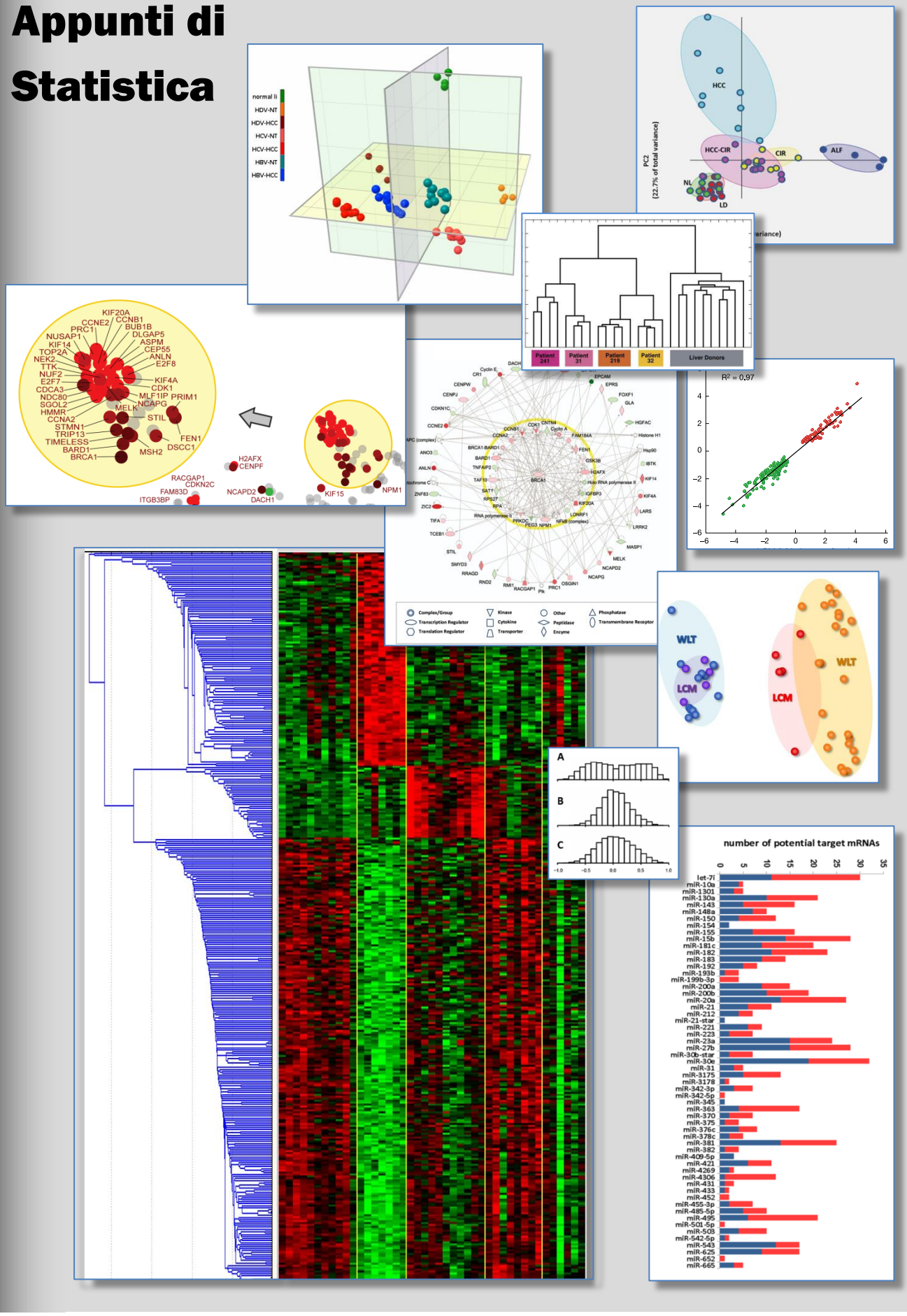
Giacomo Diaz

Appunti di Statistica



Indice

Note. Le funzioni di Excel riportate in rosso in alcuni punti sono state verificate in Excel 2010. Probabilmente funzionano senza modifiche anche in versioni successive ma non è certo. Gli argomenti delle funzioni di Excel in lingua italiana sono separati dal punto e virgola. Quelli di Excel in lingua inglese sono separati dalla virgola. Un problema per chi ha Office in lingua italiana è il fatto che le macro di Excel e l'ambiente di sviluppo VBA hanno comunque i nomi delle funzioni in inglese, che non corrispondono ai nomi in italiano delle funzioni inseribili nelle celle dello spreadsheet.

Scopi principali dei metodi statistici

- rappresentare sinteticamente i dati
- valutare la variabilità
- fare confronti
- verificare modelli, fare previsioni
- analizzare relazioni tra variabili e tra oggetti
- individuare tipologie, discriminare gruppi, classificare oggetti, ecc.

Campione e popolazione

Il campione è formato da un gruppo limitato di oggetti/soggetti scelti a rappresentare l'intera popolazione, e da cui speriamo di ottenere informazioni valide sulla popolazione. Notare che la definizione di popolazione in statistica non corrisponde alla definizione di popolazione naturale. In statistica, le popolazioni si riferiscono alle variabili, come i valori dei parametri molecolari, biochimici, fisici, comportamentali, ecc. E a differenza dalle popolazioni biologiche, le popolazioni statistiche... non si riproducono. La popolazione è spesso costituita da un numero infinito di individui, oppure anche da un numero fisicamente finito ma praticamente irraggiungibile (es., tutti i soggetti con una certa malattia, tutti i globuli rossi di un soggetto, ecc.) per cui è quasi sempre impossibile disporre dei dati di tutti i componenti della popolazione. I campioni possono essere osservazionali quando ci si limita ad osservare i fenomeni naturali, senza intervenire su di essi (es., animali cresciuti nel loro ambiente), oppure sperimentali quando si interviene tentando di modificarli (es., animali allevati in particolari condizioni o con particolari trattamenti farmacologici). E' comunque fondamentale che il campione sia univocamente definito. Se ad es., si parla di animali di laboratorio occorre specificare la varietà, l'età, il peso, il sesso e le condizioni di allevamento come la dieta, il ritmo giorno notte, ecc. Analogamente, in campo clinico, occorre specificare età, sesso, provenienza, eventuali commorbidità dei pazienti e dei controlli sani, ecc.. A seconda dei problemi, i metodi statistici possono prendere contemporaneamente in considerazione uno, due o più campioni da confrontare, e di ciascun campione possono valutare una, due o più variabili (statistiche mono, bi, o multivariate).

Fonti di variabilità e campionamento. Imprecisione e vizio

Come variabilità statistica bisogna intendere la somma di due fattori: (a) la variabilità naturale dei fenomeni [i frutti di una stessa pianta hanno pesi diversi] e (b) la variabilità dovuta all'imprecisione dell'acquisizione dei dati [uno stesso frutto pesato più volte con la stessa bilancia - difettosa - mostra pesi diversi]. L'imprecisione in senso stretto determina un aumento della variabilità, ma di per sé non altera la media. L'imprecisione può essere valutata, ad es. facendo diverse misurazioni di uno stesso oggetto, e quindi in parte compensata, da un maggior numero di misurazioni e facendone la media. Un certo grado di imprecisione è comunque sempre presente; l'importante è che essa sia così piccola da essere trascurabile in rapporto alla variabilità intrinseca del fenomeno da valutare.

Il vizio invece è invece un tipo di errore molto più subdolo in quanto causa una alterazione costante dei dati (vedi una bilancia starata) ma non accresce la variabilità. In assenza di campioni di riferimento, detti anche standard o gold standard (nel caso della bilancia, un oggetto di peso noto) il vizio è impossibile da riconoscere e correggere, e questo ha conseguenze ben più negative della semplice imprecisione, in quanto determina sottostime o sovrastime, e spesso differenze tra i dati

ottenuti in diversi laboratori. Analizzeremo questo secondo caso quando parleremo di correlazioni spurie e correlazioni parziali.

L'approssimazione numerica è un buon esempio di controllo della precisione e di assenza di vizio. Infatti si eliminano le cifre meno significative che non sono rilevanti, eventualmente frutto di imprecisione, ma d'altro canto un semplice troncamento porterebbe ad una diminuzione costante del valore (vizio di sottostima). Quindi si arrotonda il valore dell'ultima cifra mantenuta a in base al valore della successiva cifra eliminata. Invece un tipico esempio di vizio dovuto a troncamento è quello dell'età. L'età infatti è aggiornata solo allo scadere del compleanno. Pertanto risulta che l'età è mediamente sottostimata di ½ anno. Ma questa è una convenzione e non crea problemi. Se si vuole considerare la vera età fisiologica media basta sommare 0.5 all'età anagrafica media.

Vizi di misure e vizi di campionamento

Imprecisione e vizio non risiedono solo nella fase di misurazione o di rappresentazione numerica dei dati. Il loro pericolo è presente soprattutto nella fase del campionamento. Se ad es. occorre formare un gruppo di soggetti da sottoporre a trattamento ed un altro gruppo da utilizzare come controllo, è sempre bene estrarre a sorte quali soggetti andranno a formare l'uno o l'altro gruppo. E' classico l'esempio del ricercatore che destina al trattamento i primi animali che riesce a prendere dalla gabbia e lascia gli altri come controllo. E' assai probabile che la condizione fisica degli animali che si lasciano prendere per primi sia differente da quella degli altri che si sottraggono alla cattura. Oppure l'esempio dei pazienti dell'illustre medico con onorari a quattro zeri. Si può suppore che i pazienti di tale medico abbiano condizioni economiche e quindi di vita, alimentazione, ecc. piuttosto speciali. Lo stesso vale per indagini sulla popolazione utilizzando referti diagnostici (es. radiografie). E' ovvio che i soggetti che si sottopongono a diagnosi soffrono di qualche disturbo e non lo fanno per passatempo, e quindi non possono onestamente rappresentare l'intera popolazione. E' quindi chiaro che il risultato dell'indagine può essere deviato o influenzato da cause estranee introdotte con il campionamento. La randomizzazione (o campionamento casuale) è in genere un buon criterio. Ma migliori sono quei metodi attuano un campionamento sistematico o stratificato, che consiste nella scelta equilibrata e bilanciata dei soggetti considerando le diverse condizioni ambientali. Dovendo ad es. estrarre un campione di soggetti che rappresentino gli abitanti di una città è senz'altro meglio scegliere individui di diversa età, sesso, professione, quartiere di residenza, ecc. anziché procedere in modo del tutto randomizzato.

I tipi principali di variabili

Capire i diversi tipi di variabili è essenziale per la scelta dei metodi statistici da applicare.

(a) Il semplice conteggio. Ciò che interessa è solo il numero degli individui o degli eventi, ovviamente riferiti ad un certo ambito di spazio o di tempo (es., numero di batteri in un certo volume di terreno, numero di cellule per campo microscopico, numero di impulsi elettrici in un certo intervallo di tempo, numero di abitanti per Km^2, ecc.). La numerosità riferita all'unità di spazio è in genere detta densità; quella riferita all'unità di tempo è in genere detta frequenza. I valori sono ovviamente discreti, cioè rappresentati da numeri interi. Ma in ogni caso è importante che l'intervallo di spazio o di tempo non ponga limiti al numero di individui o eventi osservabili al suo interno, o per lo meno, i numeri osservati siano ben lontani dalla condizione di saturazione. Ad es. il numero di persone presenti in un autobus è limitato dalle dimensioni dell'autobus e quindi la distribuzione del numero di viaggiatori in un autobus è fortemente condizionata. Idem per le cellule in coltura che a confluenza subiscono inibizione da contatto.

(b) Il conteggio di oggetti, questa volta distinti per tipo, carattere, categoria, qualità, modalità, attributo, ecc. Ciò che interessa non è più il semplice numero di individui, ma il numero di diversi tipi o classi di soggetti. Esiste anche questa volta il riferimento ad un certo ambito spaziale o temporale, ma giusto per conoscere il contenitore da cui provengono i soggetti. L'interesse principale sta ora nella ripartizione degli individui con diverse caratteristiche (es., quante cellule normali e non-normali in una sezione di tessuto, quanti linfociti e monociti in uno striscio di sangue, quanti individui di gruppo A, B, AB e 0 o quanti soggetti maschi e quanti femmine in una certa popolazione, ecc.). Tali conteggi sono spesso espressi come proporzioni o rapporti percentuali (es., il 5% dei globuli bianchi è rappresentato da monociti, ecc.) ma ai fini statistici deve essere sempre noto il numero assoluto reale. Attenzione: le modalità' o classi delle variabili qualitative devono sempre essere mutualmente esclusive, senza sovrapposizioni o ambiguità, ed esaustive comprendere tutte le possibili tipologie del campione. Il caso più semplice di modalità mutualmente esclusive ed esaustive è quello della presenza/assenza di un determinato carattere (es., malattia presente/malattia assente), oppure la presenza di un determinato carattere contro tutti gli altri (es., oggetti di colore rosso/oggetti di qualsiasi altro colore).

(c) Il conteggio di oggetti distinti per tipo come sopra, ma ordinabili su una certa scala o grado di apprezzamento. L'esempio classico è quello degli scores o punteggi (es., il grado di una patologia, assente/lieve/moderata/grave o il merito insufficiente/sufficiente/buono/ottimo, ecc.). Gli scores nominali sono spesso sostituiti da valori all'interno di scale convenzionali (es., grading tumorale, punteggi scolastici, quoziente di intelligenza, ecc.) rappresentati in genere da piccoli interi. E' interessante a questo punto notare come molte scale di qualità (es., affidabilità di un venditore, qualità di un alimento, categoria di un albergo, ecc.) siano in effetti ottenute dalla combinazione di varie caratteristiche misurabili singolarmente.

(d) Le misure. Riguardano grandezze fisiche (es. densità, peso, grandezza, durata, massa, distanza, forza, ecc. ecc.). Sono valutabili con un grado di precisione teoricamente infinito anche se praticamente limitato dal livello di accuratezza dello strumento di misura e comunque adeguato alle nostre esigenze. I valori sono rappresentati da numeri su scala continua. Per praticità si usa approssimare i dati alle prime 3 o 4 cifre significative, capaci in tal modo di distinguere una gamma di mille o diecimila valori diversi. Esiste la possibilità di trasformare le variabili continue nel tipo precedente, cioè come variabili qualitative ordinabili sostituendo al valore dei dati il rango, cioè la posizione occupata nel loro ordine crescente. Con alcuni vantaggi e svantaggi.

Esempio

dati	disposti in ordine crescente	ranghi
1.5	1.5	1
6.3	2.9	2
3.2	3.2	3
9.1	6.3	4
2.9	9.1	5

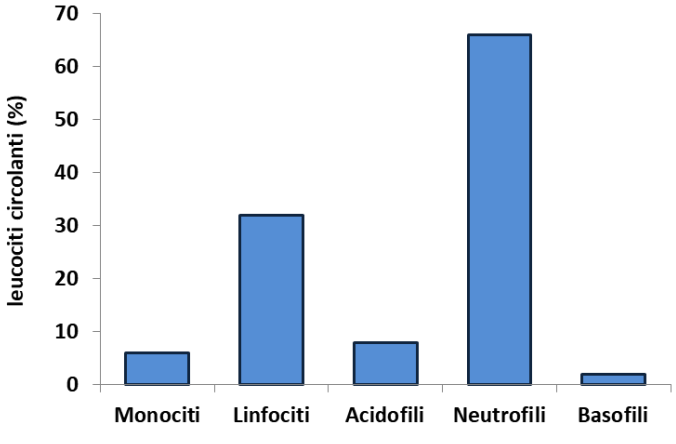
La trasformazione in rango comporta necessariamente una perdita di informazione (aspetto negativo) ma elimina anche eventuali condizionamenti della distribuzione che potrebbero invalidare diverse analisi statistiche basate sull'assunto che i dati abbiano una distribuzione normale (aspetto positivo). L'argomento sarà ripreso quando si parlerà dei cosiddetti metodi non-parametrici.

Istogramma di frequenze

Una utile forma di rappresentazione dei dati consiste nell'istogramma delle frequenze dei dati, che possono essere ripartiti per categorie (variabili nominali) o per classi di intervalli (variabili quantitative o nominali ordinabili). Ad es. consideriamo le frequenze di diversi tipi di leucociti (variabile nominale non ordinabile) presenti in uno striscio di sangue:

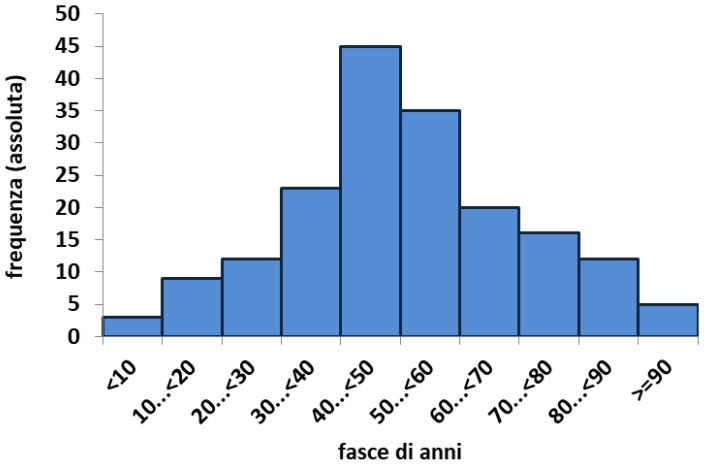
Categoria	frequenza
Monociti	6
Linfociti	32
Granulociti acidofili	8
Granulociti neutrofili	66
Granulociti basofili	2

L'informazione della tabella può essere meglio presentata in forma grafica come istogramma di frequenze:



Vediamo ora le frequenze di una variabile quantitativa, ad es. l'età di un campione di 180 soggetti (dati di pura fantasia), ripartiti per fasce di età di 10 anni:

età	frequenza
<10	3
10-<20	9
20-<30	12
30-<40	23
40-<50	45
50-<60	35
60-<70	20
70-<80	16
80-<90	12
=>90	5



Per fare un istogramma di una variabile quantitativa è importante che le classi non siano né troppe (si otterrebbe un istogramma 'sdentato' irregolare) né troppo poche (si otterrebbe un istogramma concentrato senza dinamica).

C'è una formula che suggerisce quante classi rappresentare in base al numero di dati (n):

n° ottimale di classi = $1 + 3.3 \log_{10}(n)$

Es.:
se n = 100, n° ottimale di classi = $1 + 3.3\log_{10}(100) \approx 8$
se n = 500, n° ottimale di classi = $1 + 3.3\log_{10}(500) \approx 10$

Ma comunque non va bene seguire alla lettera questa formula. Prima di tutto è bene fare in modo che i confini degli intervalli delle classi corrispondano a valori interi. E inoltre se nel campo specifico esistono valori critici o convenzioni consolidate sugli intervalli delle classi, è bene rispettarli per consentire confronti con i dati di altri laboratori.

Alla ricerca di un parametro di riferimento...

che rappresenti la variabilità dei dati di una popolazione o di un campione, e che sia:

1. prodotto con il contributo di tutti i dati
2. indipendente dalla numerosità dei dati
3. espresso nella stessa unità di misura dei dati

Quando la distribuzione è normale o per lo meno simmetrica è estremamente importante associare alla media un parametro che esprima la dispersione o variabilità dei dati. Partiamo da un esempio concreto.

campione A: 3, 6, 4, 7
campione B: 0, 1, 9, 10

I due campioni hanno la stessa media (m = 5). Ma è evidente che la distribuzione del campione B, che va da 0 a 10, è più dispersa di quella del campione A, che va da 3 a 7. La media quindi è del tutto indipendente dalla variabilità dei dati. Per arrivare ad un parametro che esprima la variabilità cominciamo a considerare le differenze tra i dati e la media del campione A.

