Università degli studi di Padova

Facoltà di scienze statistiche

Corso di laurea in Statistica e tecnologie informatiche

Tesi di laurea

INDIVIDUAZIONE DI GENI DIFFERENZIALMENTE ESPRESSI: UNO STUDIO DI SIMULAZIONE

Relatore: Prof.ssa MONICA CHIOGNA

Laureando: Hasko Ervin

Anno accademico 2004/2005

Indice

Introduzione 5

Capitolo I

1 Microarray

1.1 Breve rassegna delle scoperte riguardanti la molecola di DNA 9

1.2 Alcune nozioni di biologia 11

1.3 Le nuove tecnologie: i microarray 13

1.3.1 Come funziona l'esperimento 14

1.3.2 Tipi di distorsioni 16

1.4 Rassegna bibliografica 18

Capitolo II

2 Metodi statistici 21

2.1 Il test t-Student 21

2.2 Il modello semiparametrico e le sue ipotesi 22

2.3 Il test Wilcoxon-Mann-Whitney 24

Capitolo III

3.1 Simulazione 29

3.2 Risultati 29

3.3 Attendibilità dei test 40

Appendice 43

Bibliografia 47

Introduzione

Ormai da tempo, lo studio del DNA umano da parte di studiosi e scienziati sta prendendo piede in quanto, dopo una lunga analisi delle espressioni genetiche, si è ipotizzata la teoria secondo la quale gran parte delle malattie derivano da alterazioni del codice genetico.
La tecnologia del *DNA microarray* può essere un utile strumento per identificare mutazioni presenti nei geni e per comprendere, attraverso l'analisi simultanea di migliaia di geni, la patogenesi sia delle malattie genetiche vere e proprie (quelle cioè che si trasmettono in famiglia), sia di quelle multifattoriali come diabete, osteoporosi, arterosclerosi. Questa strumentazione offre quindi un ottimo mezzo per cinque principali obbiettivi biologici:

1. l'identificazione di geni con livelli di espressione diversa sotto diverse condizioni sperimentali o tra soggetti che presentano varie forme della stessa patologia;
2. l'identificazione di gruppi di geni che con buona probabilità sono correlati tra loro;
3. la caratterizzazione genomica della cellula malata attraverso la classificazione di campi biologici (soggetti sani vs soggetti affetti da una determinata patologia);
4. l'identificazione di geni il cui valore di espressione è biologicamente utile per determinare un particolare gruppo o fenotipo (tali geni sono detti marcatori);

L'identificazione di mutazioni è fondamentale per la prevenzione delle malattie genetiche, per la diagnostica precoce dei tumori, nonché in microbiologia per la identificazione di ceppi batterici o virali. Un altro settore di applicazione è quello dell'analisi funzionale simultanea di decine di migliaia di geni e, in un futuro prossimo, di tutti i geni che costituiscono il nostro patrimonio genetico. Inoltre i risultati che ci si aspetta di ottenere con questa nuova tecnologia saranno fondamentali per sviluppare nuovi farmaci, e per meglio utilizzare quelli attualmente disponibili dando al medico la possibilità di adattare la terapia sulla base delle caratteristiche genetiche di ognuno di noi. Gli studi sul cancro costituisco una delle maggiori aree di ricerca in campo medico. Un'accurata distinzione dei diversi tipi di tumori ha un'importanza fondamentale nel fornire trattamenti più mirati e rendere minima l'esposizione del paziente alla tossicità delle terapie. Fino a poco tempo fa, la classificazione delle varie forme di cancro ha sempre avuto basi morfologiche e cliniche; i metodi convenzionali però soffrono di diverse limitazioni soprattutto per quanto riguarda la capacità diagnostica. E' stato recentemente suggerito che la differenziazione dei trattamenti in accordo con la differenziazione dei tipi di cancro, potrebbe avere un impatto positivo sull'efficacia della terapia, infatti le diverse classi tumorali presentano caratteristiche molecolari differenti oltre a distinti decorsi clinici.
Il livello di espressione genica contiene la chiave per affrontare problemi legati alla

prevenzione e alla cura di alcune malattie, per comprendere i meccanismi di evoluzione biologica e per scoprire adeguati trattamenti farmacologici. Il recente avvento della tecnologia del DNA ha permesso di manipolare simultaneamente migliaia di geni, motivando lo sviluppo della classificazione di tumore con l'utilizzo dei dati d'espressione genica.
Nel presente elaborato ci si propone un studio di alcuni metodi per l'individuazione di geni differenzialmente espressi effettuato mediante uno studio di simulazione. Lo scopo sarà pertanto quello di individuare il metodo efficiente per distinguere i geni differenzialmente espressi ("*controlli-tessuti cancerogeni*"), saranno utilizzati dati simulati e diverse numerosità campionarie.
In letteratura sono stati applicati diversi metodi per la distinzione per esempio, di diverse forme di cancro, ma esistono alcuni problemi che rendono questo compito tutt'altro che banale quando effettuato con dati di espressione genica. I dati di espressione genica sono infatti diversi da quelli con cui è abituato a trattare normalmente lo statistico. A questo proposito, Gordon K. Smith *et al.* Rif[l] hanno stilato una rassegna dei problemi statistici.

Il primo problema che pone il *dataset* è la grande dimensionalità (qualche migliaia di geni) a cui si contrappongono campioni molto limitati (di solito una centinaia di unità). In letteratura si fa riferimento a questo problema con l'espressione *"large p and small n"*. Oltre a comportare spesso tempi di elaborazione piuttosto lunghi, questa caratteristica espone al rischio di sovraparametrizzazione del modello tanto per l'alta dimensionalità quanto per la limitata numerosità campionaria. In secondo luogo la maggior parte dei geni nel *dataset* sono irrilevanti ai fini della classificazione e costituiscono un *rumore* che interferisce con il potere discriminante degli altri geni. Questo accresce non solo i tempi di calcolo, ma anche la difficoltà di classificazione. E' evidente che i metodi di classificazione esistenti non sono concepiti per essere applicati a questo tipo di dati. Alcuni ricercatori propongono il raggruppamento di geni in classi omogenee come operazione preliminare alla classificazione dei soggetti, in quanto tale operazione ha la capacità di ridurre la dimensianalità , i tempi di calcolo ed eliminare i geni irrilevanti che comportano minor accuratezza nella classificazione. Una terza questione riguarda la natura stessa dei dati che sono caratterizzati dalla massiccia presenza di rumore di tipo *biologico* o *tecnico.*

I problemi sopra menzionati riguardano l'ambito statistico, ma esistono diverse questioni derivanti dal contesto biologico e dall'importanza dei risultati in campo medico. Una questione riguarda la corrispondenza tra rilevanza biologica e statistica di uno stesso gene differenzialmente espresso come classificatore: la rilevanza biologica è un criterio da tenere in forte considerazione in quanto ogni informazione rilevata durante l'analisi può essere utile per

la scoperta delle funzioni specifiche di un certo gene, per la determinazione di gruppi di geni che concorrono allo sviluppo di cellule o tessuti cancerogeni, per la scoperta di interazione tra i geni o per altri studi biologici come l'individuazione di geni marcatori. Nel presente lavoro, verranno confrontati tre metodi: uno parametrico, uno semi-parametrico e uno completamente non-parametrico.

Capitolo I

I Microarray

1.1 *Breve rassegna delle scoperte riguardanti la molecola di DNA*

Ogni essere vivente possiede un programma genetico, cioè un insieme di istruzioni che specificano le sue caratteristiche e dirigono le sue attività metaboliche. Questo insieme di istruzioni costituisce l'informazione biologica, cioè è ereditaria ed è trasferita da una generazione all'altra attraverso la riproduzione. Le caratteristiche trasmesse sono dette caratteri ereditari.

L'informazione biologica è organizzata in unità fondamentali, dette geni, ciascuna delle quali interviene nella determinazione di un carattere ed è ereditata dai genitori.

Già con le prime ipotesi riguardanti l'evoluzione, si era cercato di comprendere come i caratteri peculiari di un organismo venissero trasmessi e come le specie evolvessero. Alcuni, come il noto scienziato Lamarck (1787), avevano ipotizzato che i caratteri acquisiti durante la vita fossero trasmissibili di padre in figlio: è il caso del famoso esempio della giraffa e del suo collo. Molte furono però le critiche rivolte a questa teoria, dovute al fatto che molti caratteri acquisiti durante la vita non sono ereditari.

Il primo a dedicarsi con metodo scientifico allo studio dell'ereditarietà dei caratteri fu un abate austriaco Gregor Mendel, Bateson W. (1909). A quei tempi Mendel, non aveva nessuna conoscenza della struttura intrinseca del DNA, tuttavia aveva intuito alcuni caratteri che ricomparivano regolarmente nelle popolazioni. Le regole fondamentali che mettevano in connessione questi eventi non erano ancora chiare. I primi esperimenti che Mendel condusse furono su piante di piselli odorosi caratterizzate dalla capacità di effettuare l'autofecondazione e caratterizzate da cicli vitali non troppo lunghi.

Le ricerche di Mendel non furono prese immediatamente in considerazione, ma le basi della genetica erano comunque scoperte, senza avere ancora idea di come fosse strutturato il DNA. La prova decisiva che il depositario dell'informazione è il DNA fu fornita nel 1952 da AD. Harshey e M. Chase, i quali dimostrarono che i batteriofagi, per introdurre nel batterio ospite il loro materiale ereditario, iniettano una molecola di DNA. L'importante scoperta fatta sul DNA suscitò la curiosità degli scienziati sulla struttura di tale molecola. Agli inizi degli anni '50, un giovane scienziato americano, James Watson, si recò a Cambridge, in Inghilterra, con una borsa di studio per lavorare sui problemi di struttura molecolare e, al Cavendish

Laboratory, incontrò il fisico Francis Crick. Entrambi si interessavano di DNA e ben presto cominciarono a lavorare insieme per cercare di capire come fosse strutturata tale molecola. Essi non eseguirono veri e propri esperimenti, ma intrapresero, piuttosto, un esame razionale dei dati allora noti sul DNA, cercando di organizzarli in modo logico. Le informazioni che essi avevano su tale molecola riguardavano le sue grosse dimensioni e la sua struttura lunga e filiforme formata da nucleotidi. Inoltre nel 1950 Linus Pauling aveva dimostrato che le proteine sono spesso disposte in maniera elicoide e vengono mantenuti in questa disposizione da legami idrogeno-idrogeno che si formano sulle spire adiacenti all'elica. Questa dimostrazione risultò utile ai fini della ricerca in quanto la molecola di DNA si comporta in modo simile alla molecola delle proteine.

Studi intrapresi da Maurice Wilkins e Rosalind Frankling ai raggi X dimostrarono la forma a grande elica del DNA. Infine Chargaff verificò l'esattezza di due proporzioni che dimostravano l'impossibilità di legare chimicamente due basi purine (a due anelli) o due basi pirimidine (ad una unico anello), e quindi all'assunzione che la timina si può legare solamente alla adenina, e la citosina solamente alla guanina.

L'insieme di tutte queste scoperte portò la formulazione della struttura definitiva della molecola di DNA: doppia elica lunga e spiralizzata in cui le due spirali sono formate da molecole alternate di zucchero e di fosfato e vengono tenute insieme da una coppia di basi azotate (ogni base è legata in modo covalente alla subunità glucidica posta nel tratto montante adiacente ad essa).

Sulla molecola di DNA sono state formulate numerose ipotesi e nel corso degli anni si è scoperto quasi con completezza come essa trasferisce l'informazione biologica da un individuo all'altro.

Numerosi scienziati si sono occupati e si occupano di individuare le particolarità del DNA tramite lo studio descrittivo dei geni e la loro classificazione.

L'applicazione più interessante è nello studio di malattie: molti studiosi concordano oramai da tempo sulla teoria secondo la quale alcune patologie derivino da piccole alterazioni del codice genetico. Ciò che distingue un individuo sano da un malato, sono delle differenze nell' espressione dei geni, ossia nel modo con cui essi sono utilizzati e nelle proteine a cui danno origine. La disciplina metodologica che si occupa di questi problemi è la biostatistica la quale assiste il ricercatore biologo nel disegno e nella valutazione probabilistica di variazione di espressione genetica. Il problema è quello di caratterizzare le anomalie genetiche della cellula malata, ossia ciò che la differenzia da quella sana, in modo tale che una volta noto il profilo genetico di un paziente, risulti possibile identificarlo come sano o affetto da malattia.

1.2 *Alcune nozioni di biologia*

Per comprender meglio la trattazione della fase sperimentale, è opportuno fissare alcuni concetti base di biologia molecolare.
Le cellule sono le unità funzionali e strutturali biologiche di base. Sono separate dall'ambiente esterno da una membrana che, oltre a garantire l'integrità funzionale della cellula, regola il passaggio delle sostanze dall'interno verso l'esterno e viceversa.
All'interno si trova il citoplasma, una soluzione acquosa concentrata, attraversata e suddivisa da un elaborato sistema di membrane, il reticolo *endoplasmatico,* e contenente enzimi, ioni e molecole disciolte oltre ad un certo numero di organuli con funzioni specifiche. Tra questi organuli rivestono particolare interesse i *ribosomi* che sono i siti in cui ha luogo l'assemblaggio e la sintesi proteica.
Essi possono ricoprire il *reticolo endoplasmatico* oppure trovarsi liberi nel citoplasma. Oltre ai ribosomi, nel citoplasma hanno sede anche i *mitocondri,* in cui avvengono le reazioni chimiche che forniscono energia per le attività cellulari, *l'apparato di Golgi,* dove sono immagazzinate le molecole sintetizzate nella cellula, i *lisosomi* e i *perossisomi,* che sono delle vescicole in cui le molecole vengono scomposte in elementi più semplici che possono essere usati dalla cellula oppure eliminati. Il citoplasma è inoltre fornito di un *citoscheletro,* che determina la forma della cellula, le consente di muoversi e fissa i suoi organuli.
Ma la struttura più grossa ed importante presente nella cellula è il nucleo, che interagendo con il citoplasma, aiuta a regolare le attività che si svolgono nella cellula. All'interno dell'involucro nucleare, formato da una doppia membrana, ha sede il *nucleolo,* il sito di formazione delle subunità ribosomiali nonché della *cromatina,* sostanza formata da un complesso di proteine e di DNA. Essa è la sostanza costituitiva dei cromosomi, è presente in tutto il nucleo e prende questo nome quando si trova in forma disciolta. Il DNA (acido deossiribonucleico) è una lunga molecola costituita da due filamenti avvolti l'uno sull'altro e uniti da ponti infinitesimali detti *ponti idrogeno.* I due filamenti sono costituiti da subunità ripetute di un gruppo fosfato e dello zucchero deossiribosio a cinque atomi di carbonio, mentre i ponti sono formati da una coppia di basi azotate. Uno zucchero deossiribosio, un gruppo fosfato e una base azotata costituiscono un *nucleotide.*

Esistono quattro tipi di basi: *adenina, timina*, *citosina, guanina* ed hanno la caratteristica di accoppiarsi sempre nello stesso modo, adenina con timina e citosina con guanina. Esse sono una sorta di alfabeto con il quale viene scandito il messaggio genetico: a seconda di come si

presentano e si organizzano le triplette, si ha la formazione di un particolare gene, che è per l'appunto un segmento di DNA in grado di trasmettere messaggi per la sintesi delle proteine ed altre sequenze regolative. Quando una molecola di DNA si duplica, i due filamenti si separano grazie alla rottura dei legami idrogeno e ciascuno, con le proprie basi azotate, funge da stampo per la formazione di un nuovo filamento complementare. E' così che l'informazione ereditaria si trasmette fedelmente da una cellula madre alla cellula figlia in quella che viene detta *duplicazione semiconservativa.*

La sequenza dei nucleotidi presenti nella molecola di DNA determina una sequenza degli amminoacidi, ossia delle subunità necessarie per la sintesi proteica: una serie di tre nucleotidi (detta *codone)* codifica per un aminoacido.

