb{R}^+$.

Tavola 12.1.4. Numero di scintillazioni (in intervalli di 72*sec*) del polonio.

Numero scintillazioni	Frequenze osservate
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
≥ 12	2

Fonte: Rutheford e Geiger (1910)

Analogamente all'Esempio 12.1.3, il parametro θ può essere stimato col metodo della massima verosimiglianza. Tuttavia, in questo caso f_{13} non è nullo, per cui la funzione di log-verosimiglianza risulta

$$\log L(\theta) = c + \sum_{i=1}^{13} f_i \log \pi_i = c + \log\theta \sum_{i=1}^{12} (i-1) f_i - n\theta + f_{13} \log(1 - \sum_{i=1}^{12} e^{-\theta}\theta^{i-1}/(i-1)!)\,, \theta > 0\,,$$

che può essere massimizzata solo numericamente. Per via numerica si ottiene che il massimo è raggiunto per $\widehat{\theta} = 3.8678$. Sostituendo, si ha $\pi_1(\widehat{\theta}) = 0.0209$, $\pi_2(\widehat{\theta}) = 0.0808$, $\pi_3(\widehat{\theta}) = 0.1563$, $\pi_4(\widehat{\theta}) = 0.2016$, $\pi_5(\widehat{\theta}) = 0.1950$, $\pi_6(\widehat{\theta}) = 0.1508$, $\pi_7(\widehat{\theta}) = 0.0972$, $\pi_8(\widehat{\theta}) = 0.0537$, $\pi_9(\widehat{\theta}) = 0.0260$, $\pi_{10}(\widehat{\theta}) = 0.0112$, $\pi_{11}(\widehat{\theta}) = 0.0043$, $\pi_{12}(\widehat{\theta}) = 0.0015$, $\pi_{13}(\widehat{\theta}) = 0.0004$, da cui $q \simeq 14.6580$. Di conseguenza, si ha

$$0.10 < \Pr(Q \geq 14.6580) \simeq \Pr(\chi^2_{11} \geq 14.6580) < 0.25\,,$$

quindi la significatività osservata risulta $0.10 < \alpha_{oss} < 0.25$. In questo caso si può accettare H_0, ovvero che il campione proviene da una variabile casuale di Poisson, ad ogni livello di significatività $\alpha \leq 0.10$. ◁

• **Esempio 12.1.5.** In un indagine sono state contate le bombe cadute in un area di 36km^2 nella parte sud di Londra durante la Seconda Guerra Mondiale. L'area è stata suddivisa in sottoaree ognuna delle quali misurava 1/16 di km^2, per un totale di 576 sottoaree in cui è stato contato il numero di bombe cadute. I dati relativi sono forniti nella Tavola 12.1.5. Si vuole verificare se le bombe sono cadute casualmente o se sono state lanciate su precisi obiettivi. Se le bombe sono cadute casualmente, allora dalla teoria dei processi stocastici di punto sul piano è noto che i dati provengono da una $Po(\theta)$. Si vuole verificare l'ipotesi $H_0 : \pi_i = e^{-\theta}\theta^{i-1}/(i-1)!, i = 1, \ldots, 5, \pi_6 = 1 - \sum_{i=1}^{5} e^{-\theta}\theta^{i-1}/(i-1)!$, dove $\theta \in \mathbb{R}^+$.

Tavola 12.1.5. Numero di bombe (in aree di 1/4 di km^2).

Numero bombe	Frequenze osservate
0	229
1	211
2	93
3	35
4	7
≥ 5	1

Fonte: Clarke (1946)

Analogamente all'Esempio 13.1.4, lo stimatore di massima verosimiglianza di θ viene ottenuto massimizzando numericamente la log-verosimiglianza, da cui si ha $\widehat{\theta} = 0.9275$. Sostituendo si ha $\pi_1(\widehat{\theta}) = 0.3955$, $\pi_2(\widehat{\theta}) = 0.3669$, $\pi_3(\widehat{\theta}) = 0.1701$, $\pi_4(\widehat{\theta}) = 0.0526$, $\pi_5(\widehat{\theta}) = 0.0122$, $\pi_6(\widehat{\theta}) = 0.0027$, da cui $q \simeq 1.1878$. Di conseguenza, si ha

$$0.50 < \Pr(Q \geq 1.1878) \simeq \Pr(\chi^2_4 \geq 1.1878) < 0.90\,,$$

e quindi la significatività osservata risulta $0.50 < \alpha_{oss} < 0.90$. In questo caso, si può accettare H_0, ovvero che il campione proviene da una variabile casuale di Poisson, ad ogni livello di significatività $\alpha \leq 0.50$. ◁

13.2. Il test Chi-quadrato per la bontà di adattamento con *k* campioni. Siano $(X_{j1}, \ldots, X_{jn_j})$, per $j = 1, \ldots, k$, k campioni casuali indipendenti, ognuno dei quali proviene rispettivamente da una variabile casuale discreta X_j, per $j = 1, \ldots, k$. Sia inoltre $n = \sum_{j=1}^{k} n_j$. Supponiamo che le X_j abbiano identico supporto finito e che la relativa funzione di probabilità sia data da

$$p(x_i) = \Pr(X_j = x_i) = \pi_{ji}\,,\ i = 1, \ldots, r\,,$$

Sia inoltre f_{ji} la frequenza osservata di determinazioni del valore x_i nel j-esimo campione, per $i = 1, \ldots, r$, $j = 1, \ldots, k$. Si ha $\sum_{i=1}^{r} f_{ji} = n_j$ per $j = 1, \ldots, k$. Inoltre, si indica con $f_{+i} = \sum_{j=1}^{k} f_{ji}$, per $i = 1, \ldots, r$. In questo caso, il sistema di ipotesi è dato da $H_0 : \pi_{1i} = \ldots = \pi_{ki} = \pi_i, i = 1, \ldots, r$, contro $H_1 : \pi_{ji} = \pi_{li}, \exists j \neq l = 1, \ldots, k, i = 1, \ldots, r$. Si vuole dunque verificare l'omogeneità delle distribuzioni delle X_j, per $j = 1, \ldots, k$. Dal momento che le probabilità π_{ji}, per $i = 1, \ldots, r, j = 1, \ldots, k$, specificano completamente le funzioni di ripartizione delle X_j, la precedente ipotesi è a tutti gli effetti una ipotesi funzionale. Una statistica test conveniente in questo caso è la statistica Chi-quadrato per la bontà d'adattamento con k campioni data da

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(f_{ji} - n_j \widehat{\pi}_i)^2}{n_j \widehat{\pi}_i}\,,$$

dove $\widehat{\pi}_i = f_{+i}/n$, per $i = 1, \ldots, r$. La statistica Q può essere alternativamente espressa come

$$Q = n \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{f_{ji}^2}{n_j f_{+i}} - n\,.$$

Se le frequenze osservate si discostano molto dalle frequenze attese stimate si ottengono realizzazioni elevate di Q, che conseguentemente portano a respingere l'ipotesi di base. La distribuzione esatta di Q è proibitiva da calcolare e quindi è conveniente impiegare il seguente risultato per grandi campioni.

Teorema 12.2.1. *Siano $(X_{j1}, \ldots, X_{jn_j})$, per $j = 1, \ldots, k$, k campioni casuali indipendenti, ognuno dei quali proviene rispettivamente da una variabile casuale discreta X_j con funzione di probabilità $p(x_i) = \pi_i$ per $i = 1, \ldots, r, j = 1, \ldots, k$. Allora, per $n \to \infty$*

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(f_{ji} - n_j \widehat{\pi}_i)^2}{n_j \widehat{\pi}_i} \xrightarrow{d} \chi^2_{rk-k-r+1}\,,$$

dove $\widehat{\pi}_i = f_{+i}/n$, per $i = 1, \ldots, r$.

Dimostrazione. Si veda Serfling (1980). □

Un'approssimazione per grandi campioni della regione critica del test è quindi data dall'insieme

$$\{q : q \geq \chi^2_{rk-k-r+1,1-\alpha}\}\,.$$

• **Esempio 12.2.1.** Quando i dati vengono raccolti, il rilevatore frequentemente arrotonda l'ultima cifra del dato a certe cifre convenienti, quali ad esempio 0 e 5. Questo fenomeno di “accatastamento” di cifre è stato

osservato nella rilevazione dei dati più disparati. Si considerino ad esempio i dati della Tavola 12.2.1, che si riferiscono alle frequenze della cifra finale delle misurazioni delle temperature massime e minime giornaliere ufficiali negli Stati Uniti negli anni 1922-1924.

Tavola 12.2.1. Cifra finale delle temperature minime e massime.

Cifra	max	min
0	194	177
1	148	149
2	108	113
3	72	70
4	29	35
5	51	50
6	32	27
7	54	68
8	102	98
9	210	213

Fonte: Preece (1981)

Si vuole determinare se i dati provengono dalla stessa distribuzione, ovvero si vuole verificare $H_0 : \pi_i = \pi_{1i} = \pi_{2i}, i = 1, \ldots, 10$. Si ha dunque $q \simeq 3.6276$, da cui

$$0.90 < \Pr(q \geq 3.6276) \simeq \Pr(\chi^2_9 \geq 3.6276) < 0.95 \,,$$

quindi la significatività osservata risulta $0.90 < \alpha_{oss} < 0.95$. Quindi, si deve accettare che i dati provengono dalla stessa distribuzione, in quanto si può accettare H_0 ad ogni livello di significatività $\alpha \leq 0.90$. ◁

• **Esempio 12.2.2.** Nel 1861 il periodico New Orleans Crescent pubblicò 10 lettere a firma Quinto Curzio Snodgrass. Sebbene gli eventi discussi in queste lettere sembrano essere accaduti realmente, non esiste alcuna traccia di uno scrittore con questo nome. Si è supposto che il vero autore delle lettere sia stato Mark Twain. Un modo per verificare statisticamente la paternità degli scritti è quella di comparare le distribuzioni delle lunghezze delle parole. I dati della Tavola 12.2.2 si riferiscono alle distribuzioni del numero di lettere delle parole di due brani scelti casualmente dagli scritti di Mark Twain e Quinto Curzio Snodgrass.

Tavola 12.2.2. Lunghezza delle parole di due brani.

Lunghezza	Twain	Q.C.S.
1	312	424
2	1146	2685
3	1394	2752
4	1177	2302
5	661	1431
6	442	992
7	367	896
8	231	638
9	181	465
10	109	276
11	50	152
12	24	101
> 13	12	61

Fonte: Brinegar (1963)

Si vuole determinare se i due brani sono stati scritti dalla medesima persona, ovvero si vuole verificare $H_0 : \pi_i = \pi_{1i} = \pi_{2i}, i = 1, \ldots, 13$. Si ha $q \simeq 101.4938$, da cui

$$\Pr(Q \geq 101.4938) \simeq \Pr(\chi^2_{12} \geq 101.4938) < 0.001 \,,$$

e quindi la significatività osservata risulta $\alpha_{oss} < 0.001$. Dunque, si deve respingere l'ipotesi che Twain e Snodgrass siano la stessa persona, in quanto si può respingere H_0 ad ogni livello di significatività $\alpha \geq 0.001$. ◁

Quando le probabilità π_i, per $i = 1, \ldots, r$, sono note il test si basa sulla statistica

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(f_{ji} - n_j \pi_i)^2}{n_j \pi_i} \, .$$

La distribuzione esatta di Q è proibitiva da calcolare e quindi è conveniente impiegare il seguente risultato per grandi campioni.

Teorema 12.2.2. *Siano $(X_{j1}, \ldots, X_{jn_j})$, per $j = 1, \ldots, k$, k campioni casuali indipendenti, ognuno dei quali proviene rispettivamente da una variabile casuale discreta X_j con funzione di probabilità $p(x_i) = \pi_i$, per $i = 1, \ldots, r$, $j = 1, \ldots, k$. Allora, per $n \to \infty$*

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(f_{ji} - n_j \pi_i)^2}{n_j \pi_i} \xrightarrow{d} \chi^2_{rk-k} \, .$$

Dimostrazione. Si veda Serfling (1980). □

Un'approssimazione per grandi campioni della regione critica del test è quindi data dall'insieme

$$\{q : q \geq \chi^2_{rk-k,1-\alpha}\} \, .$$

• **Esempio 12.2.3.** Mediante un elaborazione con un sistema di calcolo simbolico sono state determinate le prime 10000 cifre nell'espansione decimale dei numeri trascendenti π ed e. Per ognuno di questi numeri è stata contata la frequenza di ogni cifra $\{0, 1, \ldots, 9\}$ e si sono ottenuti i dati di Tavola 12.2.3.

Tavola 12.2.3. Cifre dei numeri trascendenti π ed e.

Cifra	π	e
0	968	974
1	1026	989
2	1021	1004
3	974	1008
4	1012	982
5	1046	992
6	1021	1079
7	970	1008
8	948	996
9	1014	968

Dalla teoria dei numeri è noto che π ed e sono normali con probabilità 1 (un numero è detto normale se ogni cifra si presenta con la medesima frequenza nella sua espansione decimale). Si vuole ottenere una conferma statistica dell'uniformità della distribuzione delle cifre di π ed e, ovvero si vuole verificare $H_0 : \pi_i = 1/10, i = 1, \ldots, 10$. Si ha $q \simeq 17.928$, da cui

$$0.25 < \Pr(Q \geq 17.928) \simeq \Pr(\chi^2_{18} \geq 17.928) < 0.50 \, ,$$

quindi la significatività osservata risulta $0.25 < \alpha_{oss} < 0.50$. Si può dunque accettare l'ipotesi di uniformità della distribuzione delle cifre dell'espansione decimale di π e di e ad ogni livello di significatività $\alpha \leq 0.25$⋉

12.3. La statistica di Kolmogorov. In questo paragrafo viene considerata la principale statistica per costruire test per la verifica di ipotesi funzionali quando la variabile casuale d'interesse è assolutamente continua.

Definizione 12.3.1. Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, allora si definisce come funzione di ripartizione empirica

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i)\,,\, x \in \mathbb{R}\,. \qquad \triangle$$

Se $(X_{(1)}, \ldots, X_{(n)})$ è la statistica ordinata relativa a $(X_1, \ldots, X_n)$, allora una rappresentazione alternativa della funzione di ripartizione empirica è data da

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} i\mathbf{1}_{[X_{(i)},X_{(i+1)})}(x)\,,$$

dove con abuso di notazione si è assunto che $X_{(0)} = -\infty$ e $X_{(n+1)} = \infty$.

Definizione 12.3.2. Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, la statistica di Kolmogorov è data da

$$D = \sup_x |\widehat{F}(x) - F(x)|\,. \qquad \triangle$$

La statistica di Kolmogorov è una misura della discrepanza fra la funzione di ripartizione e la funzione di ripartizione empirica. Questa statistica è "distribution-free", nel senso che non dipende dalla struttura funzionale di F, come viene dimostrato nel seguente teorema.

Teorema 12.3.3. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, allora la statistica D di Kolmogorov è "distribution-free" su $\mathcal{C}_F$.*

Dimostrazione. Si ha

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(0,F(x)]}(F(X_i))\,,\, x \in \mathbb{R}\,,$$

essendo F una funzione monotona crescente. Inoltre, tenendo presente la trasformata dell'Integrale di Probabilità si ha $F(X_i) \stackrel{d}{=} Y_i$, dove Y_i ha distribuzione Uniforme $U(0,1)$ per $i = 1, \ldots, n$. Se $H(y) = y\mathbf{1}_{(0,1]}(y) + \mathbf{1}_{(1,\infty)}(y)$ rappresenta la funzione di ripartizione di una variabile casuale Uniforme $U(0,1)$ e se $\widehat{H}(y)$ rappresenta la funzione di ripartizione empirica relativa al campione casuale trasformato $(Y_1, \ldots, Y_n)$, tenendo presente il Teorema 1.1.2, si ha

$$D = \sup_x |\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(0,F(x)]}(F(X_i)) - F(x)| \stackrel{d}{=} \sup_{0\le y\le 1} |\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(0,y]}(Y_i) - y| = \sup_y |\widehat{H}(y) - H(y)|\,,$$

ovvero D è distribuito come una variabile casuale che non dipende da F. $\triangleleft$

Il Teorema di Glivenko-Cantelli (vedi Serfling, 1980) implica $D \xrightarrow{p} 0$ per $n \to \infty$. Per quanto riguarda la distribuzione di D si ha il seguente teorema.

Teorema 12.3.4. *Se $(X_1, \ldots, X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, la funzione di ripartizione della statistica D di Kolmogorov è data da*

$$G(d) = n!(\int_{1/n-d}^{d}\int_{2/n-d}^{1/n+d}\ldots\int_{1-d}^{(n-1)/n+d} \mathbf{1}_A(x_1, \ldots, x_n)\,dx_1dx_2\ldots dx_n)\,\mathbf{1}_{(1/(2n),1)}(d) + \mathbf{1}_{[1,\infty)}(d)\,,$$

dove $A = \{(x_1, \ldots, x_n) : 0 < x_1 < \ldots < x_n < 1\}$.