Es.:

dato x	differenza x-media
3	3-5= -2
6	6-5= +1
4	4-5= -1
7	7-5= +2
somma = 20 media = 5	somma = 0

Le differenze tra i dati e la media sono dette in gergo deviazioni o scarti. Notiamo subito che la somma degli scarti, positivi e negativi, è necessariamente nulla in quanto per definizione la media è data dalla somma dei dati. Pertanto la somma degli scarti non serve a niente. Potrebbe solo servire per verificare se abbiamo calcolato bene le differenze. Le cose cambiano se eleviamo al quadrato gli scarti. In questo modo tutti gli scarti acquistano segno positivo.

dato x	differenza x-media	differenza $(x\text{-media})^2$
3	3-5= -2	$(3-5)^2= 4$
6	6-5= +1	$(6-5)^2= 1$
4	4-5= -1	$(4-5)^2= 1$
7	7-5= +2	$(7-5)^2= 4$
somma = 20 media = 5	somma = 0	somma = devianza (S) = 10

In gergo, la somma delle delle differenze al quadrato tra ciascun dato e la media è detta semplicemente '**somma dei quadrati**' o **devianza**. Il primo termine, somma dei quadrati, sarebbe da preferire a quello di devianza perché è quello più utilizzato in letteratura. Il simbolo più frequente è una **S** maiuscola, ma talora si trova anche il simbolo **SS** (Sum of Squares) o in italiano **SQ** (Somma dei Quadrati). In simboli algebrici:

$S = \Sigma(x-m)^2$

In Excel, somma dei quadrati =DEVSQ(...)

La devianza è un potenziale indice di dispersione. Ma ha un grosso difetto: aumenta con l'aumentare del numero dei dati. Se ad es., raddoppiamo il campione, la media non cambia, ma la devianza raddoppia:

dato x	differenza x-media	differenza (x-media)²
3	3-5= -2	(3-5)²= 4
6	6-5= +1	(6-5)²= 1
4	4-5= -1	(4-5)²= 1
7	7-5= +2	(7-5)²= 4
3	3-5= -2	(3-5)²= 4
6	6-5= +1	(6-5)²= 1
4	4-5= -1	(4-5)²= 1
7	7-5= +2	(7-5)²= 4
media = 5		devianza = 20

Quindi la devianza aumenta in funzione di n. Per ottenere un parametro stabile, a prescindere dalla numerosità del campione (tecnicamente: invariante rispetto ad n), possiamo dividere la devianza per il numero di dati. Il risultato di tale operazione è **detto scarto quadratico medio** o meglio **varianza**, indicata col simbolo **s²** (s minuscolo al quadrato):

$$s^2 = \frac{\sum (x - m)^2}{n}$$

In Excel, varianza =VAR.P(...) [P sta per population]

Purtroppo nei campioni poco numerosi la varianza così stimata è inferiore (sottostimata) rispetto alla vera varianza della popolazione. Per correggere tale vizio è preferibile dividere la devianza per n-1 anziché per n.

$$s^2 = \frac{\sum (x - m)^2}{n - 1}$$

In Excel, varianza =VAR.S(...) [S sta per sample]

Con campioni di grandi campioni (n>100) tale correzione è irrilevante. E quindi è meglio applicarla sempre. Nel nostro caso dobbiamo comunque applicarla in quanto n = 4,

$s^2 = 10/(4-1) = 3.333$.

Al valore n-1 si da il nome di **gradi di libertà** (**gdl**, o **degrees of freedom, DF**). Perché n-1? Perché i dati che concorrono a determinare la varianza in modo indipendente sono n-1 e non n. Infatti un dato è stato già 'speso' per calcolare la media (la varianza l'abbiamo calcolata partendo dagli scarti dalla media). Per chiarire meglio questo punto diciamo che qualsiasi media può essere modificata a piacere modificando anche un solo dato:
media di 6, 4, 8, 10, **2** = 6
media di 6, 4, 8, 10, **27** = 11
media di 6, 4, 8, 10, **52** = 16
ecc.
Quindi, se un solo dato è in grado di condizionare la media, allora, ragionando a rovescio, una volta stabilita una certa media, i dati effettivamente liberi di variare non sono più n ma n-1. In seguito

capiterà spesso di considerare altri parametri statistici. In tutti i casi, quei parametri avranno anch'essi dei gradi di libertà che corrisponderanno al numero di dati liberi di variare indipendentemente l'uno dall'altro una volta stabilito il valore di quel determinato parametro.

Tornando alla nostra storia, anche la varianza ha però un problema: quello di essere espressa in una unità di scala quadratica (essendo ottenuta dai quadrati degli scarti). Ad es., la varianza della statura è espressa in m^2, la varianza dell'età in $anni^2$, ecc. Per ottenere un parametro di dispersione nella stessa scala di misura dei dati estraiamo la radice quadrata. Il parametro ottenuto da tale operazione è detto **deviazione standard**, indicata col simbolo **s** (minuscolo):

$$s = \sqrt{\frac{\sum (x - m)^2}{n - 1}}$$

Nel nostro caso $s = \sqrt{3.333} = 1.83$

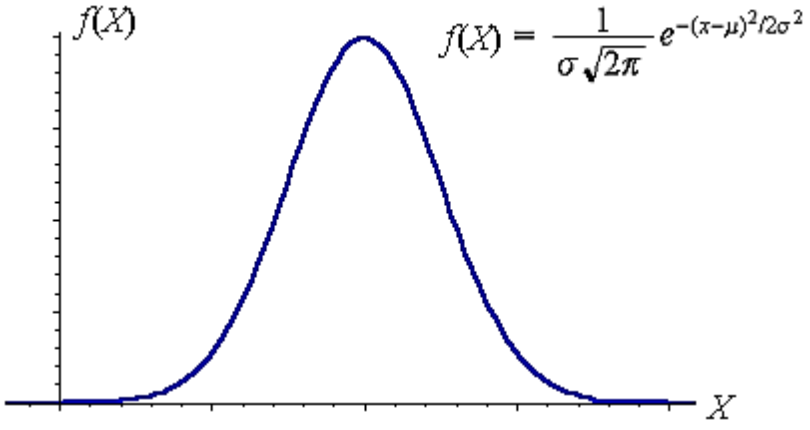
In Excel, deviazione standard =STDEV.S(...)

Con la deviazione standard abbiamo finalmente trovato un parametro che
- rappresenta la dispersione dei dati attorno alla media
- tiene conto di tutti i dati del campione (a differenza del range minimo - massimo)
- non è influenzato dalla numerosità del campione (a differenza della devianza)
- è valutato nella stessa scala dei dati originari (a differenza della varianza).

Media e deviazione standard definiscono esaustivamente una distribuzione normale.
Media e deviazione standard del campione sono le migliori stime della media e deviazione standard, non note, della popolazione, talvolta indicate con le lettere greche μ e σ. Pertanto, sulla base della media e deviazione standard del campione possiamo stimare la distribuzione della popolazione e da questa fare valutazioni di probabilità e varie altre cose.

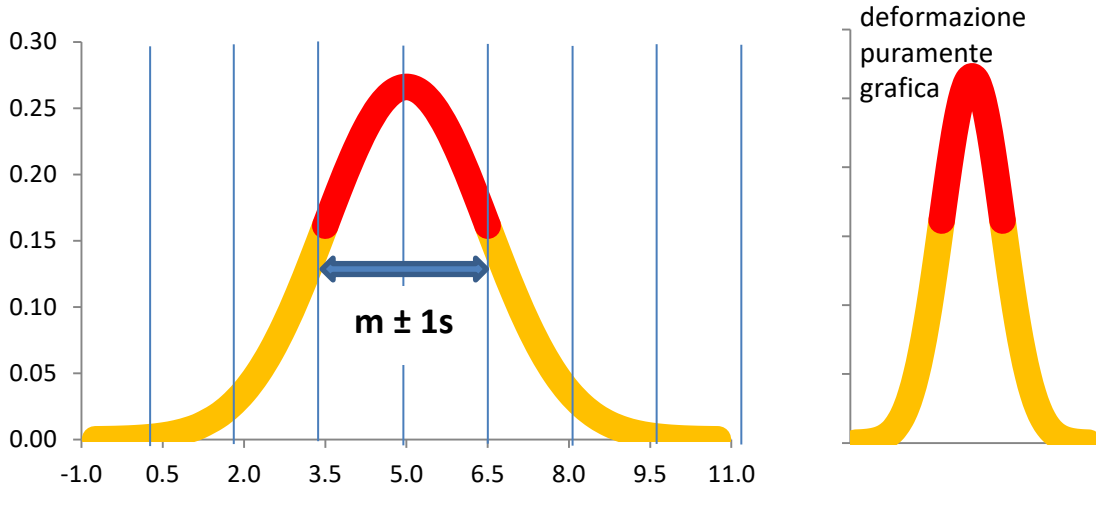
Dall'istogramma delle frequenze alla curva normale

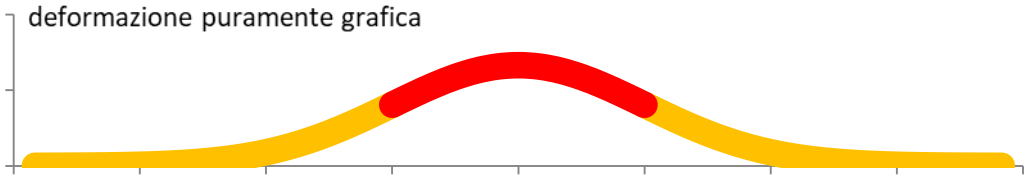
Ritornando all'istogramma visto in precedenza, ed ipotizzando di avere infiniti dati a disposizione, la formula indicherebbe di creare un numero infinito di classi. In questo modo il profilo dell'istogramma diventerebbe equivalente a quello di una curva continua. La curva della distribuzione di una popolazione infinita può pertanto considerarsi come un istogramma suddiviso in infinite classi di intervallo infinitesimo. Se la distribuzione è 'normale' (è la distribuzione più frequente in natura ma non necessariamente sempre) il grafico e la funzione sono questi:



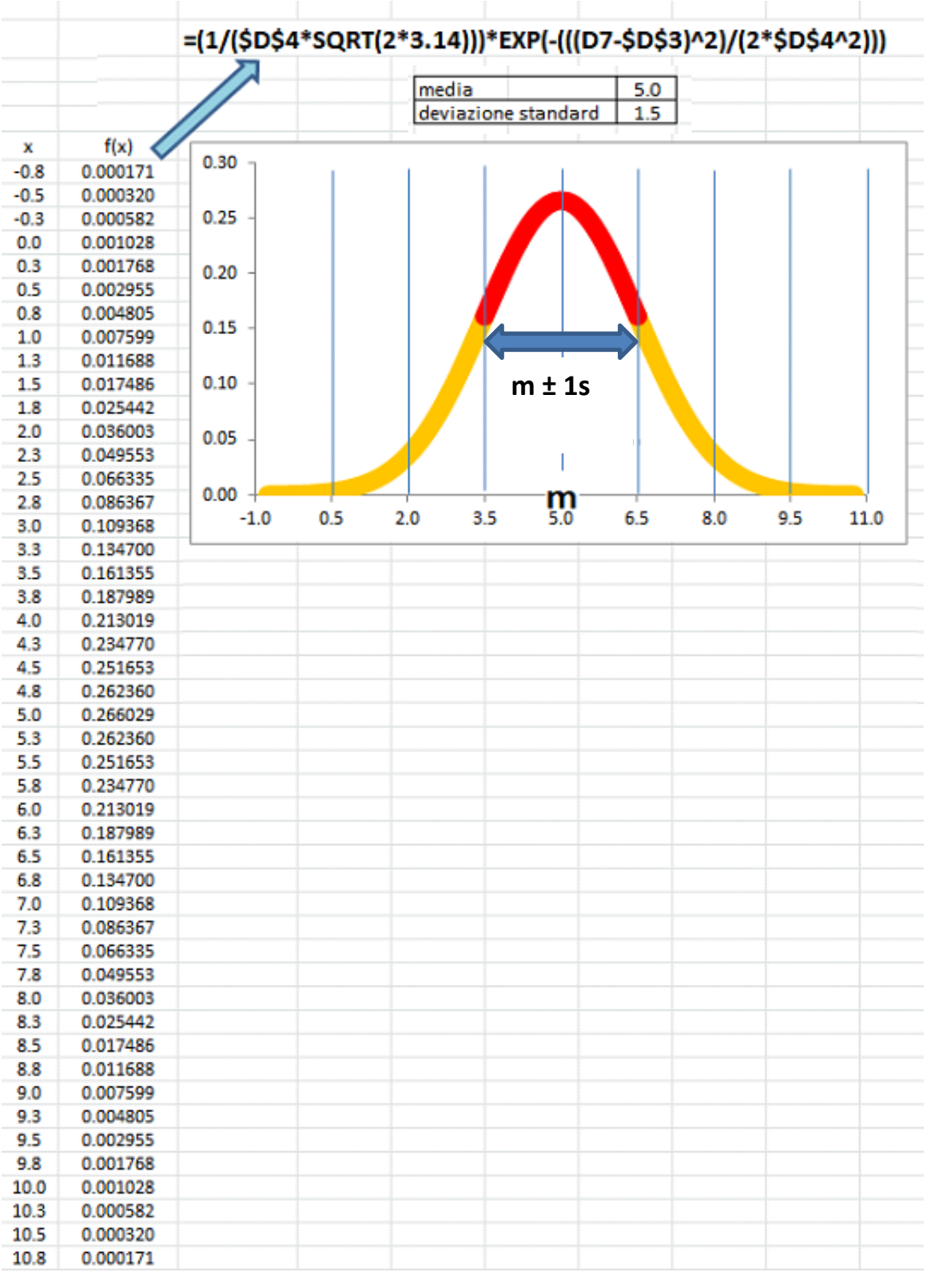
La funzione calcola l'altezza della curva per ogni valore di x, conoscendo due soli parametri: la media (μ) e la deviazione standard (σ). Le lettere greche sono sempre riferite alla popolazione. Noi useremo i valori di media e deviazione standard calcolati dai campioni.

Il grafico seguente mostra una distribuzione normale con media = 5 e deviazione standard = 1. Deformazioni puramente grafiche degli assi non alterano la curva, come a priva vista sembrerebbe. Questo fa capire come a occhio sia difficile valutare variazioni della kurtosis, parametro di cui parleremo tra breve.





Il grafico è stato ottenuto con Excel utilizzando la funzione vista in precedenza. Notare il calcolo va ripetuto per ogni valore x in ascissa, come mostrato sotto.

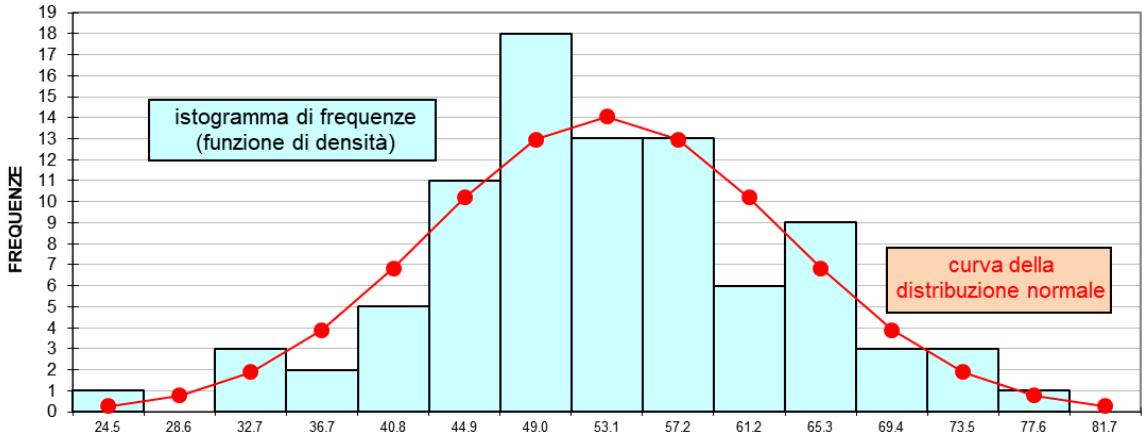


x	f(x)
-0.8	0.000171
-0.5	0.000320
-0.3	0.000582
0.0	0.001028
0.3	0.001768
0.5	0.002955
0.8	0.004805
1.0	0.007599
1.3	0.011688
1.5	0.017486
1.8	0.025442
2.0	0.036003
2.3	0.049553
2.5	0.066335
2.8	0.086367
3.0	0.109368
3.3	0.134700
3.5	0.161355
3.8	0.187989
4.0	0.213019
4.3	0.234770
4.5	0.251653
4.8	0.262360
5.0	0.266029
5.3	0.262360
5.5	0.251653
5.8	0.234770
6.0	0.213019
6.3	0.187989
6.5	0.161355
6.8	0.134700
7.0	0.109368
7.3	0.086367
7.5	0.066335
7.8	0.049553
8.0	0.036003
8.3	0.025442
8.5	0.017486
8.8	0.011688
9.0	0.007599
9.3	0.004805
9.5	0.002955
9.8	0.001768
10.0	0.001028
10.3	0.000582
10.5	0.000320
10.8	0.000171

Comunque in Excel troviamo una funzione che semplifica il calcolo della curva normale:
f(x) = NORM.DIST(x, mean, stdev, FALSE) [FALSE sta per non cumulativa]

Attenzione, una media ed una deviazione standard si possono sempre calcolare per qualsiasi campione di dati, ma questo non vuol dire che ciò abbia sempre senso. Se la distribuzione non è normale, o per lo meno se non è simmetrica, media e deviazione standard non rappresentano niente. Quindi media e deviazione standard si calcolano sempre assumendo che la distribuzione sia normale. La condizione di normalità è molto importante anche per andare avanti, perché molti test statistici si basano su questo assunto. Nel caso in cui i dati non siano distribuiti normalmente bisogna ricorrere ad altri test che vedremo in seguito. Per cui prima di effettuare test statistici occorre verificare le loro assunzioni circa la distribuzione dei dati. Esistono numerosi test per verificare se i dati sono distribuiti normalmente. Alcuni li vedremo tra breve.

In Excel possiamo anche sovrapporre la curva normale teorica all'istogramma delle frequenze - si tratta di sovrapporre un grafico a colonne ad uno a linee. In questo grafico possiamo notare che i dati si adattano abbastanza bene alla distribuzione normale, e quindi possiamo ritenere che essi siano distribuiti normalmente.

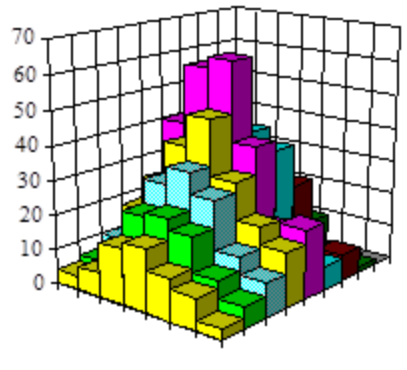


Se poi consideriamo due variabili contemporaneamente (ad es. peso e statura) possiamo anche fare tabelle di frequenza a due entrate, come la seguente

peso (kg)	<80	80-<100	100-<120	120-<140	140-<160	160-<180	180-<200	>=200
>=120	4	6	10	17	21	19	13	2
100-<120	7	11	15	27	43	35	18	4
80-<100	16	23	30	39	60	37	25	6
60-< 80	18	24	35	48	63	41	32	18
40-< 60	12	21	29	32	40	37	25	12
20-< 40	9	11	15	21	18	9	3	0
<20	3	6	10	16	20	7	8	2

statura (cm)

che possono ancora essere rappresentate graficamente. Il grafico 3D è di un certo effetto ma poco informativo perché numerose classi risultano nascoste.



Approfittiamo di questo istogramma per dire che esso rappresenta una distribuzione bivariata definita cioè da 2 variabili. Oltre questo non è possibile rappresentare (decentemente) tabelle né istogrammi di frequenze di più di due variabili. E' tuttavia possibile estrarre importanti informazioni attraverso i cosiddetti metodi multivariati di cui ci occuperemo verso la fine del corso.