Il processo secondo la quale il DNA viene tradotto in proteine consta in due fasi fondamentali: *trascrizione* e *traduzione.* Durante la prima fase, l'informazione viene *trascritta* da un filamento singolo di DNA in un filamento singolo di RNA detto messaggero o rnRNA. L'RNA messaggero è una molecola del tutto simile al DNA, la sola differenza è che al posto della timina si trova un'altra base azotata: *l'uracile.* Una volta trascritto, l'rnRNA esce dal nucleo e si sposta sui ribosomi, dove ha luogo la sintesi proteica o *traduzione.* I ribosomi sono costituiti da subunità formate da RNA ribosomiale e proteine (o rRNA). A questo punto interviene l'RNA di trasporto (o tRNA), una molecola che assume la forma di un trifoglio e provvede al trasporto degli aminoacidi. Il tRNA è munito di una tripletta di basi, detta *anticodone,* specifica per l'aminoacido che trasporta. Durante la sintesi, il tRNA mette in corrispondenza ciascuna tripletta di basi *(codone)* dell'rnRNA con il suo anticodone, in modo che ogni molecola di tRNA apporti l'aminoacido specifico relativo al codone dell' rnRN A a cui si attacca. In questo modo, in base allea sequenza dettata inizialmente dal DNA, le unità aminoacidiche vengono allineate una dopo l'altra andando ad assemblare la catena polipeptidica ossia la *proteina.* Le mutazioni non sono altro che cambiamenti nella sequenza o nel numero di nucleotidi nell' acido nucleico della cellula, dovuti all'aggiunta, alla delezione o alla sostituzione di un nucleotide con un altro. Molte malattie genetiche sono il risultato della mancanza o inattività di enzimi o altre proteine. Queste, a loro volta, sono provocate da mutazione dei geni che codificano per tali proteine.

Per comprendere la tecnica dei *microarray* chip è fondamentale notare che, nella fase di trascrizione, ciascuna cellula produce RNA solamente per quei geni (ossia quei segmenti di DNA) che sono attivi in quel momento; pertanto un modo per indagare quali sono i geni attivi e quali quelli inattivi in un determinato istante sarà quello di analizzare l'RNA prodotto dalla cellula, ed è da questo punto che parte l'intuizione della *DNA microarray tecnology.*

1.3 *Le nuove tecnologie: i microarray*

Le nuove tecnologie di studio del DNA, *microarray,* descritte per la prima volta nel 1995, stanno rapidamente trovando applicazione in molti ambiti di ricerca, che vanno dalla fisiologia cellulare, all'oncologia, alla farmacogenomica. Numerose sono anche le applicazioni di questa tecnologia in ambito microbico-virologico, come la genotipizzazione e lo studio della biologia dei microrganismi e delle interazioni ospite-patogeno. Il sistema di indagine basato sui *microarray* permette di misurare contemporaneamente molte sequenze diverse, e quindi di analizzare l'intero patrimonio genetico di diversi organismi.
Attualmente sono state sviluppate due principali piattaforme tecnologiche per la produzione dei *microarray:*

- *Microarray di cloni di DNA micropipettati:* diversi singoli geni vengono depositati in anticipo sui vetrini opportunamente trattati con agenti chimici che favoriscono il legame del DNA utilizzando apparecchiature automatizzate;
- *Microarray di oligonucleoti disintetizzati in situ:* utilizza chip di silicio su cui sono direttamente sintetizzati oligonucleotidi rappresentativi della sequenza bersaglio.

I vantaggi dei primi sono l'alta automazione delle procedure sperimentali, una elevata riproducibilità dovuta al costo relativamente basso e il fatto che non sia necessario conoscere la sequenza del DNA da stampare; mentre i vantaggi dei microarray di oligonucleotidi sono l'alta densità e l'opportunità di disegnare la sequenza bersaglio dall'utente e quindi adattato alle diverse situazioni sperimentali.
Il *microarray* consente di verificare quanti e quali geni sono attivi in un tipo cellulare o in un tessuto, qual è il loro livello di espressione e quali variazioni accadono in condizioni patologiche. In tal modo e' possibile identificare i geni con potenziale attività oncogena che sono attivi nelle cellule tumorali di un paziente rispetto ad un altro, o rispetto al tessuto normale. Allo stesso modo si possono valutare quali geni differenziano il tumore primario dalla relativa metastasi.
Tutto ciò, oltre a costituire un ulteriore approccio sperimentale per l'identificazione di geni collegati al fenomeno della trasformazione e progressione neoplastica, ha permesso di classificare i tumori in base ai loro profili di espressione e di preparare lo sviluppo di una tassonornia molecolare, dei tumori che consenta di complementare, aggiungendo nuove e più rilevanti informazioni, quella tradizionale di tipo isto-morfologico.

Le innovazioni che hanno reso possibile la tecnologia dei *microarray* sono l'uso di supporti solidi non porosi come vetro, molto versatile ai fini della miniaturizzazione e dell' individuazione dei marcatori fluorescente, e la sintesi ad alta densità spaziale di oligonucleotidi su vetrini sottilissimi con tecniche che utilizzano maschere fotolitografiche, impiegate nell'industria dei semiconduttori. Tra le numerose applicazioni della tecnologia dei microarray, le principali sono l'analisi su larga scala dell'espressione genetica e la ricerca di variazioni della sequenza del DNA.
La tecnologia basata sui *microarray,* rappresenta un mezzo di indagine straordinariamente innovativo, in quanto permette di analizzare con un singolo esperimento l'intero patrimonio genetico di un organismo.

1.3.1 *Come funziona l'esperimento*

La realizzazione del *microarray* consta in due fasi: la preparazione del *microchip* e quella del *target.* Ad un vetrino *(microchip)* si fissano delle sonde *(probe)* costituiti da segmenti di cDNA sintetico che riproducono i geni che in qualche modo sono notoriamente correlati con la patologia oggetto di studio. A questo scopo esistono speciali robot in grado di dispensare goccioline dell' ordine di nano litri attraverso tubi con punte eccezionalmente sottili.
Per preparare il *target,* si estrae l*'mRNA* totale prodotto dai due tipi di cellule in analisi. Per mezzo di una reazione biochimica l*'mRNA* viene retrotrascitto dando luogo al *cDNA* che, come ricordato precedentemente, presenta una molecola più stabile *dell'mRNA.* Durante questa fase nella catena di *cDNA* di ciascun gene vengono introdotte particolari molecole dette recettori in grado di legarsi a sostanze fluorescenti. Successivamente il *cDNA* dei due tipi di cellule viene etichettato con due colori (rosso e verde) mediante dei marcatori fluorescenti che vanno a legarsi ai ricettori: Cy3 (verde) per cellule sane e Cy5 (rosso) per quelle malate. Infine il *cDNA* delle due cellule viene mescolato e depositato *sull'array* affinché possa ibridizzare con le sonde. Durante 1'ibridazione i segmenti di *cDNA* target riconoscono le sonde complementari e si legano ad esse.
Una volta completata 1'ibridazione il *microchip* viene levato e successivamente eccitato con un laser affinché i marcatori fluorescenti emettano un segnale luminoso. Una -specie di *scanner* legge *1'array* illuminando ciascuno *spot* (ossia ciascun puntino che rappresenta un singolo gene) e misurando la fluorescenza emessa per ciascun colore separatamente, in modo da fornire una misura della quantità relativa di *mRNA* prodotto da ciascun gene nei due tipi di cellula. L'intensità degli spot verdi misura la quantità di *cDNA* contrassegnato con Cy3, e

quindi *mRNA* prodotto da cellule sane; mentre quella degli spot rossi misura la quantità relativa di *cDNA* contrassegnato con Cy5, e quindi di *mRNA* prodotto da cellule malate. Queste misure forniscono informazioni sul livello relativo d'espressione di ciascun gene nelle due cellule. Le due immagini monocromatiche (rossa e verde) vengono poi sovrapposte in modo da fornire una visione d'insieme: ciascuno spot corrisponde ad un gene ed il colore alla sua condizione nella cellula malata o in quella sana. Così il rosso corrisponde ad un gene molto attivo nella cellula malata e inattivo in quella sana, il nero ad un gene inattivo in entrambe le cellule, il giallo ad un gene ugualmente attivo nei due tipi di cellula, ed infine il verde ad un gene attivo nella cellula sana e inattivo in quella malata.

E' necessario che queste misure vengano aggiustate per considerare un disturbo di fondo causato, ad esempio, dall' alta concentrazione di sale e detergente durante l'ibridazione o la contaminazione del target o da altri problemi che si possono presentare nell'esecuzione dell' esperimento.

L'ibridazione del target alle sonde determina una reazione chimica che viene catturata in un' immagine digitale da uno scanner laser. Il passo successivo è quello di tradurre l'intensità del segnale luminoso emesso da ciascun gene, in un coefficiente numerico. S'intuisce pertanto l'importanza della qualità dell'immagine ai fini di un'accurata interpretazione dei dati. I passi principali delle immagini prodotte da *cDNA microarray* sono:

1. grigliatura *(gridding)*
2. estrazione di intensità
3. segmentazione

La grigliatura ritrova nell'immagine la posizione degli spot che corrispondono alle sonde. Essendo nota la posizione degli spot nel microarray, questa operazione non risulta particolarmente complessa, sebbene si renda necessaria la stima di alcuni parametri per tener conto ad esempio di *shift* (o rotazioni) del *microarray* nell'immagine o di piccole traslazioni degli spot.

L'estrazione di intensità calcola invece l'intensità della fluorescenza rossa e verde, l'intensità del background ed alcune misure di qualità.

La segmentazione consiste infine nel separare il segnale emesso dai marcatori fluorescenti *(foreground)* rispetto al disturbo di fondo *(background),* in modo da isolare le quantità di interesse. Può succedere che questa correzione abbia l'effetto indesiderato di introdurre valori negativi(ciò accade quando l'intensità del background è più forte rispetto a quella di foreground). In tal caso questi spot vengono trascurati oppure il loro segnale è sostituito, con un valore arbitrariamente piccolo e positivo.

1.3.2 *Tipi di distorsioni*

Al fine di rendere comparabili i risultati ottenuti su array diversi o anche all'interno dello stesso *array,* è necessaria la rimozione di alcune distorsioni sistematiche introdotte nella fase di preparazione dell' *array* stesso, di esecuzione dell'esperimento, nonché nel processo di ibridizzazione e nella scansione con il laser. La procedura di normalizzazione si riferisce proprio al trattamento statistico dei dati finalizzato alla rimozione di tali effetti distorsivi e i più noti sono:

1. *dye-effect* (o effetto colore);
2. *print-tip* (o deposito irregolare);
3. *array-effect(o* effetto intensità).

Ad esempio, un diffuso problema nell'interpretazione dei dati derivanti da *microarray,* noto come *dye-effect,* è la diversa intensità di fluorescenza dei due marcatori Cy3 (verde) e Cy5 (rosso), cosicché l'emissione di fluorescenza del verde è sistematicamente meno intensa di quella del rosso. Il modo più immediato per rimuovere questo tipo di distorsione, sarebbe quello di ripetere due volte l'esperimento scambiando l'assegnazione dei marcatori tra i due target, cosa che però renderebbe la tecnica ancora più dispendiosa.

Un'altra fonte di distorsione, nota come *print-tip,* è dovuta alla diversa quantità di materiale genetico (probe) depositata sul vetrino a causa delle microscopiche differenze della conformazione delle puntine del rabor che stampa *1'array.*

Infine, il terzo tipo di alterazione, *l'array-effect* può derivare da differenze di intensità tra un *array* e l'altro legate a diverse condizioni di preparazione (usura delle puntine, qualità di conservazione e quantità dei reagenti), estrazione (differenti quantità di mRN A usate per creare il target o quantità di marcatore fluorescente), ibridizzazione *(coss-ibridation)* e scansione (bilanciamenti dei laser, diversi parametri di scansione).

Ai problemi sopra esposti si cerca di dare soluzione mediante il processo di normalizzazione. La normalizzazione prevede che si calcolino fattori di standardizzazione per ciascuno dei tre effetti sopra menzionati. Si tratta di sottrarre al segnale una (i) media generale di *array,* la (ii) differenza tra le medie degli spot stampati da ciascun *print-tip* e la media generale, ed infine la (iii) differenza tra la media delle intensità con fluorescenza rossa e verde.

Anzitutto il ricercatore deve scegliere quali geni usare nel processo di standardizzazione. Questa decisione è influenzata da alcune considerazioni come la proporzione attesa di geni differenziaI mente espressi e la possibilità di controllare le sequenze di DNA. Tre gli approcci

principali. Il primo si fonda sull'assunzione che solo una piccola parte dei geni sia differenzialmente espressa. I restanti geni hanno pertanto un livello di espressione costante e possono essere usati come indicatori dell'intensità relativa ai due colori. In alti termini, quasi tutti i geni dell'array possono essere utilizzati per la normalizzazione quando si può ragionevolmente assumere che solo una piccola porzione di essi vari significativamente la propria espressione da un campione all'altro, oppure che esista simmetria nei livelli di espressione dei geni sovra e sotto espressi. In pratica è però molto difficile trovare un gruppo di spot con un segnale costante su cui trarre un fattore di correzione. Si preferisce quindi, quando il numero di geni differenzialmente espressi è limitato rispetto al numero totale dei geni indagati, usare tutti gli spot dell' array nel processo di normalizzazione dei dati. Il secondo approccio si basa sull' assunto che da proporzione di geni differenzialmente espressi sia un' altra e quindi suggerisce l'uso della restante porzione *(housekeeping genes)* che si crede abbia un livello di espressione costante nelle due condizioni. Questa piccola porzione di geni però, oltre ad essere difficilmente identificabile, spesso risulta poco rappresentativa rispetto ai geni di interesse essendo costituita per lo più da geni con alto livello di espressione. Il terzo approccio necessita dell' appoggio del laboratorio e prevede di realizzare un microarray per un solo campione di mRNA (prelevato da un'unica cellula) diviso in due porzioni uguali, ciascuna marcata con colori differenti. Trattandosi dello stesso campione di materiale genetico, in seguito all'ibridizzazione si dovrebbe avere la stessa intensità degli spot per il rosso e per il verde: eventuali differenze possono essere usate come fattore di normalizzazione.

Un altro trattamento dei dati preliminare all'analisi è la cosiddetta filtrazione. Essa è finalizzata alla riduzione della variabilità e della dimensionalità dei dati. Il primo obiettivo viene raggiunto rimuovendo quei geni le cui misure non sono sufficientemente accurate, il secondo con 1'imitazione dei geni che prevedono un livello di espressione molto piccolo o negativo (prima o dopo la normalizzazione) .

In pratica, tutti gli spot la cui differenza tra l'intensità di foreground e quella di background non supera un valore soglia di 1.4 fold (una misura dell'intensità luminosa) vengono eliminati o sostituiti con un valore piccolo arbitrario. Questa procedura è giustificata dall' evidenza empirica che livelli di espressione più piccoli di 1.4 fold sono solitamente frutto di errori di misura. Si noti che qualsiasi operazione di filtrazione introduce arbitrari età nella scelta delle soglie che determinano se un valore è troppo grande o troppo piccolo oppure se la variabilità delle misure è troppo elevata.

1.4 *Rassegna bibliografica*

La letteratura non è molto datata e tende a svilupparsi a pari passo con le scoperte in ambito biologico.

Nel 2001 Rocke e Durbin rif [2] introducono un modello di misura per gli errori dei dati da *microarray* come funzione del livello di espressione dei geni.