Dimostrazione. In conseguenza del Teorema 12.3.3 si può assumere senza perdita di generalità che F sia la funzione di ripartizione di una variabile casuale con distribuzione Uniforme $U(0,1)$ e che $\widehat{F}(x)$ sia la funzione di ripartizione empirica relativa a un campione casuale proveniente dalla medesima distribuzione. Si ha $|\widehat{F}(x) - F(x)| = 0$ per $x \notin (0,1)$, per cui $\sup_x|\widehat{F}(x) - F(x)|$ deve essere ottenuto per qualche $x \in (0,1)$, ovvero la statistica di Kolmogorov può essere espressa come

$$D = \sup_{0<x<1} |\widehat{F}(x) - x| \, .$$

Il supporto di D risulta $(0,1)$, ovvero si deve determinare $\Pr(D \leq d)$ solamente per $0 < d < 1$. Tenendo presente le precedenti considerazioni, si ha

$$G(d) = \Pr(D \leq d) = \Pr(\sup_{0<x<1} |\widehat{F}(x) - x| \leq d) = \Pr(|\widehat{F}(x) - x| \leq d, \forall x \in (0,1)) \, , \, 0 < d < 1 \, .$$

Considerando la rappresentazione di $\widehat{F}(x)$ in termini della statistica ordinata, dall'espressione precedente si ha

$$\begin{aligned} G(d) &= \Pr(|i/n - x| \leq d, \forall x \in [X_{(i)}, X_{(i+1)}), \forall i = 0, 1, \ldots, n) \\ &= \Pr(i/n - d \leq x \leq i/n + d, \forall x \in [X_{(i)}, X_{(i+1)}), \forall i = 0, 1, \ldots, n) = \Pr(\bigcap_{i=0}^{n} E_i) \, , \, 0 < d < 1 \, , \end{aligned}$$

dove con abuso di notazione si definisce $X_{(0)} = 0$ e $X_{(n+1)} = 1$, mentre

$$E_i = \{i/n - d \leq x \leq i/n + d, \forall x \in [X_{(i)}, X_{(i+1)})\} \, , \, 0 < d < 1 \, .$$

L'insieme di valori di x comune agli eventi E_i e E_{i+1}, per $i = 0, 1, \ldots, n-1$, è dato da

$$\begin{aligned} &\{i/n - d \leq x \leq i/n + d\} \cap \{(i+1)/n - d \leq x \leq (i+1)/n + d\} = \\ &\quad = \{(i+1)/n - d \leq x \leq i/n + d\} \, , \, i = 0, 1, \ldots, n-1 \, , \, 1/(2n) \leq d < 1 \, , \end{aligned}$$

dove il vincolo $d \geq 1/(2n)$ deriva dalla condizione $i/n + d \geq (i+1)/n - d$. Inoltre, si noti che la variabile casuale $X_{(i+1)}$ è comune solo agli eventi E_i e E_{i+1}, per cui

$$E_i \cap E_{i+1} = \{(i+1)/n - d \leq X_{(i+1)} \leq i/n + d\} \, , \, i = 0, 1, \ldots, n-1 \, , \, 1/(2n) \leq d < 1 \, .$$

Di conseguenza, si ha

$$\bigcap_{i=0}^{n} E_i = \bigcap_{i=0}^{n-1} (E_i \cap E_{i+1}) = \bigcap_{i=0}^{n-1} \{(i+1)/n - d \leq X_{(i+1)} \leq i/n + d\} \, , \, 1/(2n) \leq d < 1 \, ,$$

ovvero risulta

$$\begin{aligned} G(d) &= \Pr(\bigcap_{i=0}^{n-1} \{(i+1)/n - d \leq X_{(i+1)} \leq i/n + d\}) \\ &= n! \int_{1/n-d}^{d} \int_{2/n-d}^{1/n+d} \ldots \int_{1-d}^{(n-1)/n+d} \mathbf{1}_A(x_1, \ldots, x_n) \, dx_1 dx_2 \ldots dx_n \, , \, 1/(2n) \leq d < 1 \, . \end{aligned}$$

Inoltre, per $d \leq 1/(2n)$ si ha $\Pr(D \leq d) = 0$, dal momento che in questo caso $\bigcap_{i=0}^{n} E_i = \emptyset$. Infine, per $d \geq 1$ si ha $\Pr(D \leq d) = 1$. □

• **Esempio 12.3.1.** Dal Teorema 12.3.4, per $n = 2$ si ha

$$G(d) = 2(\int_{1/2-d}^{d} \int_{1-d}^{1/2+d} \mathbf{1}_A(x_1 x_2) \, dx_1 dx_2) \, \mathbf{1}_{(1/4,1)}(d) + \mathbf{1}_{[1,\infty)}(d) \, ,$$

dove $A = \{(x_1, x_2) : 0 < x_1 < x_2 < 1\}$. Per quanto riguarda il precedente integrale si deve distinguere due casi. Per $1/4 < d < 1/2$ si ha

$$2 \int_{1/2-d}^{d} \int_{1-d}^{1/2+d} \mathbf{1}_A(x_1 x_2) \, dx_1 dx_2 = 2 \int_{1/2-d}^{d} \int_{1-d}^{1/2+d} dx_1 dx_2 = 8d^2 - 4d + \frac{1}{2} \, ,$$

mentre per $1/2 < d < 1$ si ha

$$2\int_{1/2-d}^{d}\int_{1-d}^{1/2+d}\mathbf{1}_A(x_1x_2)\,dx_1dx_2 = 2\int_0^{1-d}\int_{1-d}^{1}dx_1dx_2 + 2\int_{1-d}^{d}\int_{x_1}^{1}dx_1dx_2 = -2d^2+4d-1\,.$$

Dunque, risulta

$$G(d) = (8d^2-4d+1/2)\mathbf{1}_{(1/4,1/2)}(d) + (-2d^2+4d-1)\mathbf{1}_{(1/2,1)}(d) + \mathbf{1}_{[1,\infty)}(d). \qquad \triangleleft$$

Per quanto riguarda la distribuzione per grandi campioni di $D = D_n$ si ha il seguente teorema.

Teorema 12.3.5. *Se $(X_1,\ldots,X_n)$ è un campione casuale con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, per $n \to \infty$ si ha*

$$\lim_n \Pr(\sqrt{n}D \le d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2d^2}\,,\ d \ge 0\,.$$

Dimostrazione. Si veda Billingsley (1968). □

• **Esempio 12.3.2.** Nella Tavola 12.3.1 sono riportati i quantili di ordine 0.95 della distribuzione di D e le relative approssimazioni ottenute mediante il Teorema 12.3.5 per alcuni valori di n.

Tavola 12.3.1. Quantili $d_{n,0.95}$ di D e relative approssimazioni.

n	$d_{0.95}$	approssimazione
2	0.8419	0.9612
5	0.5633	0.6685
10	0.4087	0.4864
20	0.2939	0.3524
30	0.2417	0.2898
40	0.2101	0.2521
50	0.1884	0.2260

Dunque, D converge abbastanza lentamente alla relativa distribuzione per grandi campioni. $\triangleleft$

12.4. Il test di Kolmogorov. Sia $(X_1,\ldots,X_n)$ un campione casuale da una variabile casuale assolutamente continua X con funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$. Il sistema di ipotesi da verificare risulta $H_0 : F(x) = F_0(x), \forall x \in \mathbb{R}$, contro $H_1 : F(x) \ne F_0(x), \exists x \in \mathbb{R}$, dove $F_0(x)$ è una funzione di ripartizione completamente specificata. Il test si basa sulla statistica

$$D = \sup_x |\widehat{F}(x) - F_0(x)|\,,$$

ovvero sulla statistica di Kolmogorov. La statistica D può essere convenientemente espressa come

$$D = \max_{1\le i\le n}(\max(|i/n - F_0(X_{(i)})|, |(i-1)/n - F_0(X_{(i)})|))\,.$$

Se la funzione di ripartizione empirica si discosta molto da quella ipotizzata, allora si hanno valori elevati di D, che conseguentemente portano a respingere l'ipotesi di base in favore dell'ipotesi alternativa. Se dunque $d_{n,\alpha}$ rappresenta il quantile di ordine α della distribuzione di D, la regione critica del test risulta dunque

$$\mathcal{T}_1 = \{d : d \ge d_{n,1-\alpha}\}\,.$$

• **Esempio 12.4.1.** Su un campione di dieci ragnatele è stato misurato l'angolo fra l'asse della ragnatela e la perpendicolare alla superficie terrestre, e si sono ottenuti i dati della Tavola 12.4.1. Si vuole verificare se i dati provengono da una distribuzione di von Mises, che è una distribuzione circolare adatta a modellare questi dati. La funzione di densità di una variabile casuale di von Mises è data da

$$g(x) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\mu)} \mathbf{1}_{(0,2\pi]}(x) \, , \, 0 \leq \mu < 2\pi \, , \, \kappa > 0 \, ,$$

dove $I_0(\kappa)$ rappresenta la funzione di Bessel del primo tipo e ordine 0 e dove μ e κ sono rispettivamente la direzione media e il parametro di condentrazione. La relativa funzione di ripartizione è data da

$$G(x) = \frac{\mathbf{1}_{(0,2\pi]}(x)}{2\pi I_0(\kappa)} \int_0^x e^{\kappa \cos(t-\mu)} dt + \mathbf{1}_{(2\pi,\infty)}(x) \, .$$

Dal momento che si sospetta che le ragnatele siano state costruite da ragni della specie *Araneus rufipalpus* è noto da molte osservazioni precedenti che i parametri della distribuzione di von Mises sono in questo caso $\mu = 0.27$ per quanto riguarda la direzione media e $\kappa = 37.94$ per quanto riguarda il parametro di concentrazione.

Tavola 12.4.1. Angoli fra ragnatele e asse terrestre (in radianti).

i	$x_{(i)}$	$G(x_{(i)})$	$\lvert i/10 - G(x_{(i)})\rvert$	$\lvert (i-1)/10 - G(x_{(i)})\rvert$
1	0.1745	0.2297	0.1297	0.2297
2	0.2094	0.3057	0.1057	0.2057
3	0.2269	0.3464	0.0464	0.1464
4	0.2618	0.4307	0.0307	0.1307
5	0.2793	0.4735	0.0265	0.0735
6	0.2967	0.5159	0.0841	0.0159
7	0.4189	0.7702	0.0702	0.1702
8	0.4363	0.7968	0.0032	0.0968
9	0.4538	0.8209	0.0791	0.0209
10	0.5411	0.9022	0.0978	0.0022

Fonte: Gadsen e Kanji (1981)

Si vuole verificare l'ipotesi $H_0 : F(x) = G(x), \forall x \in \mathbb{R}$, contro $H_1 : F(x) \neq G(x), \exists x \in \mathbb{R}$. Dalla Tavola 12.4.1 si ricava che $d = 0.2297$, per cui

$$\Pr(D \geq 0.2297) > 0.20$$

e quindi la significatività osservata risulta $\alpha_{oss} > 0.20$. L'evidenza empirica porta dunque ad accettare che i dati provengono da una distribuzione di von Mises, dal momento che si può accettare H_0 ad ogni livello di significatività $\alpha \leq 0.20$. ◁

• **Esempio 12.4.2.** E' stato fatto un esperimento al fine di verificare la resistenza alla rottura di alcune fibre di poliestere. Più esattamente sono stati determinati i carichi da applicare a un campione di queste fibre al fine di provocarne il cedimento. Si sospetta che la distribuzione dei carichi segua una distribuzione log-Normale. I dati della Tavola 12.4.2 sono stati ottenuti trasformando quelli originali mediante la trasformazione dell'integrale di probabilità, ovvero se l'ipotesi di log-Normalità è valida allora il campione trasformato dovrebbe provenire da una distribuzione Uniforme $U(0,1)$. Dal momento che

$$F_0(x) = x\mathbf{1}_{(0,1)}(x) + \mathbf{1}_{[1,\infty)}(x) \, ,$$

si vuole dunque verificare l'ipotesi $H_0 : F(x) = F_0(x), \forall x \in \mathbb{R}$, contro $H_1 : F(x) \neq F_0(x), \exists x \in \mathbb{R}$. Dalla Tavola 12.4.2 si ricava $d = 0.238$, per cui

$$0.05 < \Pr(D \geq 0.238) < 0.10$$

e quindi la significatività osservata risulta $0.05 < \alpha_{oss} < 0.10$. In questo caso, vi è qualche dubbio ad accettare H_0, ovvero che il campione trasformato proviene da una distribuzione Uniforme $U(0,1)$ e che quindi il campione originale proviene da una variabile casuale log-Normale, in quanto si potrebbe respingere questa ipotesi ad ogni livello di significatività $\alpha \geq 0.10$. ◁

Tavola 12.4.2. Carichi di rottura delle fibre di poliestere.

fibra	$x_{(i)}$	$F_0(x_{(i)})$	$\lvert i/30 - F_0(x_{(i)})\rvert$	$\lvert (i-1)/30 - F_0(x_{(i)})\rvert$
1	.023	.023	.010	.023
2	.032	.032	.035	.001
3	.054	.054	.046	.013
4	.069	.069	.064	.031
5	.081	.081	.086	.052
6	.094	.094	.106	.073
7	.105	.105	.128	.095
8	.127	.127	.140	.106
9	.148	.148	.152	.119
10	.169	.169	.164	.131
11	.188	.188	.178	.145
12	.216	.216	.184	.151
13	.255	.255	.178	.145
14	.277	.277	.190	.156
15	.311	.311	.189	.156
16	.361	.361	.172	.139
17	.376	.376	.191	.170
18	.395	.395	.205	.172
19	.432	.432	.201	.168
20	.463	.463	.204	.170
21	.481	.481	.219	.186
22	.519	.519	.214	.181
23	.529	.529	.238	.204
24	.567	.567	.233	.200
25	.642	.642	.191	.158
26	.674	.674	.193	.159
27	.752	.752	.148	.115
28	.832	.832	.110	.077
29	.887	.887	.080	.046
30	.926	.926	.074	.041

Fonte: Quesenberry e Hales (1980)

12.5. La statistica di Kolmogorov-Smirnov. In questo paragrafo viene considerata una modifica della statistica di Kolmogorov che permette di costruire test per la verifica di ipotesi funzionali quando la variabile casuale d'interesse è continua e si dispone di due campioni.

Definizione 12.5.1. Siano $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$ due campioni casuali indipendenti, tali che il campione misto abbia funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, dove $n = n_1 + n_2$. Siano inoltre $\widehat{F}_1(x)$ e $\widehat{F}_2(x)$ le funzione di ripartizione empiriche relative a ciascuno dei due campioni. La statistica di Kolmogorov-Smirnov è data da

$$D = \sup_x |\widehat{F}_1(x) - \widehat{F}_2(x)| \, . \qquad \triangleleft$$

La statistica di Kolmogorov-Smirnov è una misura della discrepanza fra le due funzione di ripartizione empiriche. Si può dimostrare in modo del tutto analogo a quanto fatto nel Teorema 12.3.3 per la statistica di Kolmogorov, che la statistica di Kolmogorov-Smirnov è "distribution-free" su $\mathcal{C}_F$.

Se $(V_{(1)}, \ldots, V_{(n)})$ è la statistica ordinata relativa al campione misto $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$, allora la statistica D può essere espressa come

$$D = \max_{1 \le i \le n} |\widehat{F}_1(V_{(i)}) - \widehat{F}_2(V_{(i)})| \, .$$

Il Teorema di Glivenko-Cantelli (vedi Serfling, 1980) implica che $D \xrightarrow{p} 0$ per $n_1, n_2 \to \infty$. La distribuzione di D può essere determinata in modo semplice per $n_1 = n_2 = n$ mediante alcune proprietà delle passeggiate aleatorie.

Teorema 12.5.2. *Siano $(X_1,\ldots,X_n)$ e $(Y_1,\ldots,Y_n)$ due campioni casuali indipendenti, tali che il campione misto abbia funzione di ripartizione congiunta $F_{2n}\in\mathcal{C}_F$. La distribuzione della statistica D di Kolmogorov-Smirnov risulta*

$$G(d)=1+2\binom{2n}{n}^{-1}\sum_{i=1}^{r}(-1)^i\binom{2n}{n+i(nd+1)},$$

dove $d=0,1/n,2/n,\ldots,1$ e $r=\lfloor n/(nd+1)\rfloor$.

