SUI RANGHI (E NON SOLO)

Spesso nelle analisi statistiche si parte dall'assunzione che la caratteristica oggetto di studio abbia nelle popolazioni una distribuzione normale. L'ipotesi di normalità, è un'assunzione estremamente importante ed è essenziale capire quando sia o meno lecito ritenerla valida.

A volte l'ipotesi di normalità poggia su un modello teorico come nel caso in cui si effettuino misurazioni ripetute di una stessa grandezza. Se escludiamo la presenza di errori sistematici, potremmo assumere $Y_i=\mu+\varepsilon_i$, dove la nostra osservazione Y_i sarà somma della vera misura della grandezza in esame μ e di un errore casuale ε_i descritto da una densità normale a media nulla.

Nel caso di un campione di numerosità elevata, potremmo facilmente verificare l'ipotesi di distribuzione normale, utilizzando i dati empirici per la costruzione di un istogramma. Anche in presenza di deviazioni dalla normalità, potremmo godere del sostegno offertoci dal TEOREMA CENTRALE DEL LIMITE, secondo il quale partendo da popolazioni con distribuzione non perfettamente normale, le loro medie campionarie tenderanno alla normale asintoticamente. Nel caso in cui l'istogramma ci mostrasse, invece, distribuzioni molto asimmetriche la media non rappresenterebbe più il valore centrale e quindi non potrà esserci d'aiuto.

L'assunzione di un modello ci permette di avere delle conoscenze a priori, che precedono lo svolgimento dell'esperimento, a queste si aggiungeranno quelle ricavate dai dati a esperimento concluso. Se non fossimo in grado di partire dal modello dovremo affidarci alla sola informazione empirica.

La situazione diventa più difficile nel caso in cui il campione sia di piccole dimensioni. In questo caso, a meno che non siamo supportati da un precedente studio che abbia verificato l'ipotesi di distribuzione normale per la caratteristica oggetto di studio, la costruzione di un istogramma non potrà esserci di aiuto, perché risulterebbe impreciso e poco affidabile. Non godremo inoltre della protezione del teorema centrale del limite.

In una condizione del genere, al fine di non ottenere risultati errati, saremo costretti ad allontanarci dalla STATISTICA PARAMETRICA fondata sull'assunzione di un modello (nel nostro caso normale) e capace di tradurre l'incertezza associata all'esperimento attraverso un numero *finito* di parametri (μ e σ nle caso normale). Ci sposteremo, quindi, verso la STATISTICA NON PARAMETRICA rinunciando a fare ipotesi in merito alla distribuzione della nostra variabile risposta nella popolazione che stiamo studiando. Muovendoci in questa direzione non saremo in grado di tradurre l'incertezza attraverso un numero finito di parametri, che diventeranno, così, un numero *infinito.*

Va sottolineato che il fatto di aver rinunciato all'assunzione di un modello non significa che non esista nella popolazione una precisa distribuzione, ma traduce esclusivamente la nostra incapacità nell'indagarla. Se infatti potessimo osservare tutti i soggetti della popolazione e costruire un istogramma, la forma di questa distribuzione sarebbe chiara.

Partiamo da una situazione sperimentale prettamente OSSERVAZIONALE in cui abbiamo due popolazioni: una costituita da soggetti sani e una costituita da soggetti affetti da una data patologia . La nostra

attenzione è focalizzata su un particolare gene, di cui misureremo il livello di espressione con lo scopo di verificare se questo possa essere identificato come marcatore di rischio per la patologia.

Il primo passo consiste nel raccogliere i campioni rappresentativi delle due popolazioni di interesse. Il prerequisito fondamentale del campionamento è dato dall'equiprobabilità, secondo la quale ogni soggetto della popolazione deve avere la stessa probabilità di entrare a far parte del campione. Il campionamento è un passaggio delicato e fondamentale e va affrontato con estrema attenzione. Partiamo dalla popolazione di soggetti malati: se lo studio è svolto in Italia, possiamo immaginare di disporre di una lista dei centri nazionali specializzati nella patologia, a partire dalla quale estrarremo casualmente alcuni centri, ai quali chiederemo a loro volta una lista dei soggetti in cura da cui estrarremo casualmente i soggetti che entreranno a far parte del campione. E' importante valutare, a seconda del tipo di gene e patologia, se sia il caso di reclutare soggetti già sottoposti a cura o reclutare soggetti alla prima diagnosi e mai sottoposti a terapia (naive). In questo secondo caso i pazienti dovranno essere reclutati in modo sequenziale a partire da un tempo t_0, man a man che si presenteranno al centro specializzato. Questa modalità è utile ad esempio nel caso in cui il gene in esame è in grado di mutare a seguito di trattamento farmaceutico, cosa che falserebbe la valutazione in merito al livello di espressione originario.