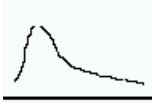
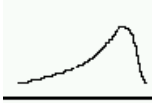
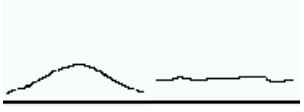
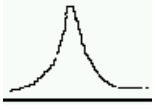
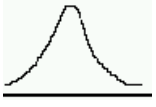
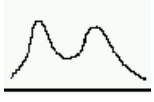
Riprendiamo il discorso dell'istogramma più semplice, cioè riferito ad una sola variabile (statistica monovariata). Dall'esame del grafico è possibile valutare tre importanti proprietà della forma della distribuzione: la simmetria (skewness), il grado di appiattimento o curtosi (kurtosis) e la modalità.

a) Simmetria (skewness). Riguarda il grado di regolarità orizzontale o specularità della curva. La distribuzione normale è del tutto simmetrica. Ma la sola simmetria non è condizione sufficiente per dire che una distribuzione sia normale. Comunque quando la distribuzione è simmetrica, media e mediana coincidono. Il valore di skewness standardizzata è zero se la distribuzione è simmetrica, negativo se vi è una coda a sinistra (per classi con valori molto bassi) e positivo se vi è una coda a destra (per classi con valori molto alti).

b) Grado di appiattimento (kurtosis). Riguarda il grado di regolarità verticale (basso-alto) della curva. La curva di distribuzione normale ha una pendenza che varia in un modo ben definito in ogni suo punto. Questa proprietà non varia anche se dilatiamo o contraiamo la scala dell'ordinata come giocando con un elastico: la curva potrà apparire molto appiattita oppure slanciata, ma i rapporti di pendenza tra i suoi intervalli resteranno invariati. Per la distribuzione normale la kurtosis standardizzata è pari a zero. La kurtosis è negativa (leptocurtica) se aumenta la concavità mentre è positiva (platicurtica) se aumenta la convessità. Ad es., se rubiamo frequenze alle code e le aggiungiamo verso il centro della curva, la kurtosis diminuisce, mentre aumenta nel caso contrario. Il caso più estremo di kurtosis positiva è quello di una distribuzione uniforme visualizzabile come un istogramma piatto.

c) Modalità. La distribuzione normale (come la stragrande maggioranza delle distribuzioni statistiche) è monomodale, cioè presenta un singolo picco di maggior frequenza che poi degrada verso destra e sinistra. Se invece l'istogramma dimostra la presenza di due o più picchi probabilmente si tratta di un campione di un mix di due o più popolazioni. Ad es. campioni comprendenti maschi e femmine producono distribuzioni bimodali. Ma per osservare i due picchi occorre un gran numero di dati, tale da avere un intervallo di classe inferiore alla distanza tra i due picchi, o la differenza tra le medie delle due popolazioni.

Nei software più completi i valori di skewness e kurtosis sono calcolati e testati per verificare se differiscono significativamente dai valori attesi per una distribuzione normale.

	distribuzione non-normale	distribuzione normale	distribuzione non-normale
skewness	 distribuzione asimmetrica coda a destra skewness positiva es. peso di neonati alla nascita	 distribuzione simmetrica skewness nulla	 distribuzione asimmetrica coda a sinistra skewness negativa es. valori del visus corretto
kurtosis	 distribuzione convessa o uniforme kurtosis positiva platicurtica	 distribuzione normalmente flessa kurtosis nulla	 distribuzione concava kurtosis negativa leptocurtica
modalità		 distribuzione monomodale	 distribuzione bimodale

Media e mediana

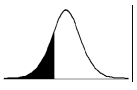
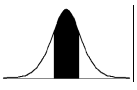
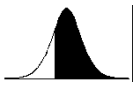
Per giudicare se la distribuzione dei dati sia compatibile con la distribuzione normale un criterio pratico, anche se piuttosto approssimativo, è il confronto tra media e mediana. Sappiamo come la media è calcolata. La mediana invece riporta il valore del dato di mezzo, cioè del dato con rango uguale a n/2+0.5 se il numero dei dati è dispari, oppure alla media dei due dati centrali con rango n/2 e n/2+1 se il numero dei dati è pari (il rango di un dato è la sua posizione, ponendo i dati in ordine crescente). Quando la distribuzione è simmetrica media e mediana coincidono. Quanto più la distribuzione è asimmetrica, tanto più media e mediana differiscono. Nel caso in cui media e mediana differiscano notevolmente è senz'altro meglio prendere come riferimento la mediana anziché la media, perché la distribuzione non è normale. Es., i voti degli esami universitari superati sono chiaramente distribuiti in modo non-normale (vanno da 18 a 30, il 30 è un limite, ecc.). In tal caso, al posto della media è meglio riferire la mediana, cioè il voto raggiunto dal 50% degli studenti. La scelta della media o della mediana anticipa la scelta tra test parametrici e non-parametrici che affronteremo in seguito. La mediana è un dato tipicamente non-parametrico. Vedremo in seguito un semplice test di normalità basato sulla mediana. Comunque, se anche media e mediana coincidono non è detto che la distribuzione sia necessariamente normale.

Un altro modo di indicare la mediana è quello di 50° percentile. L'uso del percentile è utile in quanto è generalizzabile. Ad esempio, il 5° percentile è quel valore-soglia che separa il 5% dei dati più piccoli dal restante 95%. Allo stesso modo, il valore-soglia che separa il 95% dei dati più piccoli dal restante 5% è detto 95° percentile. Una certa scarsa accuratezza di certi programmi e delle vecchie versioni di Excel sta nel fatto che non specificano se il valore del percentile includa o non includa la soglia.

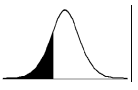
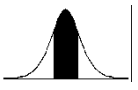
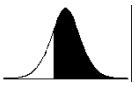
Dati standardizzati

Abbiamo visto come la curva gaussiana sia interamente definita dalla media e deviazione standard. Tuttavia, per un determinato campione, sarebbe piuttosto laborioso calcolare valori di probabilità in base alla Gaussiana definita dalla sua media e deviazione standard. E' molto più semplice ricorrere alla Gaussiana standardizzata, cioè alla distribuzione normale con media = 0 e deviazione standard = 1.

CURVA DI DISTRIBUZIONE NORMALE
media = 0
deviazione standard = 1
area totale sotto la curva = 1

Ascissa valori negativi	Area della curva a sinistra dell'ascissa 	Area della curva compresa tra ±ascissa 	Area della curva a destra dell'ascissa 
z	P(z)	R(z)	Q(z)
-4.000	0.0000	0.9999	1.0000
-3.500	0.0002	0.9995	0.9998
-3.250	0.0006	0.9988	0.9994
-3.000	0.0013	0.9973	0.9987
-2.750	0.0030	0.9940	0.9970
-2.576	**0.0050**	**0.9900**	**0.9950**
-2.500	0.0062	0.9876	0.9938
-2.250	0.0122	0.9756	0.9878
-2.000	0.0227	0.9545	0.9772
-1.960	**0.0250**	**0.9500**	**0.9750**
-1.950	0.0256	0.9488	0.9744
-1.900	0.0287	0.9426	0.9713
-1.850	0.0322	0.9357	0.9678
-1.800	0.0359	0.9281	0.9641
-1.750	0.0401	0.9199	0.9599
-1.700	0.0446	0.9109	0.9554
-1.650	0.0495	0.9011	0.9505
-1.645	**0.0500**	**0.9000**	**0.9500**
-1.600	0.0548	0.8904	0.9452
-1.550	0.0606	0.8789	0.9394
-1.500	0.0668	0.8664	0.9332
-1.450	0.0735	0.8529	0.9265
-1.400	0.0808	0.8385	0.9192
-1.350	0.0885	0.8230	0.9115
-1.300	0.0968	0.8064	0.9032
-1.250	0.1056	0.7887	0.8944
-1.200	0.1151	0.7699	0.8849
-1.150	0.1251	0.7499	0.8749
-1.100	0.1357	0.7287	0.8643
-1.050	0.1469	0.7063	0.8531
-1.000	**0.1587**	**0.6827**	**0.8413**
-0.950	0.1711	0.6579	0.8289
-0.900	0.1841	0.6319	0.8159
-0.850	0.1977	0.6047	0.8023
-0.800	0.2119	0.5763	0.7881
-0.750	0.2266	0.5467	0.7734
-0.700	0.2420	0.5161	0.7580
-0.650	0.2578	0.4843	0.7422
-0.600	0.2743	0.4515	0.7257
-0.550	0.2912	0.4177	0.7088
-0.500	0.3085	0.3829	0.6915
-0.450	0.3264	0.3473	0.6736
-0.400	0.3446	0.3108	0.6554
-0.350	0.3632	0.2737	0.6368
-0.300	0.3821	0.2358	0.6179
-0.250	0.4013	0.1974	0.5987
-0.200	0.4207	0.1585	0.5793
-0.150	0.4404	0.1192	0.5596
-0.100	0.4602	0.0797	0.5398
-0.050	0.4801	0.0399	0.5199
0.000	**0.5000**	**0.0000**	**0.5000**

CURVA DI DISTRIBUZIONE NORMALE
media = 0
deviazione standard = 1
area totale sotto la curva = 1

Ascissa valori positivi	Area della curva a sinistra dell'ascissa 	Area della curva compresa tra ±ascissa 	Area della curva a destra dell'ascissa 
z	P(z)	R(z)	Q(z)
0.000	**0.5000**	**0.0000**	**0.5000**
0.050	0.5199	0.0399	0.4801
0.100	0.5398	0.0797	0.4602
0.150	0.5596	0.1192	0.4404
0.200	0.5793	0.1585	0.4207
0.250	0.5987	0.1974	0.4013
0.300	0.6179	0.2358	0.3821
0.350	0.6368	0.2737	0.3632
0.400	0.6554	0.3108	0.3446
0.450	0.6736	0.3473	0.3264
0.500	0.6915	0.3829	0.3085
0.550	0.7088	0.4172	0.2912
0.600	0.7257	0.4515	0.2743
0.650	0.7422	0.4843	0.2578
0.700	0.7580	0.5161	0.2420
0.750	0.7734	0.5467	0.2266
0.800	0.7881	0.5763	0.2119
0.850	0.8023	0.6047	0.1977
0.900	0.8159	0.6319	0.1841
0.950	0.8289	0.6579	0.1711
1.000	**0.8413**	**0.6827**	**0.1587**
1.050	0.8531	0.7063	0.1469
1.100	0.8643	0.7287	0.1357
1.150	0.8749	0.7499	0.1251
1.200	0.8849	0.7699	0.1151
1.250	0.8944	0.7887	0.1056
1.300	0.9032	0.8064	0.0968
1.350	0.9115	0.8230	0.0885
1.400	0.9192	0.8385	0.0808
1.450	0.9265	0.8529	0.0735
1.500	0.9332	0.8664	0.0668
1.550	0.9394	0.8789	0.0606
1.600	0.9452	0.8904	0.0548
1.645	**0.9500**	**0.9000**	**0.0500**
1.650	0.9505	0.9011	0.0495
1.700	0.9554	0.9109	0.0446
1.750	0.9599	0.9199	0.0401
1.800	0.9641	0.9281	0.0359
1.850	0.9678	0.9357	0.0322
1.900	0.9713	0.9426	0.0287
1.950	0.9744	0.9488	0.0256
1.960	**0.9750**	**0.9500**	**0.0250**
2.000	0.9772	0.9545	0.0228
2.250	0.9878	0.9756	0.0122
2.500	0.9938	0.9876	0.0062
2.576	**0.9950**	**0.9900**	**0.0050**
2.750	0.9970	0.9940	0.0030
3.000	0.9987	0.9973	0.0013
3.250	0.9994	0.9988	0.0006
3.500	0.9998	0.9995	0.0002
4.000	1.0000	0.9999	0.0000

z è il risultato della standardizzazione di un certo valore x della nostra variabile secondo la semplice formula (sottrargli la media e dividere poi per la deviazione standard):

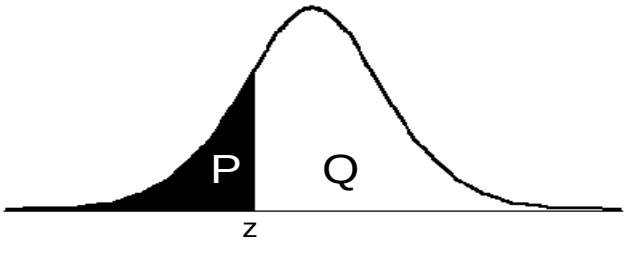
$$z = \frac{x - m}{s}$$

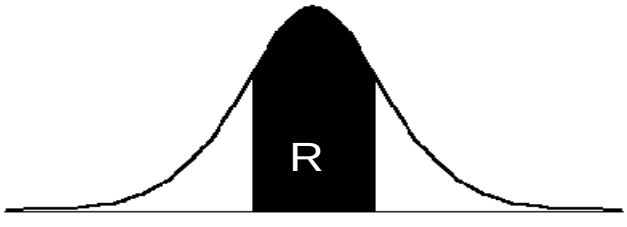
Per convenzione z è il simbolo generalmente utilizzato per i dati standardizzati.

In Excel, z =STANDARDIZE(...)

Ovviamente i dati standardizzati hanno media = 0 e deviazione standard = 1.
Se da z vogliamo calcolare il valore attuale basta invertire la relazione:
$x = (z \cdot s) + m$

P(z) R(z) e Q(z) sono spiegati qui sotto.





Dato un certo z (ascissa)
P(z) è l'area che si trova a sinistra di z (z incluso)
Q(z) è l'area che si trova a destra di z (z incluso)

Ovviamente,
P(z) e Q(z) sono complementari:
P(z) + Q(z) = 1
P(z) = 1 - Q(z)

R(z) invece è l'area compresa tra ± z:
R(z) = 1-2P(z) se z <= 0
R(z) = 1-2Q(z) se z >= 0

Da notare che 'area sotto la curva' significa probabilità.
P rappresenta la probabilità di trovare valori inferiori o uguali a z.
Q rappresenta la probabilità di trovare valori superiori o uguali a z.
R rappresenta la probabilità di trovare valori compresi tra -z e +z.

Siccome i dati sono standardizzati, la tabella è applicabile a qualsiasi tipo di dato, con qualsiasi media e qualsiasi deviazione standard (pesi di moscerini, pesi di balene, altezza di montagne, altezza di formiche, ecc.) purché la distribuzione dei dati sia normale, o assunta come normale.

In Excel,
P(z) =NORM.S.DIST(z, TRUE) [TRUE sta per cumulativa]
R(z) =1 - 2*NORM.S.DIST(z, TRUE)
Q(z) =1 - (NORM.S.DIST(z, TRUE))

all'inverso, per ottenere z
z =NORM.S.INV(P(z))

P(z) si può esprimere come **percentile**. Ad es., se z = -1.750, P(z) sarà ≈ 0.04 (vedi tabella). Vuol dire che abbiamo 4% di probabilità di trovare valori inferiori o uguali a -1.750. Diremo allora che -1.750 rappresenta il 4° percentile della distribuzione. In altre parole, il 4% dei soggetti non supera il valore standardizzato di -1.750.
Similmente, se z = 0.850, P(z) sarà ≈ 0.8. Vuol dire che 0.850 rappresenta l'80° percentile: l"80% dei soggetti non supera il valore standardizzato di 0.850.

Un importante riferimento della distribuzione è il range interquartile, cioè l'intervallo compreso tra il primo quartile ed il terzo quartile, cioè il 25° ed il 75° percentile. Il range interquartile comprende il 50% di tutti i dati.

Se poi vogliamo valutare le probabilità relative ad un certo intervallo di scala compreso tra due valori qualsiasi x_1 e x_2, dopo aver standardizzato questi in z_1 e z_2 troviamo in tabella i valori di P(z) di ciascuno e calcoliamo la differenza.

Corrispondenza tra deviate standard e percentili... se la distribuzione è normale

E' importante ricordare alcuni valori critici di riferimento della distribuzione normale.
L'intervallo

media ± 1 s	comprende il	68%	dei dati	tra il 16° e l'84° percentile
media ± 1.645 s	comprende il	90%	dei dati	tra il 5° ed il 95° percentile
media ± 1.96 s (~ 2 s)	comprende il	95%	dei dati	tra il 2.5° ed il 97.5° percentile
media ± 3 s	comprende il	99.73%	dei dati	tra lo 0.1° ed il 99.9° percentile
media ± 4 s	comprende il	99.99%	dei dati	...

Ma se la distribuzione non è normale, e non è riconducibile mediante trasformazioni ad una distribuzione normale, è scorretto non solo calcolare media e deviazione standard, ma anche calcolare percentili e probabilità utilizzando la curva normale. Se la distribuzione non è normale i percentili si devono calcolare dai dati, nel modo cosiddetto empirico.

Outliers

Non esiste una definizione univoca degli outliers. In una distribuzione normale, si può dire che un outlier è un dato che dista più di 1.96 deviazioni standard sotto o sopra la media. Che vuol dire un dato sotto o uguale al 2.5° percentile oppure uguale o sopra il 97.5° percentile. Nel caso in cui la distribuzione non sia normale, e utilizzando i percentili, un outlier è solitamente definito come un dato che dista dal 25° percentile - 1.5 volte il valore del range interquartile oppure dista dal 75° percentile + 1.5 volte il valore del range interquartile. Non è possibile stabilire a priori la probabilità di questi outliers.

I test statistici

Tutti i test, non solo i test statistici, ma anche i test diagnostici, ecc. si basano sulla verifica di una certa condizione. La verifica non si può fare in modo diretto, ma solo attraverso la valutazione di fenomeni strettamente associati alla condizione ipotizzata. Manca quindi l'evidenza diretta, cioè la certezza che la condizione sussista. Si potrà avere solo una certa fiducia o probabilità che la condizione sussista.

Ad esempio, con il test di gravidanza
(a) non siamo in grado di osservare direttamente l'embrione (l'ecografia non è sufficiente perché è ai primissimi stadi di sviluppo o perché non abbiamo l'ecografo)
(b) ma sappiamo che l'embrione produce l'ormone specifico HCG
(c) e quindi possiamo prevedere con alta probabilità una gravidanza attraverso la positività per l'HCG.
Il fatto che si tratti di probabilità e non di certezza dipende dal fatto che anche il test per l'HCG - per quanto affidabilissimo - può essere falsato dalle condizioni dei reagenti (es., mal conservati), dalle condizioni di conservazione del campione biologico, dal grado di accuratezza con cui il protocollo del test viene seguito, ecc.

Il test prevede quindi 4 situazioni.
C+ e C- indicano rispettivamente gravidanza sì e gravidanza no
T+ e T- indicano rispettivamente risultato del test positivo e negativo

		risultato del test	
		T+	**T-**
condizione reale (non nota)	**C-**	falso-positivo **T+/C-** **errore del 1° tipo** **α**	vero-negativo **T-/C-** **livello di protezione** **1- α** **specificità**
	C+	vero-positivo **T+/C+** **potenza del test** **1- β** **sensibilità**	falso-negativo **T-/C+** **errore del 2° tipo** **β**

complementari

complementari

L'espressione T+/C- significa test positivo in assenza di gravidanza, e così via. L'errore dei falsi-positivi è detto errore α, o P_α in termini di probabilità. L'errore dei falsi-negativi è detto errore β, o P_β in termini di probabilità.

Sembrerebbe strano, ma veri-negativi e falsi-positivi (prima riga) sono dati complementari. Infatti [esempio di pura fantasia] se il test per la gravidanza fosse sempre negativo sulle donne non gravide (100% di veri-negativi) non sarebbe mai positivo sulle stesse donne (0% di falsi-positivi). Allo stesso modo (seconda riga) veri positivi e falsi negativi sono complementari. Infatti se il test di gravidanza fosse sempre positivo sulle donne gravide (100% di veri positivi) non sarebbe mai negativo sulle stesse donne (0% di falsi negativi).

Dire che un test ha una bassa probabilità di falsi-positivi (α piccolo) e quindi alta probabilità di veri-negativi è come dire che è molto specifico. Dire che un test ha una bassa probabilità di falsi-negativi (β piccolo) e quindi alta probabilità di veri-positivi è come dire che è molto sensibile.

Un test per essere buono dovrebbe possedere sia un'alta specificità sia un'alta sensibilità. Non ha alcun senso un test altamente sensibile ma niente specifico, come dire un test sempre positivo nel caso di gravidanza ma anche sempre positivo nel caso di non gravidanza. Analogamente è assurdo un test altamente specifico ma niente sensibile, come un test sempre negativo in caso di non gravidanza, ma anche negativo in caso di gravidanza.

Vediamo un altro tipo di test. Consideriamo le indagini di polizia a carico di un sospetto. E immaginiamo che le indagini raccolgano diversi indizi ma non delle prove schiaccianti sulla colpevolezza o non-colpevolezza dell'indagato. Interessante notare che in inglese il termine 'trial' significa sia processo che esperimento controllato.

Le differenze rispetto all'esempio precedente sono:

- le conseguenze se si sbaglia: si tratta di lasciare in libertà un delinquente o mandare in prigione un innocente
- la difficile ripetibilità del test: l'indagine di polizia non può essere ripetuta con disinvoltura come un semplice test diagnostico, può durare mesi e costare molti soldi e a distanza di tempo l'efficacia di nuove indagini diminuisce, le testimonianze sono più confuse, ecc.