Dimostrazione. Sia $(V_{(1)},\ldots,V_{(2n)})$ la statistica ordinata relativa al campione misto $(X_1,\ldots,X_n,Y_1,\ldots,Y_n)$ e sia T_i una variabile casuale tale che vale $1/n$ se $V_{(i)}$ appartiene al primo campione e vale $-1/n$ se $V_{(i)}$ appartiene al secondo campione per $i=1,\ldots,2n$. Si consideri inoltre l'ulteriore variabile casuale $S_i=T_1+\ldots+T_i$, per $i=1,\ldots,2n$. L'insieme delle variabili casuali S_i, per $i=1,\ldots,2n$, costituisce una passeggiata aleatoria in $2n$ passi i cui percorsi (i,S_i) sono tali che iniziano in $(0,0)$ e finiscono in $(2n,0)$. Vi sono $\binom{2n}{n}$ di questi percorsi, ognuno dei quali corrisponde ad una possibile sequenza di X_i e di Y_i nel campione misto ordinato. Dal momento che i due campioni provengono dalla stessa distribuzione ogni percorso è ugualmente probabile. Si deve osservare che dalla definizione di D si ha

$$D=\max(|\max_{1\le i\le 2n}S_i|,|\min_{1\le i\le 2n}S_i|)\,,$$

per cui la determinazione di $G(d)$ è equivalente alla determinazione del numero di percorsi compresi fra la retta L^+, data da $s=d+1/n$, e la retta L^-, data da $s=-d-1/n$. Si ottenga innanzitutto il numero di percorsi che intersecano la retta L^+. Esiste un punto A in cui il percorso interseca per la prima volta la retta L^+. Per la parte di percorso da A a $(2n,0)$, esiste un percorso simmetrico rispetto alla retta L^+, da A a $(2n,2d+2/n)$. Quindi il numero di percorsi da $(0,0)$ che intersecano la retta L^+ in A e finiscono in $(2n,0)$ è uguale al numero di percorsi da $(0,0)$ che intersecano la retta L^+ in A e procedono specularmente rispetto ai percorsi originari fino a $(2n,2d+2/n)$. Questo numero è dato da $\binom{2n}{n+nd+1}$, ovvero il numero in cui $(n+nd+1)$ salti positivi e $(n-nd-1)$ salti negativi possono essere permutati. Si denoti con P_0^+ l'insieme di percorsi che intersecano L^+ e con P_0^- l'insieme dei percorsi che intersecano L^-. Sia inoltre $\#(E)$ il numero di percorsi in un insieme E di percorsi. Si ha dunque

$$\#(P_0^+\cup P_0^-)=\#(P_0^+)+\#(P_0^-)-\#(P_0^+\cap P_0^-)\,.$$

Per quanto affermato in precedenza si ha

$$\#(P_0^+)=\#(P_0^-)=\binom{2n}{n+nd+1}.$$

Si noti che $(P_0^+\cap P_0^-)=(P_1^+\cup P_1^-)$, dove P_1^+ è l'insieme di tutti i percorsi che contengono almeno una parte di percorso che va da L^+ a L^- e in maniera simile P_1^- è l'insieme di tutti i percorsi che contengono almeno una parte di percorso che va da L^- a L^+. Dunque, si ha

$$\#(P_0^+\cap P_0^-)=\#(P_1^+)+\#(P_1^-)-\#(P_1^+\cap P_1^-)\,,$$

dove, in maniera analoga a quanto visto in precedenza, si ha

$$\#(P_1^+)=\#(P_1^-)=\binom{2n}{n+2(nd+1)}.$$

Definendo con P_i^+ l'insieme di tutti i percorsi che contengono almeno i parti di percorso che vanno da L^+ a L^- e con P_i^- l'insieme di tutti i percorsi che contengono i parti di percorso che vanno da L^- a L^+, procedendo in maniera iterativa si ha

$$\#(P_{i-1}^+\cap P_{i-1}^-)=\#(P_i^+)+\#(P_i^-)-\#(P_i^+\cap P_i^-)\,,\ i=0,1,\ldots,r-1\,,$$

e

$$\#(P_i^+) = \#(P_i^-) = \binom{2n}{n + (i+1)(nd+1)}, i = 0, 1, \ldots, r-1 ,$$

dove $r = \lfloor n/(nd+1) \rfloor$. Ovviamente, si ha $\#(P_{r-1}^+ \cap P_{r-1}^-) = 0$. Dunque, si ottiene

$$\#(P_0^+ \cup P_0^-) = \sum_{i=0}^{r-1} (-1)^i (\#(P_i^+) + \#(P_i^-)) = 2\sum_{i=1}^{r} (-1)^{i-1} \binom{2n}{n + i(nd+1)} ,$$

per cui il numero di percorsi che non intersecano mai le rette L^+ e L^- è dato da

$$\binom{2n}{n} - \#(P_0^+ \cup P_0^-) = \binom{2n}{n} - 2\sum_{i=1}^{r} (-1)^{i-1} \binom{2n}{n + i(nd+1)} .$$

Tenendo presente questo risultato, si ha

$$\begin{aligned} G(d) &= \binom{2n}{n}^{-1} \left(\binom{2n}{n} - 2\sum_{i=1}^{r} (-1)^{i-1} \binom{2n}{n + i(nd+1)} \right) \\ &= 1 + 2\binom{2n}{n}^{-1} \sum_{i=1}^{r} (-1)^i \binom{2n}{n + i(nd+1)} , \end{aligned}$$

che è quanto si voleva dimostrare. □

Per quanto riguarda infine la distribuzione per grandi campioni di $D = D_{n_1,n_2}$ si ha il seguente risultato.

Teorema 12.5.3. *Per $n, n_2 \to \infty$ si ha*

$$\lim_{n_1,n_2} \Pr(\sqrt{n_1 n_2/(n_1+n_2)} D \le d) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2} , d \ge 0 .$$

Dimostrazione. Si veda Billingsley (1968). □

12.6. Il test di Kolmogorov-Smirnov. Siano $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$ due campioni casuali indipendenti, tali che il campione misto abbia funzione di ripartizione congiunta $F_n \in \mathcal{C}_F$, dove $n = n_1 + n_2$. Il sistema di ipotesi da verificare risulta $H_0 : F_1(x) = F_2(x) = F(x), \forall x \in \mathbb{R}$, contro $H_1 : F_1(x) \ne F_2(x), \exists x \in \mathbb{R}$. Il test si basa sulla statistica

$$D = \sup_x |\widehat{F}_1(x) - \widehat{F}_2(x)| ,$$

ovvero sulla statistica di Kolmogorov-Smirnov. Se le due funzione di ripartizione empiriche si discostano molto fra di loro, allora si hanno realizzazioni elevate di D che conseguentemente portano a respingere l'ipotesi di base in favore dell'ipotesi alternativa. Se dunque $d_{n_1,n_2,\alpha}$ rappresenta il quantile di ordine α della distribuzione di D, la regione critica del test risulta

$$\mathcal{T}_1 = \{d : d \ge d_{n_1,n_2,1-\alpha}\} .$$

• **Esempio 12.6.1.** In un esperimento per la verifica delle capacità percettive sono stati bendati 24 soggetti. I soggetti hanno poi percorso un tracciato irregolare di forma approssimativamente circolare. Ad un certo punto del tracciato ad ognuno dei soggetti è stato chiesto di stimare l'angolo formato dalla loro attuale posizione rispetto alla posizione di partenza. Dodici di questi soggetti hanno percorso il tracciato in senso orario, mentre gli altri lo hanno percorso in senso anti-orario. I dati della Tavola 12.6.1 riportano gli errori commessi (in gradi) nelle stime degli angoli dai due gruppi. Si vuole verificare se i dati provengono dalla stessa distribuzione. Il procedimento per il calcolo della realizzazione campionaria di D per questi dati è riportato nella Tavola 12.6.2. Si ha dunque $d = 4/12$, per cui

$$\Pr(D \geq 4/12) > 0.20$$

e la significatività osservata risulta $\alpha_{oss} > 0.20$. Si è portati ad accettare che le due distribuzioni sono identiche, in quanto si può accettare H_0 ad ogni livello di significatività $\alpha \leq 0.20$. ◁

Tavola 12.6.1. Errori nelle stime degli angoli (in gradi).

soggetto	gruppo percorso senso orario	gruppo percorso senso anti-orario
1	50	23
2	22	0
3	16	9
4	10	7
5	8	− 4
6	14	24
7	− 18	− 17
8	− 2	− 13
9	− 3	− 32
10	− 11	− 23
11	− 12	− 47
12	− 31	− 53

Fonte: Lederman, Klatsky e Barber (1985)

Tavola 12.6.2.

soggetto	$z_{(i)}$	$\widehat{F}_1(z_{(i)})$	$\widehat{F}_2(z_{(i)})$	$\lvert\widehat{F}_1(z_{(i)}) - \widehat{F}_2(z_{(i)})\rvert$
1	− 53	0	1/12	1/12
2	− 47	0	2/12	2/12
3	− 32	0	3/12	3/12
4	− 31	1/12	3/12	2/12
5	− 23	1/12	4/12	3/12
6	− 18	2/12	4/12	2/12
7	− 17	2/12	5/12	3/12
8	− 13	2/12	6/12	4/12
9	− 12	3/12	6/12	3/12
10	− 11	4/12	6/12	2/12
11	− 4	4/12	7/12	3/12
12	− 3	5/12	7/12	2/12
13	− 2	6/12	7/12	1/12
14	0	6/12	8/12	2/12
15	7	6/12	9/12	3/12
16	8	7/12	9/12	2/12
17	9	7/12	10/12	3/12
18	10	8/12	10/12	2/12
19	14	9/12	10/12	1/12
20	16	10/12	10/12	0
21	22	11/12	10/12	1/12
22	23	11/12	11/12	0
23	24	11/12	12/12	1/12
24	50	1	1	0

Appendice

A.1. Alcune distribuzioni e relative caratteristiche. Di seguito vengono introdotte alcune variabili casuali di frequente uso, insieme ad alcune loro caratteristiche. Una generica variabile casuale assolutamente continua Z viene considerata nella sua forma standard, e le relative funzioni di ripartizione e densità vengono rispettivamente indicate con F ed f. Quando si considera la variabile casuale non standard $X = \delta Z + \lambda$, dove λ è un parametro di posizione e δ è un parametro di scala, allora la funzione di ripartizione è data da $G(x) = F((x-\lambda)/\delta)$, mentre la funzione di densità risulta $g(x) = f((x-\lambda)/\delta)/\delta$. Si ha $\mathrm{E}(X) = \mu = \delta \mathrm{E}(Z) + \lambda$ e $\mathrm{Var}(X) = \sigma^2 = \delta^2 \mathrm{Var}(Z)$. In particolare, se $\mathrm{E}(Z) = 0$ e $\mathrm{Var}(Z) = 1$, allora $\mu = \lambda$ e $\sigma = \delta$.

A.1.1. Distribuzione Uniforme. Una variabile casuale Uniforme Z ammette funzione di densità

$$f(z) = \mathbf{1}_{(0,1)}(z) \, .$$

Per questa distribuzione risulta

$$\mathrm{E}(Z) = \frac{1}{2} \, , \, \mathrm{Var}(Z) = \frac{1}{12} \, ,$$

e

$$\alpha_3 = 0 \, , \, \alpha_4 = \frac{9}{5} \, .$$

La mediana è data $z_{0.5} = 1/2$. Inoltre, si ha

$$f(0) = 1 \, ,$$

$$\int_0^1 f(z)^2 \, dz = 1 \, ,$$

$$\int_0^1 z(F(z) - 1/2) f(z)^2 \, dz = \frac{1}{12} \, ,$$

$$\int_{1/2}^1 z f(z)^2 \, dz - \int_0^{1/2} z f(z)^2 \, dz = \frac{1}{4} \, .$$

La distribuzione Uniforme non standard viene indicata con la notazione $U(\lambda, \delta)$.

A.1.2. Distribuzione Normale. Una variabile casuale Normale Z ammette funzione di densità

$$f(z) = \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2} z^2} \, .$$

La funzione di ripartizione di questa variabile casuale viene indicata con il simbolo Φ. Inoltre, il quantile di ordine α viene indicato con z_α. Per questa distribuzione risulta

$$\mathrm{E}(Z) = 0 \, , \, \mathrm{Var}(Z) = 1 \, ,$$

e

$$\alpha_3 = 0 \, , \, \alpha_4 = 3 \, .$$

La mediana è data $z_{0.5} = 0$. Inoltre, risulta

$$f(0) = \frac{1}{\sqrt{2\pi}} \, ,$$

$$\int_{-\infty}^{\infty} f(z)^2 \, dz = \frac{1}{2\sqrt{\pi}} \, ,$$

$$\int_{-\infty}^{\infty} z(\Phi(z) - 1/2) f(z)^2 \, dz = \frac{1}{4\pi\sqrt{3}} \, ,$$

$$\int_{0}^{\infty} z f(z)^2 \, dz - \int_{-\infty}^{0} z f(z)^2 \, dz = \frac{1}{2\pi} \, .$$

La distribuzione Normale non standard viene indicata con la notazione $N(\mu, \sigma^2)$.

A.1.3. Distribuzione Logistica. Una variabile casuale Logistica Z ammette funzione di densità

$$f(z) = \frac{e^{-z}}{(1 + e^{-z})^2} \, .$$

Per questa distribuzione risulta

$$\mathrm{E}(Z) = 0 \, , \, \mathrm{Var}(Z) = \frac{\pi^2}{3} \, ,$$

e

$$\alpha_3 = 0 \, , \, \alpha_4 = \frac{21}{5} \, .$$

La mediana è data $z_{0.5} = 0$. Inoltre, risulta

$$f(0) = \frac{1}{4} \, ,$$

$$\int_{-\infty}^{\infty} f(z)^2 \, dz = \frac{1}{6} \, ,$$

$$\int_{-\infty}^{\infty} z(F(z) - 1/2) f(z)^2 \, dz = \frac{1}{24} \, ,$$

$$\int_{0}^{\infty} z f(z)^2 \, dz - \int_{-\infty}^{0} z f(z)^2 \, dz = \frac{4 \log 2 - 1}{12} \, .$$

La distribuzione Logistica non standard viene indicata con $Lo(\mu, \delta)$.

A.1.4. Distribuzione Esponenziale. Una variabile casuale Esponenziale Z ammette funzione di densità

$$f(z) = e^{-z} \mathbf{1}_{(0,\infty)}(z) \, .$$

Per questa distribuzione risulta

$$\mathrm{E}(Z) = 1 \,,\, \mathrm{Var}(Z) = 1 \,,$$

e

$$\alpha_3 = 2 \,,\, \alpha_4 = 9 \,.$$

La mediana è data $z_{0.5} = \log 2$. Inoltre, risulta

$$f(\log 2) = \frac{1}{2} \,,$$

$$\int_0^\infty f(z)^2 \, dz = \frac{1}{2} \,,$$

$$\int_0^\infty z(F(z) - 1/2) f(z)^2 \, dz = \frac{1}{72} \,.$$

$$\int_{\log 2}^\infty z f(z)^2 \, dz - \int_0^{\log 2} z f(z)^2 \, dz = \frac{2 \log 2 - 1}{8} \,.$$

La distribuzione Esponenziale non standard viene indicata con la notazione $E(\lambda, \sigma)$.

A.1.5. Distribuzione di Laplace. Una variabile casuale di Laplace Z ammette funzione di densità

$$f(z) = \frac{1}{2} \, e^{-|z|} \,.$$

Per questa distribuzione risulta

$$\mathrm{E}(Z) = 0 \,,\, \mathrm{Var}(Z) = 2 \,,$$

e

$$\alpha_3 = 0 \,,\, \alpha_4 = 6 \,.$$

La mediana è data $z_{0.5} = 0$. Inoltre, risulta

$$f(0) = \frac{1}{2} \,,$$

$$\int_{-\infty}^\infty f(z)^2 \, dz = \frac{1}{4} \,,$$

$$\int_{-\infty}^\infty z(F(z) - 1/2) f(z)^2 \, dz = \frac{5}{144} \,,$$

$$\int_0^\infty z f(z)^2 \, dz - \int_{-\infty}^0 z f(z)^2 \, dz = \frac{1}{8} \,.$$

La distribuzione di Laplace non standard è denotata con $L(\mu, \delta)$.

A.1.6. Distribuzione di Cauchy. Una variabile casuale di Cauchy Z ammette funzione di densità

$$f(z) = \frac{1}{\pi(1 + z^2)} \,.$$

Questa distribuzione non possiede momenti e la mediana è data $z_{0.5} = 0$. Inoltre, risulta

$$f(0) = \frac{1}{\pi}\,,$$

$$\int_{-\infty}^{\infty} f(z)^2\,dz = \frac{1}{2\pi}\,,$$

$$\int_{-\infty}^{\infty} z(F(z) - 1/2) f(z)^2\,dz = \frac{1}{4\pi^2}\,,$$

$$\int_{0}^{\infty} z f(z)^2\,dz - \int_{-\infty}^{0} z f(z)^2\,dz = \frac{1}{\pi^2}\,.$$

La distribuzione di Cauchy non standard è denotata con $C(\lambda, \delta)$.