Una volta raccolto il campione di soggetti malati procederemo all'estrazione del DNA. Anche a tal proposito è utile domandarci se sia meglio procedere all'estrazione in loco o portare tutti i campioni di DNA ad un unico centro per la valutazione del livello di espressione oppure procedere localmente. Nel primo caso per quanto riguarda macchinari, operatori, e metodologie, eviteremmo sicuramente il sommarsi di errori aggiuntivi e fattori di confondimento, ma, centralizzando l'estrazione, altereremo la rappresentatività del campione, prerequisito fondamentale per poter espandere il risultato all'intera popolazione. Il problema nasce dal fatto che, quando un singolo soggetto della popolazione vorrà indagare il livello di espressione del gene in questione, per valutare la presenza o meno di quella patologia, si rivolgerà al centro più vicino. Questo avrà una propria variabilità associata, che sarà diversa da quella del centro di fiducia in cui abbiamo svolto l'estrazione. Centralizzando, quindi, non ci confrontiamo con l'extra-variabilità associata ad ogni singolo centro, che però si presenterà quando espanderemo il risultato all'intera popolazione.

Ora che abbiamo disegnato l'esperimento per quanto riguarda i soggetti malati dobbiamo reclutare il campione di soggetti sani che fungeranno da controllo negativo (perché ci aspettiamo che il gene che vogliamo dimostrare essere marcatore della patologia non sia presente nella popolazione dei sani).

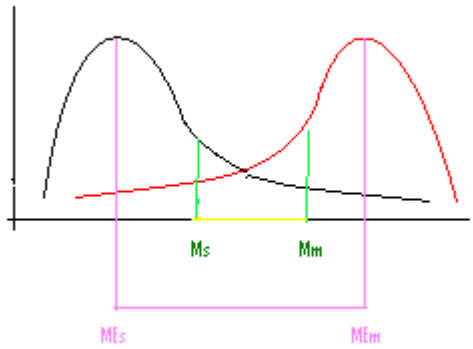
Il reclutamento dei soggetti sani ci pone davanti ad una situazione altrettanto delicata che riguarda la loro adesione alla sperimentazione: sarà certamente più facile convincerli a collaborare nel caso si possa escludere la patologia tramite una banale analisi, come per esempio nel caso del diabete in cui basterà la misurazione della glicemia. La situazione, ovviamente, si complica nel caso in cui siano necessarie analisi più complicate; in questo caso ci aspettiamo che molti preferiscano non aderire. Per ovviare a questo problema potremmo rivolgerci agli stessi centri specializzati dai quali avevamo attinto i soggetti malati, e ci faremo consegnare una lista di tutti i soggetti sani, che si sono recati nella struttura sospettando la patologia che in seguito è stata esclusa grazie ad accertamenti specifici.

A questo punto immaginiamo di avere due campioni rispettivamente rappresentativi delle due popolazioni di soggetti sani e malati e procediamo nell'analisi.

Sotto l'ipotesi di un modello normale la formulazione del sistema di ipotesi era centrata su μ , parametro fondamentale intorno al quale ci aspettavamo di trovare i valori relativi alla maggior parte della

popolazione. μ veniva poi stimato con la media campionaria. Nel caso in questione però, poiché abbiamo rinunciato all'assunzione di un modello, il parametro μ perde centralità poiché la media, sensibile ai valori estremi, tende a seguirli.