Comunque, al termine delle indagini il giudice o la giuria emetteranno un verdetto che potrà essere compreso tra due situazioni estreme. Da un lato un giudizio cosiddetto garantista che tende ad emettere condanna solo nel caso in cui esistano prove provate o una convergenza di gravissimi indizi. Questo comporta un maggior numero di colpevoli assolti (più falsi negativi, cioè maggiore errore β). Dall'altro lato il giudizio cosiddetto sommario che invece tende ad emettere condanna anche nei casi in cui gli indizi siano semplici sospetti. Questo comporta un maggior numero di innocenti condannati (più falsi positivi, cioè maggiore errore α).

Ora occorre considerare che gli errori α e β sono solo in parte vincolati tra loro. Nell'esempio delle indagini, i metodi per dimostrare la colpevolezza (o minimizzare l'errore β) sono diversi dai metodi per dimostrare l'innocenza (o minimizzare l'errore α). Infatti una impronta sull'arma del delitto dimostra la colpevolezza ma l'assenza dell'impronta non dimostra l'innocenza (potrebbe essere stata cancellata); l'assenza dal luogo del delitto dimostra l'innocenza ma la presenza nel luogo del delitto non dimostra per se la colpevolezza, ecc. Questo è un punto cruciale. Potremmo paradossalmente avere sia α che β grandi, oppure sia α che β piccoli. Il paradosso di un soggetto che ha un alibi di ferro ma anche gravi indizi è una condizione tipica delle trame dei romanzi gialli. Maestra in questo era Agatha Christie.

Analizziamo ora i risultati di un esperimento.

		risultato dell'esperimento	
		positivo (il trattamento sembra efficace)	**negativo (il trattamento non sembra efficace)**
condizione reale (non nota)	**il trattamento non è efficace**	falso-positivo	vero-negativo
	il trattamento è efficace	vero-positivo	falso-negativo

In campo scientifico, una falsa notizia - indipendentemente dal fatto che sia una notizia inventata di sana pianta o il frutto di esperimenti approssimativi - è particolarmente grave in quanto inganna altri ricercatori che perderanno tempo e risorse a riprodurre quei risultati e soprattutto, se le applicazioni vanno in porto, gli utilizzatori finali illusi di produrre/utilizzare ad es. farmaci che non curano, con rischi per la salute. A prescindere dalle frodi, in ambito scientifico il risultato positivo di un esperimento è considerato significativo solo quando la probabilità α (che il risultato sia positivo per caso, quindi falso-positivo) sia minore del 5%. Se la probabilità di α è uguale o maggiore del 5% si deve rinunciare al riconoscimento del risultato, per quanto positivo possa sembrare. Quindi si privilegia un atteggiamento di sano scetticismo, che assume a priori che il risultato osservato, per quanto apparentemente positivo, dipenda dal caso, cioè da errori del campionamento o delle misurazioni o errori dovuti ad altri fattori non controllabili. Questa ipotesi è detta ipotesi zero (H0) o ipotesi nulla, e deve essere mantenuta sino a quando la sua probabilità scende sotto il 5%. Solo allora possiamo rifiutarla ed accettare l'ipotesi alternativa (H1) che il dato positivo sia 'significativo', non attribuibile al caso.

valori di α ottenuti dal test

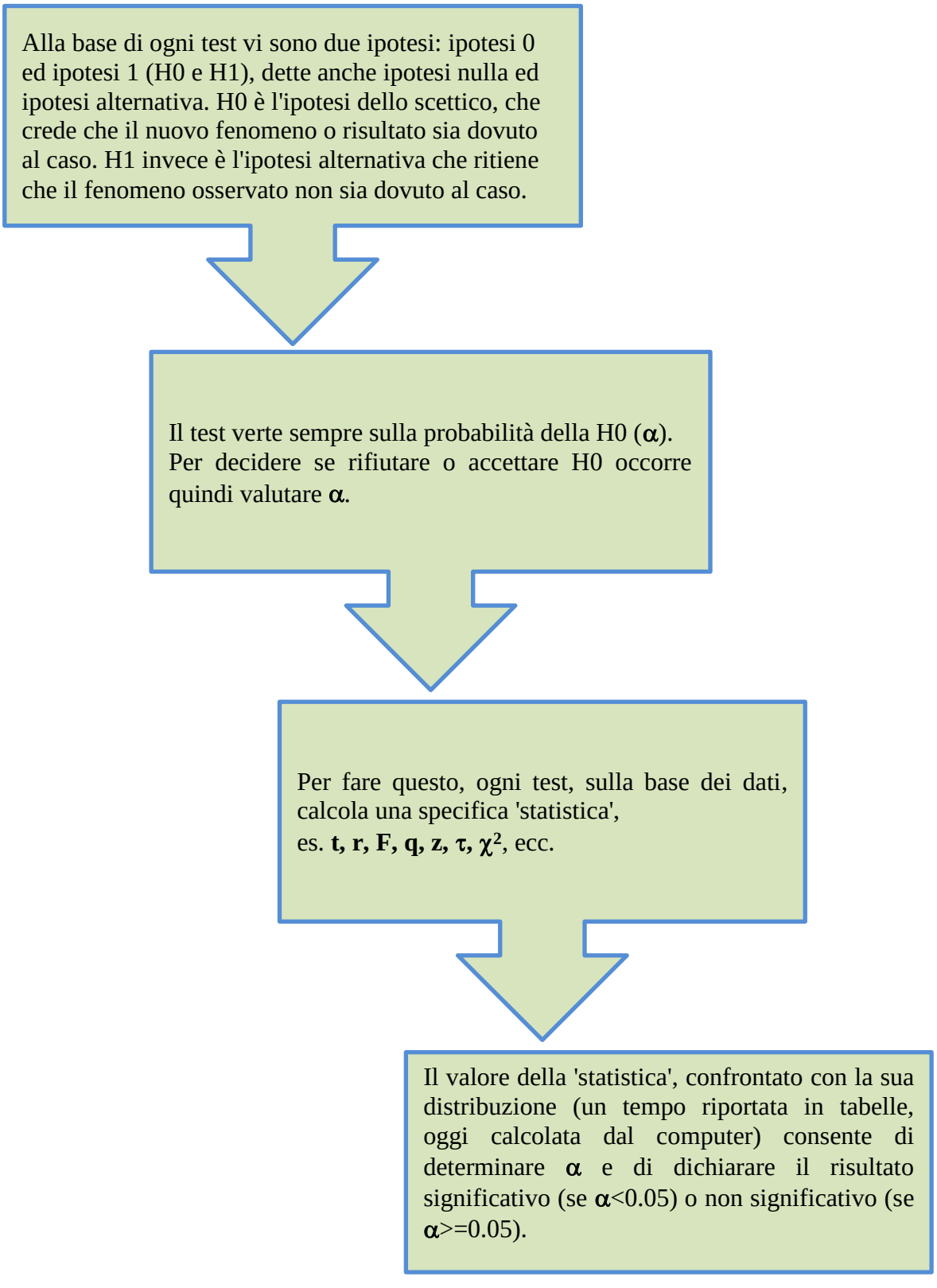
<0.05 (<**5%**)	=>0.05 (=>**5%**)
H0 rifiutata	**H0 conservata**
risultato significativo	**risultato non significativo**

Quindi l'ipotesi nulla va conservata fino a che le prove o i dati in nostro possesso non siano tali da consentirci di rifiutarla, analogamente al principio che ogni soggetto deve essere ritenuto innocente sino a prova contraria. Facendo un altro giro di parole, α può anche essere definita come la probabilità di sbagliarsi dicendo che il risultato è significativo. Quanto più piccola è α, tanto più piccola è la probabilità di sbagliarsi. C'è anche la possibilità di distinguere tra risultato significativo ($\alpha<0.05$) e altamente significativo (se $\alpha<0.005$). Questo succedeva soprattutto in passato, all'incirca prima del 1980, quando esistevano solo le tabelle di significatività nei testi di statistica e non i personal computer. Ora i programmi di statistica calcolano α in modo esatto per cui si può ottenere ad es. $\alpha=0.037$ (significativo) oppure $\alpha=0.00053$ (altamente significativo).

Se α valuta la probabilità che un risultato sia falsamente positivo, β valuta quella che un risultato sia falsamente negativo. Così come è importante che α sia piccolo, è altrettanto importante che β sia piccolo. Infatti se β è alto rischiamo di scartare risultati che appaiono negativi mentre in realtà non lo sono. β dipende da più fattori: in parte dal valore critico di α (soglia generalmente prefissata a 0.05), in parte dal numero di dati (sample size) e in parte dalla entità del risultato stesso (effect size - vedremo avanti). Per stabilire questi parametri al fine di ottenere un valore di β basso si fa quella che in gergo è chiamata **analisi della potenza**. Esistono diversi programmi che compiono questa analisi. Uno ottimo e del tutto gratuito è G*Power: Statistical Power Analyses (www.gpower.hhu.de).

Infine si può aggiungere che in relazione ad effetti negativi importanti, come quelli che riguardano la salute dell'uomo, inclusi gli effetti avversi dei farmaci, occorre prendere in considerazione soglie di significatività ben inferiori al 5%. Questo è il cosiddetto criterio di precauzione. Se dobbiamo attraversare una strada non ci basta certo una probabilità minore del 5% di essere investiti. Una probabilità del 5% non è certamente accettabile. Questo è un problema complesso che richiede una attenta valutazione dei potenziali benefici e danni di ogni nuovo farmaco e nuova terapia attraverso rigorose fasi di sperimentazione.

Riassumendo:



La variabilità delle medie e l'errore standard

Abbiamo visto che:

- la variabilità dei dati di un campione rispetto alla media è rappresentata dalla deviazione standard (s)

- la variazione standardizzata di un singolo dato rispetto alla media è rappresentata dal suo valore standardizzato (z)

Ora saliamo di un piano e facciamo lo stesso ragionamento rispetto alle medie. Facendo diversi campioni inevitabilmente le loro medie sono diverse, pur di poco ma diverse. Le medie dei campioni variano rispetto alla media vera, non nota, della popolazione. E' quindi chiaro che anche le medie, non solo i singoli dati, hanno una certa variabilità. Il parametro che stima la variabilità delle medie è la deviazione standard della media o errore standard, scritto col simbolo s_m (deviazione standard della media) o SE o meglio ancora SEM (standard error of the mean). Bisogna assolutamente specificare deviazione standard della media per non confondersi con la semplice deviazione standard, che si riferisce invece ai dati. Da questo punto di vista il termine di errore standard è meno ambiguo anche se meno appropriato.

Analogamente a quanto visto per i singoli dati, a livello di medie:

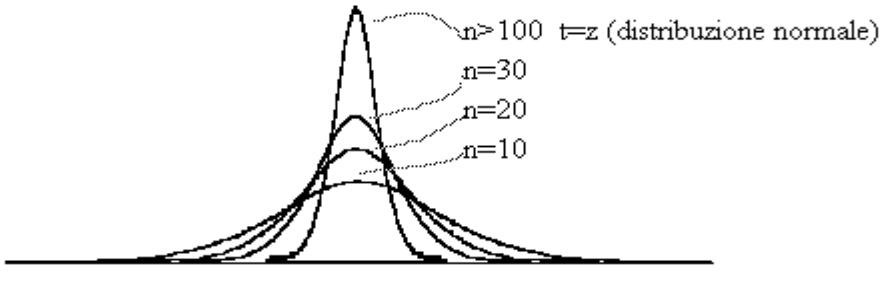
- la variabilità delle medie di diversi campioni rispetto alla media vera non nota della popolazione è rappresentata dalla 'deviazione standard delle medie' o 'errore standard' (SEM)

- la variazione standardizzata di una singola media rispetto alla media vera della popolazione è rappresentata da

$$t = \frac{\text{media del campione} - \text{media vera della popolazione}}{\text{deviazione standard della media}}$$

Purtroppo non conosciamo la media vera della popolazione, ma per calcolare l'errore standard o per confrontare due medie, come vedremo, non sarà necessario.

Semmai un problema deriva dal fatto che, se anche i dati hanno una distribuzione normale, la distribuzione di t non è perfettamente normale per cui non possiamo utilizzare le tabelle di z viste prima. La distribuzione di t per campioni piccoli è più schiacciata, cioè le medie di campioni piccoli sono più disperse. Come la grandezza dei campioni cresce, la distribuzione di t tende a diventare normale. Con oltre 100 dati la distribuzione di t può essere assimilata ad una distribuzione normale.



Se avessimo diversi campioni possiamo calcolare la deviazione standard delle loro medie esattamente come calcoliamo la deviazione standard di un campione di dati. Ma questo non

succede quasi mai. Per fortuna la deviazione standard della media può essere calcolata anche se disponiamo di un solo campione, mediante questa formula:

$$s_m = \frac{s}{\sqrt{n}}$$

In pratica, la deviazione standard della/e media/e si può ottenere dalla deviazione standard dei dati divisa per la radice quadrata della numerosità del campione.

A questo punto dobbiamo considerare bene questi concetti:

- La deviazione standard tout court, intesa come dei singoli dati, (s), è una caratteristica fissa della popolazione, invariante rispetto alla numerosità o grandezza del campione (n). In altre parole, la deviazione standard non cambia se abbiamo campioni piccoli o grandi. Cambia solo in precisione.

  mentre

- La deviazione standard della media (s_m) o errore standard è una caratteristica dello specifico campione, poiché varia in base alla numerosità del campione (n). Le medie di campioni piccoli e grandi estratti dalla stessa popolazione hanno deviazioni standard diverse. Come i dati aumentano di numero, la deviazione standard della media diminuisce, e tende a zero. Possiamo quindi considerare la deviazione standard della media come parametro di inaffidabilità della media.

6 campioni grandi
n= 49

53.6	51.5	56.2	56.0	46.9	49.2
52.4	53.5	49.3	51.8	56.5	50.1
53.2	53.1	48.4	53.6	57.0	42.2
47.1	51.0	49.1	44.7	52.2	54.4
51.4	41.1	48.6	49.5	58.1	50.8
54.5	50.4	51.0	43.9	55.1	48.6
43.5	58.5	47.4	46.7	48.8	46.4
57.1	50.9	47.3	47.2	48.4	56.0
45.7	47.8	48.9	48.2	47.1	44.6
55.2	58.2	50.9	56.0	58.3	49.6
48.5	46.8	48.9	54.1	46.9	45.8
46.5	49.1	56.3	54.5	44.2	45.2
46.5	50.6	45.3	47.5	44.0	46.1
46.5	46.4	43.4	57.5	56.5	54.5
49.1	53.0	45.3	55.5	47.8	46.2
47.9	52.2	46.5	53.3	50.9	50.7
54.3	45.9	59.8	49.2	49.7	46.0
49.0	54.3	56.6	53.9	45.9	49.6
54.6	50.7	51.7	49.7	55.4	55.2
47.6	48.0	49.3	53.7	50.4	52.1
45.7	45.5	49.4	54.3	51.9	51.1
49.5	54.7	54.1	41.5	55.2	50.4
49.2	45.5	47.2	52.6	52.1	46.3
48.5	47.2	47.0	46.7	50.9	64.2
54.2	42.5	55.3	51.2	43.9	49.1
53.2	46.4	47.4	46.0	48.3	49.3
48.5	49.9	45.3	42.8	53.2	47.8
57.8	51.3	42.1	51.2	40.3	51.3
50.7	49.4	49.7	49.2	52.2	57.4
54.4	50.1	51.2	50.5	52.6	50.6
50.0	43.6	44.1	48.7	44.6	50.9
52.9	50.9	52.6	46.3	51.4	50.3
45.8	49.8	50.7	53.2	56.0	53.2
45.3	48.6	45.1	52.3	41.7	41.7
51.4	45.5	54.9	48.7	56.2	51.9
56.4	48.8	48.0	47.3	48.3	45.8
51.2	44.6	52.8	52.8	49.8	47.2
54.6	52.0	46.9	47.1	48.1	50.1
49.4	49.0	46.4	45.5	48.3	44.0
50.5	51.4	52.5	48.1	50.5	49.8
47.0	52.1	45.7	53.4	50.0	48.5
49.3	46.4	48.3	46.3	57.9	47.1
55.5	56.3	44.6	51.6	51.3	52.4
52.9	49.2	51.9	44.6	52.1	52.7
51.1	50.9	56.1	47.5	47.3	49.6
51.6	43.6	53.2	51.9	57.3	51.8
48.8	47.8	50.4	49.1	51.1	49.9
54.0	49.1	49.0	41.1	42.7	52.2
48.4	58.4	46.8	43.6	45.5	46.4

media	50.65	49.67	49.58	49.62	50.43	49.73
SD	3.49	3.93	3.91	4.04	4.61	4.00
SEM	0.50	0.56	0.56	0.58	0.66	0.57
n	49	49	49	49	49	49

media delle medie	49.97
deviazione standard delle medie	0.564

diverse serie di campioni di dati estratti da una distribuzione normale con
media
m = 50
e deviazione standard
s = 4

in Excel:
=NORMINV(RAND();50; 4)

6 campioni piccoli
n = 9

50.9	57.8	45.0	52.8	44.8	44.2
48.2	43.8	50.5	50.0	53.4	53.6
53.7	58.1	53.4	49.5	43.4	44.1
50.0	44.5	55.4	49.6	47.7	49.7
43.3	49.5	50.8	52.4	53.6	46.7
48.4	51.4	49.0	47.7	53.6	52.1
52.0	48.4	55.2	53.6	45.5	48.7
50.0	47.0	49.8	46.4	55.0	49.5
47.4	55.6	52.5	46.9	53.7	52.7

media	49.33	50.70	51.29	49.87	50.06	49.04
SD	3.00	5.44	3.27	2.60	4.64	3.50
SEM	1.00	1.81	1.09	0.87	1.55	1.17
n	9	9	9	9	9	9

media delle medie	50.09
deviazione standard delle medie	1.346

Si può notare come:

- **media e deviazione standard non dipendono dalla grandezza del campione**
- **l'errore standard dipende (anche) dalla grandezza del campione**
- **l'errore standard si può calcolare anche da un singolo campione**
- **i rapporti tra gli errori standard stanno come i rapporti tra le radici di n**

$\sqrt{49}/\sqrt{7}$		2.333
SEM campioni grandi / SEM campioni piccoli	1.346/0.564	2.389

Test t di Student per campioni non appaiati e campioni appaiati

Il test t di Student saggia la differenza tra due medie. Alla base del test:

- l'ipotesi nulla (H0 o HN) sostiene che le due medie provengano da campioni estratti dalla stessa popolazione e quindi la loro differenza sia attribuibile a cause accidentali inerenti al campionamento e/o alle misurazioni.
  H0: differenza tra le due medie = 0

- l'ipotesi alternativa (H1 o HA) sostiene che le due medie siano diverse in quanto provenienti d campioni di popolazioni diverse (naturali o sperimentali).
  H1: differenza tra le due medie $\neq 0$

Il test calcola la probabilità dell'H0, e quindi la probabilità di falsi-positivi, utilizzando la deviazione standard della media e la statistica t. Come abbiamo visto, nei campioni numerosi (n>=100) t è distribuito in modo quasi normale ed ha quindi gli stessi valori critici di z (circa ±1.96 di ascissa includono il 95% di probabilità). Nei campioni più piccoli la distribuzione è più dispersa e quindi per ottenere il 95% di probabilità occorrono valori di t > 1.96 (trascurando il segno). La tabella, a scopo illustrativo, riporta i valori critici di t per i diversi gradi di libertà. I programmi calcolano la probabilità esatta associata al valore di t trovato.

gradi di libertà	α (rischio di falsi positivi)					
	zona della non-significatività		soglia critica		zona della significatività	
	.20	**.10**	**.05**	**.02**	**.01**	**.001**
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.819
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Si dicono campioni indipendenti (o non appaiati) i campioni costituiti da dati di diversi oggetti/soggetti. Sono invece detti appaiati i campioni costituiti da dati degli stessi oggetti/soggetti, valutati o osservati in tempi diversi (prima e dopo un certo trattamento o in condizioni diverse). Il disegno sperimentale che utilizza campioni appaiati è senz'altro più efficace di quello basato su campioni non appaiati. Tuttavia non sempre è possibile applicarlo, sia per problemi pratici di fattibilità (se un esperimento modifica irreversibilmente l'oggetto, l'esperimento non può essere ripetuto sullo stesso oggetto) sia per problemi etici nel caso di sperimentazioni cliniche.