Nel 1999 Lausen rif [3] si concentra sulle misure di distanza allineando sequenze di dati secondo diversi criteri, propone poi un grafico *(dot-matrix plot)* come possibile test sulla bontà dell'allineamento. Nello stesso anno Golub *et al* rif [4] applicano su un campione di dati derivanti da leucemie di tipo acuto l'analisi *cluster* e l'analisi discriminante. Jean Clavarie rif [5] rivede invece l'approccio teorico e computazionele utilizzato fino ad allora per identificare i geni differenzialmente espressi, per selezionare geni co-regolati attraverso un insieme di condizioni e per creare *cluster* di geni che raggruppino in modo coerente caratteristiche di espressione simili. Nell'ottobre dello stesso anno Gloub *et al* rif [6] applicano due procedure di classificazione *(class discovery* e *class prediction)* per distinguere diversi tipi di cancro per leucemie acute. Platt rif [7]

mette appunto la *"sequential minimal optimisation"* che permette l'implementazione delle *support vector machines* (SVM) per affrontare problemi di classificazione che coinvolgono grandi *dataset.*

L'anno successivo Brown *et al* rif [8] testano diverse SVM usando varie misure di sorveglianza su dati da microarray trovando le SVM garantiscono prestazioni migliori rispetto ad altre tecniche nel riconoscere geni coinvolti nelle comuni funzioni biologiche. Ben Dor (2000) *et al* rif [9] descrivono un'applicazione di SVM con nuclei lineare e quadratico che ha classificato con successo tessuti normali e tumorali del colono Alizadeh *et al* sempre nel 2000 rif [lO] analizzano *dataset* sul cancro ed usano regole di raggruppamento gerarchico per studiare l'espressione genetica nelle tre prevalenti forme di tumore linfoide che colpisce gli adulti. Golub *et al* (2000) rif [Il] partendo da un campione di 6817 geni e 38 pazienti creano una regola per distinguere tra leucemie ALL ed AML formando dei *cluster* in cui raggruppano geni simili. Veer *et al* (2000) rif [12] studiano un *dataset* di 78 pazienti con il cancro al seno. Partendo da 5000 geni si restringono a 231 esaminando il coefficiente di correlazione di ciascun gene con il risultato della prognosi. Sempre nello stesso anno Bem-Dor *et* al rif [13] verificano che mentre le SVM hanno maggiore accuratezza sui dati da leucemie e i metodi basati sul *clustering* funzionano meglio su dati di tumori al colon, il metodo *nearest neighbor* dà buoni risultati in entrambi i casi. Keller *et* al (2000) rif [14]

comparano il metodo bayesiano semplice con il metodo *weighted voting* e nell'agosto dello stesso anno presentano il primo metodo per la classificazione di tipi di tessuto con dati da microarray usando una tecnica basata sulla massima verosimiglianza per selezionare i geni più utili alla classificazione.

Applicando questa tecnica ad un *dataset* con due tipi di tessuti riscontrano un'eccezionale accuratezza e fanno notare che è facilmente estendibile ad una classificazione con più di due classi fornendo ottimi risultati se applicati a *dataset* con tre tipi di tessuto. Gen Rori *et al* (2000) rif [15] dimostrano l'applicazione del metodo ICA *(indipendent component analysis)* che è in grado di classificare un vasto insieme di dati di espressione genica in gruppi significativi dal punto di vista biologico. In particolare dimostrano che geni la cui espressione è campionata a diversi istanti temporali possono essere classificati in gruppi differenti e che questi gruppi hanno una buona somiglianza con quelli che si determinano solo sulla base delle conoscenze biologiche; questo suggerisce anche che il metodo ICA può essere uno strumento potente per la scoperta di ignote funzioni biologiche dei geni. L'anno successivo Zhang *et al* rif e Kerr *et al* usano il metodo *boatstrap* per valutare la qualità dell' analisi di raggruppamento i primi assumendo che i livelli di espressione hanno distribuzione normale, i secondi usando il modello ANOVA per generare campioni *bootsrapped.*

Nel maggio del 2001 David B., Allison *et al* sviluppano una sequenza di procedure che comprendono modelli misti e inferenza *bootstrap* per affrontare problemi (come *large p and small n)* che sorgono nel trattamento dei dati che coinvolgono l'espressione di migliaia di geni. Nel luglio dello stesso anno Lorenz Wernisch rif propone una rassegna dei principali metodi di trattamento dei dati da microarray. Tibshirani *et al* (2001) propongono una quantità per la stima del numero di *cluster* in un *dataset:* tale quantità suggerisce quanti gruppi devono essere formati e quanto affidabile è la previsione. Nel corso dell' analisi sviluppano anche una nuova nozione di distorsione e di varianza per dati senza variabile risposta.

Nel 2002 Chris Fraley e Adrian E. Raftlery rif rivedono una metodologia generale dell'analisi di raggruppamento che fornisce un approccio statistico a problemi come il numero di *cluster* da formare, il trattamento dei dati anomali *(oulyers),* il tipo di legame da usare ecc. Dimostrano anche che questa metodologia può essere utile nei problemi di analisi multivariata come l'analisi discriminante o la stima di densità multivariate. Sempre nello stesso anno Gengxin Che *et al* applicano diversi algoritmi di analisi di raggruppamento su un *dataset* di espressioni geniche di cellule embrionali. Propongono diversi indici basati sull' omogeneità interna, sulla separabilità, sulle *silouette,* sui geni in eccedenza in un dato gruppo ecc. I risultati dimostrano che il dataste pone effettivamente dei problemi per l'analisi *cluster,* gli

autori valutano vantaggi e svantaggi dei vari algoritmi. Lo studio fornisce quindi una linea generale su come scegliere tra diversi algoritmi e può aiutare ad estrarre dal *dataset* le informazioni biologiche più significative.

Nel febbraio del 2003 Romualdi, Campanaro *et al* comparano diverse tecniche di *supervised clustering* sulla base della capacità di classificare correttamente diversi tipi di cancro usando inizialmente l'approccio della simulazione per controllare la grande variabilità tra ed entro i pazienti. Mettono a confronto diverse tecniche di riduzione della dimensionalità che andranno poi ad aggiungersi all'analisi discriminante e verranno comparate sulla base della loro capacità di catturare l'informazione genetica principale. I risultati della simulazione sono poi stati vagliati applicando gli algoritmi a due *dataset* di espressioni geni che di pazienti malati di cancro, misurando il corrispondente tasso di errata classificazione. Nel marzo dello stesso anno Erich Hungan *et al* analizzano un campione di 89 pazienti con tumore della mammella usando tecniche non lineari allo scopo di mettere in luce modelli d' interazioni di gruppi di geni che hanno valore predittivo per singoli pazienti relativamente alla presenza di linfonodi con metastasi e ricaduta nella malattia. Trovano dei *pattern* in grado di fare previsioni con accuratezza del 90%. Nell'aprile del 2003 Michael O'Neil e Li Song studiano le reti artificiali applicate su dati da microarray da pazienti con linfoma di tipo DLCL e, per la prima volta, prevedono con accuratezza del 100% il tempo di sopravvivenza, restringendo il profilo genico a meno di tre dozzine di geni per ogni differenziazione espressa. Identificano le reti artificiali come miglior strumento sia per l'individuazione di gruppi di geni sia per evidenziare i geni più importanti che producono una corretta differenziazione.

Capitolo II

2. Metodi statistici

L'analisi dei dati forniti dagli esperimenti di microarray può essere condotta utilizzando tecniche statistiche note, sviluppate in altri contesti. E tuttavia, necessario porre attenzione alle peculiarità dei dati, e individuare opportuno aggiustamenti. Si possono utilizzare diverse metodologie, per identificare geni differenzialmente espressi (si veda esempio, Dudoit et al, 2002), qui ci si concentra sulla verifica d'ipotesi. Per ogni gene si dispone di disporre di un campione di livelli di espressione per *m* casi e di un campione di livelli di espressione di *n* controlli. Si vuole effettuare un test di uguaglianza delle medie di espressione nelle due popolazioni. Vari studiosi hanno affrontato l'argomento. Vediamo di seguito gli approcci addottati, in particolare prestando attenzione ai modelli proposti, ai test e alle distribuzioni delle statistiche utilizzate.

2.1 Il test t-Student

Per un fissato gene, sia $y = (y_1,...,y_m)$ il campione osservato di livelli di espressione genica per gli *m* casi e sia $x = (x_1,...,x_n)$ il campione relativo agli controlli si supponga inoltre che *y* sia un campione casuale semplice (c.c.s.) da una distribuzione normale di parametri μ_y e σ_y^2 (N(μ_y, σ_y^2)) e *x* sia un c.c.s. da una distribuzione normale $N(\mu_x, \sigma_x^2)$.

Ammettiamo di voler accertare se i livelli di espressione medi nelle 2 popolazioni μ_x e μ_y da cui i campioni sono stati estratti differiscono o meno. Tale ipotesi si traduce nell'ipotesi statistica

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases}$$

che viene verificata utilizzando la statistica test

$$t = \frac{\bar{y} - \bar{x}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

con $\bar{y} = \frac{\sum_{i=1}^{m} y_i}{m}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $s = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \bar{y}) + \sum_{i=1}^{n}(x_i - \bar{x})}{n + m - 2}}$

Sotto H_0 , la statistica test si distribuisce come una *t-Student* con $(m+n-2)$ gradi di liberta.

A seconda dei gradi di liberta ed il valore α di probabilità di errore del I tipo prescelto, in un tabulato sono forniti i quantili della statistica test t-student che consentono di discriminare fra l'accettazione e rifiuto dell'ipotesi nulla. Se di *t* osservato è maggiore di quello tabulato l'ipotesi nulla viene respinta, e si afferma che non vi è una differenza significativa tra le medie dei gruppi . Se viceversa il *t* osservato è inferiore a quello tabulato, non vi sono elementi sufficienti per respingere l'ipotesi nulla, e si afferma che non vi e una differenza statisticamente significativa fra le due medie in esame.

Una condizione necessaria per la validità del t-Student è che la variabile in esame nelle due popolazioni, segua una distribuzione normale. Ricordiamo comunque che la distribuzione di t e approssimativamente valide anche per marcate deviazioni della normalità.

2.2 Il modello semiparametrico e le sue ipotesi

Siano X ed Y due variabili casuali per le quali si dispone dei due campioni, $x = (x_1, ..., x_n)$ ed $y = (y_1, ..., y_m)$ con n>m. Si supponga che la variabile Y abbia una distribuzione, $F_y(Y, \theta)$ nota, mentre su X non si facciano assunzioni distributive. Nel seguito, si supponga che la variabile Y sia distribuita normalmente, cosicché $\theta = (\mu_y, \sigma_y^2)$. L'ipotesi da verificare:

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases}$$

equivale a verificare che:

$$\begin{cases} H_0 : \Pr[X > Y; \vartheta] = \Pr[X < Y; \vartheta] \\ H_1 : \Pr[X > Y; \vartheta] \neq \Pr[X < Y; \vartheta] \end{cases}$$

Indicata con ρ la quantità $\Pr[X > Y; \vartheta]$, il sistema d'ipotesi diventa quindi:

$$\begin{cases} H_0 : \rho = \rho_0 = 0.5 \\ H_1 : \rho \neq \rho_0 \end{cases}$$

Si tratta ora di stimare correttamente i due parametri ϑ e ρ. Detta $\hat{\vartheta}$ la stima di massima verosimiglianza di ϑ e basata sul $y = (y_1, ..., y_m)$, questa è data da:

$$\hat{\vartheta} = (\hat{\mu}_y, \hat{\sigma}_Y^2) = (\frac{1}{m}\sum_{i=1}^{m} y_i, \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{\mu}_y)^2) .$$

Una stima di ρ e data da

$$\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} S(x_i;\vartheta) = \frac{1}{n}\sum_{i=1}^{n}\{1 - F_Y(x_i;\vartheta)\} = \frac{1}{n}\sum_{i=1}^{n}\{1 - \phi(\frac{x_i - y_i}{\sigma_y})\}$$

dove $\phi(\cdot)$ rappresenta la funzione di ripartizione di una normale standard. Asintoticamente, si ha che $\sqrt{n}(\hat{\rho} - \rho_0) \sim N(0, \varpi^2)$

Pertanto, sotto l'ipotesi nulla, la statistica test

$t = \dfrac{\hat{\rho} - \rho_0}{\varpi / \sqrt{n}}$ segue una distribuzione normale standard.

Una stima del parametro ϖ^2, si può ottenere attraverso la seguente espressione:

$\hat{\varpi}^2 = \hat{\varpi}_s^2 + \dfrac{n}{m}\hat{\beta}^T \Omega \hat{\beta}$ dove $\hat{\varpi}_s^2 = \sum_{i=1}^{n}[S(x_i;\hat{\rho}]^2$ e Ω è la matrice di varianze e covarianze di

$\sqrt{n}(\vartheta - \vartheta_0)$, che nel nostro caso risulta essere $\Omega = \begin{bmatrix} \sigma^2 0 \\ 0 \ \ \sigma^2 \end{bmatrix}$ e $\hat{\beta}$ è definito come:

$$\hat{\beta} = \begin{bmatrix} \dfrac{1}{n}\hat{\sigma}_y \sum_{i=1}^{n} \phi\left(\dfrac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}\right) \\ \dfrac{1}{2n\hat{\sigma}_y^2}\sum_{i=1}^{n}\left(\dfrac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}\right)\phi\left(\dfrac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}\right) \end{bmatrix}$$

con $\phi(\cdot)$ funzione di densità di una normale standard.

Una volta fissato il livello di significatività α, è possibile calcolare i limiti superiore ed inferiore del intervallo di confidenza per ρ.

Essendo ρ definito come $\Pr[X > Y]$, si ha, per definizione $\rho \in [0,1]$. Il metodo presentato sopra può pero fornire dei limiti superiori od inferiori che cadono al di fuori di questo intervallo. Per ovviare tale problema, e possibile utilizzare una trasformazione, per esempio di tipo logit. Si definisce pertanto la quantità τ.

$$\tau = \log\left(\frac{\rho}{1-\rho}\right)$$

Il sistema da verificare diventa pertanto

$$\begin{cases} H_0 : \tau = \tau_0 \\ H_0 : \tau \neq \tau_0 \end{cases}$$

Sotto H_0, la variabile $\hat{\tau}$, definita come $\hat{\tau} = \log\left(\frac{\hat{\rho}}{1-\hat{\rho}}\right)$, ha distribuzione asintotica normale

con media 0 e varianza $\varpi^2/(np^2(1-p)^2)$ cioè.

$$\hat{\tau} \sim N\left(0, \frac{\varpi^2}{np^2(1-p)^2}\right)$$

Mediante la standardizzazione, è possibile quindi riportarsi ad una statistica che è asintoticamente distribuita come una normale standard. Anche in questo caso, una volta fissato il livello di significatività α, è possibile calcolare i limiti superiore ed inferiore dell'intervallo di confidenza associato al parametro τ. Applicando la formula inversa a tali intervalli, è possibile ottenere gli intervalli di confidenza per ρ. Tali intervalli non fuoriescono per costruzione dall'intervallo [0,1].

2.3 Il test di Wilcoxon-Mann-Whitney

Tale test serve per verificare se due campioni indipendenti provengono dalla stessa popolazione quando, per le variabili studiate, si raggiunge almeno un livello di misurazione ordinale. Questo è uno dei test "non parametrici" più potenti, e rappresentata inoltre, un'alternativa molto valida al test parametrico *t-Student*, quando il ricercatore vuole evitare i postulati del test *t* oppure quando la scala di misura è più debole di una scala ad intervalli. Quando si applica il test di Wilcoxon a dati che potrebbero essere analizzati in modo idoneo con il più potente test parametrico $t-Student$, la sua potenza – efficienza si avvicina a $3/\pi = 0.955$ all'aumentare di N ed è vicina al 95% anche per i campioni di dimensioni modeste. Esso è, quindi un alternativa eccellente al test t e non ha naturalmente tutti i requisiti associati al test *t-Student*.

Si supponga, quindi di avere, x e y relativamente indipendenti e identicamente distribuiti di due variabili casuali X e Y. L'ipotesi nulla è che X e Y hanno la stessa distribuzione. Per test a

due code, trattandosi della previsione di differenze di cui non si stabilisce la direzione . come nei casi precedenti, sia che $x = (x_1,...,x_n)$ e $y = (y_1,...,y_m)$. Sia inoltre $N=n+m$.

Per applicare il test di Wilcoxon, bisogna disporre le osservazioni, di entrambi i gruppi, in ordine di grandezza crescente, e sostituire ad ogni osservazione il rango che essa riceve in questo ordinamento. Si definisca con W_x e W_y la somma dei ranghi delle osservazioni appartenenti al campione dei casi *(x)* e *(y)*. La somma dei ranghi per i due gruppi è uguale alla somma dei primi *N* numeri interi, cioè:

$$W_X + W_Y = \frac{N(N+1)}{2}$$

Se H_0 è vera ci si aspetterebbe che in media le somme dei ranghi in ognuno dei due gruppi, siano quasi uguali. Se la somma dei ranghi per un gruppo è molto grande (o molto piccola), potremmo aver ragione di sospettare che i campioni non siano tratti dalla stessa popolazione. La distribuzione campionaria $W_x(W_y)$, quando H_0 e vera è nota. E, a partire da questa, possiamo determinare la probabilità associata al verificarsi dell'evento - con H_0 vera – di qualsiasi $W_x(W_y)$

Nel seguito si considera solo W_x.