Vediamo un esempio:



Nel grafico la curva nera rappresenta la distribuzione dei soggetti sani, mentre la curva rossa rappresenta la distribuzione dei soggetti malati, come possiamo osservare entrambe sono asimmetriche. La media dei soggetti sani (Ms) tende a spostarsi a destra perché, sebbene la maggior parte della popolazione (picco) mostrerà valori bassi di espressione del gene, ci sarà una piccola parte di essa (coda) costituita da soggetti sani ma esprimenti il gene. Queste poche osservazioni con valori, però, molto lontani da quelli della maggior parte della popolazione, saranno avvertite dalla media che tenderà a seguirle, spostandosi. Nel caso della media dei soggetti malati (Mm),invece, avremo che la maggior parte della popolazione esprimerà alti livelli del gene, ma sarà sempre presente una piccola parte, che, al contrario, non esprimerà il gene anche in presenza di malattia. La presenza di queste poche osservazioni, in cui il gene non è espresso influenzeranno la media spostandola, questa volta a sinistra. In una distribuzione asimmetrica la media perde centralità e di conseguenza non rappresenta l'effettiva distanza tra le due popolazioni. Piu' coerente con lo scopo della ricerca è la distanza tra le mediane calcolate rispettivamente sulle due popolazioni (MEs e MEm). Infatti la mediana si colloca al centro comunque siano fatte le distribuzioni, è robusta rispetto agli estremi e ci permette di valutare le effettive distanze esistenti tra le due popolazioni.

Scegliamo allora di esprimere il sistema d'ipotesi come:

$$\begin{cases} H_0 & \delta_S = \delta_M \\ H_1 & \delta_S \neq \delta_M \end{cases}$$

Dove con δ_S e δ_M indichiamo rispettivamente la mediana nella popolazione dei sani e la mediana nella popolazione dei malati.

Il passo successivo sarà la costruzione della relativa statistica test. Possiamo immaginare di assumere come stimatori di δ_S e δ_M le corrispondenti mediane campionarie e, dal momento che stiamo trattando con un parametro di posizione, il loro confronto potrà essere basato sulla differenza. I problemi sorgono al momento di derivare la distribuzione della statistica test sotto H_0. Nella teoria classica del T-TEST la statistica test è rappresentata dalla differenza tra le medie campionarie, con distribuzione normale poiché

calcolata a partire da osservazioni normali. Nel nostro caso, se utilizzassimo la differenza tra mediane campionarie, la sua distribuzione dipenderebbe dalla densità delle singole osservazioni che non conosciamo e non saremmo in grado di procedere oltre. Ricordiamo infatti che le nostre osservazioni sono variabili aleatorie i.i.d. (indipendenti e identicamente distribuite) con $X_i \sim f$(x) dove f(x) è la distribuzione del livello di espressione del gene nella popolazione, a noi sconosciuta. Per procedere nel confronto tra le due popolazioni abbiamo bisogno di costruire una statistica test la cui distribuzione sotto H_0 non dipenda dalla forma di *f.* Proviamo a partire, dall'unica informazione di cui disponiamo, cioè quella empirica. Avremo

$\Pr\{X_1 = x_1, \dots, X_n = x_n\} = \prod_{i=1}^{n} \Pr\{X_i = x_i\} = \prod_{i=1}^{n} f(x_i)$

La probabilità di osservare un campione caratterizzato dai valori (x_1,…, x_n) è uguale alla probabilità dell'intersezione dei singoli eventi che, sfruttando l'indipendenza, può essere scritta come la produttoria della probabilità relative alle singole osservazioni che essendo state estratte tutte dalla stessa popolazione, erediteranno da essa la distribuzione *f*, indipendente dall'indice. E' importante capire come i campioni osservabili non sono equiprobabili; la probabilità di osservare un certo campione nello spazio dei campioni possibili, dipende dalla forma di *f* e, come è logico attendersi, saranno più probabili quei campioni che contengono valori molto frequenti nella popolazione d'origine.

Il metodo che andiamo a descrivere prevede che l'informazione empirica di cui disponiamo sia scissa in due parti, al fine di utilizzare quella la cui distribuzione è indipedente da *f*.

Da un lato definiamo come STATISTICA D'ORDINE l'insieme dei valori osservati ordinati in senso crescente (non decrescente). In simboli $(X_{(1)}, \dots, X_{(n)})$. E' importante notare che gli indici sono scritti tra parentesi per indicare l'ordinamento delle singole osservazioni: $X_{(1)}$rappresenta la più piccola delle osservazioni, mentre X_1 indica la prima osservazione estratta casualmente dalla popolazione di partenza. Attuando quest'ordinamento, perdiamo l'identità (label) del soggetto, legata all'ordine di estrazione. E' diverso dire che il primo soggetto estratto, "Marco", ha un livello di espressione del gene pari a 3 ($X_1 = 3$) dal dire che il più piccolo livello di espressione del gene osservato è 3 ($X_{(1)} = 3$).