A.2. Alcuni risultati matematici. Il seguente teorema consente di determinare la somma di alcune potenze dei primi n interi.

Teorema A.2.1. *Si ha*

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}\,,$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}\,,$$

$$\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4}\,,$$

$$\sum_{i=1}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}\,.$$

Dimostrazione. Si veda Randles e Wolfe (1979). □

A.3. Alcuni risultati di teoria della probabilità. In questo paragrafo vengono richiamati alcuni risultati sulla convergenza di successioni di variabili casuali. Una completa trattazione di questo argomento è dato in Serfling (1980).

Teorema A.3.1. (Legge Debole dei Grandi Numeri di Khintchine) *Si consideri una successione* $(X_n)_{n\geq 1}$ *di variabili casuali indipendenti ed ugualmente distribuite con* $\mathrm{E}(X_1) = \mu < \infty$. *Se* $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ *per* $n \geq 1$, *si ha*

$$\bar{X}_n \xrightarrow{p} \mu\,.$$

Dimostrazione. Vedi Serfling (1980). □

• **Esempio A.3.1.** Sia $(X_n)_{n\geq 1}$ una successione di variabili casuali indipendenti ed ugualmente distribuite con $\mathrm{E}(X_1) = \mu < \infty$ e $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Si consideri la successione di variabili casuali $(Z_n)_{n\geq 1}$, dove $Z_n = (X_n - \mu)^2$ per $n \geq 1$. Si noti che $(Z_n)_{n\geq 1}$ è una successione di variabili casuali indipendenti ed ugualmente distribuite con $\mathrm{E}(Z_1) = \sigma^2 < \infty$. Di conseguenza, per la Legge Debole dei Grandi Numeri di Khintchine (Teorema A.3.1), si ha

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{p} \sigma^2 \,. \qquad \triangleleft$$

Teorema A.3.2. (Legge Debole dei Grandi Numeri di Markov) *Per ogni n sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = \mu_{in} < \infty$, con $i = 1, \ldots, n$. Si supponga inoltre che esista un $\delta \in (0, 1]$, per cui si ha $\mathrm{E}(|X_{in} - \mu_{in}|^{1+\delta}) < \infty$ e*

$$\lim_n \frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathrm{E}(|X_{in} - \mu_{in}|^{1+\delta}) = 0 \,.$$

Se $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_{in}$ per $n \geq 1$, si ha

$$\overline{X}_n - \mathrm{E}(\overline{X}_n) \xrightarrow{p} 0 \,.$$

Dimostrazione. Vedi Serfling (1980). □

• **Esempio A.3.2.** Sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = c/\sqrt{n}$ e $\mathrm{Var}(X_{in}) = \sigma^2$ con c costante, per $i = 1, \ldots, n$. Per $\delta = 1$, si ha $\mathrm{E}(|X_{in} - \mu_{in}|^2) = \sigma^2 < \infty$ e

$$\lim_n \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \lim_n \frac{\sigma^2}{n} = 0 \,,$$

e dalla Legge Debole dei Grandi Numeri di Markov (Teorema A.3.2) si ottiene

$$\overline{X}_n - c/\sqrt{n} \xrightarrow{p} 0 \,. \qquad \triangleleft$$

Teorema A.3.3. (Legge Forte dei Grandi Numeri di Khintchine) *Per ogni n sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = \mu < \infty$, con $i = 1, \ldots, n$. Se $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_{in}$ per $n \geq 1$, si ha*

$$\overline{X}_n \xrightarrow{qc} \mu \,.$$

Dimostrazione. Vedi Serfling (1980). □

• **Esempio A.3.3.** Sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = c/\sqrt{n}$ e $\mathrm{Var}(X_{in}) = \sigma^2 < \infty$, dove c è costante, per $i = 1, \ldots, n$. Si consideri il vettore trasformato $(Z_{1n}, \ldots, Z_{nn})$, dove $Z_{in} = (X_{in} - c/\sqrt{n})^2$ per $i = 1, \ldots, n$, che risulta ancora un vettore di variabili casuali indipendenti con $\mathrm{E}(Z_{in}) = \sigma^2 < \infty$. Di conseguenza, dalla Legge Forte dei Grandi Numeri di Khintchine (Teorema A.3.3) si ha

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n}(X_{in} - c/\sqrt{n})^2 \xrightarrow{qc} \sigma^2 \,,$$

che implica

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n}(X_{in} - c/\sqrt{n})^2 \xrightarrow{p} \sigma^2 \,. \qquad \triangleleft$$

Teorema A.3.4. (Teorema di Sverdrup) *Se $(X_n)_{n\geq 1}$ è una successione di variabili casuali e $g : \mathbb{R} \to \mathbb{R}$ è una funzione continua, allora*

$$X_n \xrightarrow{p} \theta \Rightarrow g(X_n) \xrightarrow{p} g(\theta) \,,$$

$$X_n \xrightarrow{qc} \theta \Rightarrow g(X_n) \xrightarrow{qc} g(\theta) ,$$

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X) .$$

Dimostrazione. Vedi Serfling (1980). □

• **Esempio A.3.4.** Sia $(X_n)_{n\geq 1}$ una successione di variabili casuali indipendenti ed ugualmente distribuite con $\mathrm{E}(X_1) = \mu < \infty$ e $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Sia inoltre

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 ,$$

dove $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Con la notazione dell'Esempio A.3.1, S_n^2 può essere espresso come

$$S_n^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - (\bar{X}_n - \mu)^2\right) = \frac{n}{n-1} \left(\bar{Z}_n - (\bar{X}_n - \mu)^2\right) .$$

Dalla Legge Debole dei Grandi Numeri di Khintchine (Teorema A.3.1) si ha $\bar{X}_n \xrightarrow{p} \mu$, mentre dall'Esempio A.3.1 si ha $\bar{Z}_n \xrightarrow{p} \sigma^2$. Si noti che $g(x, z) = z - (x - \mu)^2$ è una funzione continua. Tenendo dunque presente che $\lim_n n/(n-1) = 1$, per la generalizzazione al caso multivariato del Teorema di Sverdrup (Teorema A.3.4) si ha $S_n^2 \xrightarrow{p} \sigma^2$. ◁

• **Esempio A.3.5.** Sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = c/\sqrt{n}$ e $\mathrm{Var}(X_{in}) = \sigma^2 < \infty$ con c costante, per $i = 1, \ldots, n$. Sia inoltre

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{in} - \bar{X}_n)^2 ,$$

dove $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_{in}$. Con la notazione dell'Esempio A.3.3, S_n^2 può essere espresso come

$$S_n^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} (X_{in} - c/\sqrt{n})^2 - (\bar{X}_n - c/\sqrt{n})^2\right) = \frac{n}{n-1} \left(\bar{Z}_n - (\bar{X}_n - c/\sqrt{n})^2\right) .$$

Dall'Esempio A.3.2 si ha $\bar{X}_n - c/\sqrt{n} \xrightarrow{p} 0$, mentre dall'Esempio A.3.3 si ha $\bar{Z}_n \xrightarrow{p} \sigma^2$. Tenendo presente che $g(x, z) = z - x^2$ è una funzione continua e che $\lim_n n/(n-1) = 1$, per la generalizzazione al caso multivariato del Teorema A.3.4 si ha infine $S_n^2 \xrightarrow{p} \sigma^2$. ◁

Teorema A.3.5. (Teorema di Slutsky) *Se $(X_n)_{n\geq 1}$ e $(Y_n)_{n\geq 1}$ sono due successioni di variabili casuali tali che $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{p} \theta$ con θ costante, allora*

$$X_n + Y_n \xrightarrow{d} X + \theta ,$$

$$Y_n X_n \xrightarrow{d} \theta X ,$$

$$X_n / Y_n \xrightarrow{d} X/\theta , \theta \neq 0 .$$

Dimostrazione. Vedi Serfling (1980). □

Teorema A.3.6. (Teorema Fondamentale del Limite Classico) *Sia $(X_n)_{n\geq 1}$ una successione di variabili casuali indipendenti ed ugualmente distribuite con $\mathrm{E}(X_1) = \mu < \infty$ e $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Se $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ per $n \geq 1$, si ha*

$$\frac{\sqrt{n}}{\sigma}\,(\bar{X}_n - \mu) \xrightarrow{d} N(0,1)\,.$$

Dimostrazione. Vedi Serfling (1980). □

• **Esempio A.3.6.** Sia $(X_n)_{n\geq 1}$ una successione di variabili casuali indipendenti ed ugualmente distribuite con $\mathrm{E}(X_1) = \mu$ e $\mathrm{Var}(X_1) = \sigma^2$. Si assuma inoltre che $\mathrm{E}(X_1^4) < \infty$ e $\mathrm{E}((X_1 - \mu)^4) = \mu_4$, da cui $\mathrm{Var}((X_1 - \mu)^2) = \gamma^2 = \mu_4 - \sigma^4$. Tenendo presente la notazione e i risultati dell'Esempio A.3.4 si ha

$$\begin{aligned}\frac{\sqrt{n}}{\gamma}\,(S_n^2 - \sigma^2) &= \frac{\sqrt{n}}{\gamma}\,(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{n-1}\,(\bar{X}_n - \mu)^2 - \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1}) \\ &= \frac{n}{n-1}\,\frac{\sqrt{n}}{\gamma}\,(\bar{Z}_n - \sigma^2) - \frac{n}{n-1}\,\frac{1}{\gamma}\,(\sqrt{n}(\bar{X}_n - \mu)^2 - \sigma^2/\sqrt{n})\,.\end{aligned}$$

Dal momento che $\mathrm{E}((\bar{X}_n - \mu)^2) = \sigma^2/n$, allora

$$\lim_n\,(\sqrt{n}\mathrm{E}((\bar{X}_n - \mu)^2) - \sigma^2/\sqrt{n}) = \lim_n 2\sigma^2/\sqrt{n} = 0\,,$$

ovvero $(\sqrt{n}(\bar{X}_n - \mu)^2 - \sigma^2/\sqrt{n})$ converge in media a 0, che implica

$$\sqrt{n}(\bar{X}_n - \mu)^2 - \sigma^2/\sqrt{n} \xrightarrow{p} 0\,.$$

Inoltre, dal Teorema Fondamentale del Limite Classico (Teorema A.3.6) si ha

$$\frac{\sqrt{n}}{\gamma}\,(\bar{Z}_n - \sigma^2) \xrightarrow{d} N(0,1)\,.$$

Quindi per il Teorema di Slutsky (Teorema A.3.5) si può concludere che

$$\frac{\sqrt{n}}{\gamma}\,(S_n^2 - \sigma^2) \xrightarrow{d} N(0,1)\,. \qquad \triangleleft$$

Teorema A.3.7. (Teorema Fondamentale del Limite di Lindberg) *Per ogni n sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = \mu_{in} < \infty$ e $\mathrm{Var}(X_{in}) = \sigma_{in}^2 < \infty$ per $i = 1, \ldots, n$, e sia inoltre $\mu_n = \sum_{i=1}^n \mu_{in}$ e $\sigma_n^2 = \sum_{i=1}^n \sigma_{in}^2$. Se per ogni $\epsilon > 0$ si ha*

$$\lim_n \frac{1}{\sigma_n^2}\sum_{i=1}^{n}\mathrm{E}((X_{in} - \mu_{in})^2 \mathbf{1}_{(-\infty,\mu_{in}-\epsilon\sigma_n]\cup[\mu_{in}+\epsilon\sigma_n,\infty)}(X_{in})) = 0\,,$$

e se $\bar{X}_n = n^{-1}\sum_{i=1}^n X_{in}$, si ha

$$\frac{1}{\sigma_n}\,(n\bar{X}_n - \mu_n) \xrightarrow{d} N(0,1)\,.$$

Dimostrazione. Vedi Serfling (1980). □

Corollario A.3.8. *Per ogni n sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = \mu_n < \infty$ e $\mathrm{Var}(X_{in}) = \sigma_n^2 < \infty$ per $i = 1, \ldots, n$. Se $\bar{X}_n = n^{-1}\sum_{i=1}^n X_{in}$, si ha*

$$\frac{\sqrt{n}}{\sigma_n}\,(\bar{X}_n - \mu_n) \xrightarrow{d} N(0,1)\,.$$

Dimostrazione. Vedi Serfling (1980). □

Teorema A.3.9. (Teorema Fondamentale del Limite di Lyapunov) *Per ogni n sia $(X_{1n}, \ldots, X_{nn})$ un vettore di variabili casuali indipendenti con $\mathrm{E}(X_{in}) = \mu_{in} < \infty$ e $\mathrm{Var}(X_{in}) = \sigma_{in}^2 < \infty$ per $i = 1, \ldots, n$, e sia inoltre $\mu_n = \sum_{i=1}^n \mu_{in}$ e $\sigma_n^2 = \sum_{i=1}^n \sigma_{in}^2$. Se esiste un $\delta > 0$ tale che*

$$\lim_n \frac{1}{\sigma_n^{2+\delta}} \sum_{i=1}^n \mathrm{E}(|X_{in} - \mu_{in}|^{2+\delta}) = 0 \,,$$

e se $\bar{X}_n = n^{-1} \sum_{i=1}^n X_{in}$ si ha

$$\frac{1}{\sigma_n} (n\bar{X}_n - \mu_n) \xrightarrow{d} N(0,1) \,.$$

Dimostrazione. Vedi Serfling (1980). □

Teorema A.3.10. (Metodo Delta) *Se $(X_n)_{n\geq 1}$ è una successione di variabili casuali tale che $\sqrt{n}(X_n - \mu)/\sigma \xrightarrow{d} N(0,1)$ e se $g : \mathbb{R} \to \mathbb{R}$ è una funzione continua tale che $g'(\mu)$ esiste e $g'(\mu) \neq 0$, allora*

$$\sqrt{n}\, \frac{g(X_n) - g(\mu)}{\sigma g'(\mu)} \xrightarrow{d} N(0,1) \,.$$

Analogamente, se $(\boldsymbol{X}_n)_{n\geq 1}$ è un vettore di variabili casuali tale che $\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu}) \xrightarrow{d} N_k(\mathbf{0}, \boldsymbol{\Sigma})$ e se $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^l$ è un vettore di funzioni continue tale che $\mathbf{D} = \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\boldsymbol{x}=\boldsymbol{\mu}}$ esiste e $\mathbf{D} \neq \mathbf{0}$, si ha

$$\sqrt{n}(\mathbf{g}(\boldsymbol{X}_n) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{d} N_l(\mathbf{0}, \mathbf{D}^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{D}) \,.$$

Dimostrazione. Vedi Serfling (1980). □

• **Esempio A.3.7.** Sia $(X_n)_{n\geq 1}$ una successione di variabili casuali indipendenti ed ugualmente distribuite per cui si ha $\mathrm{E}(X_1) = \mu$, $\mathrm{Var}(X_1) = \sigma^2$ e $\mathrm{E}((X_1 - \mu)^4) = \mu_4 < \infty$. Nell'Esempio A.3.3 è stato dimostrato che $\sqrt{n}(S_n^2 - \sigma^2)/\gamma \xrightarrow{d} N(0,1)$. Se si considera la trasformata $S_n = g(S_n^2) = \sqrt{S_n^2}$ si ha $g'(\sigma^2) = 1/(2\sqrt{\sigma^2})$, per cui applicando il Metodo Delta si ha

$$\sqrt{n}\, \frac{S_n - \sigma}{\gamma/(2\sigma)} \xrightarrow{d} N(0,1) \,. \qquad \triangleleft$$

Teorema A.3.11. (Teorema di Cochran) *Sia $(\boldsymbol{X}_n)_{n\geq 1}$ una successione di vettori casuali tale che $\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu}) \xrightarrow{d} N_k(\mathbf{0}, \boldsymbol{\Sigma})$. Sia inoltre $\mathbf{C}$ una matrice quadrata simmetrica di ordine k. Si ha*

$$n(\boldsymbol{X}_n - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{C} (\boldsymbol{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \chi_q^2$$

se e solo se $\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}$, dove $\mathrm{tr}(\mathbf{C}\boldsymbol{\Sigma}) = q$.