Se $n > 10$ o se $m > 10$, è stato dimostrato che per *n* e *m* crescenti, la distribuzione campionaria W_x si avvicina rapidamente alla distribuzione normale con

$$media = \mu_{W_x} = \frac{n(N+1)}{2} \qquad \text{e} \qquad \text{var}\,ianza = \sigma^2_{W_x} = \frac{nm(N+1)}{12}$$

cioè possiamo determinare la significatività di un valore osservato di W_x mediante la statistica :

$$z_x = \frac{W_x \pm 0.5 - \mu_{W_x}}{\sigma^2_{W_x}} = \frac{W_x \pm 0.5 - n(N+1)}{\sqrt{nm(N+1)/12}}$$

che è distribuita asintoticamente in modo normale con media 0 e varianza 1. Pertanto, la probabilità associata al manifestarsi - quando H_0 è vera – di un valore estremo, osservata può essere determinato consultando il tabulato della distribuzione normale.

2.3.1 Ranghi ripetuti(ties)

Il test di Wilcoxon postula che i punteggi siano campionati da una distribuzione continua. Con una misurazione molto precisa di una variabile continua, la probabilità che vi siano valori uguali (ties) è pressoché nulla. Tuttavia, con le misurazioni un po' sommarie spesso impiegate nei campi sperimentali, si possono verificare punteggi di valore uguali. Quindi presumiamo che due (o più) osservazioni con ranghi ripetuti siano in realtà diverse, ma questa differenza è troppo piccola per essere rilevata con gli abituali strumenti di misura disponibili. Quando si verifica, ad ognuna di tali osservazioni viene assegnata la media dei ranghi che avrebbe avuto se non fossero emerse repliche di valori identici.

Se i ranghi ripetuti si verificano tra due o più osservazioni nello stesso gruppo, il valore di W_x non si modifica. Ma le repliche tra due o più osservazioni concernerti entrambi i gruppi, determinano cambiamenti sul valore di $W_x(W_y)$. Sebbene l'effetto sia generalmente trascurabile, la correzione per i ranghi identici è valida e deve essere usata anche quando si impegna l'approssimazione per grandi campioni alla distribuzione campionaria di W_x.

L'effetto dei ranghi identici è quello di cambiare la variabilità della serie dei ranghi. Quindi, la correzione per i ranghi ripetuti deve essere applicata alla varianza della distribuzione di campionamento di W_x. Corretta per le repliche dei ranghi identici, la varianza diventa:

$$\sigma^2_{W_x} = \frac{nm}{N(N-1)}\left(\frac{N^3 - N}{12} - \sum_{j=1}^{g}\frac{t_j^3 - t_j}{12}\right)$$

dove $N = n + m$, g è il numero di raggruppamenti di ranghi replicati e t_j è il numero dei ranghi uguali nel raggruppamento *j-esimo*. Usando questa correzione, per i ranghi replicati , la statistica test z_x diventa

$$z_x = \frac{W_X \pm 0.5 - n(N+1)}{\sqrt{[nm/N(N-1)]\left[(N^3 - N)/12 - \sum_{j=1}^{g}(t_j^3 - t_j)/12\right]}}$$

Si può vedere se non ci sono "ties" l'espressione riportata sopra si riduce direttamente a quella presentata originariamente nella statistica test per i ranghi non ripetuti.

Quando viene impegnata, la correzione *aumenta* leggermente il valore di z_x rendendolo più significativo. Quando non si corregge per i ranghi ripetuti, il nostro test è "*conservativo*" in quando la probabilità associata è leggermente inflazionata nel confronto con la z_x corretta. In altre parole, il valore della probabilità associata ai dati osservati – quando H_0 e vera – tende a essere leggermente più grande di quello che si sarebbe riscontrato se si fossero applicate le

correzioni. Si raccomanda di correggere sempre i ranghi ripetuti, sia se la proporzione dei ranghi ripetuti sia abbastanza grande, sia che alcuni ranghi ripetuti siano grandi, sia se le probabilità ottenuta senza correzione è molto vicina al valore α del livello di significatività.

Capitolo III

3 Simulazione

In questo studio di simulazione si vuole vedere l'efficienza dei test statistici del tipo parametrico (t-Student), semiparametrico (semiparametric.test) e non parametrico (Wilcoxon) nell'individuare, i geni cancerogeni.

Sia Y il livello d'espressione di un gene nel caso e X il livello d'espressione nel controllo. Siano inoltre $y=(y_1,\ldots,y_m)$ e $x=(x_1,\ldots,x_n)$ i campioni osservati. Nella simulazione si è immaginato di osservare un microarray di 2000 geni, di cui 100 differenzialmente espressi nei casi. Si è assunto che i geni fossero distribuiti come normali $N(0,1)$ standard e che i geni differenzialmente espressi nei casi fossero distribuiti come normali $N(1,1)$ *(Caso A)* e come normali $N(1,2)$ *(Caso B)* .

Nello studio, si è fissato $n=m$. Si è inoltre scelto di considerare 4 possibili numerosità: $n=(15,25,35,100)$. Si sono quindi condotti otto esperimenti ottenuti incrociando i due casi (A e B) con le quattro numerosità.

Utilizzando il pacchetto R e la libreria *"multest"* è stato possibile generare i dati e fare il confronto tra i tipi di geni differenzialmente espressi. I dati sono stati classificati in 16 matrici $2000 \times n$ dove n varia a seconda dell'esperimento. Ai dati simulati tramite la funzione *"sim" (Appendice)* sono stati applicati i test parametrico, semiparametrico e non parametrico, e sono stati salvati i valori delle statistiche test e il valore di significatività osservato. Di seguito sono riportati per la statistica test gli istogrammi, i grafici quantile-quantile e le densità stimate. Per vedere la potenza dei test è stato usato la funzione "*sel" (Appendice)* la quale identifica i geni differenzialmente espressi con un livello di significatività α e calcola tabelle delle decisioni.

3.2 *Risultati*

Figura 1 numerosità 15 nell'istogramma *(1)* caso A si nota che sulle code non c'è un andamento regolare, ed è proprio sulle code della distribuzione che si concentra l'attenzione, perché più ci si allontana dal valore centrale, più si è portati a ritenere che i valori assunti dalla statistica test identifichino geni in cui livello d'espressione differisce tra i due gruppi.

Nel secondo l'istogramma *(2)* si nota l'irregolarità nella parte destra dell'istogramma questo confrontato con il grafico quantile-quantile ha un adeguatezza migliore, cioè è poco efficiente

nel individuare i geni differenzialmente espressi rispetto al primo *(1)* istogramma. Anche per il terzo istogramma si vedono delle code irregolari nella parte sinistra e bassi valori centrali. Nell'caso della numerosità delle espressioni geniche *n=15* si vede che il test t-student ha un efficienza nell'individuare i geni differenzialmente espressi. Questo si vede anche nella coda iniziale nel grafico dei quantili che è ben differenziata, rispetto ai grafici dei quantili per le statistiche semiparametriche e non parametriche.

Figura 2 numerosità 25 e Figura 3 numerosità 35

Si vede che negli istogrammi l'andamento irregolare dovuto ai geni differenzialmente espressi. Anche in questo caso si vede che le code sono abbastanza irregolari questo perché è stata aumentata la numerosità dei campioni per cui c'è una maggiore precisione nell'individuare il valore medio dell'espressione.

Figura 4 numerosità 100

In questo caso si nota subito che c'è una bimodalità nelle code, ovvero i geni differenzialmente espressi sono allineati nella coda, a sinistra del istogramma *(10)*, a destra dell'istogramma *(11)* e a sinistra dell'istogramma *(12)*. Vedendo anche i grafici dei quantili e difficile identificare il test migliore; però si vede che il test semiparametrico e il test non parametrico hanno un andamento simile e meno irregolare.

Caso B

Nel caso di basse numerosità non si nota un cambiamento apprezzabile nelle distribuzioni delle statistiche generate dai differenti test. Si notano sempre sulle code gli andamenti irregolari che portano ad individuare i geni falsi positivi e i falsi negativi. Vedendo l'istogramma *(13)* si vede che i dati si distaccano dal valore medio, ma non troppo rispetto ai grafici *(14)* e *(15)*. Facendo un confronto Q-Q plot, si nota che il grafico per il test parametrico e più irregolare rispetto ai due grafici per i test semiparametrico e non parametrico.

Osservando l'andamento dei geni differenzialmente espressi quando la numerosità della sequenza aumenta, si vedono delle code più decise ovvero i geni che hanno una espressione diversa sono allineati nelle code che vanno verso destra o sinistra. Anche in questo caso l'istogramma appartenente al test *t* individua sempre delle code irregolari con andamenti verso sinistra.

Nel caso *n=100* si aspetta di vedere l'efficienza dei test nell'individuare i dati differenzialmente espressi, perché è stato quasi triplicato il numero delle unita rispetto ai casi

precedenti. Si nota la bimodalità formata dai dati che sono pesati in media per quando riguarda i geni che sono espressi differenzialmente. Si nota anche in questo caso che i test sono simili tra di loro e non individuare il test più potente nel individuare le sequenze significative (espresse differenzialmente). Si vede che il test t fornisce una bimodalita decisiva nell'individuare i geni differenzialmente espressi.

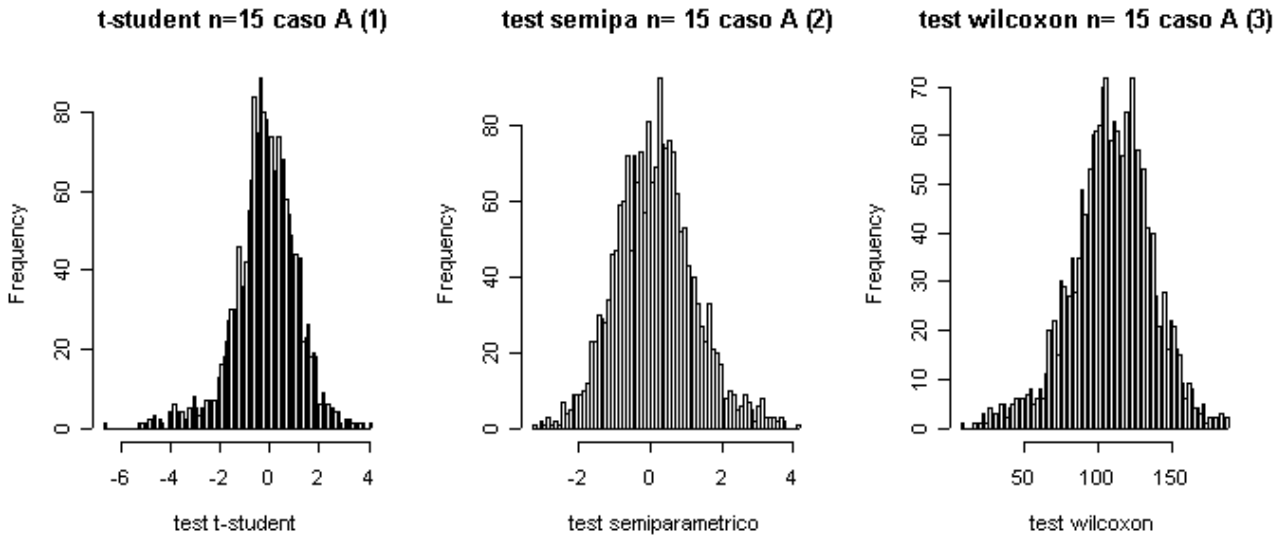


Figura 1: numerosità 15 (Caso A)

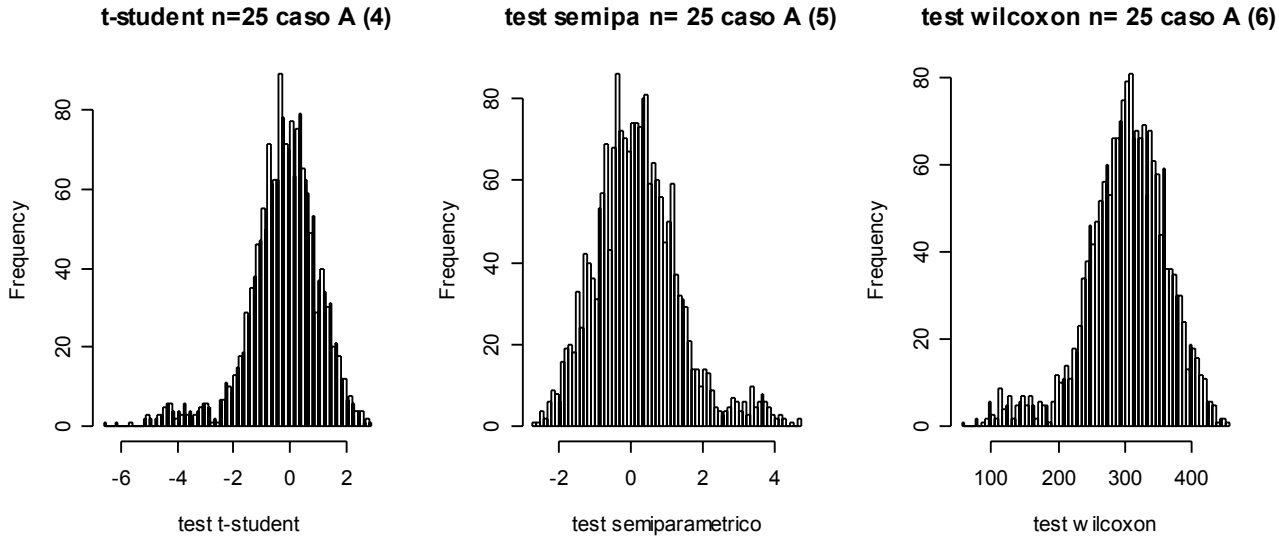


Figura 2: numerosità 25 (Caso A)

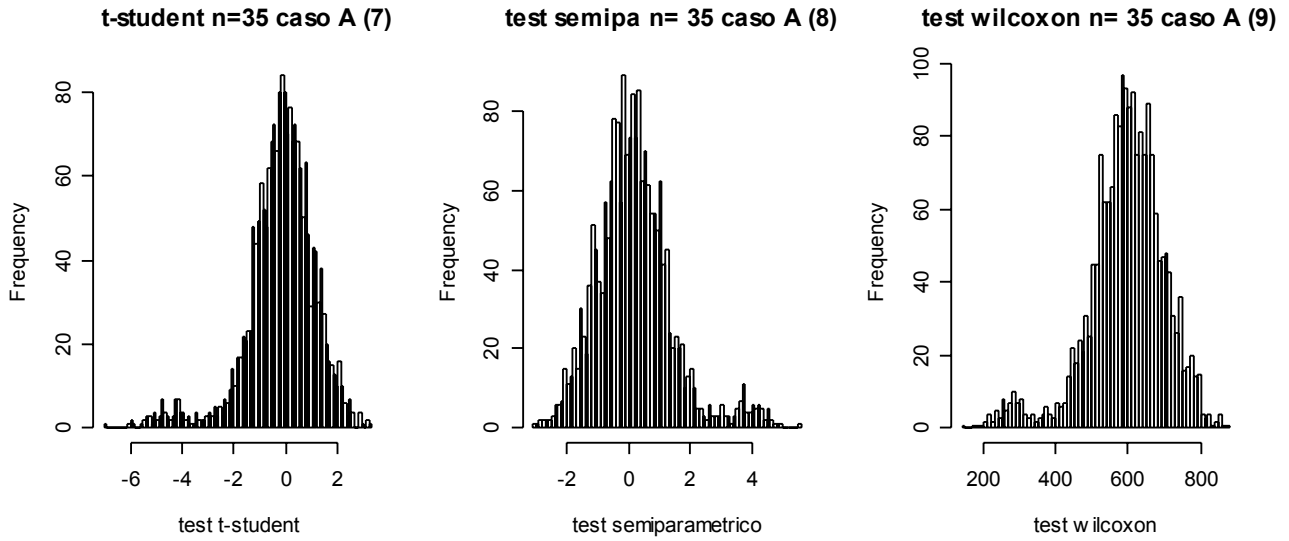


Figura 3 : numerosità 35 (Caso A)