Accanto alla statistica d'ordine definiamo la STATISTICA RANGO $(R_1, \dots, R_n)$ dove $R_i = j \leftrightarrow X_i = X_{(j)}$. In altri termi il rango R_i dell'*i*-esimo soggetto estratto è la posizione da lui occupata nella statistica d'ordine.

Esiste una stretta analogia tra i ranghi e l' "ordine di arrivo" come la intendiamo ad esempio nelle comperizioni sportive. Proviamo a costruire un esempio pratico che possa aiutarci a campire quanto detto utilizzando i risultati ottenuti dagli atleti durante la competizione finale dei 200 metri maschili ai giochi della XXX olimpiade di Londra 2012. Elenchiamo gli atleti immaginando di conoscere l'ordine in cui si sono iscritti alla competizione e accanto i risultati da essi ottenuti:

ORDINE ISCRIZIONE	NOMINATIVI PARTECIPANTI	TEMPI OTTENUTI
1	Richard Thompson	9"98
2	Asafa Powel	11"99
3	Churandy Martina	9"94
4	Usain Bolt	9"63
5	Tyson Gay	9"80
6	Yohan Blake	9"75
7	Justin Gatin	9"79
8	Ryan Bailey	9"88

A questo punto costruiamo la STATISTICA D'ORDINE ordinando in modo crescente i tempi ottenuti:

STATISTICA D'ORDINE	TEMPI OTTENUTI
$x_{(1)}$	9"63
$x_{(2)}$	9"75
$x_{(3)}$	9"79
$x_{(4)}$	9"80
$x_{(5)}$	9"88
$x_{(6)}$	9"94
$x_{(7)}$	9"98
$x_{(8)}$	11"99

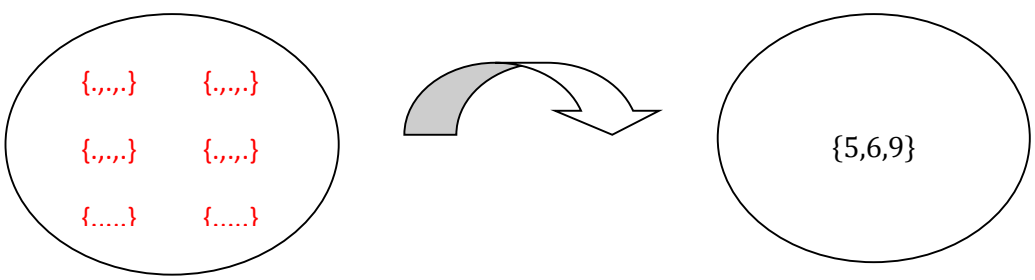
Ora costruiamo la STATISTICA RANGO dove il rango R_i rappresenta la posizione occupata dall'*i*-esimo atleta nella statistica d'ordine o, in altri termini, il suo ordine di arrivo:

STATISTICA RANGO	GRADUTORIA FINALE
r_1	7
r_2	8
r_3	6
r_4	1
r_5	4
r_6	2
r_7	3
r_8	5

Notiamo che la statistica rango si libera immediatamente dalle quantità osservate (tempi ottenui), che sono variabili quantitative continue, trasformando l'informazione di partenza in un numero naturale che va da 1 ad n (variabile discreta). Seguendo il nostro esempio ci limiteremo ad avere la classifica finale dei soggetti tralasciando la distanza effettiva tra le osservazioni che può essere valutata esclusivamente considerando i tempi ottenuti dai singoli. Considerando R_3=6 ed R_1=7 sappiamo che questi due atleti distano una posizione all'interno della classifica generale, ma non conosciamo più la distanza effettiva in termini di tempi ottenuti. Perdiamo l'informazione legata al valore continuo del tempo perché essendo la sua distribuzione dipendente da f non siamo in grado di utilizzarla.