Dimostrazione. Vedi Serfling (1980). □

Teorema A.3.12. *Sia $(F_n)_{n\geq 1}$ una successione di funzioni di ripartizione di variabili casuali assolutamente continue che converge uniformemente alla funzione di ripartizione G. Se $(c_n)_{n\geq 1}$ è una successione in $\mathbb{R}$, allora*

$$\lim_n F_n(c_n) = \alpha \,,\ 0 < \alpha < 1 \,,$$

se e solo se $\lim_n c_n = x_\alpha$, dove x_α è tale che $G(x_\alpha) = \alpha$.

Dimostrazione. Vedi Serfling (1980). □

Tavole

Tavola 1. Gli elementi della tavola danno le probabilità di coda sinistra di una variabile casuale con distribuzione Normale $N(0, 1)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995

Tavola 2. Gli elementi della tavola danno i quantili di una variabile casuale con distribuzione Chi-quadrato χ^2_n per alcune probabilità di coda sinistra P e gradi di libertà $n = 1, 2, \ldots, 30$.

$n \setminus P$	0.05	0.10	0.50	0.75	0.90	0.95	0.975	0.99	.995	.999
1	0.004	0.016	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.10	0.21	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	0.35	0.58	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.71	1.06	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	1.15	1.61	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	1.64	2.20	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	2.17	2.83	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	2.73	3.49	7.34	10.22	12.36	15.51	17.54	20.09	21.96	26.12
9	3.33	4.17	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	3.94	4.87	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	4.57	5.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	5.23	6.30	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	5.89	7.04	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	6.57	7.79	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	7.26	8.55	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	7.96	9.31	15.34	19.37	23.54	26.30	28.84	32.00	34.27	39.25
17	8.67	10.09	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	9.39	10.86	17.34	21.60	25.99	28.87	31.53	34.81	37.16	43.31
19	10.12	11.65	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	10.85	12.44	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	11.59	13.24	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	12.34	14.04	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	13.09	14.85	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	13.85	15.66	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	14.61	16.47	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	15.38	17.29	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	16.15	18.11	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	16.93	18.94	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	17.71	19.77	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	18.49	20.60	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70

Per $n > 30$ le probabilità di coda sinistra possono essere ottenute tenendo presente che $Z = \sqrt{2\chi^2_n} - \sqrt{2n-1}$ converge in probabilità ad una variabile casuale con distribuzione Normale $N(0,1)$.

Tavola 3. Gli elementi della tavola danno le probabilità P di coda sinistra o destra della statistica B del test dei segni (a secondo che nella tavola b sia a sinistra o a destra di P), $n = 2, 3, \ldots, 20$.

n	b	P	b	n	b	P	b	n	b	P	b	n	b	P	b
2	0	.2500	2	10	2	.0547	8	14	6	.3953	8	18	2	.0007	16
	1	.7500	1		3	.1719	7		7	.6047	7		3	.0038	15
3	0	.1250	3		4	.3770	6	15	0	.0000	15		4	.0154	14
	1	.5000	2		5	.6230	5		1	.0005	14		5	.0481	13
4	0	.0625	4	11	0	.0005	11		2	.0037	13		6	.1189	12
	1	.3125	3		1	.0059	10		3	.0176	12		7	.2403	11
	2	.6875	2		2	.0327	9		4	.0592	11		8	.4073	10
5	0	.0313	5		3	.1133	8		5	.1509	10		9	.5927	9
	1	.1875	4		4	.2744	7		6	.3036	9	19	0	.0000	19
	2	.5000	3		5	.5000	6		7	.5000	8		1	.0000	18
6	0	.0156	6	12	0	.0002	12	16	0	.0000	16		2	.0004	17
	1	.1094	5		1	.0032	11		1	.0003	15		3	.0022	16
	2	.3438	4		2	.0193	10		2	.0021	14		4	.0096	15
	3	.6563	3		3	.0730	9		3	.0106	13		5	.0318	14
7	0	.0078	7		4	.1938	8		4	.0384	12		6	.0835	13
	1	.0625	6		5	.3872	7		5	.1051	11		7	.1796	12
	2	.2266	5		6	.6128	6		6	.2272	10		8	.3238	11
	3	.5000	4	13	0	.0001	13		7	.4018	9		9	.5000	10
8	0	.0039	8		1	.0017	12		8	.5982	8	20	0	.0000	20
	1	.0352	7		2	.0112	11	17	0	.0000	17		1	.0000	19
	2	.1445	6		3	.0461	10		1	.0001	16		2	.0002	18
	3	.3633	5		4	.1334	9		2	.0012	15		3	.0013	17
	4	.6367	4		5	.2905	8		3	.0064	14		4	.0059	16
9	0	.0020	9		6	.5000	7		4	.0245	13		5	.0207	15
	1	.0195	8	14	0	.0001	14		5	.0717	12		6	.0577	14
	2	.0898	7		1	.0009	13		6	.1662	11		7	.1316	13
	3	.2539	6		2	.0065	12		7	.3145	10		8	.2517	12
	4	.5000	5		3	.0287	11		8	.5000	9		9	.4119	11
10	0	.0010	10		4	.0898	10	18	0	.0000	18		10	.5881	10
	1	.0107	9		5	.2120	9		1	.0001	17				

Tavola 4. Gli elementi della tavola danno le probabilità P di coda sinistra o destra della statistica W^+ di Wilcoxon (a secondo che nella tavola w^+ sia a sinistra o a destra di P), $n = 2, 3, \ldots, 15$.

n	w^+	P	w^+	n	w^+	P	w^+	n	w^+	P	w^+	n	w^+	P	w^+	n	w^+	P	w^+
2	0	.2500	3	7	4	.0547	24	9	5	.0195	40	10	17	.1611	38	11	24	.2324	42
	1	.5000	2		5	.0781	23		6	.0273	39		18	.1875	37		25	.2598	41
3	0	.1250	6		6	.1094	22		7	.0371	38		19	.2158	36		26	.2886	40
	1	.2500	5		7	.1484	21		8	.0488	37		20	.2461	35		27	.3188	39
	2	.3750	4		8	.1875	20		9	.0645	36		21	.2783	34		28	.3501	38
	3	.6250	3		9	.2344	19		10	.0820	35		22	.3125	33		29	.3823	37
4	0	.0625	10		10	.2891	18		11	.1016	34		23	.3477	32		30	.4155	36
	1	.1250	9		11	.3438	17		12	.1250	33		24	.3848	31		31	.4492	35
	2	.1875	8		12	.4063	16		13	.1504	32		25	.4229	30		32	.4829	34
	3	.3125	7		13	.4688	15		14	.1797	31		26	.4609	29		33	.5171	33
	4	.4375	6		14	.5313	14		15	.2129	30		27	.5000	28	12	0	.0002	78
	5	.5625	5	8	0	.0039	36		16	.2480	29	11	0	.0005	66		1	.0005	77
5	0	.0313	15		1	.0078	35		17	.2852	28		1	.0010	65		2	.0007	76
	1	.0625	14		2	.0117	34		18	.3262	27		2	.0015	64		3	.0012	75
	2	.0938	13		3	.0195	33		19	.3672	26		3	.0024	63		4	.0017	74
	3	.1563	12		4	.0273	32		20	.4102	25		4	.0034	62		5	.0024	73
	4	.2188	11		5	.0391	31		21	.4551	24		5	.0049	61		6	.0034	72
	5	.3125	10		6	.0547	30		22	.5000	23		6	.0068	60		7	.0046	71
	6	.4063	9		7	.0742	29	10	0	.0010	55		7	.0093	59		8	.0061	70
	7	.5000	8		8	.0977	28		1	.0020	54		8	.0122	58		9	.0081	69
6	0	.0156	21		9	.1250	27		2	.0029	53		9	.0161	57		10	.0105	68
	1	.0313	20		10	.1563	26		3	.0049	52		10	.0210	56		11	.0134	67
	2	.0469	19		11	.1914	25		4	.0068	51		11	.0269	55		12	.0171	66
	3	.0781	18		12	.2305	24		5	.0098	50		12	.0337	54		13	.0212	65
	4	.1094	17		13	.2734	23		6	.0137	49		13	.0415	53		14	.0261	64
	5	.1563	16		14	.3203	22		7	.0186	48		14	.0508	52		15	.0320	63
	6	.2188	15		15	.3711	21		8	.0244	47		15	.0615	51		16	.0386	62
	7	.2813	14		16	.4219	20		9	.0322	46		16	.0737	50		17	.0461	61
	8	.3438	13		17	.4727	19		10	.0420	45		17	.0874	49		18	.0549	60
	9	.4219	12		18	.5273	18		11	.0527	44		18	.1030	48		19	.0647	59
	10	.5000	11	9	0	.0020	45		12	.0654	43		19	.1201	47		20	.0757	58
7	0	.0078	28		1	.0039	44		13	.0801	42		20	.1392	46		21	.0881	57
	1	.0156	27		2	.0059	43		14	.0967	41		21	.1602	45		22	.1018	56
	2	.0234	26		3	.0098	42		15	.1162	40		22	.1826	44		23	.1167	55
	3	.0391	25		4	.0137	41		16	.1377	39		23	.2065	43		24	.1331	54

Tavola 4. (segue)

n	w^+	P	w^+	n	w^+	P	w^+	n	w^+	P	w^+	n	w^+	P	w^+	n	w^+	P	w^+
12	25	.1506	53	13	20	.0402	71	14	9	.0020	96	14	44	.3129	61	15	26	.0277	94
	26	.1697	52		21	.0471	70		10	.0026	95		45	.3349	60		27	.0319	93
	27	.1902	51		22	.0549	69		11	.0034	94		46	.3574	59		28	.0365	92
	28	.2119	50		23	.0636	68		12	.0043	93		47	.3804	58		29	.0416	91
	29	.2349	49		24	.0732	67		13	.0054	92		48	.4039	57		30	.0473	90
	30	.2593	48		25	.0839	66		14	.0067	91		49	.4276	56		31	.0535	89
	31	.2847	47		26	.0955	65		15	.0083	90		50	.4516	55		32	.0603	88
	32	.3110	46		27	.1082	64		16	.0101	89		51	.4758	54		33	.0677	87
	33	.3386	45		28	.1219	63		17	.0123	88		52	.5000	53		34	.0757	86
	34	.3667	44		29	.1367	62		18	.0148	87	15	0	.0000	120		35	.0844	85
	35	.3955	43		30	.1527	61		19	.0176	86		1	.0001	119		36	.0938	84
	36	.4250	42		31	.1698	60		20	.0209	85		2	.0001	118		37	.1039	83
	37	.4548	41		32	.1879	59		21	.0247	84		3	.0002	117		38	.1147	82
	38	.4849	40		33	.2072	58		22	.0290	83		4	.0002	116		39	.1262	81
	39	.5151	39		34	.2274	57		23	.0338	82		5	.0003	115		40	.1384	80
13	0	.0001	91		35	.2487	56		24	.0392	81		6	.0004	114		41	.1514	79
	1	.0002	90		36	.2709	55		25	.0453	80		7	.0006	113		42	.1651	78
	2	.0004	89		37	.2939	54		26	.0520	79		8	.0008	112		43	.1796	77
	3	.0006	88		38	.3177	53		27	.0594	78		9	.0010	111		44	.1947	76
	4	.0009	87		39	.3424	52		28	.0676	77		10	.0013	110		45	.2106	75
	5	.0012	86		40	.3677	51		29	.0765	76		11	.0017	109		46	.2271	74
	6	.0017	85		41	.3934	50		30	.0863	75		12	.0021	108		47	.2444	73
	7	.0023	84		42	.4197	49		31	.0969	74		13	.0027	107		48	.2622	72
	8	.0031	83		43	.4463	48		32	.1083	73		14	.0034	106		49	.2807	71
	9	.0040	82		44	.4730	47		33	.1206	72		15	.0042	105		50	.2997	70
	10	.0052	81		45	.5000	46		34	.1338	71		16	.0051	104		51	.3193	69
	11	.0067	80	14	0	.0001	105		35	.1479	70		17	.0062	103		52	.3394	68
	12	.0085	79		1	.0001	104		36	.1629	69		18	.0075	102		53	.3599	67
	13	.0107	78		2	.0002	103		37	.1788	68		19	.0090	101		54	.3808	66
	14	.0133	77		3	.0003	102		38	.1955	67		20	.0108	100		55	.4020	65
	15	.0164	76		4	.0004	101		39	.2131	66		21	.0128	99		56	.4235	64
	16	.0199	75		5	.0006	100		40	.2316	65		22	.0151	98		57	.4452	63
	17	.0239	74		6	.0009	99		41	.2508	64		23	.0177	97		58	.4670	62
	18	.0287	73		7	.0012	98		42	.2708	63		24	.0206	96		59	.4890	61
	19	.0341	72		8	.0015	97		43	.2915	62		25	.0240	95		60	.5110	60

Tavola 5. Gli elementi della tavola danno le probabilità P di coda sinistra o destra della statistica W di Mann-Whitney-Wilcoxon (a secondo che nella tavola w sia a sinistra o a destra di P), $n_1 \leq n_2 = 1, 2, \ldots, 10$.

n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w
1	1	1	.5000	2	1	10	2	.1818	10	2	7	3	.0278	17	2	10	6	.0909	20	3	6	8	.0476	22
1	2	1	.3333	3			3	.2727	9			4	.0556	16			7	.1364	19			9	.0833	21
		2	.6667	2			4	.3636	8			5	.1111	15			8	.1818	18			10	.1310	20
1	3	1	.2500	4			5	.4545	7			6	.1667	14			9	.2424	17			11	.1905	19
		2	.5000	3			6	.5455	6			7	.2500	13			10	.3030	16			12	.2738	18
1	4	1	.2000	5	2	2	3	.1667	7			8	.3333	12			11	.3788	15			13	.3571	17
		2	.4000	4			4	.3333	6			9	.4444	11			12	.4545	14			14	.4524	16
		3	.6000	3			5	.6667	5			10	.5556	10			13	.5455	13			15	.5476	15
1	5	1	.1667	6	2	3	3	.1000	9	2	8	3	.0222	19	3	3	6	.0500	15	3	7	6	.0083	27
		2	.3333	5			4	.2000	8			4	.0444	18			7	.1000	14			7	.0167	26
		3	.5000	4			5	.4000	7			5	.0889	17			8	.2000	13			8	.0333	25
1	6	1	.1429	7			6	.6000	6			6	.1333	16			9	.3500	12			9	.0583	24
		2	.2857	6	2	4	3	.0667	11			7	.2000	15			10	.5000	11			10	.0917	23
		3	.4286	5			4	.1333	10			8	.2667	14	3	4	6	.0286	18			11	.1333	22
		4	.5714	4			5	.2667	9			9	.3556	13			7	.0571	17			12	.1917	21
1	7	1	.1250	8			6	.4000	8			10	.4444	12			8	.1143	16			13	.2583	20
		2	.2500	7			7	.6000	7			11	.5556	11			9	.2000	15			14	.3333	19
		3	.3750	6	2	5	3	.0476	13	2	9	3	.0182	21			10	.3143	14			15	.4167	18
		4	.5000	5			4	.0952	12			4	.0364	20			11	.4286	13			16	.5000	17
1	8	1	.1111	9			5	.1905	11			5	.0727	19			12	.5714	12	3	8	6	.0061	30
		2	.2222	8			6	.2857	10			6	.1091	18	3	5	6	.0179	21			7	.0121	29
		3	.3333	7			7	.4286	9			7	.1636	17			7	.0357	20			8	.0242	28
		4	.4444	6			8	.5714	8			8	.2182	16			8	.0714	19			9	.0424	27
		5	.5556	5	2	6	3	.0357	15			9	.2909	15			9	.1250	18			10	.0667	26
1	9	1	.1000	10			4	.0714	14			10	.3636	14			10	.1964	17			11	.0970	25
		2	.2000	9			5	.1429	13			11	.4545	13			11	.2857	16			12	.1394	24
		3	.3000	8			6	.2143	12			12	.5455	12			12	.3929	15			13	.1879	23
		4	.4000	7			7	.3214	11	2	10	3	.0152	23			13	.5000	14			14	.2485	22
		5	.5000	6			8	.4286	10			4	.0303	22	3	6	6	.0119	24			15	.3152	21
1	10	1	.0909	11			9	.5714	9			5	.0606	21			7	.0238	23			16	.3879	20