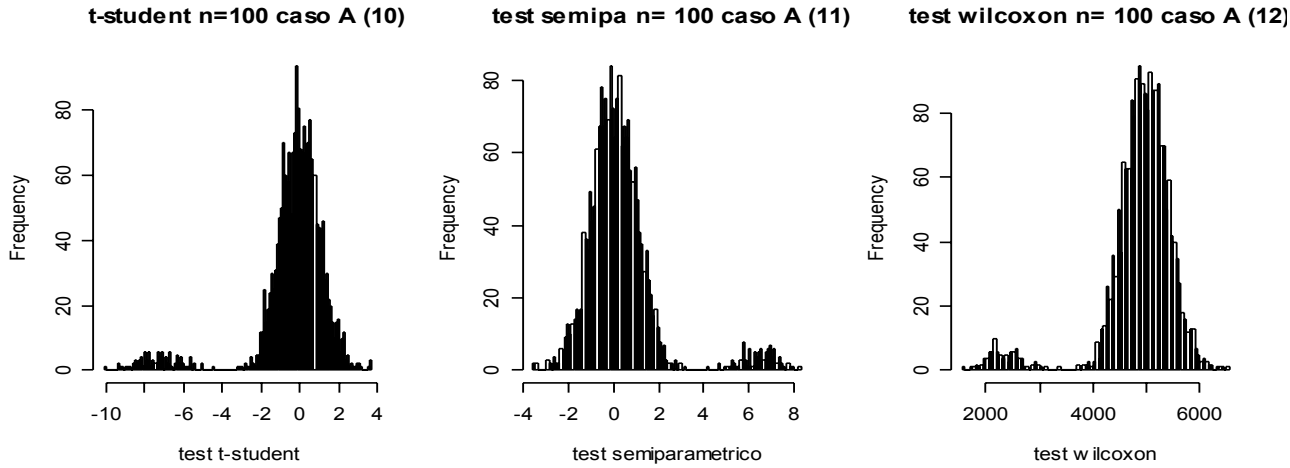


Figura 4: numerosità 100

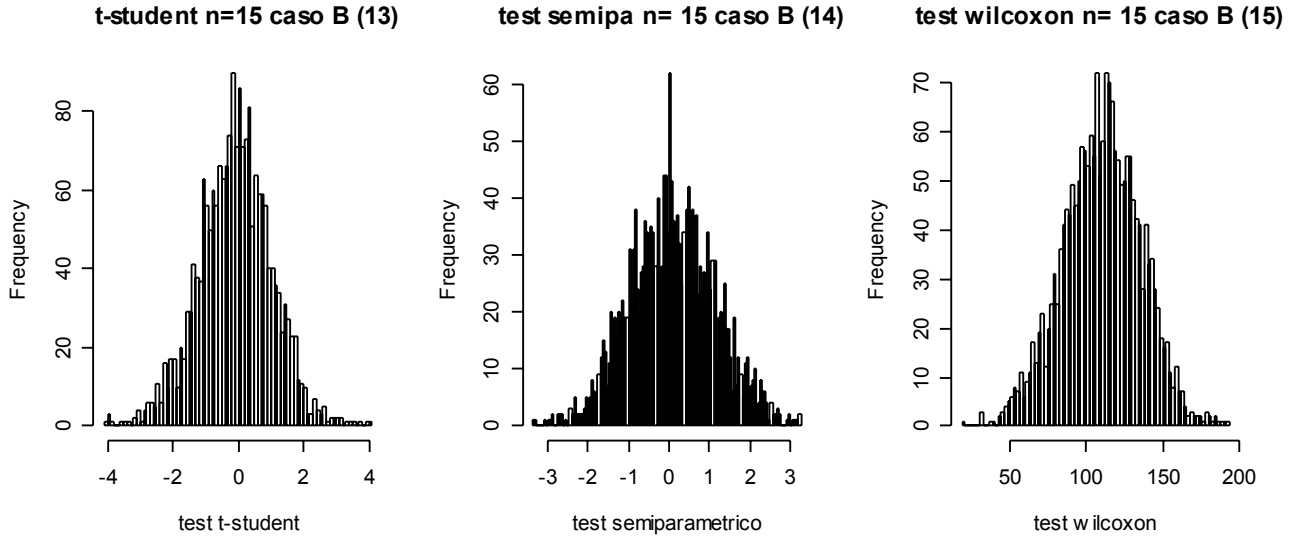


Figura 5: numerosità 15 (Caso B)

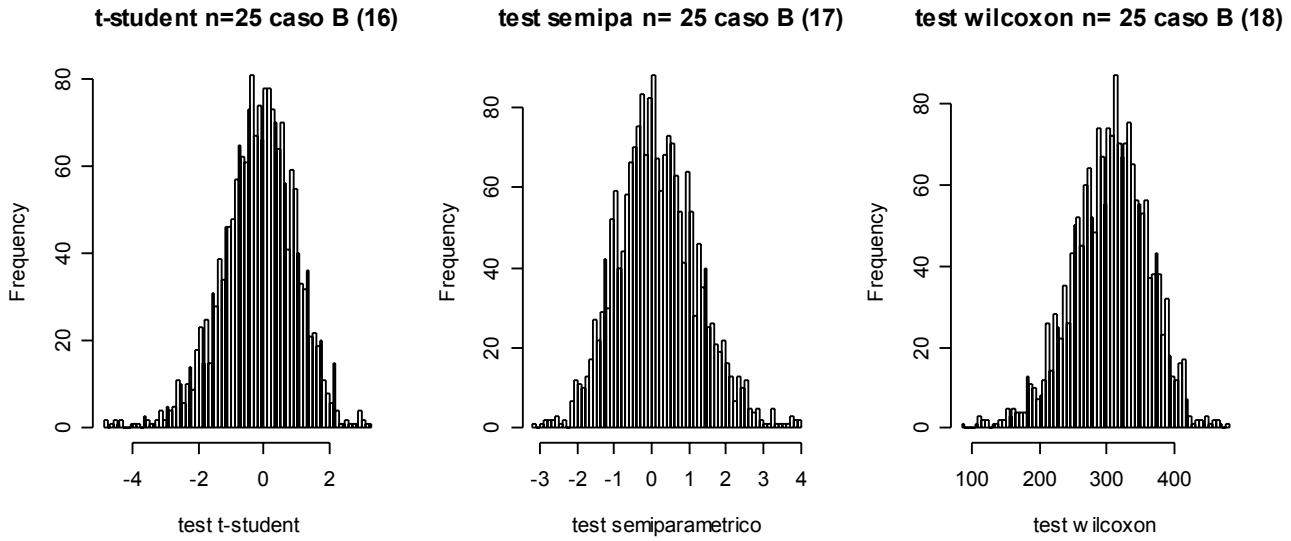


Figura 6: numerosità 25 (Caso B)

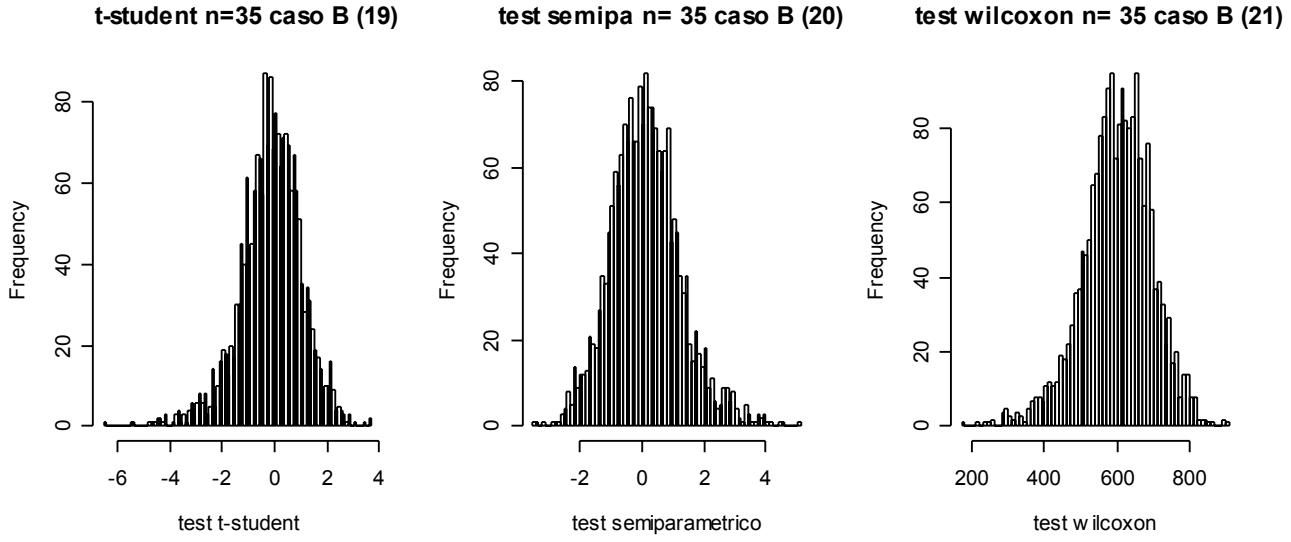


Figura 7: numerosità 35 (Caso B)

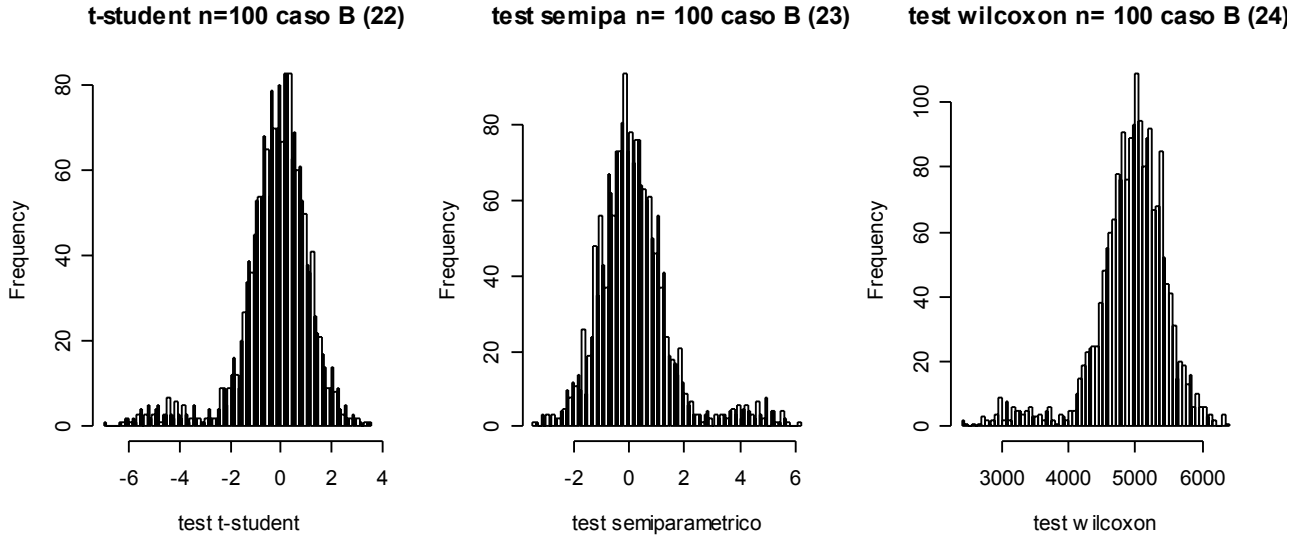
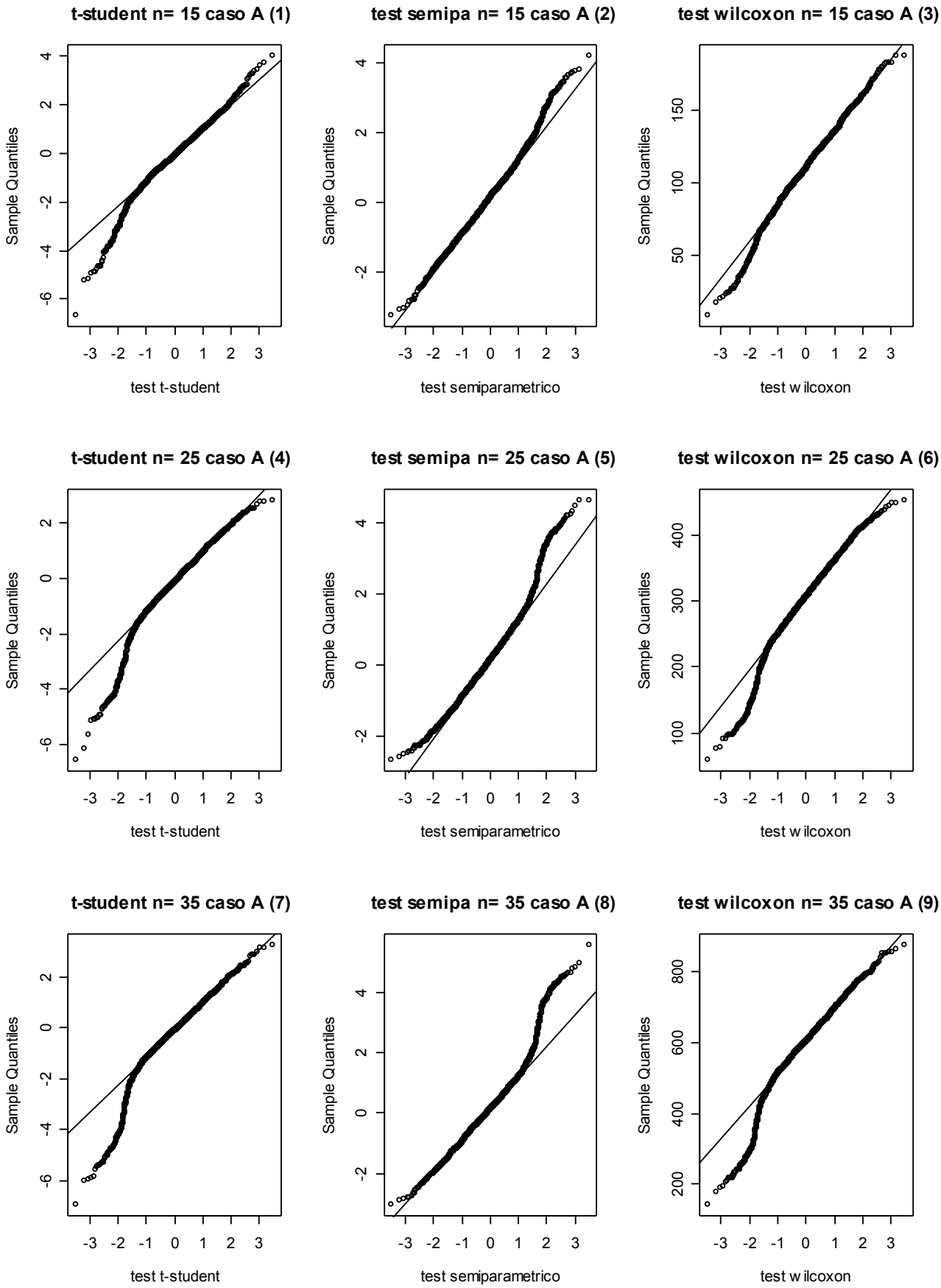
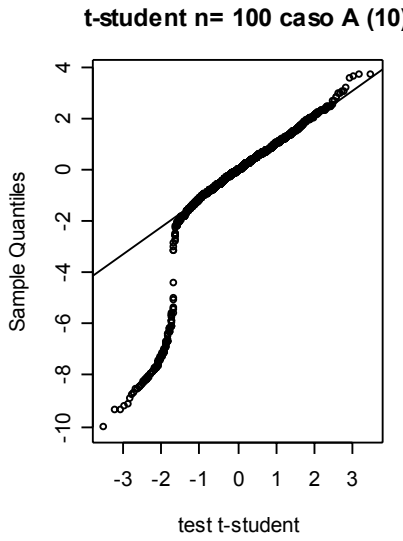
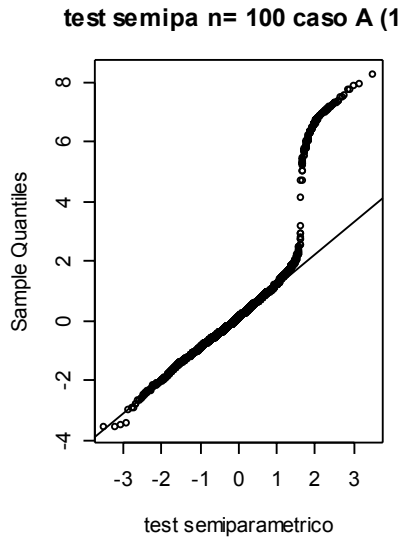
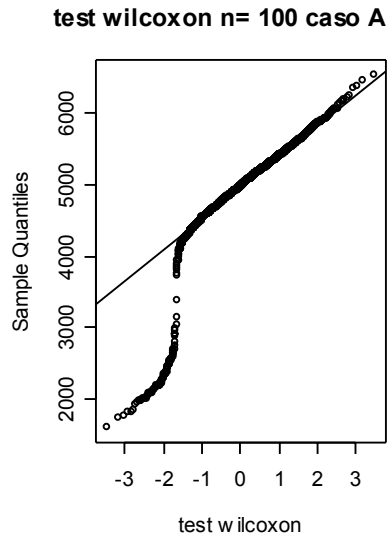