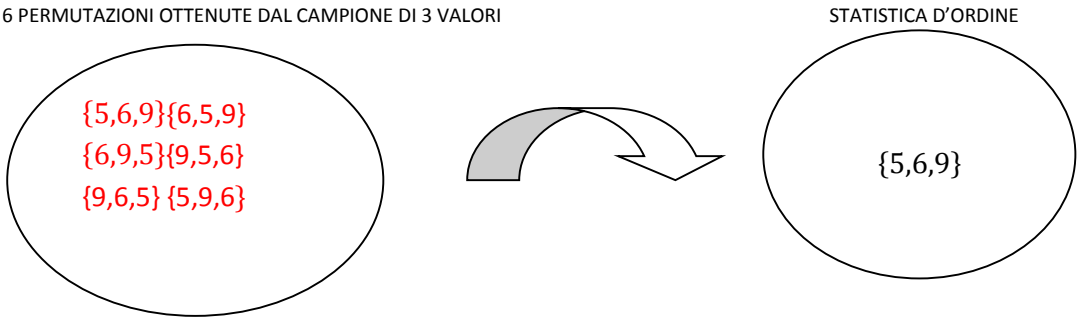
Immaginiamo di aver fatto soltanto tre osservazioni e consideriamo i due insiemi a seguire. Il primo rappresenta lo spazio dei campioni, composto da vettori di tre elementi contenenti livelli diversi di espressione del gene e l'altro contiene i vettori osservabili per la rispettiva statistica d'ordine:

SPAZIO DEI CAMPIONI OSSERVABILI Ω_X STATISTICA D'ORDINE $\Omega_{X(i)}$



Ricordiamo che la probabilità di osservare un generico campione in Ω_X è Pr $\{(\mathrm{X}_1 = x_1) \cap \ldots \cap (\mathrm{X}_\mathrm{n} = x_n)\} = \prod_{i=1}^{n} f(x_i)$

Osserviamo che come ad una particolare statistica d'ordine corrispona piu' campioni osservabili. Ad esempio la statistica d'ordine $\{5,6,9\}$ può derivare da tutti quei campioni costituiti da questi valori ma ordinati in modi diversi. Questi ultimi saranno tanti quante sono le permutazioni (n!) dei 3 valori,cioé 3!=6.



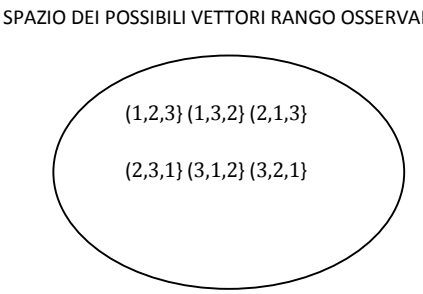
Una volta capito quali siano i campioni che conducono alla medesima statistica d'ordine, interessiamoci di capire con quale probabilità saranno osservati. La probabilità di ottenere quella determinata statistica d'ordine sarà uguale alla probabilità dell' unione dei singoli campioni osservabili e, quindi, uguale alla somma delle loro(6) probabilità essendo eventi incompatibili. Dal momento che ciascuno dei 6 possibili campioni condivide la stessa probabilità di essere osservato, $\prod_{i=1}^{n} f(x_i)$= *f(3)·f(6)·f(9)* , la probabilità di osservare quella specifica statistica d'ordine sarà data da 6 volte questo valore. In generale avremo

Pr $\{(\mathrm{X}_1 = x_1) \cap \ldots \cap (\mathrm{X}_\mathrm{n} = x_n)\} = n! \prod_{i=1}^{n} f(x_i)$

Osserviamo come la distribuzione della statistica d'ordine rimanga legata a *f* : questo perché è strettamente influenzata da quanto sia frequente, nella popolazione, ogni singolo valore che la costituisce.