Tavola 5. (segue)

n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w
3	8	17	.4606	19	3	10	20	.4685	22	4	6	18	.2381	26	4	8	20	.1838	32	4	10	14	.0120	46
		18	.5394	18			21	.5315	21			19	.3048	25			21	.2303	31			15	.0180	45
3	9	6	.0045	33	4	4	10	.0143	26			20	.3810	24			22	.2848	30			16	.0270	44
		7	.0091	32			11	.0286	25			21	.4571	23			23	.3414	29			17	.0380	43
		8	.0182	31			12	.0571	24			22	.5429	22			24	.4040	28			18	.0529	42
		9	.0318	30			13	.1000	23	4	7	10	.0030	38			25	.4667	27			19	.0709	41
		10	.0500	29			14	.1714	22			11	.0061	37			26	.5333	26			20	.0939	40
		11	.0727	28			15	.2429	21			12	.0121	36	4	9	10	.0014	46			21	.1199	39
		12	.1045	27			16	.3429	20			13	.0212	35			11	.0028	45			22	.1518	38
		13	.1409	26			17	.4429	19			14	.0364	34			12	.0056	44			23	.1868	37
		14	.1864	25			18	.5571	18			15	.0545	33			13	.0098	43			24	.2268	36
		15	.2409	24	4	5	10	.0079	30			16	.0818	32			14	.0168	42			25	.2697	35
		16	.3000	23			11	.0159	29			17	.1152	31			15	.0252	41			26	.3177	34
		17	.3636	22			12	.0317	28			18	.1576	30			16	.0378	40			27	.3666	33
		18	.4318	21			13	.0556	27			19	.2061	29			17	.0531	39			28	.4196	32
		19	.5000	20			14	.0952	26			20	.2636	28			18	.0741	38			29	.4725	31
3	10	6	.0035	36			15	.1429	25			21	.3242	27			19	.0993	37			30	.5275	30
		7	.0070	35			16	.2063	24			22	.3939	26			20	.1301	36	5	5	15	.0040	40
		8	.0140	34			17	.2778	23			23	.4636	25			21	.1650	35			16	.0079	39
		9	.0245	33			18	.3651	22			24	.5364	24			22	.2070	34			17	.0159	38
		10	.0385	32			19	.4524	21	4	8	10	.0020	42			23	.2517	33			18	.0278	37
		11	.0559	31			20	.5476	20			11	.0040	41			24	.3021	32			19	.0476	36
		12	.0804	30	4	6	10	.0048	34			12	.0081	40			25	.3552	31			20	.0754	35
		13	.1084	29			11	.0095	33			13	.0141	39			26	.4126	30			21	.1111	34
		14	.1434	28			12	.0190	32			14	.0242	38			27	.4699	29			22	.1548	33
		15	.1853	27			13	.0333	31			15	.0364	37			28	.5301	28			23	.2103	32
		16	.2343	26			14	.0571	30			16	.0545	36	4	10	10	.0010	50			24	.2738	31
		17	.2867	25			15	.0857	29			17	.0768	35			11	.0020	49			25	.3452	30
		18	.3462	24			16	.1286	28			18	.1071	34			12	.0040	48			26	.4206	29
		19	.4056	23			17	.1762	27			19	.1414	33			13	.0070	47			27	.5000	28

Tavola 5. (segue)

n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w
5	6	15	.0022	45	5	7	29	.3194	36	5	9	20	.0095	55	5	10	27	.0646	53	6	6	37	.4091	41
		16	.0043	44			30	.3775	35			21	.0145	54			28	.0823	52			38	.4686	40
		17	.0087	43			31	.4381	34			22	.0210	53			29	.1032	51			39	.5314	39
		18	.0152	42			32	.5000	33			23	.0300	52			30	.1272	50	6	7	21	.0006	63
		19	.0260	41	5	8	15	.0008	55			24	.0415	51			31	.1548	49			22	.0012	62
		20	.0411	40			16	.0016	54			25	.0559	50			32	.1855	48			23	.0023	61
		21	.0628	39			17	.0031	53			26	.0734	49			33	.2198	47			24	.0041	60
		22	.0887	38			18	.0054	52			27	.0949	48			34	.2567	46			25	.0070	59
		23	.1234	37			19	.0093	51			28	.1199	47			35	.2970	45			26	.0111	58
		24	.1645	36			20	.0148	50			29	.1489	46			36	.3393	44			27	.0175	57
		25	.2143	35			21	.0225	49			30	.1818	45			37	.3839	43			28	.0256	56
		26	.2684	34			22	.0326	48			31	.2188	44			38	.4296	42			29	.0367	55
		27	.3312	33			23	.0466	47			32	.2592	43			39	.4765	41			30	.0507	54
		28	.3961	32			24	.0637	46			33	.3032	42			40	.5235	40			31	.0688	53
		29	.4654	31			25	.0855	45			34	.3497	41	6	6	21	.0011	57			32	.0903	52
		30	.5346	30			26	.1111	44			35	.3986	40			22	.0022	56			33	.1171	51
5	7	15	.0013	50			27	.1422	43			36	.4491	39			23	.0043	55			34	.1474	50
		16	.0025	49			28	.1772	42			37	.5000	38			24	.0076	54			35	.1830	49
		17	.0051	48			29	.2176	41	5	10	15	.0003	65			25	.0130	53			36	.2226	48
		18	.0088	47			30	.2618	40			16	.0007	64			26	.0206	52			37	.2669	47
		19	.0152	46			31	.3108	39			17	.0013	63			27	.0325	51			38	.3141	46
		20	.0240	45			32	.3621	38			18	.0023	62			28	.0465	50			39	.3654	45
		21	.0366	44			33	.4165	37			19	.0040	61			29	.0660	49			40	.4178	44
		22	.0530	43			34	.4716	36			20	.0063	60			30	.0898	48			41	.4726	43
		23	.0745	42			35	.5284	35			21	.0097	59			31	.1201	47			42	.5274	42
		24	.1010	41	5	9	15	.0005	60			22	.0140	58			32	.1548	46	6	8	21	.0003	69
		25	.1338	40			16	.0010	59			23	.0200	57			33	.1970	45			22	.0007	68
		26	.1717	39			17	.0020	58			24	.0276	56			34	.2424	44			23	.0013	67
		27	.2159	38			18	.0035	57			25	.0376	55			35	.2944	43			24	.0023	66
		28	.2652	37			19	.0060	56			26	.0496	54			36	.3496	42			25	.0040	65

Tavola 5. (segue)

n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w
6	8	26	.0063	64	6	9	31	.0248	65	6	10	33	.0280	69	7	7	39	.0487	66	7	8	44	.0946	68
		27	.0100	63			32	.0332	64			34	.0363	68			40	.0641	65			45	.1159	67
		28	.0147	62			33	.0440	63			35	.0467	67			41	.0825	64			46	.1405	66
		29	.0213	61			34	.0567	62			36	.0589	66			42	.1043	63			47	.1678	65
		30	.0296	60			35	.0723	61			37	.0736	65			43	.1297	62			48	.1984	64
		31	.0406	59			36	.0905	60			38	.0903	64			44	.1588	61			49	.2317	63
		32	.0539	58			37	.1119	59			39	.1099	63			45	.1914	60			50	.2679	62
		33	.0709	57			38	.1361	58			40	.1317	62			46	.2279	59			51	.3063	61
		34	.0906	56			39	.1638	57			41	.1566	61			47	.2675	58			52	.3472	60
		35	.1142	55			40	.1942	56			42	.1838	60			48	.3100	57			53	.3894	59
		36	.1412	54			41	.2280	55			43	.2139	59			49	.3552	56			54	.4333	58
		37	.1725	53			42	.2643	54			44	.2461	58			50	.4024	55			55	.4775	57
		38	.2068	52			43	.3035	53			45	.2811	57			51	.4508	54			56	.5225	56
		39	.2454	51			44	.3445	52			46	.3177	56			52	.5000	53	7	9	28	.0001	91
		40	.2864	50			45	.3878	51			47	.3564	55	7	8	28	.0002	84			29	.0002	90
		41	.3310	49			46	.4320	50			48	.3962	54			29	.0003	83			30	.0003	89
		42	.3773	48			47	.4773	49			49	.4374	53			30	.0006	82			31	.0006	88
		43	.4259	47			48	.5227	48			50	.4789	52			31	.0011	81			32	.0010	87
		44	.4749	46	6	10	21	.0001	81			51	.5211	51			32	.0019	80			33	.0017	86
		45	.5251	45			22	.0002	80	7	7	28	.0003	77			33	.0030	79			34	.0026	85
6	9	21	.0002	75			23	.0005	79			29	.0006	76			34	.0047	78			35	.0039	84
		22	.0004	74			24	.0009	78			30	.0012	75			35	.0070	77			36	.0058	83
		23	.0008	73			25	.0015	77			31	.0020	74			36	.0103	76			37	.0082	82
		24	.0014	72			26	.0024	76			32	.0035	73			37	.0145	75			38	.0115	81
		25	.0024	71			27	.0037	75			33	.0055	72			38	.0200	74			39	.0156	80
		26	.0038	70			28	.0055	74			34	.0087	71			39	.0270	73			40	.0209	79
		27	.0060	69			29	.0080	73			35	.0131	70			40	.0361	72			41	.0274	78
		28	.0088	68			30	.0112	72			36	.0189	69			41	.0469	71			42	.0356	77
		29	.0128	67			31	.0156	71			37	.0265	68			42	.0603	70			43	.0454	76
		30	.0180	66			32	.0210	70			38	.0364	67			43	.0760	69			44	.0571	75

Tavola 5. (segue)

n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w
7	9	45	.0708	74	7	10	43	.0277	83	8	8	45	.0074	91	8	9	42	.0012	102	8	9	72	.5187	72
		46	.0869	73			44	.0351	82			46	.0103	90			43	.0019	101	8	10	36	.0000	116
		47	.1052	72			45	.0439	81			47	.0141	89			44	.0028	100			37	.0000	115
		48	.1261	71			46	.0544	80			48	.0190	88			45	.0039	99			38	.0001	114
		49	.1496	70			47	.0665	79			49	.0249	87			46	.0056	98			39	.0002	113
		50	.1755	69			48	.0806	78			50	.0325	86			47	.0076	97			40	.0003	112
		51	.2039	68			49	.0966	77			51	.0415	85			48	.0103	96			41	.0004	111
		52	.2349	67			50	.1148	76			52	.0524	84			49	.0137	95			42	.0007	110
		53	.2680	66			51	.1349	75			53	.0652	83			50	.0180	94			43	.0010	109
		54	.3032	65			52	.1574	74			54	.0803	82			51	.0232	93			44	.0015	108
		55	.3403	64			53	.1819	73			55	.0974	81			52	.0296	92			45	.0022	107
		56	.3788	63			54	.2087	72			56	.1172	80			53	.0372	91			46	.0031	106
		57	.4185	62			55	.2374	71			57	.1393	79			54	.0464	90			47	.0043	105
		58	.4591	61			56	.2681	70			58	.1641	78			55	.0570	89			48	.0058	104
		59	.5000	60			57	.3004	69			59	.1911	77			56	.0694	88			49	.0078	103
7	10	28	.0001	98			58	.3345	68			60	.2209	76			57	.0836	87			50	.0103	102
		29	.0001	97			59	.3698	67			61	.2527	75			58	.0998	86			51	.0133	101
		30	.0002	96			60	.4063	66			62	.2869	74			59	.1179	85			52	.0171	100
		31	.0004	95			61	.4434	65			63	.3227	73			60	.1383	84			53	.0217	99
		32	.0006	94			62	.4811	64			64	.3605	72			61	.1606	83			54	.0273	98
		33	.0010	93			63	.5189	63			65	.3992	71			62	.1852	82			55	.0338	97
		34	.0015	92	8	8	36	.0001	100			66	.4392	70			63	.2117	81			56	.0416	96
		35	.0023	91			37	.0002	99			67	.4796	69			64	.2404	80			57	.0506	95
		36	.0034	90			38	.0003	98			68	.5204	68			65	.2707	79			58	.0610	94
		37	.0048	89			39	.0005	97	8	9	36	.0000	108			66	.3029	78			59	.0729	93
		38	.0068	88			40	.0009	96			37	.0001	107			67	.3365	77			60	.0864	92
		39	.0093	87			41	.0015	95			38	.0002	106			68	.3715	76			61	.1015	91
		40	.0125	86			42	.0023	94			39	.0003	105			69	.4074	75			62	.1185	90
		41	.0165	85			43	.0035	93			40	.0005	104			70	.4442	74			63	.1371	89
		42	.0215	84			44	.0052	92			41	.0008	103			71	.4813	73			64	.1577	88

Tavola 5. (segue)

n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w	n_1	n_2	w	P	w
8	10	65	.1800	87	9	9	63	.0252	108	9	10	52	.0005	128	9	10	82	.2745	98	10	10	76	.0144	134
		66	.2041	86			64	.0313	107			53	.0007	127			83	.3019	97			77	.0177	133
		67	.2299	85			65	.0385	106			54	.0011	126			84	.3304	96			78	.0216	132
		68	.2574	84			66	.0470	105			55	.0015	125			85	.3598	95			79	.0262	131
		69	.2863	83			67	.0567	104			56	.0021	124			86	.3901	94			80	.0315	130
		70	.3167	82			68	.0680	103			57	.0028	123			87	.4211	93			81	.0376	129
		71	.3482	81			69	.0807	102			58	.0038	122			88	.4524	92			82	.0446	128
		72	.3809	80			70	.0951	101			59	.0051	121			89	.4841	91			83	.0526	127
		73	.4143	79			71	.1112	100			60	.0066	120			90	.5159	90			84	.0615	126
		74	.4484	78			72	.1290	99			61	.0086	119	10	10	55	.0000	155			85	.0716	125
		75	.4827	77			73	.1487	98			62	.0110	118			56	.0000	154			86	.0827	124
		76	.5173	76			74	.1701	97			63	.0140	117			57	.0000	153			87	.0952	123
9	9	45	.0000	126			75	.1933	96			64	.0175	116			58	.0000	152			88	.1088	122
		46	.0000	125			76	.2181	95			65	.0217	115			59	.0001	151			89	.1237	121
		47	.0001	124			77	.2447	94			66	.0267	114			60	.0001	150			90	.1399	120
		48	.0001	123			78	.2729	93			67	.0326	113			61	.0002	149			91	.1575	119
		49	.0002	122			79	.3024	92			68	.0394	112			62	.0002	148			92	.1763	118
		50	.0004	121			80	.3332	91			69	.0474	111			63	.0004	147			93	.1965	117
		51	.0006	120			81	.3652	90			70	.0564	110			64	.0005	146			94	.2179	116
		52	.0009	119			82	.3981	89			71	.0667	109			65	.0008	145			95	.2406	115
		53	.0014	118			83	.4317	88			72	.0782	108			66	.0010	144			96	.2644	114
		54	.0020	117			84	.4657	87			73	.0912	107			67	.0014	143			97	.2894	113
		55	.0028	116			85	.5000	86			74	.1055	106			68	.0019	142			98	.3153	112
		56	.0039	115	9	10	45	.0000	135			75	.1214	105			69	.0026	141			99	.3421	111
		57	.0053	114			46	.0000	134			76	.1388	104			70	.0034	140			100	.3697	110
		58	.0071	113			47	.0000	133			77	.1577	103			71	.0045	139			101	.3980	109
		59	.0094	112			48	.0001	132			78	.1781	102			72	.0057	138			102	.4267	108
		60	.0122	111			49	.0001	131			79	.2001	101			73	.0073	137			103	.4559	107
		61	.0157	110			50	.0002	130			80	.2235	100			74	.0093	136			104	.4853	106
		62	.0200	109			51	.0003	129			81	.2483	99			75	.0116	135			105	.5147	105

Tavola 6. Gli elementi della tavola danno le probabilità P di coda destra della statistica A di Ansari-Bradley, $n_1 \leq n_2 = 2, 3, \ldots, 10$.

n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P
2	2	2	1.0000	2	7	7	.3056	2	10	10	.1515	3	6	11	.1429	3	9	7	.9273
		3	.8333			8	.1389			11	.0758			12	.0595			8	.8636
		4	.1667			9	.0556			12	.0152			13	.0119			9	.7636
2	3	2	1.0000	2	8	2	1.0000	3	3	4	1.0000	3	7	4	1.0000			10	.6364
		3	.9000			3	.9778			5	.9000			5	.9833			11	.5000
		4	.5000			4	.8889			6	.7000			6	.9500			12	.3636
		5	.2000			5	.7778			7	.3000			7	.8667			13	.2364
2	4	2	1.0000			6	.6000			8	.1000			8	.7500			14	.1364
		3	.9333			7	.4000	3	4	4	1.0000			9	.5833			15	.0727
		4	.6667			8	.2222			5	.9429			10	.4167			16	.0273
		5	.3333			9	.1111			6	.8286			11	.2500			17	.0091
		6	.0667			10	.0222			7	.5714			12	.1333	3	10	4	1.0000
2	5	2	1.0000	2	9	2	1.0000			8	.3429			13	.0500			5	.9930
		3	.9524			3	.9818			9	.1429			14	.0167			6	.9790
		4	.7619			4	.9091			10	.0286	3	8	4	1.0000			7	.9441
		5	.5238			5	.8182	3	5	4	1.0000			5	.9879			8	.8951
		6	.2381			6	.6727			5	.9643			6	.9636			9	.8182
		7	.0952			7	.5091			6	.8929			7	.9030			10	.7168
2	6	2	1.0000			8	.3273			7	.7143			8	.8182			11	.5979
		3	.9643			9	.2000			8	.5000			9	.6909			12	.4755
		4	.8214			10	.0909			9	.2857			10	.5455			13	.3497
		5	.6429			11	.0364			10	.1071			11	.3939			14	.2413
		6	.3571	2	10	2	1.0000			11	.0357			12	.2606			15	.1503
		7	.1786			3	.9848	3	6	4	1.0000			13	.1455			16	.0839
		8	.0357			4	.9242			5	.9762			14	.0727			17	.0420
2	7	2	1.0000			5	.8485			6	.9286			15	.0303			18	.0175
		3	.9722			6	.7273			7	.8095			16	.0061			19	.0035
		4	.8611			7	.5909			8	.6548	3	9	4	1.0000	4	4	6	1.0000
		5	.7222			8	.4091			9	.4643			5	.9909			7	.9857
		6	.5000			9	.2727			10	.2857			6	.9727			8	.9286