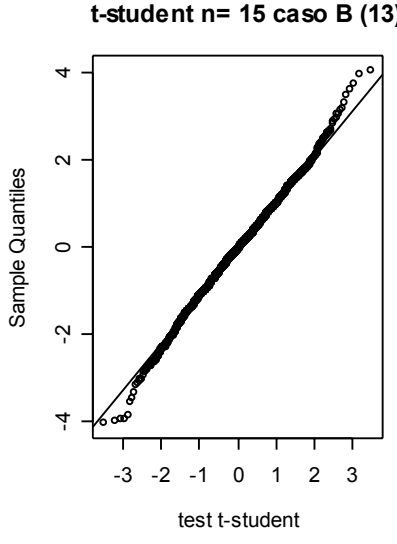
Figura 8: numerosità 100 (Caso B)

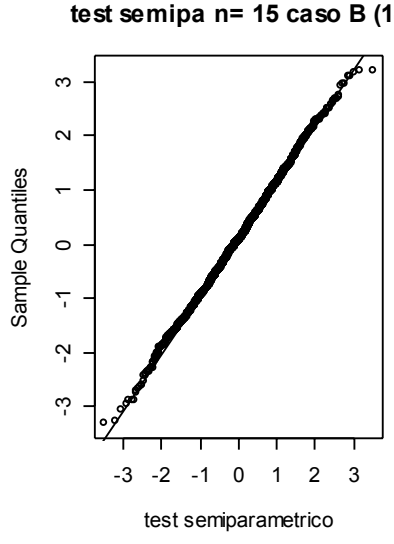

t-student n= 15 caso A (1)
test semipa n= 15 caso A (2)
test wilcoxon n= 15 caso A (3)
t-student n= 25 caso A (4)
test semipa n= 25 caso A (5)
test wilcoxon n= 25 caso A (6)
t-student n= 35 caso A (7)
test semipa n= 35 caso A (8)
test wilcoxon n= 35 caso A (9)
Sample Quantiles
test t-student
test semiparametrico
test w ilcoxon