1° caso: medie di due campioni indipendenti
Il test t per campioni indipendenti o non appaiati è dato dal rapporto:

$$t = \frac{\text{differenza tra due medie}}{\text{errore standard delle differenze tra le medie}} = \frac{m_a - m_b}{s_{m_a - m_b}} = \frac{m_a - m_b}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \cdot \frac{n_a + n_b}{n_a \cdot n_b}}}$$

L'espressione al numeratore non è tanto la differenza tra le due medie quanto la media delle differenze tra i dati dei due gruppi, presi a 2 a 2 (anche se le due espressioni danno risultati equivalenti). Allo stesso modo, l'espressione al denominatore non è tanto un pool dei due errori standard, ma piuttosto l'errore standard di questa media delle differenze tra i dati dei due gruppi.

La formula sfrutta la proprietà che la varianza delle (di tutte le possibili) differenze tra i dati di due popolazioni corrisponde alla somma delle due rispettive varianze:

$$s^2_{a-b} = s^2_a + s^2_b$$

Poiché nel nostro casi ci si riferisce a distribuzioni di medie:

$$s^2_{m_a - m_b} = s^2_{m_a} + s^2_{m_b}$$

da cui:

$$s_{m_a - m_b} = \sqrt{s^2_{m_a} + s^2_{m_b}}$$

Dentro la radice possiamo sostituire i due termini ponendo
varianza della media = quadrato della deviazione standard della media = quadrato della (deviazione standard del campione diviso radice di n), cioè:

$$s^2_m = (s_m)^2 = \left(\frac{s}{\sqrt{n}}\right)^2 = \frac{s^2}{n}$$

Il denominatore della formula pertanto diventa:

$$s_{m_a - m_b} = \sqrt{s^2_{m_a} + s^2_{m_b}} = \sqrt{\frac{s^2_a}{n_a} + \frac{s^2_b}{n_b}}$$

Poi si può anche osservare che secondo l'ipotesi nulla H0 i due campioni provengono dalla stessa popolazione, per cui s^2_a e s^2_b sarebbero sono stime della stessa varianza dei dati della medesima popolazione. Pertanto, è possibile sostituire ciascuna delle due deviazioni standard con una unica stima combinata:

$$s^2_{comb} = \frac{\text{somma devianze}}{\text{somma gradi di libertà}} = \frac{S_a + S_b}{n_a + n_b - 2}$$

Quindi, sostituendo e semplificando, il denominatore della formula del t diventa finalmente:

$$s_{m_a - m_b} = \sqrt{s^2_{m_a} + s^2_{m_b}} = \sqrt{\frac{s^2_a}{n_a} + \frac{s^2_b}{n_b}} = \sqrt{\frac{s^2_{comb}}{n_a} + \frac{s^2_{comb}}{n_b}} = \sqrt{\frac{s^2_{comb} \cdot (n_a + n_b)}{n_a \cdot n_b}} = \sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \cdot \frac{n_a + n_b}{n_a \cdot n_b}}$$

La formula si semplifica molto se i campioni sono **bilanciati** (quando $n_a = n_b$). Quella esposta è preferibile perché generalizzata.

I gradi di libertà del t sono $n_a + n_b - 2$.

Le ipotesi del test t per campioni indipendenti sono:
H0: differenza tra le due medie = 0
H1: differenza tra le due medie ≠ 0

test t per campioni non-appaiati (anche non bilanciati)	
H0: differenza tra medie=0	
frequenza del battito cardiaco in gruppi di animali diversi (dati di pura fantasia)	
topi bianchi	topi neri
56	79
75	73
65	85
60	82
76	73
78	-
n=6	n=5
m=68.33	m=78.40
S=429.33	S=115.20
t=2.138	
gdl=9	
p=0.061 (non significativo)	

$$t = \frac{m_{\text{topi neri}} - m_{\text{topi bianchi}}}{\sqrt{\frac{S_{\text{topi neri}} + S_{\text{topi bianchi}}}{n_{\text{topi neri}} + n_{\text{topi bianchi}} - 2} \cdot \frac{n_{\text{topi neri}} + n_{\text{topi bianchi}}}{n_{\text{topi neri}} \cdot n_{\text{topi bianchi}}}}} = \frac{68.33 - 78.4}{\sqrt{\frac{429.33 + 115.2}{6+5-2} \cdot \frac{6+5}{6 \cdot 5}}} = -\frac{10.07}{\sqrt{60.50 \cdot 0.3666}} = -2.138$$

Il computer indica immediatamente p=0.061. Prima dei computer si confrontava il t calcolato (valore assoluto) con quello tabulato per il livello minimo di significatività del 95% (α=0.05) con 9 gradi di libertà, che è pari a 2.262. Poiché il t calcolato non supera il t tabulato si conclude che non possiamo ritenere che i valori di frequenza di battito cardiaco nei topi bianchi e neri costituiscano diverse popolazioni e che quindi che la differenza riscontrata sia attribuibile alle normali fluttuazioni dei campioni. Questo vale anche se il valore di p (0.061) è molto vicino alla soglia di significatività (0.05).

2° caso: medie di due campioni appaiati (o dipendenti)
Il test t per campioni appaiati è dato dal rapporto:

$$t = \frac{\text{differenza media}}{\text{errore standard della differenza media}}$$

L'errore standard delle differenze si calcola come abitualmente..
I gradi di libertà sono n-1, ove n è il numero di coppie di dati.

Le ipotesi del test t per campioni appaiati sono:
H0: differenza media = 0

H1: differenza media ≠ 0

test t per campioni appaiati (necessariamente bilanciati)		
H0: differenza media=0		
frequenza del battito cardiaco negli stessi atleti (dati di pura fantasia)		
prima di una corsa	dopo una corsa	differenze
66	69	-3
70	77	-7
65	95	-30
76	89	-13
70	78	-8
65	74	-9
n=6 coppie di dati		
		m=-11.67
		s_m=3.896
		t=-2.995
		gdl=5
		p=0.03 (significativo)

$$t = \frac{\text{media delle differenze}}{\text{errore standard delle differenze}} = \frac{-11.67}{3.896} = -2.995$$

Il computer indica un p=0.03. Se usassimo le tabelle, il t tabulato per il livello minimo di significatività del 95% (α=0.05) con 5 gradi di libertà è 2.571. Il t calcolato supera il t tabulato. In entrambi i modi si conclude che la corsa ha modificato la distribuzione dei valori di frequenza di battito cardiaco, determinandone un incremento.

Ultime considerazioni

Nel test t l'ordine delle medie (a-b, b-a) è irrilevante in quanto la distribuzione t è simmetrica. Pertanto si considera il valore assoluto di t prescindendo dal segno.

Poiché il test t valuta differenze tra medie, è facile intuire che è applicabile solo a dati distribuiti normalmente e con varianze uguali. Tuttavia si dice anche che il test t è robusto, è cioè in grado di reggere anche in caso di piccoli scostamenti rispetto a queste condizioni. Qualora le varianze siano diverse si può utilizzare il test di Welch (vedi di seguito).

Lo schema esemplifica diversi casi:

medie diverse, varianze uguali
test t applicabile

medie uguali, varianze diverse
test t inutile (sarà t=0)

medie diverse, varianze diverse
test t normale non applicabile
è applicabile il test t di Welch

L'omogeneità delle varianze si può verificare mediante specifici test.

Infine, se le distribuzioni non sono normali si deve applicare un test non-parametrico, ad es. il test di Wilcoxon (vedi avanti).

Se nell'ambito dello stesso studio si effettuano diversi test t tra diverse medie, il rischio complessivo di falsi positivi aumenta. Per cui il test t non è adatto quando si pianifica un esperimento con molti gruppi o trattamenti da confrontare tra loro. Per questo tipo di analisi esistono specifici test che mantengono un valore complessivo di $\alpha<5\%$ per tutti i confronti pianificati. Oppure (o anche) utilizzare una procedura che controlli il false discovery rate (FDR) sotto il 5%. Questi argomenti saranno affrontati in seguito.

Test t di Welch o test t per campioni con varianza diversa

Il test di Welch è un test t adattato per campioni con varianze diseguali (condizione di eteroscedasticità). Il test di Welch può essere anche applicato ai ranghi.

Test t per campioni appaiati e non-appaiati, e con varianze uguali o diseguali (Welch) sono disponibili in Excel. La funzione ha lo stesso nome del test t
P_α =T.TEST(...)
ma cambiano i parametri dentro le parantesi. La formula calcola direttamente il valore di p.

Limiti fiduciali della media

Dopo aver calcolato una media è utile valutare quanto questa media può differire dalla media (non nota) della popolazione, per un certo livello di probabilità p scelto da noi, solitamente del 95% (p=0.05). Per far questo cerchiamo il valore di t per i gradi di libertà (n-1) e per il livello di probabilità scelto.

In Excel, t=T.INV.2T(...) [2T sta per distribuzione a due code]

A questo punto possiamo dire che, al livello di probabilità scelto, la media della popolazione è compresa nei limiti fiduciali (LF)

$$LF = m \pm t \cdot s_m$$

Questo è anche detto intervallo fiduciale o intervallo di confidenza.

Esempio. Abbiamo una media m=40, con s_m=5 e n=21. Scegliamo il livello di probabilità p=0.05. Con 21-1 gradi di libertà e p=0.05 si trova in tabella o con il PC [T.INV.2T(0.05;20)] un valore di t=2.09. Quindi avremo

$$LF = 40 \pm 2.09 \cdot 5 = \begin{cases} 50.45 \\ 29.55 \end{cases}$$

Con una probabilità di sbagliarci meno di 5 volte su 100 (p=0.05) diremo che la media vera è compresa tra 29.55 e 50.45.

Grandezza del campione o sample size

La numerosità del campione è spesso detta grandezza del campione (sample size). Dalla formula dei limiti fiduciali possiamo anche stabilire quanti dati occorre prendere per ottenere una buona media, cioè una media i cui limiti fiduciali non superino del 5% il valore della media stessa. Il problema spesso deriva dal fatto di avere troppi o troppo pochi dati da processare. Nel primo caso perderemmo una sacco di tempo a valutarli tutti. Nel secondo potrebbe costare molto aumentare il numero di esperimenti, oltre quelli minimi in triplicato. Una risposta al problema può essere fornita sempre dal t. Infatti possiamo stabilire un'eguaglianza tra l'espressione generale:

$$LF = m \pm t \cdot s_m$$

e la nostra opzione (che i LF siano non oltre il 5% della media):

$$LF = m \pm 0.05 \cdot \text{media}$$

Dalle due espressioni si evidenzia che

$$t \cdot s_m = 0.05 \cdot m$$

Ma poiché sappiamo che:

$$s_m = \frac{s}{\sqrt{n}}$$

possiamo scrivere:

$$t \cdot \frac{s}{\sqrt{n}} = 0.05 \cdot m$$

Da cui otteniamo finalmente:

$$n = \frac{t^2 \cdot s^2}{0.05^2 \cdot m^2}$$

Ovviamente il procedimento implica l'esecuzione di uno studio pilota per stimare in prima approssimazione la media e la deviazione standard. Come già detto, media e deviazione standard non sono viziate dalla dimensione del campione, ma sono meno precise se il campione è piccolo (vedi Capitolo 1, vizio ed imprecisione). Quindi l'esperimento pilota può essere condotto su un campione anche piccolo ma in grado di fornire media e deviazione standard sufficientemente precise.

Cohen's d per valutare l'ampiezza della differenza tra due medie (effect size)

La significatività non è tutto. Quando osserviamo due campioni, oltre alla significatività è importante anche valutare l'ampiezza della differenza tra le due medie. Questo è il cosiddetto effect size. Un esempio classico è quello di un ipotetico farmaco antipiretico che faccia scendere la febbre in modo significativo rispetto al controllo senza farmaco, ma che il calo sia di appena 0.1 gradi. L'effetto, pur essendo statisticamente significativo, è estremamente piccolo per essere di interesse farmacologico. Il Cohen's d è uno dei metodi più comuni per valutare l'effect size. La sua formula è semplicissima, e ricorda in parte la formula del test t:

$$d = \frac{m_a - m_b}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2}}}$$

Il denominatore è una deviazione standard 'pooled' delle due medie. Si tratta quindi di una differenza standardizzata tra due medie. Notare che esiste una formula più semplificata, meno accurata, che dà risultati differenti, utile quando abbiamo solo carta e penna. Ma purtroppo anche molti programmi ed applicazioni online utilizzano la formula semplificata.

Per apprezzare d ci si può riferire a questa scala:

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

Test di equivalenza

Si tratta di un test poco conosciuto ma estremamente importante. Come abbiamo visto, quando si è interessati a verificare la differenza tra due medie possiamo eseguire un test t e verificare se le due medie differiscono significativamente. Ma il test t non è utilizzabile quando invece si è interessati a verificare se le medie sono molto simili, cioè equivalenti. Potremmo voler verificare se l'effetto di un nuovo farmaco non differisce dal farmaco accreditato. O verificare se i dati di uno strumento diagnostico portatile non differisca da quelli dello strumento di riferimento 'gold standard'. Ecc.

In questi casi il test t non va bene perché se il t non è significativo si può dire che manca l'evidenza di una differenza, ma questo non vuol dire che c'è l'evidenza della mancanza di una differenza. Paradossalmente, se così fosse, basterebbe prendere giusto due o tre dati da ciascun gruppo e fare un test t. Con pochissimi dati il test t è quasi sempre non significativo e... voilà, le medie dei due gruppi sono uguali! Quindi se il ricercatore intende verificare se le medie sono equivalenti occorre un approccio diverso. Si parla di test di equivalenza.

La più semplice forma di test di equivalenza è il cosiddetto **tost** test (**t**wo **o**ne-**s**ided **t**est). Innanzitutto noi definiamo quanto piccola debba essere la differenza per essere trascurabile. Chiamiamo questo valore Δ. Δ è scelto da noi, sulla base della nostra esperienza oppure adottando valore per il quale dalla formula di Cohen risulti un effet size 'very small', es. Cohen's d = 0.1. Δ può essere simmetrico attorno allo zero ($\pm\Delta$) oppure non simmetrico. In tal caso si usano Δ_1 e Δ_2. Ma per ora consideriamo il caso di un Δ simmetrico. Nel test di equivalenza le ipotesi sono invertite. L'ipotesi nulla assume che la differenza tra le due medie, in valore assoluto, sia maggiore di Δ.

H0: le medie non sono equivalenti e quindi:
$|m1-m2| \Rightarrow \Delta$

H1: le medie sono equivalenti e quindi:
$|m1-m2| < \Delta$

Esempi grossolani (senza test):

m1 = 5, m2 =8, Δ = 2
le medie non sono equivalenti in quanto
$|5-8| > 2$

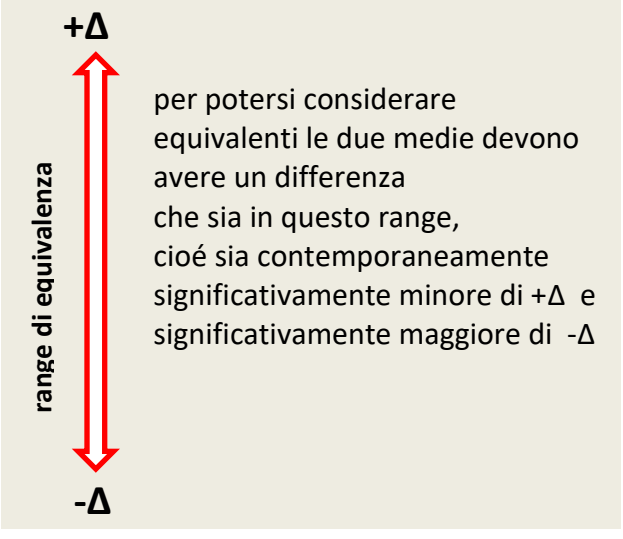
m1 = 7, m2 =11, Δ = 5
le medie sono equivalenti in quanto
$|7-11| < 5$

Naturalmente non basta considerare le medie. Dobbiamo includere la loro variabilità e quindi costruire un test così come abbiamo fatto per il test t.

Per il **tost test** si fa un normale t-test ad una coda (o ad una via, o unilaterale, one-sided) per ciascuna delle 2 ipotesi nulle
$m_a - m_b > -\Delta$
$m_a - m_b < +\Delta$
ponendo al numeratore la differenza rispetto al valore Δ positivo e negativo, oppure Δ_1 e Δ_2.

$$t_1 = \frac{m_a - m_b - (-\Delta)}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \cdot \frac{n_a + n_b}{n_a \cdot n_b}}}; \quad t_2 = \frac{m_a - m_b - (+\Delta)}{\sqrt{\frac{S_a + S_b}{n_a + n_b - 2} \cdot \frac{n_a + n_b}{n_a \cdot n_b}}}$$

Se **entrambi** i test sono significativi si rifiuta l'ipotesi nulla e si accetta quella alternativa della equivalenza. In tal caso, le due medie possono considerarsi equivalenti anche se paradossalmente potrebbero essere significativamente diverse tra loro.



Da notare che:

- se le varianze sono diseguali occorre utilizzare un test con una correzione analoga a quella del test t di Welch
- esiste un test di equivalenza anche per campioni appaiati
- esistono test di equivalenza anche per altri parametri (coefficiente di correlazione, fattore angolare della regressione, ecc.)
- se una media fosse necessariamente solo superiore (o inferiore) all'altra si applica solo uno dei due test

Purtroppo sono pochissimi i programmi che includono questo test. C'è un'ottima applicazione per il calcolo del TOST test su Excel: https://osf.io/qzjaj/download

L'analisi della varianza

Quando l'insegnante di educazione fisica vuole organizzare un piccolo torneo di calcetto tra alunni in genere tende a formare squadre equilibrate, scegliendo a caso o, meglio ancora, mettendo insieme in ogni squadra ragazzi bravi, meno bravi e scarponi. In questo modo crea squadre di composizione interna molto eterogenea ma molto simili tra loro, dando la possibilità di giocare ad armi pari. In termini statistici, il criterio che adotta è quello di avere la massima varianza **entro** le squadre e la minima **tra** squadre diverse. Se avesse lasciato l'iniziativa agli stessi alunni, tutti i più bravi si sarebbero messi assieme nella stessa squadra lasciando nelle altre i meno bravi. Questo avrebbe portato a squadre di composizione interna omogenea ma fortemente sbilanciate e senza speranza di competere. In questo modo si sarebbe verificata una minima varianza **entro** le squadre (tutti bravi in certe squadre, e tutti scarsi nelle altre) ed una massima **tra** squadre diverse. Abbandoniamo il termine di squadra e usiamo quello di gruppi. Dato un certo set di dati, la varianza entro gruppi e quella tra gruppi sono strettamente legate tra loro: se aumenta l'una diminuisce l'altra e viceversa, in base a diverse ripartizioni dei dati nei gruppi. In altre parole, la varianza entro gruppi e quella tra gruppi sono ripartizioni della varianza totale dei dati. Perciò, quando si considerano più gruppi, che siano squadre di calcio o gruppi sperimentali, se vogliamo verificare la presenza di differenze tra i gruppi è possibile scomporre la variabilità totale dei dati (quella che sarebbe mettendo tutti i dati in un solo gruppo) in due componenti: una variabilità dovuta alle differenze **tra** i gruppi ed una dovuta alle differenze **entro** i gruppi. L'ipotesi nulla assume che tutti i gruppi provengano casualmente dalla stessa popolazione e quindi la varianza totale e quella calcolata entro gruppi siano uguali, mentre la varianza tra gruppi sarà minima, tendente a zero. Se invece i gruppi provengono da popolazioni diverse la varianza tra gruppi potrà aumentare e quella entro gruppi diminuire. rispetto alla varianza totale.

Il test è basato sul rapporto

F = varianza tra gruppi / varianza entro gruppi

Quanto più F è grande, tanto più si va verso il rifiuto dell'ipotesi nulla. Per rifiutare l'ipotesi nulla, e quindi ritenere che esistano differenze significative tra le medie, occorre che le probabilità di trovare per caso un F grande come quello calcolato siano meno del 5%. Se si rifiuta l'ipotesi nulla, vale l'ipotesi alternativa che sostiene che i gruppi provengano da differenti popolazioni. Purtroppo, il test F è un test globale e non consente di precisare quale o quali gruppi differiscano tra loro. Come quando in certe trame poliziesche si è certi che tra un certo numero di persone c'è l'assassino ma non si sa chi sia. In tal senso il test F è quasi sempre solo una premessa obbligatoria per fare confronti diretti tra i gruppi, come vedremo tra breve.