Tavola 6. (segue)

n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P
4	4	9	.8000	4	7	6	1.0000	4	8	21	.0101	4	10	15	.6593	5	6	14	.8420
		10	.6286			7	.9970			22	.0020			16	.5554			15	.7446
		11	.3714			8	.9848	4	9	6	1.0000			17	.4446			16	.6147
		12	.2000			9	.9576			7	.9986			18	.3407			17	.4805
		13	.0714			10	.9091			8	.9930			19	.2458			18	.3463
		14	.0143			11	.8242			9	.9804			20	.1658			19	.2294
4	5	6	1.0000			12	.7152			10	.9580			21	.1039			20	.1342
		7	.9921			13	.5818			11	.9161			22	.0599			21	.0693
		8	.9603			14	.4424			12	.8573			23	.0300			22	.0303
		9	.8889			15	.3030			13	.7762			24	.0140			23	.0108
		10	.7778			16	.1939			14	.6783			25	.0050			24	.0022
		11	.6032			17	.1061			15	.5650			26	.0010	5	7	9	1.0000
		12	.4286			18	.0515			16	.4503	5	5	9	1.0000			10	.9975
		13	.2619			19	.0182			17	.3357			10	.9921			11	.9924
		14	.1349			20	.0061			18	.2378			11	.9762			12	.9773
		15	.0476	4	8	6	1.0000			19	.1538			12	.9286			13	.9495
		16	.0159			7	.9980			20	.0923			13	.8492			14	.9015
4	6	6	1.0000			8	.9899			21	.0490			14	.7302			15	.8333
		7	.9952			9	.9717			22	.0238			15	.5873			16	.7374
		8	.9762			10	.9394			23	.0084			16	.4127			17	.6237
		9	.9333			11	.8788			24	.0028			17	.2698			18	.5000
		10	.8571			12	.7980	4	10	6	1.0000			18	.1508			19	.3763
		11	.7333			13	.6889			7	.9990			19	.0714			20	.2626
		12	.5810			14	.5677			8	.9950			20	.0238			21	.1667
		13	.4190			15	.4323			9	.9860			21	.0079			22	.0985
		14	.2667			16	.3111			10	.9700	5	6	9	1.0000			23	.0505
		15	.1429			17	.2020			11	.9401			10	.9957			24	.0227
		16	.0667			18	.1212			12	.8961			11	.9870			25	.0076
		17	.0238			19	.0606			13	.8342			12	.9610			26	.0025
		18	.0048			20	.0283			14	.7542			13	.9156				

Tavola 6. (segue)

n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P
5	8	9	1.0000	5	9	17	.8212	5	10	24	.3044	6	7	12	1.0000	6	8	20	.8751
		10	.9984			18	.7423			25	.2268			13	.9994			21	.8139
		11	.9953			19	.6523			26	.1608			14	.9971			22	.7366
		12	.9860			20	.5514			27	.1086			15	.9918			23	.6474
		13	.9689			21	.4486			28	.0686			16	.9802			24	.5501
		14	.9386			22	.3477			29	.0406			17	.9592			25	.4499
		15	.8936			23	.2577			30	.0220			18	.9242			26	.3526
		16	.8275			24	.1788			31	.0107			19	.8735			27	.2634
		17	.7451			25	.1179			32	.0047			20	.8048			28	.1861
		18	.6457			26	.0709			33	.0017			21	.7203			29	.1249
		19	.5385			27	.0400			34	.0003			22	.6189			30	.0783
		20	.4266			28	.0200	6	6	12	1.0000			23	.5122			31	.0453
		21	.3209			29	.0090			13	.9989			24	.4038			32	.0240
		22	.2269			30	.0030			14	.9946			25	.3030			33	.0113
		23	.1507			31	.0010			15	.9848			26	.2133			34	.0047
		24	.0917	5	10	9	1.0000			16	.9632			27	.1410			35	.0017
		25	.0513			10	.9993			17	.9264			28	.0851			36	.0003
		26	.0249			11	.9980			18	.8658			29	.0484	6	9	12	1.0000
		27	.0109			12	.9940			19	.7846			30	.0239			13	.9998
		28	.0039			13	.9867			20	.6807			31	.0105			14	.9990
		29	.0008			14	.9734			21	.5649			32	.0035			15	.9972
5	9	9	1.0000			15	.9524			22	.4351			33	.0012			16	.9932
		10	.9990			16	.9197			23	.3193	6	8	12	1.0000			17	.9856
		11	.9970			17	.8761			24	.2154			13	.9997			18	.9724
		12	.9910			18	.8182			25	.1342			14	.9983			19	.9518
		13	.9800			19	.7483			26	.0736			15	.9953			20	.9215
		14	.9600			20	.6663			27	.0368			16	.9887			21	.8803
		15	.9291			21	.5771			28	.0152			17	.9760			22	.8260
		16	.8821			22	.4832			29	.0054			18	.9547			23	.7600
						23	.3916			30	.0011			19	.9217			24	.6829

Tavola 6. (segue)

n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P
6	9	25	.5984	6	10	27	.5425	7	7	31	.2652	7	8	36	.1005	7	9	37	.1477
		26	.5085			28	.4575			32	.1894			37	.0648			38	.1035
		27	.4190			29	.3754			33	.1770			38	.0393			39	.0694
		28	.3323			30	.2975			34	.0804			39	.0221			40	.0441
		29	.2543			31	.2283			35	.0466			40	.0115			41	.0266
		30	.1860			32	.1678			36	.0256			41	.0053			42	.0149
		31	.1303			33	.1188			37	.0122			42	.0022			43	.0079
		32	.0859			34	.0798			38	.0052			43	.0008			44	.0037
		33	.0539			35	.0513			39	.0017			44	.0002			45	.0016
		34	.0312			36	.0308			40	.0006	7	9	16	1.0000			46	.0005
		35	.0170			37	.0175	7	8	16	1.0000			17	.9998			47	.0002
		36	.0082			38	.0090			17	.9997			18	.9995	7	10	16	1.0000
		37	.0036			39	.0042			18	.9991			19	.9984			17	.9999
		38	.0012			40	.0017			19	.9972			20	.9963			18	.9997
		39	.0004			41	.0006			20	.9935			21	.9921			19	.9991
6	10	12	1.0000	7	7	16	1.0000			21	.9862			22	.9851			20	.9978
		13	.9999			17	.9994			22	.9744			23	.9734			21	.9954
		14	.9994			18	.9983			23	.9549			24	.9559			22	.9912
		15	.9983			19	.9948			24	.9270			25	.9306			23	.9841
		16	.9958			20	.9878			25	.8878			26	.8965			24	.9734
		17	.9910			21	.9744			26	.8375			27	.8523			25	.9574
		18	.9825			22	.9534			27	.7748			28	.7981			26	.9354
		19	.9692			23	.9196			28	.7021			29	.7336			27	.9059
		20	.9487			24	.8730			29	.6194			30	.6608			28	.8685
		21	.9202			25	.8106			30	.5324			31	.5820			29	.8221
		22	.8812			26	.7348			31	.4435			32	.5000			30	.7676
		23	.8322			27	.6463			32	.3577			33	.4180			31	.7052
		24	.7717			28	.5507			33	.2777			34	.3392			32	.6368
		25	.7025			29	.4493			34	.2075			35	.2664			33	.5637
		26	.6246			30	.3537			35	.1478			36	.2019			34	.4888

Tavola 6. (segue)

n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P
7	10	35	.4139	8	8	33	.7650	8	9	30	.9504	8	10	23	.9997	8	10	53	.0113
		36	.3421			34	.6970			31	.9262			24	.9992			54	.0065
		37	.2753			35	.6212			32	.8947			25	.9983			55	.0035
		38	.2154			36	.5413			33	.8549			26	.9965			56	.0017
		39	.1633			37	.4587			34	.8069			27	.9935			57	.0008
		40	.1199			38	.3788			35	.7508			28	.9887			58	.0003
		41	.0847			39	.3030			36	.6877			29	.9813			59	.0001
		42	.0576			40	.2350			37	.6184			30	.9704			60	.0000
		43	.0375			41	.1754			38	.5457			31	.9551	9	9	25	1.0000
		44	.0233			42	.1263			39	.4714			32	.9344			26	1.0000
		45	.0136			43	.0867			40	.3983			33	.9075			27	.9999
		46	.0075			44	.0572			41	.3281			34	.8738			28	.9996
		47	.0038			45	.0357			42	.2636			35	.8328			29	.9991
		48	.0017			46	.0211			43	.2055			36	.7847			30	.9981
		49	.0007			47	.0115			44	.1557			37	.7296			31	.9963
		50	.0003			48	.0059			45	.1139			38	.6686			32	.9932
		51	.0001			49	.0026			46	.0807			39	.6031			33	.9882
8	8	20	1.0000			50	.0011			47	.0548			40	.5347			34	.9805
		21	.9999			51	.0004			48	.0358			41	.4653			35	.9695
		22	.9996			52	.0001			49	.0221			42	.3969			36	.9540
		23	.9989	8	9	20	1.0000			50	.0131			43	.3314			37	.9332
		24	.9974			21	1.0000			51	.0072			44	.2704			38	.9062
		25	.9941			22	.9998			52	.0037			45	.2153			39	.8724
		26	.9885			23	.9994			53	.0017			46	.1672			40	.8313
		27	.9789			24	.9986			54	.0007			47	.1262			41	.7833
		28	.9643			25	.9969			55	.0002			48	.0925			42	.7283
		29	.9428			26	.9938			56	.0001			49	.0656			43	.6677
		30	.9133			27	.9886	8	10	20	1.0000			50	.0449			44	.6025
		31	.8737			28	.9804			21	1.0000			51	.0296			45	.5346
		32	.8246			29	.9680			22	.9999			52	.0187			46	.4654

Tavola 6. (segue)

n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P	n_1	n_2	a	P
9	9	47	.3975	9	10	36	.9741	9	10	66	.0008	10	10	55	.5296
		48	.3323			37	.9618			67	.0004			56	.4704
		49	.2717			38	.9453			68	.0002			57	.4119
		50	.2167			39	.9240			69	.0001			58	.3551
		51	.1687			40	.8972			70	.0000			59	.3014
		52	.1276			41	.8646	10	10	30	1.0000			60	.2514
		53	.0938			42	.8259			31	1.0000			61	.2060
		54	.0668			43	.7813			32	1.0000			62	.1656
		55	.0460			44	.7310			33	.9999			63	.1306
		56	.0305			45	.6759			34	.9998			64	.1007
		57	.0195			46	.6166			35	.9996			65	.0761
		58	.0118			47	.5548			36	.9992			66	.0560
		59	.0068			48	.4916			37	.9984			67	.0403
		60	.0037			49	.4287			38	.9971			68	.0282
		61	.0019			50	.3673			39	.9951			69	.0192
		62	.0009			51	.3092			40	.9920			70	.0126
		63	.0004			52	.2552			41	.9874			71	.0080
		64	.0001			53	.2064			42	.9808			72	.0049
		65	.0000			54	.1632			43	.9718			73	.0029
9	10	25	1.0000			55	.1262			44	.9597			74	.0016
		26	1.0000			56	.0952			45	.9440			75	.0008
		27	.9999			57	.0700			46	.9239			76	.0004
		28	.9998			58	.0500			47	.8993			77	.0002
		29	.9995			59	.0347			48	.8694			78	.0001
		30	.9990			60	.0232			49	.8344			79	.0000
		31	.9980			61	.0150			50	.7940			80	.0000
		32	.9964			62	.0093			51	.7486				
		33	.9937			63	.0056			52	.6986				
		34	.9894			64	.0031			53	.6449				
		35	.9831			65	.0017			54	.5881				

Tavola 7. Gli elementi della tavola danno le probabilità P di coda sinistra o destra della statistica S di Spearman (a secondo che nella tavola s sia a sinistra o a destra di P), $n = 3, 4, \ldots, 10$.

n	s	P	s	n	s	P	s	n	s	P	s	n	s	P	s
3	10	.167	14	6	69	.329	78	7	111	.482	113	8	150	.250	174
	11	.500	13		70	.357	77		112	.518	112		151	.268	173
4	20	.042	30		71	.401	76	8	120	.000	204		152	.291	172
	21	.167	29		72	.460	75		121	.000	203		153	.310	171
	22	.208	28		73	.500	74		122	.001	202		154	.332	170
	23	.375	27	7	84	.000	140		123	.001	201		155	.352	169
	24	.458	26		85	.001	139		124	.002	200		156	.376	168
	25	.542	25		86	.003	138		125	.004	199		157	.397	167
5	35	.008	55		87	.006	137		126	.005	198		158	.420	166
	36	.042	54		88	.012	136		127	.008	197		159	.441	165
	37	.067	53		89	.017	135		128	.011	196		160	.467	164
	38	.117	52		90	.024	134		129	.014	195		161	.488	163
	39	.175	51		91	.033	133		130	.018	194		162	.512	162
	40	.225	50		92	.044	132		131	.023	193	9	165	.000	285
	41	.258	49		93	.055	131		132	.029	192		166	.000	284
	42	.342	48		94	.069	130		133	.035	191		167	.000	283
	43	.392	47		95	.083	129		134	.042	190		168	.000	282
	44	.475	46		96	.100	128		135	.048	189		169	.000	281
	45	.525	45		97	.118	127		136	.057	188		170	.001	280
6	56	.001	91		98	.133	126		137	.066	187		171	.001	279
	57	.008	90		99	.151	125		138	.076	186		172	.002	278
	58	.017	89		100	.177	124		139	.085	185		173	.002	277
	59	.029	88		101	.198	123		140	.098	184		174	.003	276
	60	.051	87		102	.222	122		141	.108	183		175	.004	275
	61	.068	86		103	.249	121		142	.122	182		176	.005	274
	62	.088	85		104	.278	120		143	.134	181		177	.007	273
	63	.121	84		105	.297	119		144	.150	180		178	.009	272
	64	.149	83		106	.331	118		145	.163	179		179	.011	271
	65	.178	82		107	.357	117		146	.180	178		180	.013	270
	66	.210	81		108	.391	116		147	.195	177		181	.016	269
	67	.249	80		109	.420	115		148	.214	176		182	.018	268
	68	.282	79		110	.453	114		149	.231	175		183	.022	267

Tavola 7. (segue)

n	s	P	s	n	s	P	s	n	s	P	s	n	s	P	s
9	184	.025	266	9	216	.354	234	10	242	.010	363	10	274	.165	331
	185	.029	265		217	.372	233		243	.012	362		275	.174	330
	186	.033	264		218	.388	232		244	.013	361		276	.184	329
	187	.038	263		219	.405	231		245	.015	360		277	.193	328
	188	.043	262		220	.422	230		246	.017	359		278	.203	327
	189	.048	261		221	.440	229		247	.019	358		279	.214	326
	190	.054	260		222	.456	228		248	.022	357		280	.224	325
	191	.060	259		223	.474	227		249	.025	356		281	.235	324
	192	.066	258		224	.491	226		250	.027	355		282	.246	323
	193	.074	257		225	.509	225		251	.030	354		283	.257	322
	194	.081	256	10	220	.000	385		252	.033	353		284	.268	321
	195	.089	255		221	.000	384		253	.037	352		285	.280	320
	196	.097	254		222	.000	383		254	.040	351		286	.292	319
	197	.106	253		223	.000	382		255	.044	350		287	.304	318
	198	.115	252		224	.000	381		256	.048	349		288	.316	317
	199	.125	251		225	.000	380		257	.052	348		289	.328	316
	200	.135	250		226	.000	379		258	.057	347		290	.341	315
	201	.146	249		227	.000	378		259	.062	346		291	.354	314
	202	.156	248		228	.000	377		260	.067	345		292	.367	313
	203	.168	247		229	.001	376		261	.072	344		293	.379	312
	204	.179	246		230	.001	375		262	.077	343		294	.393	311
	205	.193	245		231	.001	374		263	.083	342		295	.406	310
	206	.205	244		232	.001	373		264	.089	341		296	.419	309
	207	.218	243		233	.002	372		265	.096	340		297	.433	308
	208	.231	242		234	.002	371		266	.102	339		298	.446	307
	209	.247	241		235	.003	370		267	.109	338		399	.459	306
	210	.260	240		236	.004	369		268	.116	337		300	.473	305
	211	.276	239		237	.004	368		269	.124	336		301	.486	304
	212	.290	238		238	.005	367		270	.132	335		302	.500	303
	213	.307	237		239	.007	366		271	.139	334				
	214	.322	236		240	.008	365		272	.148	333				
	215	.339	235		241	.009	364		273	.156	332				