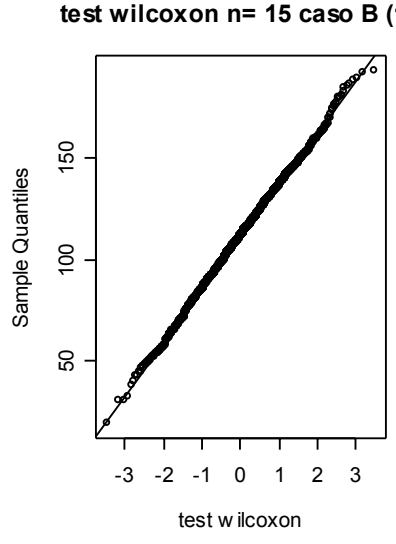

t-student n= 100 caso A (10)
Sample Quantiles
test t-student

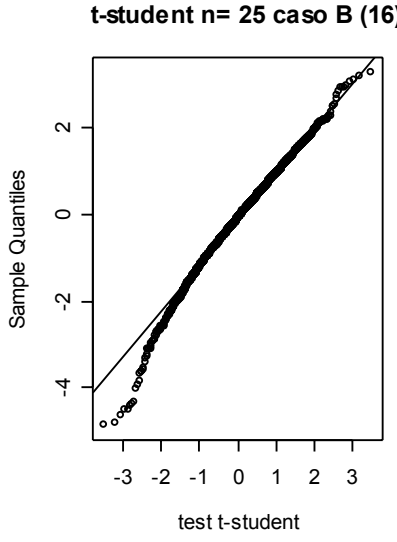

test semipa n= 100 caso A (11)
Sample Quantiles
test semiparametrico

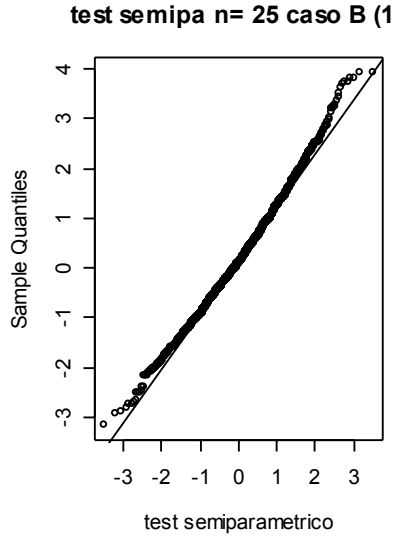

test wilcoxon n= 100 caso A (12)
Sample Quantiles
test w ilcoxon

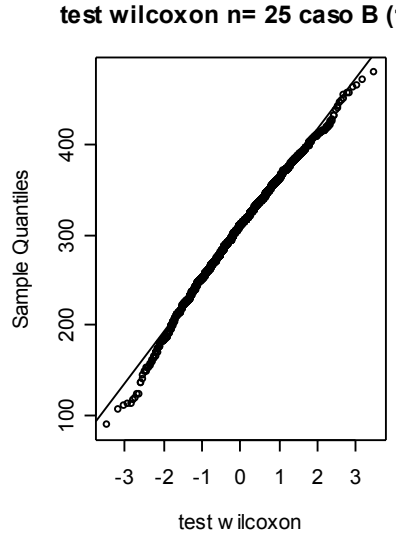

t-student n= 15 caso B (13)
Sample Quantiles
test t-student

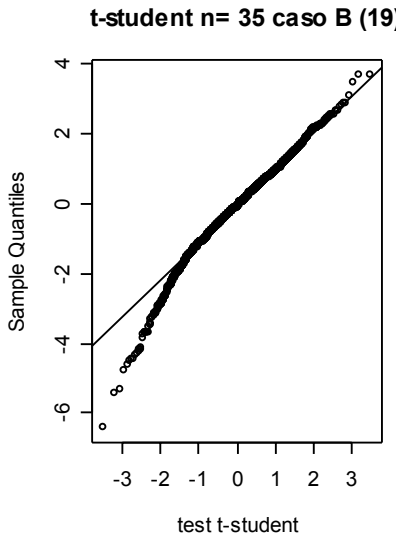

test semipa n= 15 caso B (14)
Sample Quantiles
test semiparametrico

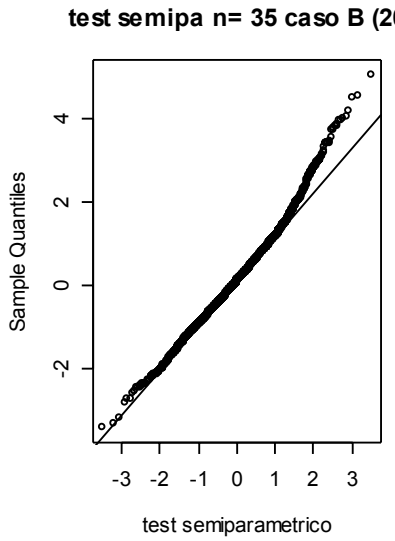

test wilcoxon n= 15 caso B (15)
Sample Quantiles
test w ilcoxon

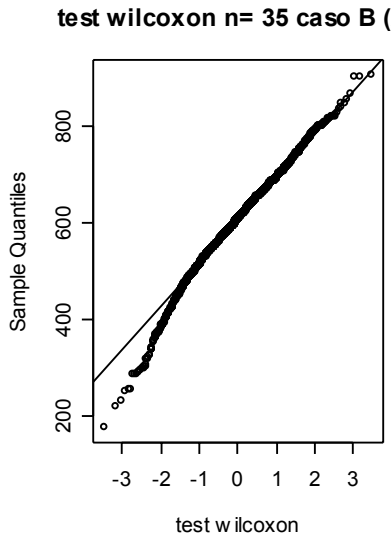

t-student n= 25 caso B (16)
Sample Quantiles
test t-student

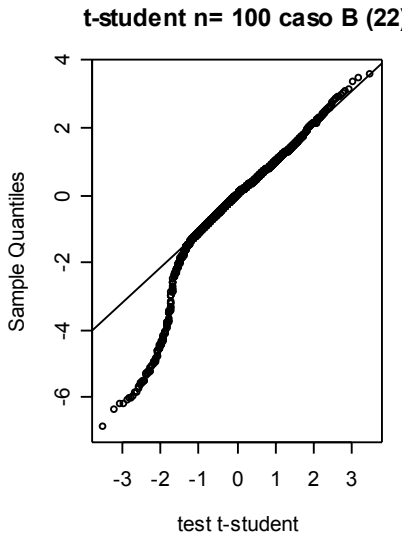

test semipa n= 25 caso B (17)
Sample Quantiles
test semiparametrico

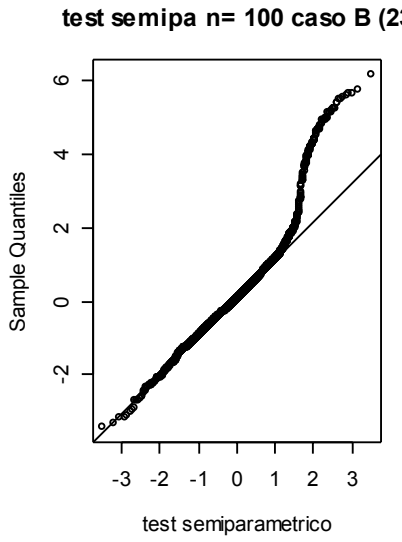

test wilcoxon n= 25 caso B (18)
Sample Quantiles
test w ilcoxon

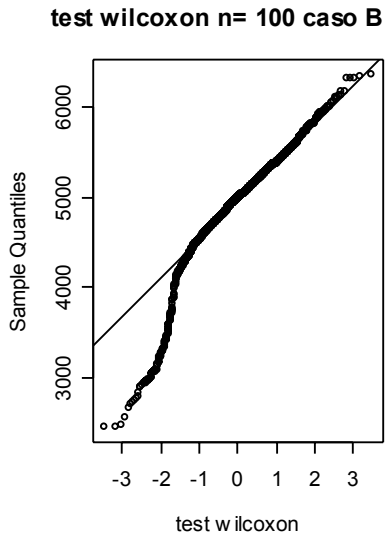

t-student n= 35 caso B (19)
Sample Quantiles
test t-student

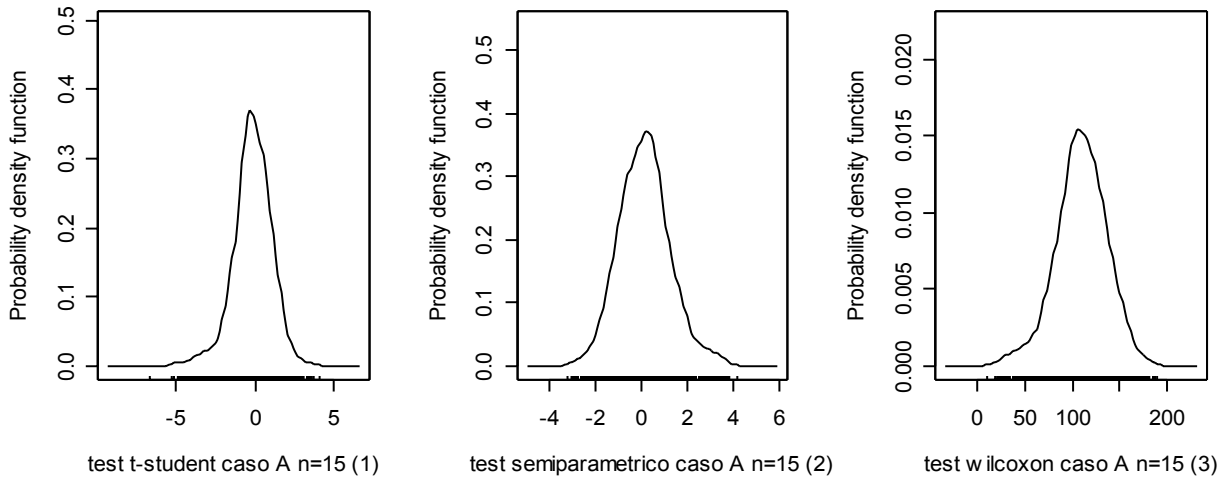

test semipa n= 35 caso B (20)
Sample Quantiles
test semiparametrico

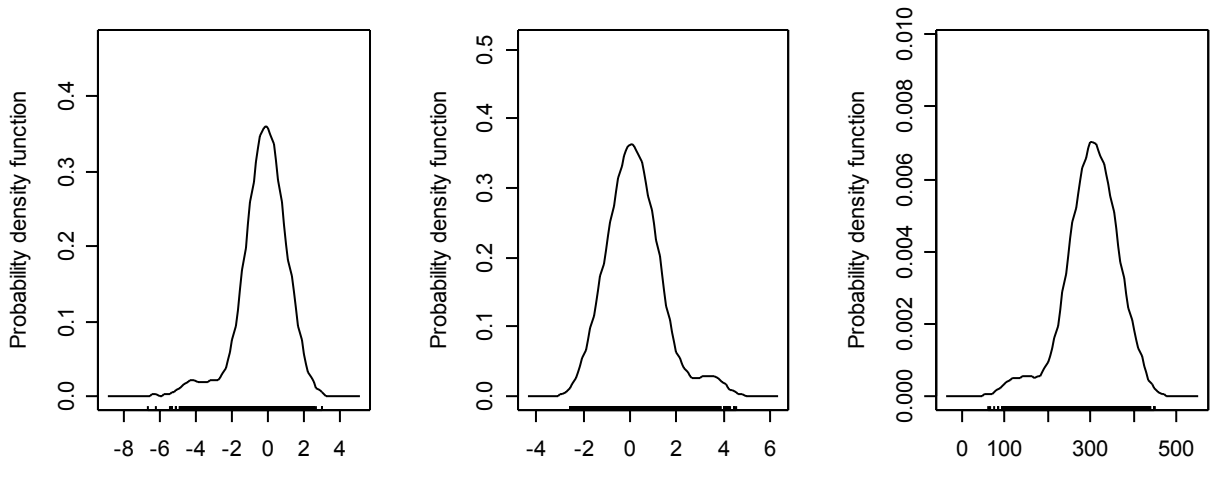

test wilcoxon n= 35 caso B (21)
Sample Quantiles
test w ilcoxon

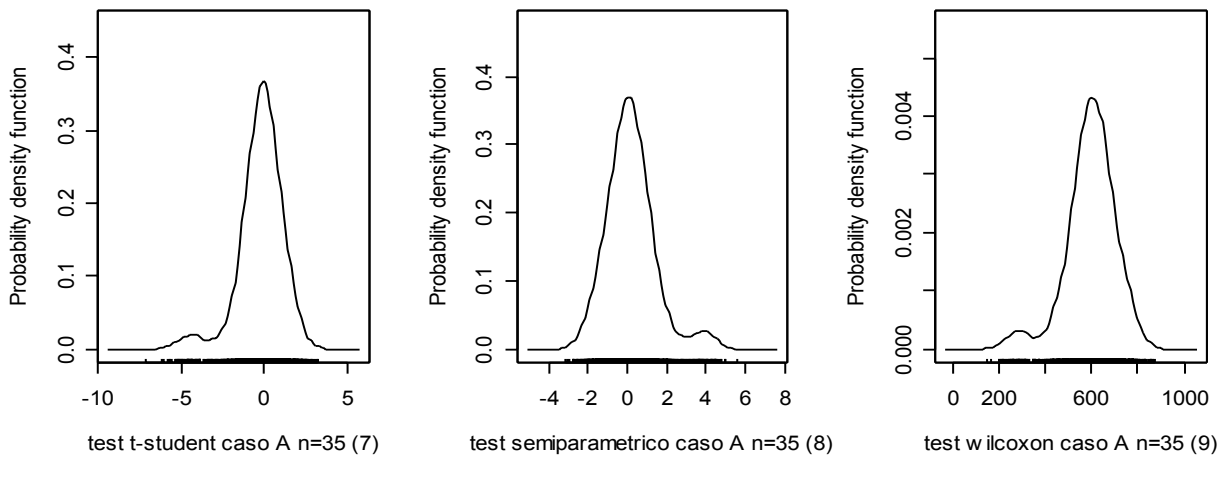

t-student n= 100 caso B (22)
Sample Quantiles
test t-student

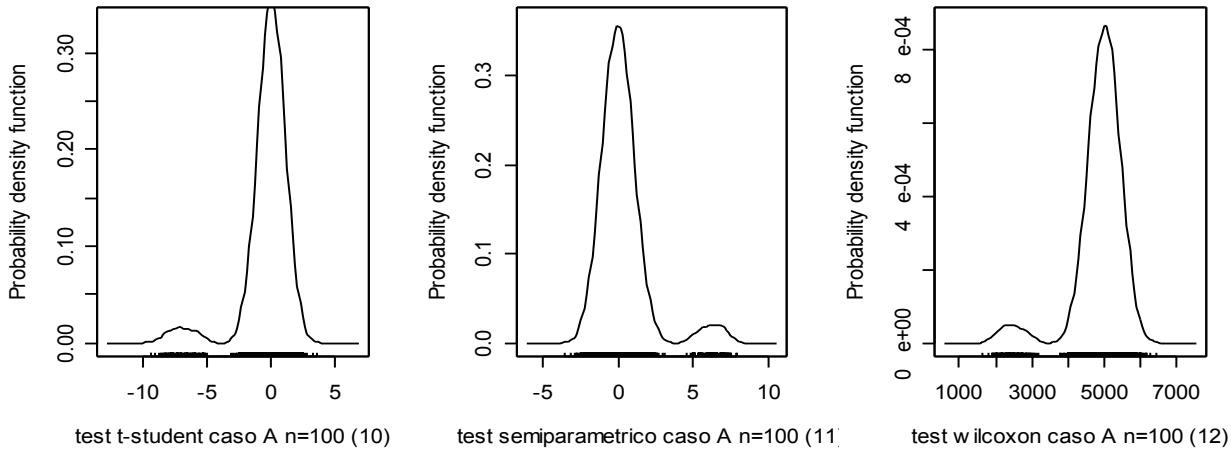

test semipa n= 100 caso B (23)
Sample Quantiles
test semiparametrico

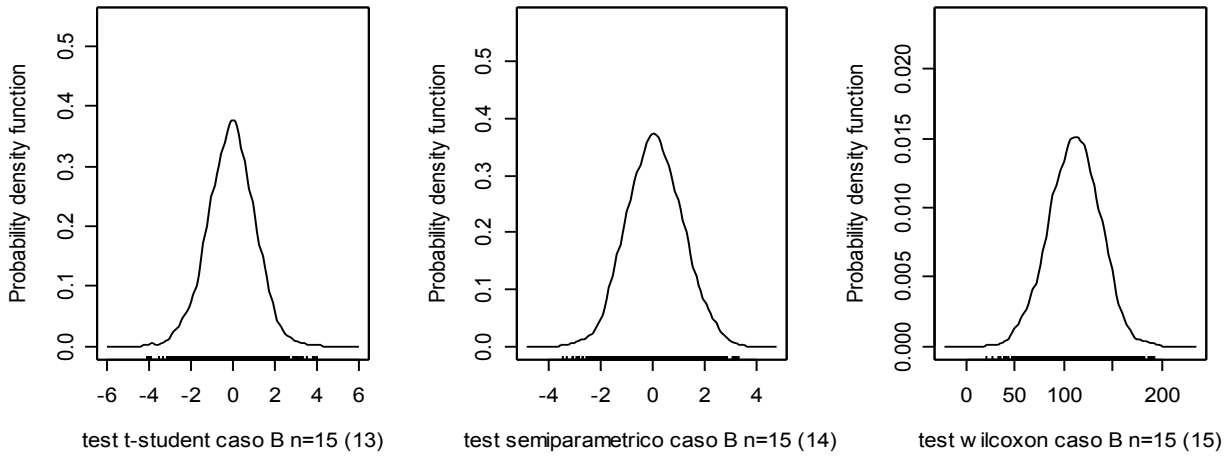

test wilcoxon n= 100 caso B (24)
Sample Quantiles
test w ilcoxon

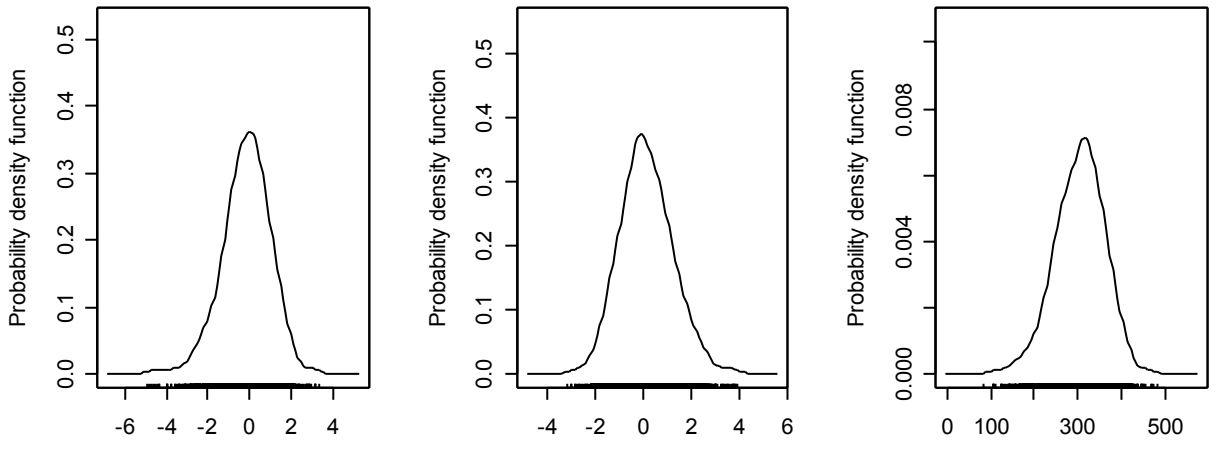

Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-5
0
5
test t-student caso A n=15 (1)
Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
6
test semiparametrico caso A n=15 (2)
Probability density function
0.000
0.005
0.010
0.015
0.020
0
50
100
200
test w ilcoxon caso A n=15 (3)

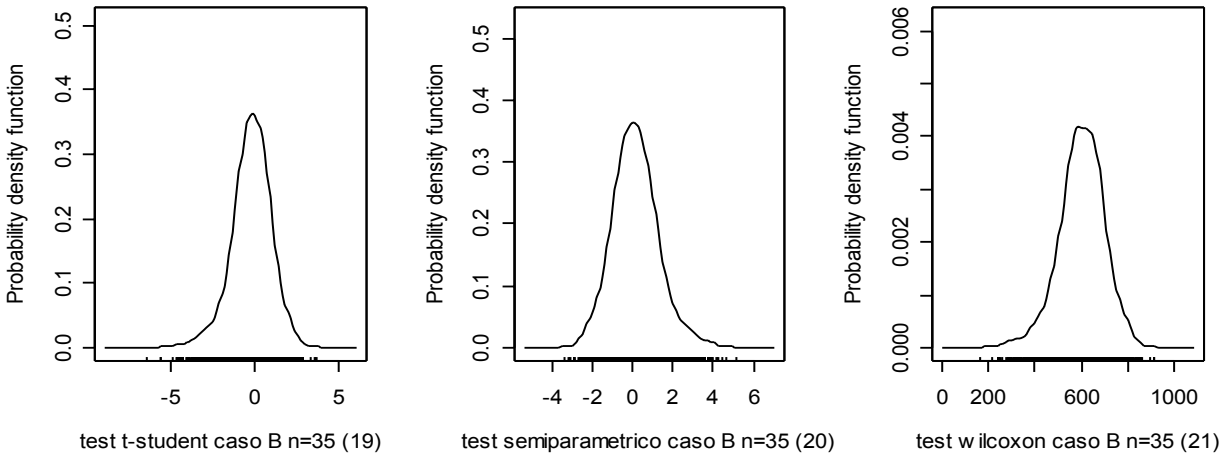

Probability density function
0.0
0.1
0.2
0.3
0.4
-8
-6
-4
-2
0
2
4
test t-student caso A n=25 (4)
Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
6
test semiparametrico caso A n=25 (5)
Probability density function
0.000
0.002
0.004
0.006
0.008
0.010
0
100
300
500
test w ilcoxon caso A n=25 (6)

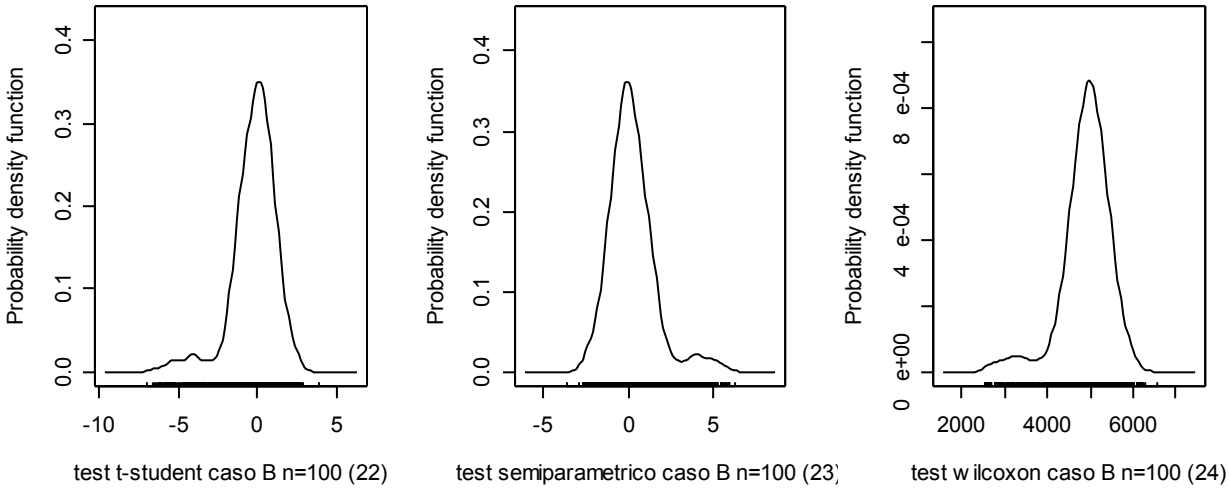

Probability density function
0.0
0.1
0.2
0.3
0.4
-10
-5
0
5
test t-student caso A n=35 (7)
Probability density function
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
8
test semiparametrico caso A n=35 (8)
Probability density function
0.000
0.002
0.004
0
200
600
1000
test w ilcoxon caso A n=35 (9)


Probability density function
0.00
0.10
0.20
0.30
-10
-5
0
5
test t-student caso A n=100 (10)
Probability density function
0.0
0.1
0.2
0.3
-5
0
5
10
test semiparametrico caso A n=100 (11
Probability density function
0 e+00
4 e-04
8 e-04
1000
3000
5000
7000
test w ilcoxon caso A n=100 (12)


Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-6
-4
-2
0
2
4
6
test t-student caso B n=15 (13)
Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
test semiparametrico caso B n=15 (14)
Probability density function
0.000
0.005
0.010
0.015
0.020
0
50
100
200
test w ilcoxon caso B n=15 (15)


Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-6
-4
-2
0
2
4
test t-student caso B n=25 (16)
Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
6
test semiparametrico caso B n=25 (17)
Probability density function
0.000
0.004
0.008
0
100
300
500
test w ilcoxon caso B n=25 (18)


Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-5
0
5
test t-student caso B n=35 (19)
Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
6
test semiparametrico caso B n=35 (20)
Probability density function
0.000
0.002
0.004
0.006
0
200
600
1000
test w ilcoxon caso B n=35 (21)


Probability density function
0.0
0.1
0.2
0.3
0.4
-10
-5
0
5
test t-student caso B n=100 (22)
Probability density function
0.0
0.1
0.2
0.3
0.4
-5
0
5
test semiparametrico caso B n=100 (23)
Probability density function
0
e+00
4
e-04
8
e-04
2000
4000
6000
test w ilcoxon caso B n=100 (24)

3.3 Attendibilità dei test

Supponiamo che si possa classificare un individuo di una grande popolazione come realmente positivo o realmente negativo a una particolare diagnosi. Quando la vera diagnosi può basarsi su metodi più raffinati di quelli applicati nel test o può derivare da informazioni che emergono col passare del tempo, per esempio l'autopsia. Per ogni classe di appartenenza dell'gene, vero positivo o vero negativo, si possono considerare le probabilità che il test dia una risposta positiva o negativa, come nella tabella che segue.

La numerosità dei geni nelle caselle: *a=veri negativi*, *b=falsi negativi*, *c=falsi positivi* e *d=veri positivi*.

REALTÀ / TEST	H_0	H_1	TOTALE
H_0	$1-\alpha$	α	1
H_1	β	$1-\beta$	1

β è la probabilità di un falso negativo, mentre $1-\beta$ è detta sensibilità del test e $1-\alpha$ è detta specificità del test.
Nell'ipotesi nulla che un individuo sia vero negativo, considerando *"significativi"* i risultati al test ,α è l'analogo del livello di significatività e $1-\beta$ della potenza, cioè la capacita la capacita di evidenziare l'ipotesi alternativa che un gene sia un vero positivo.
Chiaramente, a un test si richiede di avere piccoli valori di α e β, anche se altre considerazioni, come il costo e la facilità di applicazione, sono altamente rilevanti.
A parità di ogni considerazione, se il test A ha valori di α e β minori del test B, può essere considerato migliore il test B. Supponiamo,tuttavia che A abbia, rispetto a B, un valore minore di α ma un valore maggiore di β. A meno di non associare pesi diversi ai due tipi di errore, cioè falsi positivi e falsi negativi, non e definito un giudizio definito di probabilità. Valutando i due errori approssimativamente di uguale importanza, un metodo naturale per tenerne conto globalmente è la somma delle due probabilità di errore, $\alpha+\beta$.
Youden (1950) un indice sostanzialmente equivalente:

$J=1-(\alpha+\beta)$

Usando l'indice di Youden si assume implicitamente che la sensibilità e la specificità abbiano uguale importanza, poiché α e β hanno lo stesso peso. Esistono buone ragioni contro questo punto di vista. In primo luogo, i due tipi di errore hanno conseguenze molto differenti, e spesso è ragionevole attribuire un peso maggiore a un falso negativo piuttosto che a un falso positivo; tale considerazione suggerirebbe di dare un peso maggiore a β (e quindi fissare un $\beta<\alpha$) rispetto a α, che condurrebbe alla scelta di una procedura con un relativamente piccolo di β/α (minore di 1) . D'altra parte se la malattia e rara, i falsi negativi sono pochi e si minimizza il numero di geni differenzialmente espressi, in modo errato scegliendo $\beta>\alpha$. Queste due considerazioni chiaramente in conflitto, e bisogna pervenire a una soluzione alla luce delle conseguenze di errori diagnostici per la malattia in questione.
Finora la discussione è stata condotta in termini di probabilità che, in pratica si stimano tramite rilevazioni. Supponiamo di organizzare un'indagine mirata su un campione casuale della popolazione abbia le seguenti frequenze:

			Realtà		
		H0		H1	totale
	H0	d		b	d+b
Test					
	H1	c		a	c+a
	totale	c+d		b+a	

La probabilità di un falso negativo potrebbe essere stimata da $\hat{\beta} = b/(a+b)$; la probabilità di un falso positivo da $\hat{\alpha} = c/(c+d)$.