Ora passiamo a considerare la probabilità di osservare un certo vettore RANGO:

Partiamo dal capire il numero dei possibili vettori rango osservabili se n=3:



Ogni possibile elemento dello spazio dei campioni a dimensione *n* fissata corrisponde ad un vettore rango appartenente all' Ω_R ma ad ogni vettore rango osservabile corripsonderanno tutti quei campioni che, pur essendo costutiti da valori diversi, condivideranno lo stesso ordinamento. Ad esempio al vettore rango {1,2,3} corrisponderanno tutti i campioni che contengono valori già ordinati in senso crescente, indipendentemente da quali siano, ad esempio {7,9,12} ma anche {20,26,90}. Notiamo che il numero dei possibili vettori rango è pari al numero dei possibili ordinamenti cioè al numero delle possibili permutazioni dei primi n numeri naturali n!, nel nostro caso 6. Se le nostre osservazioni provengono dalla stessa popolazione ne condivideranno tutte la distribuzione. Ne segue che i diversi vettori rango saranno

tra loro equiprobabili poiché non esisteranno motivi per ritenere a priori che una osservazione possa essere maggiore (o minore) di un'altra. vettori rango saranno tra loro equiprobabili. Avremo quind

$Pr\{(R_1 = r_1) \cap \dots \cap (R_n = r_n)\} = \frac{1}{n!}$

Comprendiamo che, passando alla statistica rango, abbiamo compiuto un enorme semplificazione, poiché siamo passati da R^3 (se abbiamo 3 osservazioni reali possiamo mettere in corrispondenza ciascun campione osservabile con il punto nello spazio a tre dimensioni che ha come coordinatei valori osservati) ad uno spazio costituito da soli sei elementi.

Questa perdita di informazione, rappresenta il prezzo che siamo costretti a pagare per non essere in grado di fare ipotesi su *f* . La probabilità di osservare un certo vettore di ranghi, infatti, si svincola totalmente dalla forma di *f*, per dipendere esclusivamente dalla numerosità *n* del campione, che invece, è un valore noto.

Notiamo che, mentre il vettore (X_1,…,X_n) è costituito da osservazioni tra loro indipendenti, ne' il vettore rango, ne' il vettore statistica d'ordine lo sono. Se immaginiamo di conoscere i primi n-1 ranghi l'ultimo è per forza determinato.

Vediamo concretamente come utilizzare i ranghi nel nostro caso. Siamo partiti da due popolazioni costituite rispettivamente da soggetti sani e malati e cerchiamo di valutare il livello di espressione di un dato gene che speriamo possa essere indicato come marcatore della patologia in questione senza alcuna ipotesi sulla sua distribuzione nelle due popolazioni. A partire da due campioni di osservazioni il nostro obiettivo è portare a verifica il sistema di ipotesi:

$$\begin{cases} H_0 & \delta_S = \delta_M \\ H_1 & \delta_S \neq \delta_M \end{cases}$$

A questo scopo dobbiamo costruire una statistica test basata sui ranghi e sensibile all'ipotesi alternativa. Partiamo considerando congiutamente i due campioni e calcoliamo la statistica mantenendo l'informazione relativa al campione di appartenenza. Definiamo poi come statistica test la somma dei ranghi corrispondenti ad un dei due campioni, in genere quello di numerosità maggiore (WILCOXON RANK-SUM TEST). Ipotizziamo di aver osservato i seguenti campioni S={3, 7, 12, 10, 5}, M={11, 26} rispettivamente di numerosità n_S=5 e n_M=2.

La statistica d'ordine congiunta sarà $X_{()}$={ 3(A), 5(A), 7(A), 10(A), 11(B), 12(A), 26(B)}, da cui R_{1M}=5 e R_{2M}=7. Ne segue che il valore osservato della nostra statistica test sarà W=5+7=12.

Ricordiamo che una buona statistica test deve rispondere a due requisiti fondamentali: deve avere una distribuzione nota sotto H_0 e deve essere sensibile ad H_1 cioè deve essere capace di individuare ed enfatizzare eventuali differenze esistenti tra $f_S(x)$ e $f_M(x)$.

Partiamo analizzando cosa osserveremmo se fosse vera l'ipotesi H_0 che prevede uguale distribuzione per le due popolazioni. In questa situazione è come se avessimo un'unica popolazione, quindi, i ranghi che osserveremo relativamente ai soggetti malati potranno essere considerati un campione casuale dall'insieme totale dei ranghi osservabili che nel nostro caso va da 1 a 7. Ci attendiamo pertanto che la somma dei ranghi tenda ad assumere valori centrali rispetto al range dei valori osservabili

Ora vediamo cosa succederebbe se fosse vera l'ipotesi H_1 che prevede una distribuzione diversa del gene nei sani e nei malati. Consideriamo il caso in cui i soggetti malati abbiano livelli di espressione del gene

sistematicamente più alti, immaginando una netta separazione delle due popolazioni. Ci aspettiamo, quindi, che i ranghi dei soggetti malati occupino le ultime 2 posizioni e conducano a valori elevati della statistica.