Calcolo. Sappiamo che la varianza è data dalla devianza (somma dei quadrati) divisa per i gradi di libertà. Quindi ci sarà anche una devianza tra gruppi ed una entro gruppi. Allo stesso modo, ci saranno i gradi di libertà tra gruppi ed entro gruppi.

Il magico dell'analisi della varianza, è dato da questa decomposizione:

$$S_{totale} = S_{tra\ gruppi} + S_{entro\ gruppi}$$

$$GDL_{totali} = GDL_{tra\ gruppi} + GDL_{entro\ gruppi}$$

Vediamo come fare questa decomposizione.

	TOTALE	TRA GRUPPI	ENTRO GRUPPI
DEVIANZA S	S_{totale} Mettendo insieme tutti i dati e calcolando la devianza	S_{tra} Sostituendo ai dati di ciascun gruppo la media del gruppo e calcolando poi la devianza rispetto alla media totale. In tal modo si annulla la variazione entro i gruppi.	S_{entro} Calcolando le devianze entro ciascun gruppo e sommando insieme. In tal modo si annulla la variazione tra gruppi.
GDL	numero dei dati - 1	numero dei gruppi - 1	n - numero dei gruppi
	⇩	⇩	⇩
VARIANZA	s^2_{totale}	s^2_{tra}	s^2_{entro}
	$S_{totale} / GLD_{totali}$	S_{tra} / GDL_{tra}	S_{entro} / GDL_{entro}

Esempio di calcolo:

	TOTALE	**TRA GRUPPI**	**ENTRO GRUPPI**
	$(x-m_{TOTALE})^2$	$(m_{GRUPPO}-m_{TOTALE})^2$	$(x-m_{GRUPPO})^2$
Gruppo A			
1	$(1-6)^2=25$	$(2-6)^2=16$	$(1-2)^2=1$
2	$(2-6)^2=16$	$(2-6)^2=16$	$(2-2)^2=0$
3	$(3-6)^2=9$	$(2-6)^2=16$	$(3-2)^2=1$
	S=50	S=48	S=2
$m_A=2$			
Gruppo B			
4	$(4-6)^2=4$	$(6-6)^2=0$	$(4-6)^2=4$
6	$(6-6)^2=0$	$(6-6)^2=0$	$(6-6)^2=0$
8	$(8-6)^2=4$	$(6-6)^2=0$	$(8-6)^2=4$
	S=8	S=0	S=8
$m_B=6$			
Gruppo C			
9	$(9-6)^2=9$	$(10-6)^2=16$	$(9-10)^2=1$
10	$(10-6)^2=16$	$(10-6)^2=16$	$(10-10)^2=0$
11	$(11-6)^2=25$	$(10-6)^2=16$	$(11-10)^2=1$
	S=50	S=48	S=2
$m_C=10$			
$\Sigma_{TOTALE}=54$			
$m_{TOTALE}=6$			
	TOTALE	**TRA GRUPPI**	**ENTRO GRUPPI**
DEVIANZA (S)	50+8+50=**108**	48+0+48=**96**	2+8+2=**12**
GDL	numero dei dati - 1 9-1=**8**	numero dei gruppi - 1 3-1=**2**	n - numero dei gruppi 9-3=**6**
VARIANZA (s^2)	108/8=**13.5**	96/2=**48**	12/6=**2**

I dati confermano che la devianza totale è pari alla somma delle devianze tra ed entro gruppi:
108 = 96 + 12

come pure i gradi di libertà totali:
8 = 2 + 6

Notare anche che mentre le devianze sono sommabili, le varianze, in quanto rapporti, non lo sono. Il rapporto tra le varianze tra ed entro gruppi è la statistica F:

$$F_{2,6\ GDL} = \frac{48}{2} = 24$$

Nel nostro caso il valore di F ottenuto corrisponde ad un valore di p=0.0013, significativo. La tabella dei valori critici non è riportata perché piuttosto complessa. Meglio utilizzare sempre i programmi.

In Excel, P_{α} = F.DIST (...)

Si conclude quindi che le tre medie dei tre gruppi in questione potrebbero provenire dalla stessa popolazione con una probabilità circa uguale ad uno su mille. Pertanto si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa.

Un esempio di analisi della varianza applicata ad un disegno sperimentale complesso

Abbiamo visto i principi fondamentali dell'analisi della varianza. In particolare abbiamo esaminato il modello più semplice di analisi della varianza, quello che viene comunemente detto ANOVA (**AN**alysis **O**f **VA**riance) a una via (one way ANOVA). Ma l'analisi della varianza è molto flessibile ed applicabile a situazioni sperimentali anche molto complesse. In tal caso anche lo schema dell'esperimento (*disegno sperimentale*) è molto importante per ricavare il massimo dell'informazione statistica.
Un esempio di analisi della varianza più complesso (per fornire un'idea) è il seguente:
Supponiamo di voler sperimentare l'effetto di due farmaci sulla pressione arteriosa.

- Otto animali ricevono i due farmaci (A e B) in due diverse dosi (1× e 2×). Queste variabili, tipo del farmaco e dose del farmaco, sono dette variabili di controllo o fattori o... (tutti sinonimi, vedi prima colonna della tabella). Poiché diversi gruppi di animali sono trattati in modo diverso, i confronti tra farmaci e dosi sono confronti tra gruppi.
- I valori di pressione arteriosa (variabile dipendente) sono letti dopo 1, 2 e 3 ore dal trattamento. Questo consente di stabilire se l'effetto del trattamento varia in quanto tale e varia anche in funzione del tempo. Poiché le misure vengono ripetute sugli stessi animali, la valutazione dell'effetto del tempo rappresenta un confronto entro gruppi.
- Poiché si ritiene che gli animali possano essere più o meno sensibili ai due farmaci in relazione all'età ed al sesso, è possibile inserire queste variabili (covariate) nell'analisi al fine di escluderne l'interferenza sulla variabile dipendente e ridurre l'errore. Il mancato inserimento delle covariate può seriamente viziare il risultato dell'analisi, talvolta aumentando la probabilità di falsi positivi. D'altra parte è assolutamente necessario che i fattori non influiscano sulle covariate o viceversa. Nel nostro caso è assai improbabile che il trattamento influenzi l'età o il sesso degli animali.
- Infine è utile anche considerare alcuni animali di controllo. Si tratta di un gruppo al di fuori del contesto dello schema fattoriale. E' il cosiddetto gruppo isolato o appeso: *hanging group*, importante per verificare l'effetto dei farmaci rispetto alla condizione basale.

	VARIABILI (sinonimi) •DI CONTROLLO •DI RAGGRUPPAMENTO •INDIPENDENTI •FATTORI		VARIABILE (sinonimi) •DIPENDENTE •RISPOSTA CON MISURE RIPETUTE			COVARIATE	
animali	tipo di farmaco	dose del farmaco	valori di pressione arteriosa dopo 1 ora	dopo 2 ore	dopo 3 ore	età	sesso
1	A	1×	...	...	...	...	...
2	A	1×	...	...	...	...	...
3	A	2×	...	...	...	...	...
4	A	2×	...	...	...	...	...
5	B	1×	...	...	...	...	...
6	B	1×	...	...	...	...	...
7	B	2×	...	...	...	...	...
8	B	2×	...	...	...	...	...
9	controllo	controllo	...	...	...	...	...
10	controllo	controllo	...	...	...	...	...
11	controllo	controllo	...	...	...	...	...
12	controllo	controllo	...	...	...	...	...

L'analisi della varianza applicata a tale esperimento consente la valutazione

degli **effetti principali**
- tipo di farmaco *(i due farmaci hanno effetti diversi sulla pressione arteriosa ?)*
- dose *(le dosi utilizzate danno effetti diversi ?)*
- tempo *(l'effetto varia col tempo ?)*

delle **interazioni** tra variabili
- tipo di farmaco × dose *(un certo tipo di farmaco ha un effetto particolare se somministrato ad una certa dose ?)*
- tipo di farmaco × tempo *(un certo tipo farmaco ha un effetto particolare dopo un certo lasso di tempo dalla somministrazione ?)*
- dose del farmaco × tempo *(una certa dose ha un effetto particolare dopo un certo lasso di tempo dal trattamento ?)*
- farmaco × dose × tempo *(un certo tipo di farmaco ha un effetto particolare dopo un certo lasso di tempo dal trattamento e se somministrato in una certa dose ?)*

delle **interazioni** tra variabili e covariate
- tipo di farmaco × età *(un certo tipo di farmaco ha un effetto particolare su soggetti di una certa età ?)*
- dose del farmaco × sesso *(un certo tipo di farmaco ha un effetto particolare su soggetti di un certo sesso ?)*
- ecc. ecc.

e inoltre di **specifici confronti** (o contrasti) tra diversi gruppi di animali trattati ed il gruppo di animali di controllo. Ma qui il discorso si amplia troppo e lasciamo l'argomento ai più interessati.

In ultimo, quando esistono due o più variabili di controllo si parla di **ANOVA** a due o più vie; quando si utilizzano covariate si può parlare (ma non è obbligatorio) di **ANCOVA**. **MANOVA** (multivariate analysis of variance) è invece un'ANOVA con più variabili dipendenti e **MANCOVA** quella con più variabili dipendenti e covariate.

Il problema dei confronti multipli e la riduzione del false discovery rate

Il risultato dell'analisi della varianza lascia spesso poco soddisfatti. Infatti, anche se il valore di F supporta l'ipotesi alternativa, il test non dice quale o quali gruppi differiscano dagli altri. Invece spesso desideriamo saperne di più, magari facendo tutti i possibili confronti tra i gruppi presi a due a due. Questa intenzione è più che giustificata, ma non è corretto fare tanti test t contemporaneamente, per il fatto che se in una stessa analisi si considerano i risultati di diversi test, anche tutti significativi, il rischio di falsi positivi complessivamente aumenta oltre la soglia del 5%. Pertanto l'applicazione di test t multipli ha questo problema, ma esistono test studiati appositamente per eseguire confronti multipli lasciando un rischio complessivo di falsi positivi sotto il 5%. Sono anche chiamati test 'post-hoc', e ne esistono diversi tipi come il test di **Dunn**, di **Student-Newman-Keuls**, di **Scheffé**, di **Tukey**, di **Games-Howell** (non-parametrico), ecc. A proposito di questi test, non meravigliamoci troppo se vediamo che i loro risultati non sono perfettamente concordanti. Ogni test risulta più o meno conservativo in relazione alla distribuzione ed al numero dei dati. Comunque per la scelta del test è importante riferirsi al test che viene maggiormente utilizzato in ciascun campo di indagine. In biologia sperimentale è molto utilizzato il test di Student-Newman-Keuls, che fra l'altro è considerato piuttosto equilibrato, vale a dire non troppo conservativo.

Test di Student-Newman-Keuls (SNK) per confronti multipli

Per il test di SNK occorre disporre le medie da confrontare in ordine crescente. Quindi, per ogni confronto tra le medie di due gruppi **a** e **b**, occorre valutare:

n_a e n_b : la numerosità dei due gruppi

s^2_{entro} : la varianza entro gruppi, calcolata preliminarmente per tutti i gruppi

p : (da non confondere con il p di probabilità) il numero di medie comprese in graduatoria tra le due a confronto (includendo nel numero anche le due in esame; se queste sono immediatamente adiacenti p=2)

A questo punto è possibile calcolare la statistica q applicando la formula:

$$q = \frac{m_a - m_b}{\sqrt{\frac{s^2_{entro}}{2} \cdot \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}}$$

A prescindere dai test post-hoc che seguono logicamente l'analisi della varianza, in qualsiasi situazione in cui abbiamo risultati di test multipli possiamo ricorrere a metodi diretti per contrastare l'aumento di falsi positivi.

La soluzione drastica di Bonferroni

Un rimedio molto drastico è quello indicato da Bonferroni. La correzione di Bonferroni modifica la soglia α in modo inversamente proporzionale al numero di confronti che si vogliono fare. Dati N gruppi, i possibili confronti tra tutte le medie prese a 2 a 2 sono N(N-1)/2. Supponendo di avere ad es. 6 medie, i confronti possibili saranno 6(6-1)/2 = 15. Per il criterio di Bonferroni bisognerebbe innalzare la soglia di significatività da $\alpha = 0.05$ a $\alpha = 0.05/15 = 0.0033$. Tale criterio è decisamente severo e troppo **conservativo**, nel senso che conserva troppo l'ipotesi nulla, riducendo eccessivamente il rischio (α) di falsi positivi ed elevando troppo quello (β) di falsi negativi. Il motivo sta nel fatto che se anche i confronti aumentano in ragione di N^2, in quanto N(N-1)/2 = (N^2- N)/2, le medie da confrontare sono sempre le stesse N medie, per cui i confronti non sono del tutto indipendenti, cosa che il criterio di Bonferroni trascura. Per questo, nella pratica, quando è possibile si scelgono soluzioni alternative alla correzione di Bonferroni.

La procedura di Benjamini-Hochberg per il controllo del false discovery rate (FDR)

Il false discovery rate (FDR) stima la proporzione di falsi risultati tra i risultati apparentemente significativi di un certo test. Il metodo è applicabile a qualunque test e prevede questi semplici step:

1. Mettere in una colonna in ordine crescente i valori p ottenuti dai test (tutti quanti, anche quelli risultati non significativi)
2. Mettere nella colonna adiacente il rango dei valori p: il più piccolo ha rango 1, il secondo ha rango 2, ecc.
3. Infine, in una terza colonna calcolare per ogni valore p il valore critico di Benjamini-Hochberg, secondo la formula:

   BH = (R/N)×FDR

   in cui:

   R è il rango del valore p

   N è il numero totale di test (= quanti test abbiamo fatto, o quanti p)

   FDR è il false discovery rate scelto da noi, es. 0.05
4. Partendo dal basso, comparare i valori p con il valore critico di BH, e trovare il valore p più grande significativo che sia anche più piccolo del corrispondente valore critico di BH. Quel valore p e tutti i precedenti sono significativi, mentre tutti i successivi non sono significativi.

Esempio.
Da un set di variabili abbiamo calcolato N = 20 coefficienti di correlazione. 9 delle 20 correlazioni risultano significative (colonna p), ma applicando la procedura di Benjamini-Hochberg, con un FDR del 5% (0.05), solo 6 sono effettivamente significative con un false discovery rate o rischio di falsi positivi minore del 5%.

i test sono in tutto					N	20				
FDR scelto					FDR	0.05				
r	t (*)	p		r	t	p	R rango di p	BH valore critico di Benjamini-Hochberg (R/N)FDR	p è significativo ed anche più piccolo di BH?	
0.63	2.69	0.018767	riordino per p crescente	0.99	2.69	0.000000	1	0.00250	SI	
0.1	0.33	0.367168		0.9225	0.33	0.000004	2	0.00500	SI	
0.22	0.75	0.289587		0.78	0.75	0.001406	3	0.00750	SI	
0.78	4.13	0.001406		0.75	4.13	0.002735	4	0.01000	SI	
0.27	0.93	0.247635		0.74	0.93	0.003343	5	0.01250	SI	
0.62	2.62	0.021225		0.66	2.62	0.012606	6	0.01500	SI	
0.26	0.89	0.256252		0.63	0.89	0.018767	7	0.01750	NO	
0.53	2.07	0.053925		0.62	2.07	0.021225	8	0.02000	NO	
0.9225	7.93	0.000004		0.59	7.93	0.029932	9	0.02250	NO	
0.59	2.42	0.029932		0.53	2.42	0.053925	10	0.02500	NO	
0.46	1.72	0.093655		0.46	1.72	0.093655	11	0.02750	NO	
0.75	3.76	0.002735		0.39	3.76	0.144918	12	0.03000	NO	
0.66	2.91	0.012606		0.36	2.91	0.169577	13	0.03250	NO	
0.3	1.04	0.221463		0.32	1.04	0.203962	14	0.03500	NO	
0.32	1.12	0.203962		0.3	1.12	0.221463	15	0.03750	NO	
0.39	1.40	0.144918		0.27	1.40	0.247635	16	0.04000	NO	
0.74	3.65	0.003343		0.26	3.65	0.256252	17	0.04250	NO	
0.36	1.28	0.169577		0.24	1.28	0.273190	18	0.04500	NO	
0.99	23.28	0.000000		0.22	23.28	0.289587	19	0.04750	NO	
0.24	0.82	0.273190		0.1	0.82	0.367168	20	0.05000	NO	

(*) t = r/sqrt[(1—r2)/(n—2)]

9 valori di r significativi (diversi da zero) non corretti per confronti multipli

solo i primi 6 valori di r significativi (diversi da zero) corretti per un FDR < 5%

Notare che, scorrendo dal basso, una volta individuato il valore p più grande significativo che sia anche più piccolo del corrispondente valore critico di BH, tutti i precedenti sono significativi anche se alcuni sono maggiori del valore critico di BH. Questo esempio mostra i valori di p riordinati, i loro ranghi ed i valori di BH calcolati per un FDR del 5%, per 10 test.

p	Rango	BH		
0.001	1	0.005	significativo	
0.002	2	0.010	significativo	
0.007	3	0.015	significativo	
0.017	4	0.020	significativo	
0.019	5	0.025	significativo	
0.032	6	**0.030**	significativo	**p maggiore di BH ma comunque significativo**
0.033	7	0.035	significativo	
0.039	8	0.040	significativo	**p massimo e contemporaneamente minore di BH**
0.058	9	0.045	non significativo	
0.063	10	0.050	non significativo	

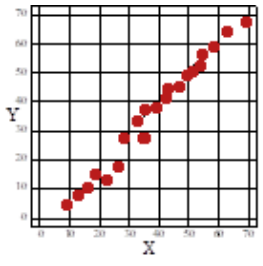
I risultati significativi con un FDR complessivo minore del 5% sono i primi 8, compreso il sesto, anche se il valore di p di questo (0.32) è maggiore del corrispondente valore di BH (0.30).

Regressione

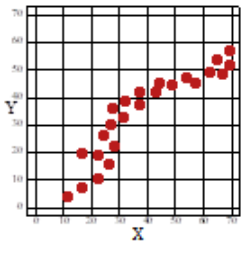
Il termine regressione ha un'origine antica e molto particolare. L'inventore è un certo Galton, genetista, che nel 1889 pubblicò un articolo in cui mostrava come "ogni caratteristica di un individuo è ereditata dalla prole, ma in media ad un livello minore". Il tema era il fatto che i figli di un genitore di statura alta sono anch'essi piuttosto alti, ma in media sono meno alti del genitore. Tale fenomeno, descritto anche graficamente, fu chiamato regressione e da allora tale termine è rimasto per definire lo studio della relazione tra due o più variabili.

Per analizzare la relazione tra due variabili occorre:
(1) assumere un modello di relazione (lineare, nel nostro caso)
(2) valutare i parametri del modello e la loro variabilità
(3) verificare la significatività dei parametri

L'analisi della regressione utilizza molto la rappresentazione grafica. Le due variabili sono rappresentate dagli assi di un sistema cartesiano e le osservazioni sono rappresentate dai punti:



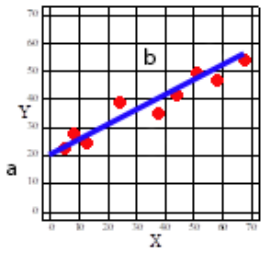
Il modello di relazione che per ora consideriamo è quello lineare. Vale a dire che la variazione tra la variabile X e la variabile Y, entro l'ambito dei valori osservati, è rappresentabile da una retta. Per relazioni non lineari come la seguente occorre trovare soluzioni specifiche che prenderemo in considerazione in seguito.



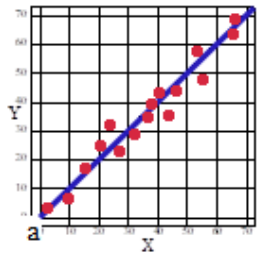
Il primo compito è quello di trovare i parametri che definiscono la retta. Poiché la retta è definita dall'equazione:

$y = \mathbf{a} + \mathbf{b}x$ (talvolta scritta con altri simboli, come per es.: $y = \mathbf{b}_0 + \mathbf{b}_1 x$)

occorre trovare i valori dei parametri **a** e **b** che meglio adattano i valori al modello.
Il parametro **a** è detto **intercetta** (valore dell'asse Y attraversato dalla retta, corrispondente a x = 0) mentre il parametro **b** è detto **pendenza** o fattore angolare o slope.