L'indice di Youden dà $\hat{J} = 1-(\hat{\alpha}+\hat{\beta})$. Gli errori campionari di queste stime seguono dalle espressioni standard delle distribuzioni binomiali.

Un punto importante da sottolineare è che le frequenze relative attese di diagnosi errate fra i soggetti apparentemente positivi e negativi non dipendono soltanto da α e β, ma dalla reale prevalenza dei geni differenzialmente espressi.

La frequenza relativa di veri positivi tra i positivi apparenti è chiamata a volte valore predittivo di un test positivo; la frequenza relativa di veri negativi tra i negativi apparenti è il valore predittivo di un test negativo.

Numerosita 15

	H_0 *Caso A*	H_0 *Caso B*	H_1 *Caso A*	H_1 *Caso B*
H_0 t-Student	1816	1805	25	38
H_0 ρ semiparametrico	1821	1816	29	38
H_0 W_x Wilcoxox	1820	1819	31	39
H_1 t-Student	84	95	75	62
H_1 ρ semiparamtrico	79	84	71	62
H_1 W_x Wilcoxon	*86*	*86*	*93*	*48*

Numerosita 25

	H_0 *Caso A*	H0 *Caso B*	H_1 *Caso A*	H_1 *Caso B*
H_0 t-Student	1819	1800	6	56
H_0 ρ semiparametrico	1818	1805	6	56
H_0 W_x Wilcoxox	1814	1814	7	52
H_1 t-Student	81	100	94	44
H_1 ρ semiparamtrico	82	95	94	44
H_1 W_x Wilcoxon	*86*	*86*	*93*	*48*

Numerosità 35

	H_0 *Caso A*	H_0 *Caso B*	H_1 *Caso A*	H_1*Caso B*
H_0 t-Student	1801	1793	0	25
H_0 ρ semiparametrico	1804	1794	3	25
H_0 W_x Wilcoxox	1799	1797	3	30
H_1 t-Student	99	107	100	75
H_1 ρ semiparamtrico	96	106	97	75
H_1 W_x Wilcoxon	*101*	*103*	*97*	*70*

Numerosità 100

	H_0 *Caso A*	H_0 *Caso B*	H_1 *Caso A*	H_1 *Caso B*
H_0 t-Student	1808	1795	0	1
H_0 ρ semiparametrico	1806	1802	0	1
H_0 W_x Wilcoxox	1806	1803	0	1
H_1 t-Student	92	105	100	99
H_1 ρ semiparamtrico	94	98	100	99
H_1 W_x Wilcoxon	*94*	*97*	*100*	*99*

Si riportano nelle tabelle successive i valori di sensibilità e specificità calcolati per i tre test considerati.

Caso A

15	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	0,75	0,013158	0,25	0,986842	0,0396616	0,2293578
semiparametrico	0,71	0,015263	0,29	0,984737	0,0375264	0,2685185
wilcoxon	0,69	0,016316	0,25	0,983684	0,0365273	0,2792793

25	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	0,94	0,003158	0,06	0,996842	0,0491375	0,0689655
semiparametrico	0,94	0,003158	0,06	0,996842	0,0491632	0,0681818
wilcoxon	0,93	0,003684	0,06	0,996316	0,0487677	0,0752688

35	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	1	0	0	1	0,0526039	0
semiparametrico	0,97	0,001579	0,03	0,998421	0,0510258	0,030303
wilcoxon	0,97	0,001579	0,03	0,998421	0,0511603	0,0288462

100	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	1	0	0	1	0,0524109	0
semiparametrico	1	0,055371	0	0,944629	0,0524659	0,0524659
wilcoxon	1	0	0	1	0,0524659	0

Caso B

15	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	0,62	0,02	0,38	0,98	0,0332084	0,2857143
semiparametrico	0,62	0,02	0,38	0,98	0,0330138	0,3114754
wilcoxon	0,61	0,020526	0,39	0,979474	0,0324468	0,325

25	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	0,56	0,023158	0,44	0,976842	0,0301724	0,3055556
semiparametrico	0,56	0,023158	0,44	0,976842	0,0300913	0,3165468
wilcoxon	0,52	0,025263	0,48	0,974737	0,0278671	0,358209

35	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	0,75	0,012019	0,25	0,987981	0,0366211	0,1893939
semiparametrico	0,75	0,013158	0,25	0,986842	0,0401284	0,1908397
wilcoxon	0,7	0,015789	0,25	0,984211	0,0374933	0,2255639

100	sensibilita	specificita	1-sensibilita	1-specificita	Valori pred +	valori pred -
t-student	0,99	0,000526	0,01	0,999474	0,0522703	0,009434
semiparametrico	0,99	0,054939	0,01	0,945061	0,0520779	0,0520779
wilcoxon	0,99	0,054908	0,01	0,945092	0,0014981	0,0520505

Appendice

```
sim<-function (n,m,s,mw,sw)
       {
       library(sm)
       test<-rep(0,2000)
       prob<-rep(0,2000)
       f<-n*1900
       o<-n*100
       x<-rep(0,f)
       h<-rep(0,o)
       L<-rep(0,f)
       G<-rep(0,o)
       w<-NULL
       p<-NULL
       c<-NULL
       r<-NULL
       wiltest<-NULL
       wilprob<-NULL
       par(mfrow=c(3,2))
       for(i in 1:f) {
                       x[i]<-rnorm(n,m,s)
                     }
       for(i in 1:o) {
                       h[i]<-rnorm(n,m,s)
                     }
       for(i in 1:f) {
                       L[i]<-rnorm(n,m,s)
                     }
       for(i in 1:o) {
                       G[i]<-rnorm(n,mw,sw)
                     }
       w<-c(x,h)
       w<-matrix(w,ncol=n,byrow=T)
       p<-c(L,G)
       p<-matrix(p,ncol=n,byrow=T)
       for(i in 1:2000){
                       test[i]<-t.test(w[i,],p[i,],var.equal=TRUE)$statistic
                       prob[i]<-t.test(w[i,],p[i,],var.equal=TRUE)$p.value
                       }

       tau<-matrix(test,ncol=1,byrow=T)
       tai<-matrix(prob,ncol=1,byrow=T)
       for(i in 1:2000){
                       wiltest[i]<-wilcox.test(w[i,],p[i,])$statistic
                       wilprob[i]<-wilcox.test(w[i,],p[i,])$p.value
                       }

       sem<-semiparametric.test(w,p)
       c<-sem$test
       r<-sem$pvalue
       c<-matrix(c,ncol=1,byrow=T)
       r<-matrix(r,ncol=1,byrow=T)
       write.table(round(tau,6),"c:\\testi\\ttest.txt",sep="
",row.names=FALSE,col.names=FALSE)
       write.table(round(tai,6),"c:\\testi\\tprob.txt",sep="
",row.names=FALSE,col.names=FALSE)
       write.table(round(w,6),"c:\\testi\\dati0101.txt",sep="
",row.names=FALSE,col.names=FALSE)
       write.table(round(p,6),"c:\\testi\\dati0111.txt",sep="
",row.names=FALSE,col.names=FALSE)
       write.table(round(wiltest,6),"c:\\testi\\wiltest.txt",sep="
",row.names=FALSE,col.names=FALSE)
       write.table(round(wilprob,6),"c:\\testi\\wilprob.txt",sep="
",row.names=FALSE,col.names=FALSE)
       write.table(round(c,6),"c:\\testi\\semtest.txt",sep="
",row.names=FALSE,col.names=FALSE)
```

```
        write.table(round(r,6),"c:\\testi\\semprob.txt",sep="
",row.names=FALSE,col.names=FALSE)
pause()
}


 semiparametric.test<-function (x,y,level=0.05)
{

# salva i quantili a seconda del livello di significatività espresso in "level"
pval<-level/2
z1<-qnorm(pval)
z2<-qnorm(1-pval)

# se le matrici vengono inserite con i pazienti sulle righe, le traspone
# inoltre salva in "m" ed "n" il numero di pazienti per ciasscuna patologia
n1<-nrow(x)
n2<-ncol(x)
m1<-nrow(y)
m2<-ncol(y)

if(n1<n2){
  x<-t(x)
  n<-ncol(x)
  }
if (n1>=n2) {
  n<-n2
   }

if(m1<m2){
  y<-t(y)
  m<-ncol(y)
  }
if(m1>=m2) {
  m<-m2
   }

ngen<-nrow(y)


# Fa corrispondere ad "x" la martice con un maggior numero di pazienti e ad "y"
# quella con un numero minore.

if(m>n){
  z<-x
  x<-y
  y<-z
  rm(z)
  n<-ncol(x)
  m<-ncol(y)
  }

# Calcolo di media, varianza, deviazione standard (relativamente ad y) e
coefficienti "rho"

mean<-apply(y,1,mean)
var<-apply(y,1,var)
var<-((m-1)/m)*var
sd<-sqrt(var)                                #mean, sd e var sono vettori p-
dimensionali
S<-1-(apply(x,2,pnorm,mean=mean,sd=sd))      #S è una matrice pxn
rho<-1/n*(apply(S,1,sum))                    #rho è un vettore p-dimensionale

#calcolo della deviazione standard dei coefficienti rho

 #calcolo di w quadro esse
 g<-(S-rho)**2
 ####w2s <- (1/n)*(apply(g,1,sum))
 w2s <- (1/(n-1))*(apply(g,1,sum))
```

```
#"matrice" omega
 omega1<-var
 omega2<-2*(var**2)

#calcolo del primo elemento del vettore beta
 zi<-(x-mean)/sd
 densxi<-apply(zi,2,dnorm)
 beta1<-(1/(n*sd))*apply(densxi,1,sum)

#calcolo del secondo elemento del vettore beta
 si<-zi*densxi
 beta2<-(1/(2*n*var))*apply(si,1,sum)
 ######beta2<-(1/(n*sd))*apply(si,1,sum)

 p1<-omega1*(beta1**2)
 p2<-omega2*(beta2**2)
 p<-(n/m)*(p1+p2)
 var.rho<-(w2s+p)


#passaggio al logit

tau<-log(rho/(1-rho))
var.tau<-(var.rho)/(rho^2*(1-rho)^2*n)
toss.tau<-(tau)/sqrt(var.tau)
tau.inf<-z1*sqrt(var.tau)
tau.sup<-z2*sqrt(var.tau)


rho<-exp(tau)/(exp(tau)+1)
rho.inf<-exp(tau.inf)/(exp(tau.inf)+1)
rho.sup<-exp(tau.sup)/(exp(tau.sup)+1)
test<-toss.tau
var.test<-var.tau


#calcolo dei valori p e conta dei geni significativi
test<-toss.tau
non.agg<-2*(1-pnorm(abs(test)))
corr<-mt.rawp2adjp(non.agg)
agg<-corr$adjp[order(corr$index),]

c1<-rep(0,ngen)
pvalue<-agg[,1]
c1[agg[,1]<level]<-1
n.pvalue<-sum(c1)

c2<-rep(0,ngen)
adjp.Bonferroni<-agg[,2]
c2[agg[,2]<level]<-1
n.adjp.Bonferroni<-sum(c2)

c3<-rep(0,ngen)
adjp.Holm<-agg[,3]
c3[agg[,3]<level]<-1
n.adjp.Holm<-sum(c3)

c4<-rep(0,ngen)
adjp.Hochberg<-agg[,4]
c4[agg[,4]<level]<-1
n.adjp.Hochberg<-sum(c4)

c5<-rep(0,ngen)
adjp.SidakSS<-agg[,5]
c5[agg[,5]<level]<-1
n.adjp.SidakSS<-sum(c5)

c6<-rep(0,ngen)
adjp.SidaskSd<-agg[,6]
c6[agg[,6]<level]<-1
```

```
n.adjp.SidaskSd<-sum(c6)

c7<-rep(0,ngen)

adjp.Bh<-agg[,7]
c7[agg[,7]<level]<-1
n.adjp.Bh<-sum(c7)

c8<-rep(0,ngen)
adjp.BY<-agg[,8]
c8[agg[,8]<level]<-1
n.adjp.BY<-sum(c8)

list(rho=rho, rho.inf=rho.inf, rho.sup=rho.sup, test=test, var.test=var.test,
pvalue=pvalue, adjp.Bonferroni=adjp.Bonferroni, adjp.Holm=adjp.Holm,
adjp.Hochberg=adjp.Hochberg, adjp.SidakSS=adjp.SidakSS,
adjp.SidaskSd=adjp.SidaskSd, adjp.Bh=adjp.Bh, adjp.BY=adjp.BY, n.pvalue=n.pvalue,
n.adjp.Bonferroni=n.adjp.Bonferroni, n.adjp.Holm=n.adjp.Holm,
n.adjp.Hochberg=n.adjp.Hochberg, n.adjp.SidakSS=n.adjp.SidakSS,
n.adjp.SidaskSd=n.adjp.SidaskSd, n.adjp.Bh=n.adjp.Bh, n.adjp.BY=n.adjp.BY)

}
sel<-function (test, yes,len)
{
    storage.mode(test) <- "logical"
    ans <- test
    nas <- is.na(test)
    if (any(test[!nas]))
        ans[test & !nas] <- rep(yes, length.out = length(ans))[test & !nas]
    ans[nas] <- NA
   ans
        print(datisignificativi<-length(ans[ans>0]))
        print(datinonsignificativi<-len-length(ans[ans>0]))
}
```

Riferimenti e bibliografia

[1] Gordon K. Smyth, Yee Hwa Yang and Terry Speed *Statistical Issues in cDNA Microarray Data Analysis (2002)*

[2] David M.Rocke and Blythe Dubrin *A Model for Measured for Gene Expression Arrays (2001)*

[3] B. Lausen *Statistical analysis of genetic distance data*

[4] Golub, T.R Slomin, D.K Tamayo, P.Huard et al *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring (1999)*

[5] Jen-Michel Clavarie *Computational methods for identification of differential coordinated gene expressiondata (1999)*

[6] T.R. Golub, D.K.Slomin. M.A.Caliguri, et al *Molecular classification by* gene *expression monitoring (oct 1999)*

[7] J.Platt *Fast training of support vector machine using sequential minimal optimization (1999)*

[8] M.P.S Brown, W.N. Grundy et al *Knowledge-based analysis of microarray gene expression data by using support vector machines(2000)*

[9] A.Brn-Dor, N.Friedman et al *tissue classification with gene expression profiles (2000)*

[10] A.A.Alizade, M.B. Elisen R.E. Davis, C.Ma, I.S. Lossos, A. Rosenvald et al. *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiles (2000)*

[11] Perou, C.M. Serlie, T.Elsen , M.B van deRijn , et al. *Molecular portraits of human breast tumors* Nature(2000)

[12] van't Veer, LJ Dai, H. vande Vijer, MJ, He, YD Hart, AM et al. *Gene expression profiling predicts clinical outcome of breast cancer* Nature(2002)

[13] Amir Ben-Dor, Laurakey Bruhn, Nir Friedman, Iftach Nachaman, Michèl Schummer, Zohar Yakhini *Tissue classification with gene expression profiles, Proceedings of the fourth annual international conference of Computational molecular Biology (apr 2000)*

[14] A . Keller , M. Shummer, L-Hood, W. Ruzzo *Bayesian classification of DNA array expression data* Technical report, University of Washington (Aug 2000)

[15] Gen Hori, Masato Inoue, Shin-ichi Nishimura and Hiroyuki Nakahara *Bild gene classification based on ICA of microarray data.*

[16] Sidney Siegel N.John Castellan JR. *Statistica non parametrica (1992)*

[17] Peter Armitage geofrey Berry *Statistica medica* (1996)

[18] www.bepress.com/uwbiostat/paper184