Resta da derivare la distribuzione della somma dei ranghi sotto H_0, cioè i valori che W potrà assumere al variare del campione osservato e le corrispondenti probabilità.

Osserviamo che nel nostro caso, poiché disponiamo in totale di 7 osservazioni, W potrà assumere valori compresi tra un minimo di 3 (1+2) ed un massimo 13 (6+7). Infatti i valori che si possono osservare per i ranghi del campione congiunto andranno da 1 a 7. Quali di questi ranghi corrisponderanno alle due osservazioni del gruppo M (R_{1M}, R_{2M})?

I ranghi osservabili per queste due osservazioni saranno:

1,2	1,3	1,4	1,5	1,6	1,7
	2,3	2,4	2,5	2,6	2,7
		3,4	3,5	3,6	3,7
			4,5	4,6	4,7
				5,6	5,7
					6,7

*ognuna di queste coppie ne contiene dentro 2 considerando che possiamo osservarle anche nell' ordine contrario, quindi per avere il totale delle coppie possibili basterà raddoppiarle.

Ciascuna delle coppie in tabella condurrà ai seguenti valori della statistica test W:

3	4	5	6	7	8
	5	6	7	8	9
		7	8	9	10
			9	10	11
				11	12
					13

Come possiamo notare siamo partiti da coppie di ranghi, tra loro equiprobabili poiché provenienti dalla stessa popolazione, da queste abbiamo ricavato le rispettive somme che, però, non sono ugualmente probabili. Come possiamo osservare, per esempio, otteniamo il valore 3 o 13 a partire da una sola coppia rispettivamente, mentre otteniamo il valore 7 a partire da 3 coppie.

E' necessario partire dalle coppie, perché lavorando direttamente sulle somme, non ci saremmo accorti che valori diversi hanno probabilità diverse di essere osservati.

A questo punto vediamo la distribuzione (discreta) della statistica test W:

W	PROBABILITA'
3	1/21
4	1/21
5	2/21
6	2/21
7	3/21
8	3/21
9	3/21
10	2/21
11	2/21
12	1/21
13	1/21

Non ci rimane che definire la regione di rifiuto del test assumendo α=0,05, regione che dovrà essere collocata sulle due code. Saremo sulla coda destra se i livelli di espressione del gene nei malati avranno valori elevati rispetto ai sani poiché in questo caso anche la somma dei ranghi nel campioni di pazienti malati sarà elevata. Saremo, invece, sulla coda sinistra se il livello di espressione del gene nei malati tenderà a valori bassi portando ad una somma dei ranghi bassa.

Come troviamo la regione di rifiuto sotto H_0? Quando eravamo in grado di ipotizzare un modello normale, avevamo una densità continua. In quel caso fissavamo α e cercavamo il percentile capace di tagliare aree pari ad $\alpha/2$. Ora che ci troviamo nel discreto invece di integrare dovremmo sommare le probabilità dei singoli valori partendo da quelli più estremi fino a raggiungere circa 0,025.

Nel nostro caso 1/21 = 0,048 > 0,025, quindi il valore W=13 si colloca già in regione di accettazione. La regione di rifiuto è, quindi, VUOTA. Il risultato ottenuto indica che in assenza di un modello e con poche osservazioni non possiamo arrivare a nessuna conclusione in merito alle due popolazioni. Il valore osservato di W=12 cade in regione di accettazione ma qualsiasi altro valore osservabile ci avrebbe condotto a non rifiutare H_0 lasciando aperta la possibiltà che entrambe le ipotesi siano vere. Ricordiamo infatti che un test non significativo non prova H_0 ma semplicemente che la nostra informazione sperimentale non è in grado di confutarla.

Il calcolo della distribuzione esatta di W al crescere della numersosità dei due campioni diventa complesso. E' allora cosigliabile standardizzare la statistica W. Questa nuova W standardizzata per $n \to \infty$ converge alla densità normale. In questo modo, per grandi campioni, si possono utilizzare le code della distribuzione normale.