Quando la retta passa per l'origine l'intercetta è zero. Per cui l'equazione si semplifica:
y = **b**x

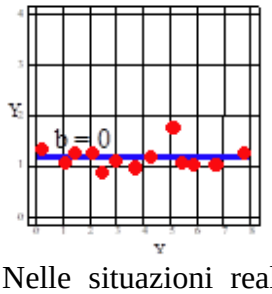


La relazione esprime in tal caso (ma solo in tal caso) un rapporto di proporzionalità, in cui il rapporto tra le due variabili è costante. Si può infatti scrivere:

$$\frac{y}{x} = \mathbf{b}$$

Notabene: c'è proporzionalità se e solo se l'intercetta è zero. Se la retta non passa per l'origine, la relazione può essere sì lineare ma non proporzionale. Ad es., non esiste proporzionalità tra gradi centigradi e gradi Fahrenheit. Per questo motivo in generale non è lecito formulare semplici rapporti, indici, ecc. tra variabili se prima non si dimostra che sono proporzionali.

Se da un lato il fatto che l'intercetta (**a**) sia zero è un fatto positivo, in quanto semplifica il modello, viceversa il fatto che la pendenza (**b**) sia zero è un fatto totalmente negativo. Infatti, se **b** è zero non sussiste alcuna relazione in quanto y è costante rispetto a x, cioè y = **a**, condizione graficamente rappresentata da una retta orizzontale.



Nelle situazioni reali, pendenza e intercetta non saranno mai perfettamente uguali a zero. Occorrerà quindi verificare se i loro valori siano significativamente diversi da zero.

Dopo queste considerazioni preliminari, le cose da fare sono due:

1. calcolare i parametri **a** e **b** (lo fanno tutti)
2. verificare se **a** e **b** sono significativamente diversi da zero (questo invece non lo fa quasi nessuno)

Come calcolare i parametri **a** e **b.**
Noi seguiamo l'approccio più propriamente statistico, che utilizza la **codevianza** ($\mathbf{S_{x,y}}$). La codevianza è una stima combinata della variabilità di due variabili. La formula è analoga a quella della devianza, e consiste nella sommatoria dei prodotti degli scarti tra i valori x e y e le rispettive medie. Diversamente dalla devianza, che è sempre positiva essendo una somma di quadrati, la codevianza è una somma di prodotti, ma non di quadrati, e quindi può essere negativa! Per ragioni di chiarezza, d'ora in avanti in questo capitolo indicheremo le medie con la notazione x_{medio} e y_{medio} anziché m_x e m_y.

$$S_{x,y} = \sum (x - x_{medio})(y - y_{medio})$$

Ad esempio:

X	Y	$(x-x_{medio})^2$	$(y-y_{medio})^2$	$(x-x_{medio})(y-y_{medio})$
1	2	$(1-2)^2=1$	$(2-4)^2=4$	(1-2)(2-4)=2
2	3	$(2-2)^2=0$	$(3-4)^2=1$	(2-2)(3-4)=0
3	7	$(3-2)^2=1$	$(7-4)^2=9$	(3-2)(7-4)=3
$x_{medio}=2$	$y_{medio}=4$	devianza $S_X=2$	devianza $S_Y=14$	codevianza $S_{X,Y}=5$

Calcolata la codevianza, il parametro **b** si ottiene come:

$$\mathbf{b} = \frac{S_{x,y}}{S_x}$$

per cui nell'esempio, **b** = 5/2 = 2.5.

La retta deve necessariamente passare per il punto di intersezione delle due medie x_{medio} e y_{medio}. Per cui in base all'equazione della retta possiamo scrivere:

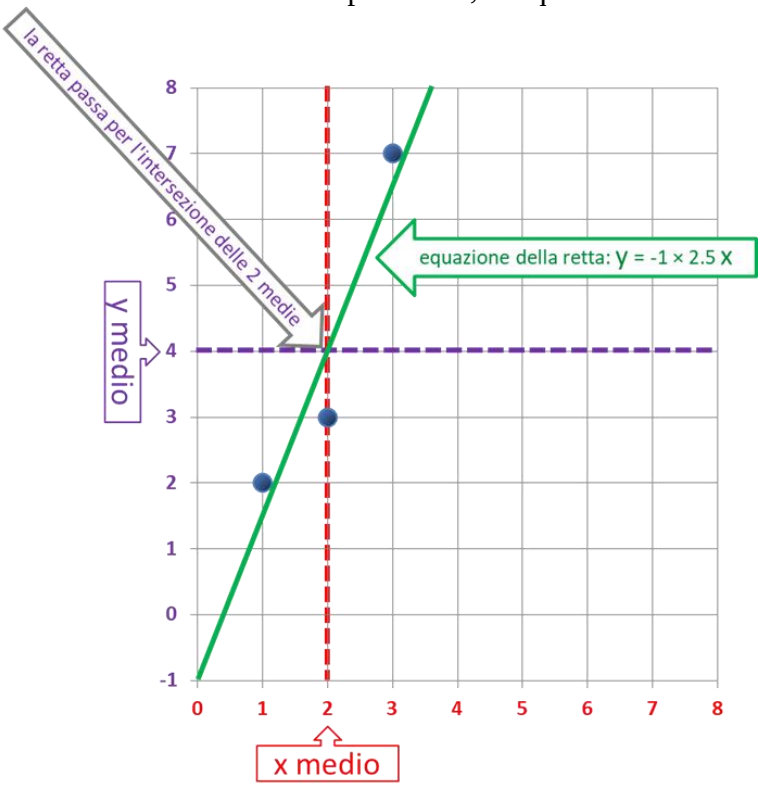
$y_{medio} = \mathbf{a} + \mathbf{b}\ x_{medio}$

da cui, conoscendo **b,** si ricava l'intercetta:

$\mathbf{a} = y_{medio} - \mathbf{b}\ x_{medio}$

Nell'esempio riportato sopra, **a** = 4 - (2.5·2) = -1

Ora che abbiamo entrambi i parametri, l'equazione della retta sarà: y = -1 + 2.5x



Prima di andare avanti, bisogna fare un'altra importante osservazione. Le variabili X e Y, poste rispettivamente in ascissa ed in ordinata, non sono intercambiabili. X è infatti la cosiddetta **variabile indipendente**, mentre Y è la cosiddetta **variabile dipendente**. Non si tratta di una dipendenza causa-effetto (su questo punto anche autorevoli testi sono spesso ambigui se non fuorvianti). Può anche esservi tra X e Y una reale dipendenza causa-effetto (come nel caso tra dose ed effetto di un farmaco), ma non è necessario. Per la statistica il termine indipendente si riferisce solamente al concetto di 'variato a priori, anche arbitrariamente'. Mentre dipendente significa 'libero di variare senza condizionamenti imposti dal campionamento'. Ad es., se decidiamo di valutare la crescita di ragazzi in rapporto all'età possiamo scegliere di prendere ragazzi di età diversa (10, 11, 12, ecc. anni) e poi di misurarne la statura (quella che risulterà). Quindi noi interveniamo sulla variabile età (ad es., prendendo 10 soggetti per ogni classe di età), mentre non interveniamo affatto sulla variabile statura, che sarà libera di variare autonomamente. Questo è il significato di variabile indipendente e variabile dipendente. Paradossalmente, in uno studio sulla maturazione dei denti è stato necessario mettere come variabile indipendente lo stadio di sviluppo del dente e come variabile dipendente l'età dei ragazzi: l'opposto esatto della relazione biologica che vuole che i denti maturino in funzione dell'età. La statistica non entra nel merito di questi fatti. In questo caso la relazione X=stadio di maturazione Y=età serviva per poter stimare l'età fisiologica in base ai dati di ortopantomografie.

A rigore, i valori della variabile dipendente dovrebbero essere distribuiti normalmente, e persino con deviazione standard costante, per ogni valore x in ascissa. Questa condizione si potrebbe controllare solo avendo un grandissimo numero di dati, suddividendoli per gruppi in base a piccoli intervalli di x. E' ovvio che quando si hanno pochi dati questo è impossibile. Comunque, dobbiamo prestare attenzione e verificare se esistono evidenti asimmetrie della distribuzione dei valori y attorno alla loro media (sospetta non-normalità) oppure grafici a forma di punta di freccia (sospetto trend della deviazione standard). In tali casi occorre introdurre dei correttivi. Vedremo qualcosa in seguito.

Come verificare la significatività dei parametri **a** e **b,** cioè verificare se siano significativamente diversi da zero. Per fare questo analizziamo la variabile dipendente Y. A fianco ai valori **osservati** (Y_{oss}) mettiamo i valori **calcolati** (Y_{calc}) in base all'equazione di regressione. Da questo schema possiamo valutare tre tipi di variabilità:

- differenze tra i valori di y osservati (y_{oss}) e il valore y_{medio}

da cui otterremo la **devianza (totale) di y**, indicata con S_y

- differenze tra i valori di y calcolati (y_{calc}) e il valore y_{medio}

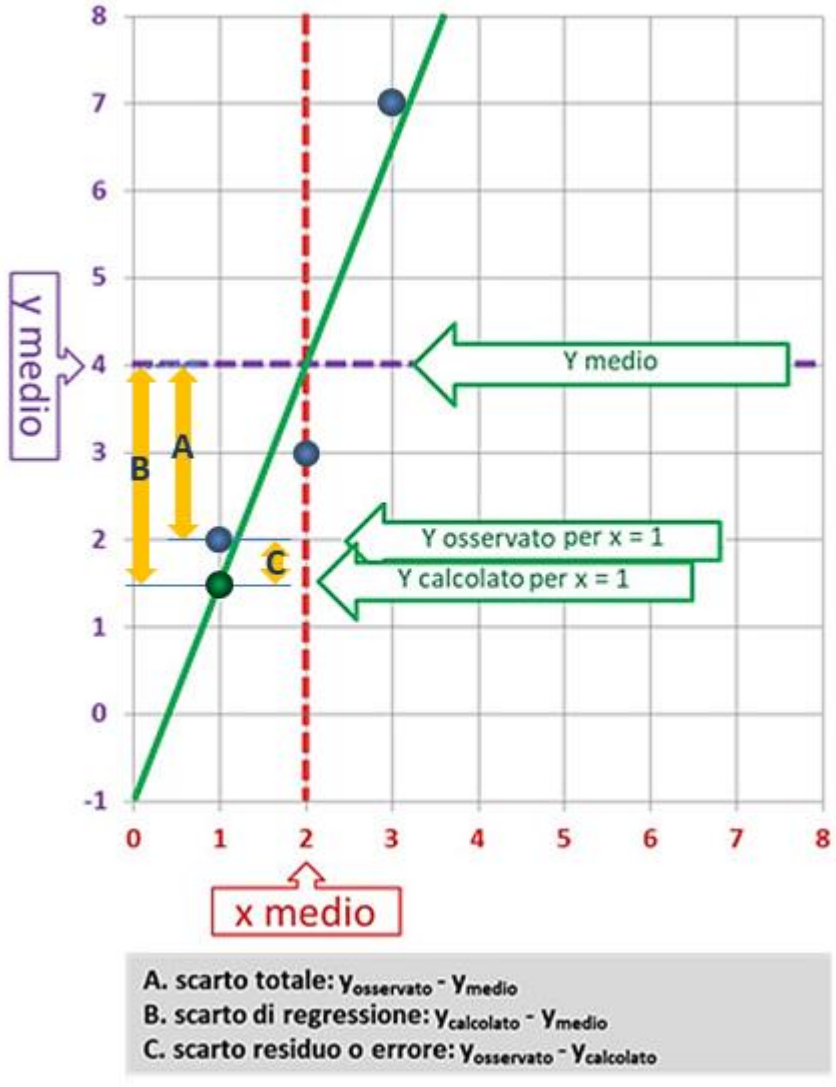
da cui otterremo la **devianza di regressione**, indicata con S_{reg}

- differenze tra i valori di y osservati (y_{oss}) e i valori calcolati (y_{calc})

da cui otterremo la **devianza di errore**, indicata con $S_{y.x}$ oppure S_{res}

					devianza o variabilità totale di y	**parte della devianza totale dovuta alla regressione**	**parte della devianza totale <u>non</u> dovuta alla regressione (devianza residua o di errore)**
	x	yoss	ycalc	yresidui (Y_{oss} - Y_{calc})	Σ (yoss - ymedio)^2	Σ (ycalc - ymedio)^2	Σ (yoss – ycalc)^2
	1	2	1.5	0.5	$(2-4)^2=4$	$(1.5-4)^2=6.25$	$(2-1.5)^2=0.25$
	2	3	4	-1	$(3-4)^2=1$	$(4-4)^2=0$	$(3-4)^2=1$
	3	7	6.5	0.5	$(7-4)^2=9$	$(6.5-4)^2=6.25$	$(7-6.5)^2=0.25$
	$x_{medio}=2$	$y_{medio}=4$			*varianze*		
codevianza	$S_{xy}=5$				$S_y=14$	$S_{reg}=12.5$	$S_{res}=1.5$
intercetta	a=-1				$GDL_y=n-1=2$	$GDL_{reg}=1$	$GDL_{res}=n-2=1$
pendenza	b=2.5				$s^2_y=7$	$s^2_{reg}=12.5$	$s^2_{res}=1.5$

Per semplicità il grafico prende in esame i valori di y_{oss} e y_{calc} della prima riga della tabella di sopra.



Come si noterà nel grafico, per ogni valore osservato di y, lo scarto totale (y_{oss}-y_{medio}: 2-4=-2 lettera A) corrisponde alla somma dello scarto di regressione (y_{calc} - y_{medio}: 1.5-4=-2.5 lettera B) più lo scarto di errore residuo (y_{oss} - y_{calc}: 2-1.5=0.5 lettera C). Questo vale anche per la somma dei quadrati degli scarti: la devianza totale di y corrisponde alla somma della devianza dovuta alla regressione più la devianza residua:

$S_y = S_{reg} + S_{res}$
$14 = 12.5 + 1.5$

Lo stesso vale per i gradi di libertà:

$GDL_y = GDL_{reg} + GDL_{res}$
$n-1 = 1 + n-2$
$2 = 1 + 1$

In questo modo abbiamo decomposto la variabilità totale di y (dispersione dei valori osservati rispetto alla media) in una variabilità dovuta alla regressione (dispersione dei valori calcolati dall'equazione rispetto alla media) ed una variabilità residua o di errore (dispersione dei valori

osservati rispetto ai valori calcolati dall'equazione). Solo quest'ultima vriabilità è interamente imputabile alla variabile Y. L'altra variabilità, quella dovuta alla regressione, è unicamente imputabile alla relazione che vincola Y ad X.

Non confondiamoci con la notazione.
La codevianza è una S maiuscola con il simbolo 'x.y' o 'x,y' (prima x e poi y) a pedice:
$S_{x,y}$ oppure $S_{x.y}$

mentre

la devianza dei residui è una S maiuscola con i simboli 'y.x' o 'y,x' (prima y e poi x) a pedice:
$S_{y,x}$ oppure $S_{y.x}$ oppure anche S_{res}

Le varianze ottenute dalla decomposizione della devianza totale di Y sono importanti per testare l'ipotesi nulla secondo cui il rapporto F tra la varianza di regressione e quella residua sia compatibile con fattori casuali, senza una relazione di dipendenza:

$$F = \frac{s^2_{reg}}{s^2_{res}}$$

Si tratta quindi di una **analisi della varianza applicata alla regressione**. Come in precedenza, valori maggiori di F riducono via via la probabilità a favore dell'ipotesi nulla. Se F risulterà significativo si potrà accettare l'ipotesi alternativa e concludere che vi è una relazione significativa tra Y ed X.

Bisogna ancora valutare se i parametri **a** e **b** sono significativi (diversi da zero) e calcolare i loro limiti fiduciali. Nota bene: poiché si tratta di parametri del campione - come nel caso delle medie - è indifferente parlare di deviazione standard o errore standard. Quindi la notazione s_a si definisce indifferentemente deviazione standard o errore standard dell'intercetta. Idem per la pendenza e per altri parametri che vedremo in seguito.

L'**errore standard della pendenza** è dato dalla formula:

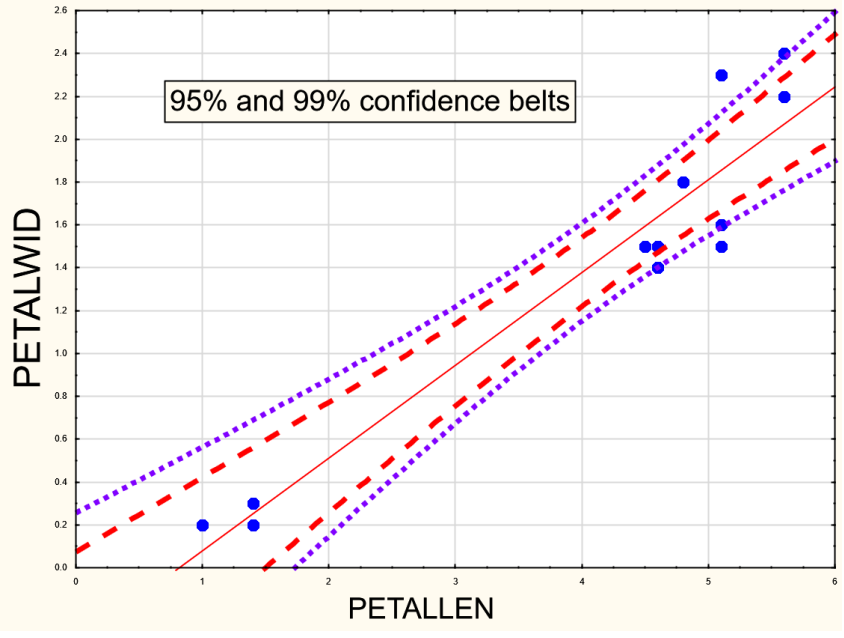
$$s_b = \sqrt{\frac{s^2_{res}}{S_x}}$$

L'**errore standard per i singoli valori di y** della retta è:

$$s_y = \sqrt{s^2_{res}\left(\frac{1}{n} + \frac{(x - x_{medio})^2}{S_x}\right)}$$

Notare che il valore dell'espressione $(x\text{-}x_{medio})^2$ è tanto minore quanto più x è vicino alla media. Per questo anche l'errore standard sarà minimo in corrispondenza del valore x medio, e andrà crescendo a destra e a sinistra, in corrispondenza di valori di x che si allontanano dalla media. Quindi, i limiti fiduciali dei valori y della retta di regressione, sulla base dell'errore standard, sono:
al 95% di probabilità, LF= $y_{calc} \pm 1.96\ s_y$.
al 99% di probabilità, LF= $y_{calc} \pm 2.576\ s_y$.

Con questi limiti fiduciali, calcolati per tutti i valori di X si ottiene una fascia attorno alla retta che è più stretta in corrispondenza del valore medio di X (dove l'errore standard è minimo) e si allarga ai lati. Questa fascia è detta cintura di confidenza (confidence belt).



L'**errore standard dell'intercetta** è semplicemente l'errore standard del valore y corrispondente a x=0. Pertanto sarà:

$$s_a = \sqrt{s_{res}^2 \left(\frac{1}{n} + \frac{x_{medio}^2}{S_x} \right)}$$

Finalmente possiamo saggiare l'ipotesi che l'intercetta sia uguale a zero (H0: a=0) mediante il test t:

$t = \frac{\mathbf{a}}{s_a}$, con n-2 gradi di libertà, gli stessi di s_{res}.

Analogamente, l'ipotesi che la pendenza sia uguale a zero (H0: b=0) si può verificare mediante il test t:

$t = \frac{\mathbf{b}}{s_b}$, sempre con n-2 gradi di libertà.

Abbiamo già visto cosa succede se a=0 e se b=0. Sono i risultati di questi test che decideranno.

Una volta calcolata l'equazione della retta, possiamo anche prevedere i valori di y per nuovi valori di x. Queste previsioni sono tuttavia consentite solo nell'ambito dell'intervallo dei valori di x utilizzati per il modello. Se i valori di x vanno ad es. da 5 a 70, non potrò fare previsioni per valori di x inferiori a 5 né per valori di x superiori a 70.

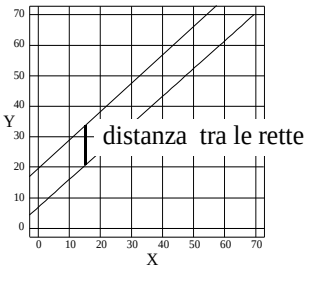
L'**errore standard di nuove previsioni**, cioè di valori y per nuovi valori x, è:

$$s_{y \text{ da nuovo } x} = \sqrt{s_{res}^2 \left(1 + \frac{1}{n} + \frac{(x - x_{medio})^2}{S_x} \right)}$$

Anche per le previsioni l'errore standard è minimo vicino alla media. E anche per le previsioni si possono calcolare le fasce dei limiti fiduciali.