Tavola 8. Gli elementi della tavola danno le probabilità P di coda sinistra o destra della statistica C di Kendall (a secondo che nella tavola c sia a sinistra o a destra di P), $n = 3, 4, \ldots, 15$.

n	c	P	c	n	c	P	c	n	c	P	c	n	c	P	c
3	0	.167	3	8	5	.016	23	10	7	.002	38	11	20	.141	35
	1	.500	2		6	.031	22		8	.005	37		21	.179	34
4	0	.042	6		7	.054	21		9	.008	36		22	.232	33
	1	.167	5		8	.089	20		10	.014	35		23	.271	32
	2	.375	4		9	.138	19		11	.023	34		24	.324	31
	3	.625	3		10	.199	18		12	.036	33		25	.381	30
5	0	.008	10		11	.274	17		13	.054	32		26	.440	29
	1	.042	9		12	.360	16		14	.078	31		27	.500	28
	2	.117	8		13	.452	15		15	.108	30	12	0	.000	66
	3	.242	7		14	.548	14		16	.146	29		1	.000	65
	4	.408	6	9	0	.000	36		17	.190	28		2	.000	64
	5	.592	5		1	.000	35		18	.242	27		3	.000	63
6	0	.001	15		2	.000	34		19	.300	26		4	.000	62
	1	.008	14		3	.000	33		20	.364	25		5	.000	61
	2	.028	13		4	.001	32		21	.431	24		6	.000	60
	3	.068	12		5	.003	31		22	.500	23		7	.000	59
	4	.136	11		6	.006	30	11	0	.000	55		8	.000	58
	5	.235	10		7	.012	29		1	.000	54		9	.000	57
	6	.360	9		8	.022	28		2	.000	53		10	.000	56
	7	.500	8		9	.038	27		3	.000	52		11	.001	55
7	0	.000	21		10	.060	26		4	.000	51		12	.002	54
	1	.001	20		11	.090	25		5	.000	50		13	.003	53
	2	.005	19		12	.130	24		6	.000	49		14	.004	52
	3	.015	18		13	.179	23		7	.000	48		15	.007	51
	4	.035	17		14	.238	22		8	.001	47		16	.010	50
	5	.068	16		15	.306	21		9	.002	46		17	.016	49
	6	.119	15		16	.381	20		10	.003	45		18	.022	48
	7	.191	14		17	.460	19		11	.005	44		19	.031	47
	8	.281	13		18	.540	18		12	.008	43		20	.043	46
	9	.386	12	10	0	.000	45		13	.013	42		21	.058	45
	10	.500	11		1	.000	44		14	.020	41		22	.076	44
8	0	.000	28		2	.000	43		15	.030	40		23	.098	43
	1	.000	27		3	.000	42		16	.043	39		24	.125	42
	2	.001	26		4	.000	41		17	.060	38		25	.155	41
	3	.002	25		5	.000	40		18	.082	37		26	.190	40
	4	.007	24		6	.001	39		19	.109	36		27	.230	39

Tavola 8. (segue)

n	c	P	c	n	c	P	c	n	c	P	c	n	c	P	c
12	28	.273	38	13	30	.153	48	14	26	.018	65	15	16	.000	89
	29	.319	37		31	.184	47		27	.024	64		17	.000	88
	30	.369	36		32	.218	46		28	.031	63		18	.000	87
	31	.420	35		33	.255	45		29	.040	62		19	.000	86
	32	.473	34		34	.295	44		30	.050	61		20	.000	85
	33	.527	33		35	.338	43		31	.063	60		21	.001	84
13	0	.000	78		36	.383	42		32	.079	59		22	.001	83
	1	.000	77		37	.429	41		33	.096	58		23	.001	82
	2	.000	76		38	.476	40		34	.117	57		24	.002	81
	3	.000	75		39	.524	39		35	.140	56		25	.003	80
	4	.000	74	14	0	.000	91		36	.165	55		26	.004	79
	5	.000	73		1	.000	90		37	.194	54		27	.006	78
	6	.000	72		2	.000	89		58	.225	53		28	.008	77
	7	.000	71		3	.000	88		39	.259	52		29	.010	76
	8	.000	70		4	.000	87		40	.295	51		30	.014	75
	9	.000	69		5	.000	86		41	.334	50		31	.018	74
	10	.000	68		6	.000	85		42	.374	49		32	.023	73
	11	.000	67		7	.000	84		43	.415	48		33	.029	72
	12	.000	66		8	.000	83		44	.457	47		34	.037	71
	13	.000	65		9	.000	82		45	.500	46		35	.046	70
	14	.001	64		10	.000	81	15	0	.000	105		36	.057	69
	15	.001	63		11	.000	80		1	.000	104		37	.070	68
	16	.002	62		12	.000	79		2	.000	103		38	.084	67
	17	.003	61		13	.000	78		3	.000	102		39	.101	66
	18	.005	60		14	.000	77		4	.000	101		40	.120	65
	19	.007	59		15	.000	76		5	.000	100		41	.141	64
	20	.011	58		16	.000	75		6	.000	99		42	.164	63
	21	.015	57		17	.001	74		7	.000	98		43	.190	62
	22	.021	56		18	.001	73		8	.000	97		44	.218	61
	23	.029	55		19	.002	72		9	.000	96		45	.248	60
	24	.038	54		20	.002	71		10	.000	95		46	.279	59
	25	.050	53		21	.003	70		11	.000	94		47	.313	58
	26	.064	52		22	.005	69		12	.000	93		48	.349	57
	27	.082	51		23	.007	68		13	.000	92		49	.385	56
	28	.102	50		24	.010	67		14	.000	91		50	.423	55
	29	.126	49		25	.013	66		15	.000	90		51	.461	54
													52	.500	53

Tavola 9. Gli elementi della tavola danno i quantili della statistica di Kolmogorov per alcune probabilità di coda sinistra P e $n = 1, 2, \ldots, 40$.

$n \setminus P$	0.80	0.90	0.95	0.98	0.99	$n \setminus P$	0.80	0.90	0.95	0.98	0.99
1	.900	.950	.975	.990	.995	21	.226	.259	.287	.321	.344
2	.684	.776	.842	.900	.929	22	.221	.253	.281	.314	.337
3	.565	.636	.780	.785	.829	23	.216	.247	.275	.307	.330
4	.493	.565	.624	.689	.734	24	.212	.242	.269	.301	.323
5	.447	.509	.563	.627	.669	25	.208	.238	.264	.295	.317
6	.410	.468	.519	.577	.617	26	.204	.233	.259	.290	.311
7	.381	.436	.483	.538	.576	27	.200	.229	.254	.284	.305
8	.358	.410	.454	.507	.542	28	.197	.225	.250	.279	.300
9	.339	.387	.430	.480	.513	29	.193	.221	.246	.275	.295
10	.323	.369	.409	.457	.489	30	.190	.218	.242	.270	.290
11	.308	.352	.391	.437	.468	31	.187	.214	.238	.266	.285
12	.296	.338	.375	.419	.449	32	.184	.211	.234	.262	.281
13	.285	.325	.361	.404	.432	33	.182	.208	.231	.258	.277
14	.275	.314	.349	.390	.418	34	.179	.205	.227	.254	.273
15	.266	.304	.338	.377	.404	35	.177	.202	.224	.251	.269
16	.258	.295	.327	.366	.392	36	.174	.199	.221	.247	.265
17	.250	.286	.318	.355	.381	37	.172	.196	.218	.244	.262
18	.244	.279	.309	.346	.371	38	.170	.194	.215	.241	.258
19	.237	.271	.301	.337	.361	39	.168	.191	.213	.238	.255
20	.232	.265	.294	.329	.352	40	.165	.189	.210	.235	.252

Tavola 10. Gli elementi della tavola danno i quantili della statistica di Kolmogorov-Smirnov per uguali campioni di numerosità n per alcune probabilità di coda sinistra P e $n = 1, 2, \ldots, 40$.

$n \setminus P$	0.80	0.90	0.95	0.98	0.99	$n \setminus P$	0.80	0.90	0.95	0.98	0.99
1						21	6/21	7/21	8/21	9/21	10/21
2						22	7/22	8/22	8/22	10/22	10/22
3	2/3	2/3				23	7/23	8/23	9/23	10/23	10/23
4	3/4	3/4	3/4			24	7/24	8/24	9/24	10/24	11/24
5	3/5	3/5	4/5	4/5	4/5	25	7/25	8/25	9/25	10/25	11/25
6	3/6	4/6	4/6	5/6	5/6	26	7/26	8/26	9/26	10/26	11/26
7	4/7	4/7	5/7	5/7	5/7	27	7/27	8/27	9/27	11/27	11/27
8	4/8	4/8	5/8	5/8	6/8	28	8/28	9/28	10/28	11/28	12/28
9	4/9	5/9	5/9	6/9	6/9	29	8/29	9/29	10/29	11/29	12/29
10	4/10	5/10	6/10	6/10	7/10	30	8/30	9/30	10/30	11/30	12/30
11	5/11	5/11	6/11	7/11	7/11	31	8/31	9/31	10/31	11/31	12/31
12	5/12	5/12	6/12	7/12	7/12	32	8/32	9/32	10/32	12/32	12/32
13	5/13	6/13	6/13	7/13	8/13	33	8/33	9/33	11/33	12/33	13/33
14	5/14	6/14	7/14	7/14	8/14	34	8/34	10/34	11/34	12/34	13/34
15	5/15	6/15	7/15	8/15	8/15	35	8/35	10/35	11/35	12/35	13/35
16	6/16	6/16	7/16	8/16	9/16	36	9/36	10/36	11/36	12/36	13/36
17	6/17	7/17	7/17	8/17	9/17	37	9/37	10/37	11/37	13/37	13/37
18	6/18	7/18	8/18	9/18	9/18	38	9/38	10/38	11/38	13/38	14/38
19	6/19	7/19	8/19	9/19	9/19	39	9/39	10/39	11/39	13/39	14/39
20	6/20	7/20	8/20	9/20	10/20	40	9/40	10/40	12/40	13/40	14/40

Bibliografia

Bibliografia di base

Billingsley, P. (1968) *Convergence of Probability Measures*, Wiley, New York.

Conover, W.J. (1980) *Practical Nonparametric Statistics*, Wiley, New York.

Daniel, W.W. (1978) *Applied Nonparametric Statistics*, Houghton Mifflin Company, Boston.

Feller W. (1971) *An Introduction to Probability Theory and its Applications*, vol. I-II, Wiley, New York.

Fraser, D.A.S. (1957) *Nonparametric Methods in Statistics*, Wiley, New York.

Gibbons, J.D. (1985) *Nonparametric Methods for Quantitative Analysis*, American Sciences Press, Syracuse, New York.

Gibbons, J.D. e Chakraborti, S. (1992) *Nonparametric Statistical Inference*, Dekker, New York.

Hettmansperger, T.P. (1991) *Statistical Inference Based on Ranks*, Krieger Publishing Company, Malabar, Florida.

Hettmansperger, T.P. e McKean, J.W. (1998) *Robust Nonparametric Statistical Methods*, Arnold, London.

Hollander, M. e Wolfe, D.A. (1973) *Nonparametric Statistical Methods*, Wiley, New York.

Hájek, J. (1969) *Nonparametric Statistics*, Holden Day, San Francisco.

Hájek, J. e Šidák, Z. (1967) *Theory of Rank Tests*, Academic Press, New York.

Kendall, M.G. (1962) *Rank Correlation Methods*, Hafner, New York.

Kendall, M.G. e Gibbons J.D. (1976) *Rank Correlation Methods*, Edward Arnold, London.

Lehmann, E.L. (1975) *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco.

Maritz, J.S. (1981) *Distribution-free Statisitical Methods*, Chapman and Hall, London.

Noether, G.E. (1967) *Elements of Nonparametric Statistics*, Wiley, New York.

Puri, M.L. e Sen, P.K. (1971) *Nonparametric Methods in Multivariate Analysis*, Wiley, New York.

Randles, R.H. e Wolfe, D.A. (1979) *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York.

Siegel, S. (1956) *Nonparametric Statistics for Behavioral Sciences*, McGraw Hill, New York.

Siegel, S. e Castellan, N.J. (1988) *Nonparametric Statistics for Behavioral Sciences*, McGraw Hill, New York.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Riferimenti bibliografici

Batschelet, E. (1981) *Circular Statistics in Biology*, Academic Press, London.

Benford, F. (1938) The law of anomalous numbers, *Proceedings of the American Philosophical Society* **78**, 551-572.

Bhatia, M.L., Manchanda, S.C. e Roy, S.B. (1969) Coronary Haemodinamic studies in chronic severe anaemia, *British Heart Journal* **31**, 365-374.

Brinegar, C.S. (1963) Mark Twain and the Q.C.S. letters - a statistical test of authorship, *Journal of the Royal Statistical Association* **58**, 85-96.

Brook, C.G.D. (1971) Determination of body composition of children from skinfold measuraments, *Archivia Disease Childhood* **46**, 182-184.

Bryson, M.C. e Siddiqui, M.M. (1969) Survival time: some criteria for aging, *Journal of the American Statistical Association* **64**, 1472-1483.

Cameron, E. e Pauling, L. (1974) Supplemental ascorbate in the supportive treatment of cancer: re-evaluation of prolongation of survival times in terminal human cancer, *Proceedings of the National Academy of Science USA* **75**, 4538-4542.

Clarke, R.D. (1946) An application of the Poisson distribution, *Journal of the Institute of the Actuaries* **72**, 48.

Darwin, C. (1876) *The effects of cross- and self-fertilization in the vegetable kingdom*, John Murray, London.

Dubois, C. (1970) *Lowie's Selected Papers in Antrophology*, University of California Press, California.

Edwards A.W. e Fraccaro M. (1960) Distribution and sequences of sexes in a selected sample of Swedish families, *Annals Human Genetics* **24**, 245-252.

Gadsen, R.J. e Kanji, G.K. (1981) Sequential analysis for angular data, *The Statistician* **30**, 119-129.

Johnson, T.E. e Graybill, F.A. (1972) An analysis of two-way model with interaction and no replication, *Journal of the American Statistical Association* **67**, 862-868.

Kendall, D.G. (1951) Some problems in the theory of queues, *Journal of the Royal Statistical Society* **B13**, 151-185.

Lederman, S.J., Klatsky, R.L. e Barber, B.P.O. (1985) Spatial and movement-based heuristics for encoding pattern information through touch, *Journal of Experimental Psycology: General* **114**, 33-49.

Lunn, A.D. e McNeil, D.R. (1991) *Computer-Interactive Data Analysis*, Wiley, New York.

Nanji, A.A. e French, S.W. (1985) Relationship between pork consumption and cirrhosis, *The Lancet* **1**, 681-683.

Preece, D.A. (1981) Distributions of final digits in the data, *The Statistician* **30**, 31-60.

Quesenberry, C.P. e Hales, C. (1980) Concentration bands for uniformity plots, *Journal of Statistical Computation and Simulation* **11**, 41-53.

Romano, A. (1977) *Applied Statistics for Science and Industry*, Allyn and Bacon, Boston.

Rutheford, E. e Geiger, M. (1910) The probability of variations in the distributions of alpha-particles, *Philosophical Magazine* **20**, 698-704.

Selvin, S. (1991) *Statistical Analysis of Epidemiological Data*, Oxford University Press, New York.

Snedecor, G.W. e Cochran, G.C. (1967) *Statistical Methods*, Iowa State University Press, Ames.

Till, R. (1974) *Statistical Methods for the Earth Scientist*, McMillan, London.

Van Oost, B.A., Veldhayzen, B., Timmermans, A.P.M. e Sixma, J.J. (1983) Increased urinary β-thromboglobulin excretion in diabets assayed with a modified RIA kit-technique, *Thrombosis and Haemostasis* **9**, 18-20.