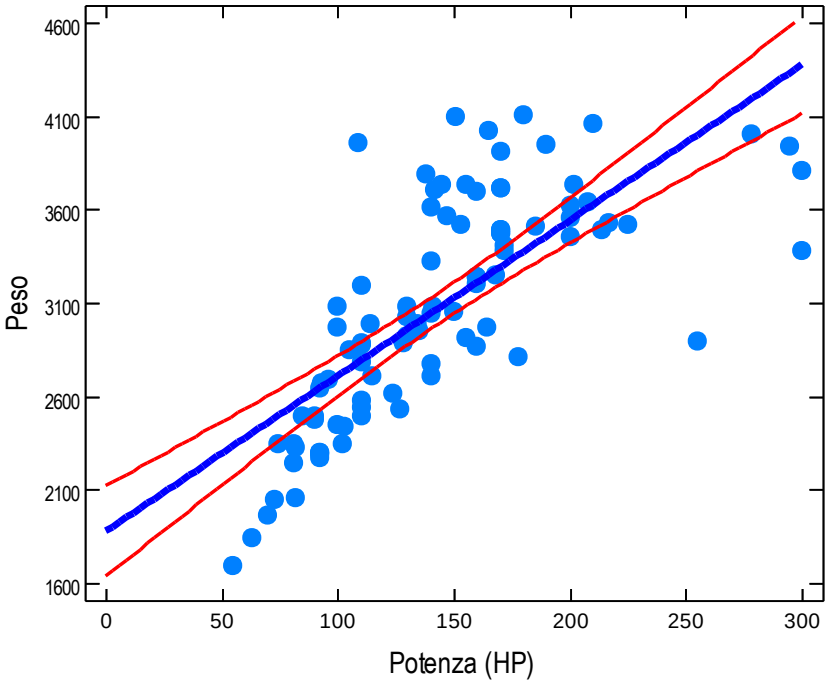
Infine, avendo per le mani 2 rette di regressione ci si può domandare se le 2 rette abbiano pendenza diversa. In pratica si può testare l'ipotesi nulla: $b_1 = b_2$ contro l'ipotesi alternativa: $b_1 \neq b_2$. Il test che saggia la probabilità a favore di tale ipotesi è impropriamente chiamato **test di parallelismo**. Meglio sarebbe chiamarlo test di non-parallelismo in quanto se t è significativo significa che le rette hanno pendenza diversa, ma se non è significativo non è detto che le rette siano parallele (vedi il discorso a proposito del test di equivalenza):

$$t = \frac{\mathbf{b}_1 - \mathbf{b}_2}{\sqrt{\frac{S_{res_1} + S_{res_2}}{n_1 + n_2 - 4} \left(\frac{1}{S_{x_1}} + \frac{1}{S_{x_2}} \right)}}$$



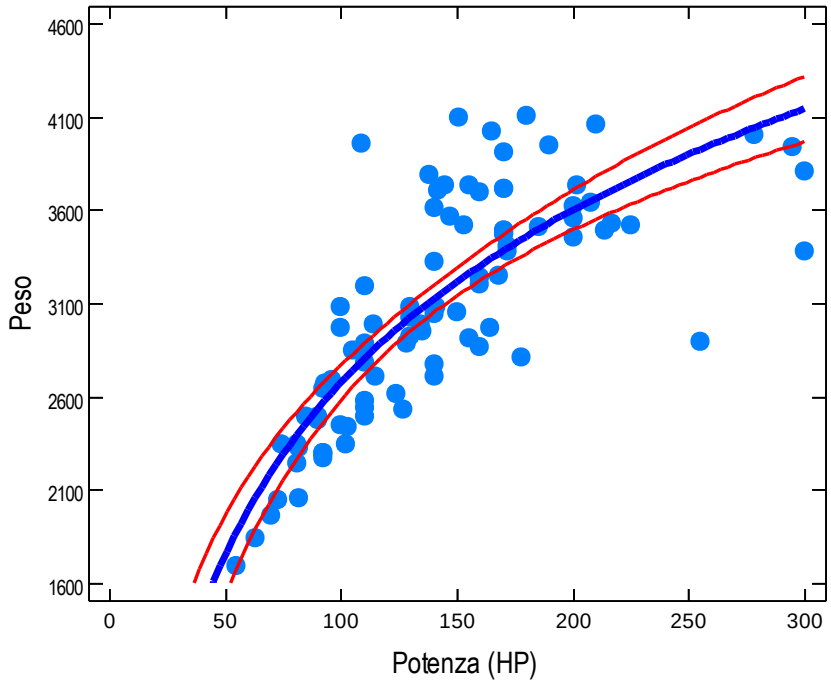
Relazioni non-lineari

Se X e Y sono mostrano una **relazione non-lineare** si può tentare di ricorrere a trasformazioni dei dati tali da restituire un andamento lineare. Classica è la trasformazione log-log che linearizza la relazione tra dimensione dello step unitario (X) e la misura del perimetro (Y) dei contorni frattali e quindi in genere delle forme naturali. Un altro classico caso di rapporto non-lineare è quello tra dose (X) ed effetto (Y) che normalmente si linearizza usando il logaritmo della concentrazione della dose. In teoria, è lecito introdurre qualsiasi trasformazione (logaritmica, esponenziale, ecc.) utile a restituire linearità ai dati. Occorre un po' di esperienza. Esistono comunque tecniche che suggeriscono il tipo trasformazione, e giudicano anche tra diverse trasformazioni quella che meglio linearizza i dati.

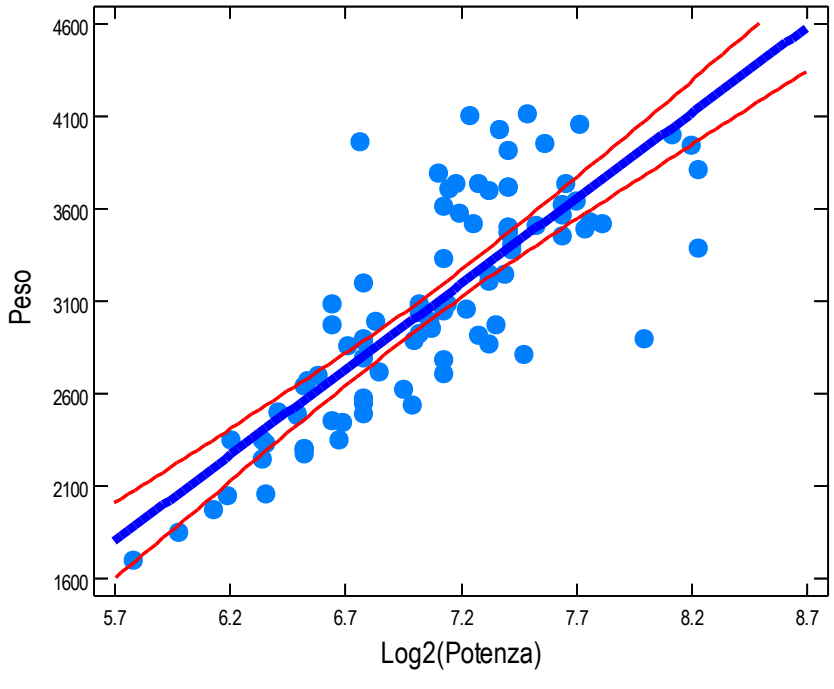


In questo caso consideriamo una trasformazione logaritmica dei valori dell'asse X. L'analisi dei dati trasformati si può presentare graficamente in due modi:

1. lasciando immutata la scala degli assi e rappresentando la retta di regressione come una curva.

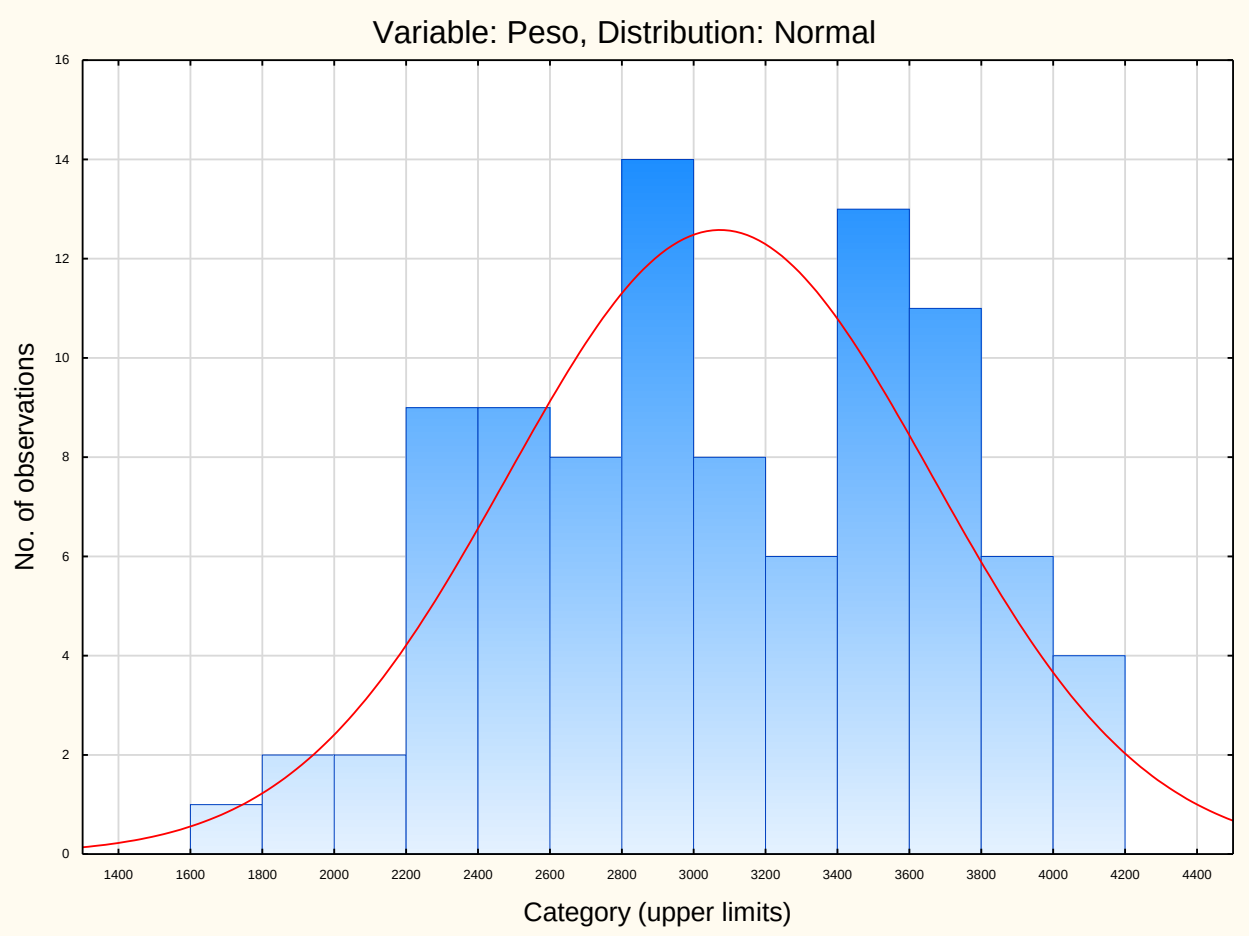


2. oppure riportando la scala degli assi trasformata.



Questo secondo metodo è forse migliore in quanto a differenza del primo muove i punti sul grafico e consente di apprezzare la linearizzazione dei dati: i punti a destra sono più ravvicinati rispetto a quelli a sinistra, così come vuole la trasformazione. L'analisi in ogni caso considera sempre i valori trasformati, a prescindere dalla scala del grafico, ed i risultati sono pertanto identici.

Un ultimo problema. L'analisi della regressione richiede che i dati della variabile dipendente, almeno complessivamente, siano distribuiti normalmente.
Quindi è opportuno fare un test di normalità ad es. mediante il goodness-of-fit test, la skewness standardizzata e la kurtosis standardizzata.



I questo esempio le tre analisi forniscono questi valori:

- goodness-of-fit chi-square: 11.40, df 7 (adjusted), p = 0.12189
- standardized skewness: -0.566, p>0.05 (in quanto minore di 1.96 in valore assoluto)
- standardized kurtosis: -1.683, p>0.05 (in quanto minore di 1.96 in valore assoluto)

Tutti e tre i test indicano che non c'è evidenza che la distribuzione dei dati non sia normale.

Correlazione

Per correlazione si intende una semplice relazione di mutua associazione bidirezionale tra due variabili, che non implica dipendenza né, pertanto, mira a fare previsioni. Al punto che alcuni testi chiamano le due variabili X_1 e X_2 anziché X e Y (la Y sembrerebbe una variabile dipendente). Noi per mantenere un collegamento con alcuni aspetti della regressione manterremo i simboli X e Y. La correlazione è valutata dal cosiddetto coefficiente di correlazione r. Il coefficiente di correlazione esprime quanto i dati sono adattabili a (o rappresentabili da) una retta (o ad una curva, nel caso di modelli non-lineari).

- r è un numero puro, adimensionale, compreso tra -1 e +1.
- r è positivo quando la pendenza è positiva (le due variabili aumentano o diminuiscono insieme) oppure negativo quando la pendenza è negativa (se una aumenta l'altra diminuisce).
- r è esattamente +1 o -1 quando i punti cadono perfettamente sulla retta.
- r è esattamente zero quando i punti formano una nuvola perfettamente omogenea e circolare (se gli assi sono dimensionati alle deviazioni standard delle due variabili) oppure un'ellissi perfettamente orizzontale o verticale.

r si calcola come:

$$r = \frac{S_{x,y}}{\sqrt{S_x S_y}} \qquad (1)$$

Poiché la codevianza (al numeratore) non è mai maggiore della radice del prodotto delle due devianze (al denominatore) r non potrà mai essere maggiore di 1 o minore di -1.

E' evidente la simmetria della formula. Significa che r non varia invertendo gli assi.

r si può calcolare anche dai dati ottenuti dall'analisi della regressione come:

$$r = \sqrt{\frac{S_{reg}}{S_y}} \qquad (2)$$

Questa formula è importante perché se eleviamo al quadrato otteniamo

$$r^2 = \frac{S_{reg}}{S_y}$$

Siccome S_{reg} (devianza dovuta alla regressione) è una parte di S_y (devianza totale), r^2 corrisponde alla quota di devianza di Y dovuta alla regressione. In pratica r^2 esprime quanta variabilità di Y è legata a X. Per questo r^2 è anche detto **coefficiente di determinazione**. Per il buon statistico il valore di r^2 è senz'altro più interessante del valore r. Quindi in un certo modo, anche se abbiamo dichiarato inizialmente che l'analisi della correlazione differisce dall'analisi della regressione per il fatto che la correlazione non comporta con una relazione di dipendenza, il valore di r^2 rientra nell'analisi della regressione, considerando Y come variabile dipendente.

Se escludiamo i valori estremi +1, 0, e -1 (che in realtà non troveremo mai), come facciamo a sapere se un certo valore di r rappresenta una reale correlazione? Come sempre dobbiamo fare un test per verificare se r è significativamente diverso da zero. E come al solito il test consiste nel rapporto tra r ed il suo errore standard.

L'errore standard di r è dato da:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Possiamo quindi fare il test t (H0: r = 0; H1: r ≠ 0):

$$t = \frac{r}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Esempio.
Riprendiamo i dati da cui abbiamo calcolato i parametri della regressione.

X	Y	$(x-x_{medio})(y-y_{medio})$
1	2	(1-2)(2-4)=2
2	3	(2-2)(3-4)=0
3	7	(3-2)(7-4)=3
		codevianza $S_{x,y}=5$
$x_{medio}=2$	$y_{medio}=4$	
$S_x=2$	$S_y=14$	

$S_y = 14$	$GDL_y = 2$
$S_{reg} = 12.5$	$GDL_{reg} = 1$
$S_{res} = 1.5$	$GDL_{res} = 1$

Calcoliamo r.

In base alla prima equazione

$$r = \frac{S_{x,y}}{\sqrt{S_x S_y}} = \frac{5}{\sqrt{2 \times 14}} = 0.945$$

In base alla seconda equazione, stesso risultato

$$r = \sqrt{\frac{S_{reg}}{S_y}} = \sqrt{\frac{12.5}{14}} = 0.945$$

Calcoliamo l'errore standard di r.

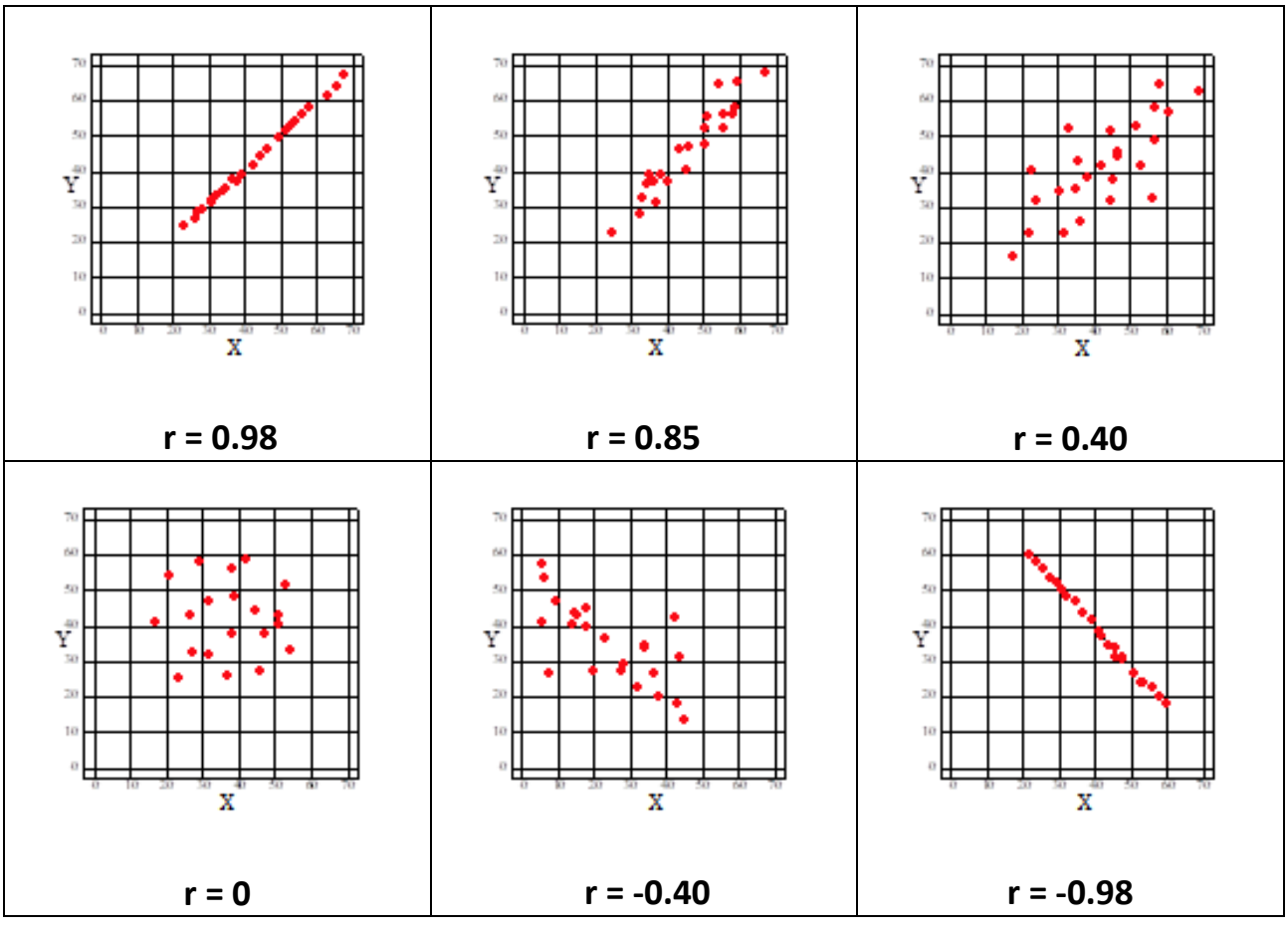
$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.945^2}{n-2}} = \sqrt{\frac{0.107}{1}} = 0.327$$

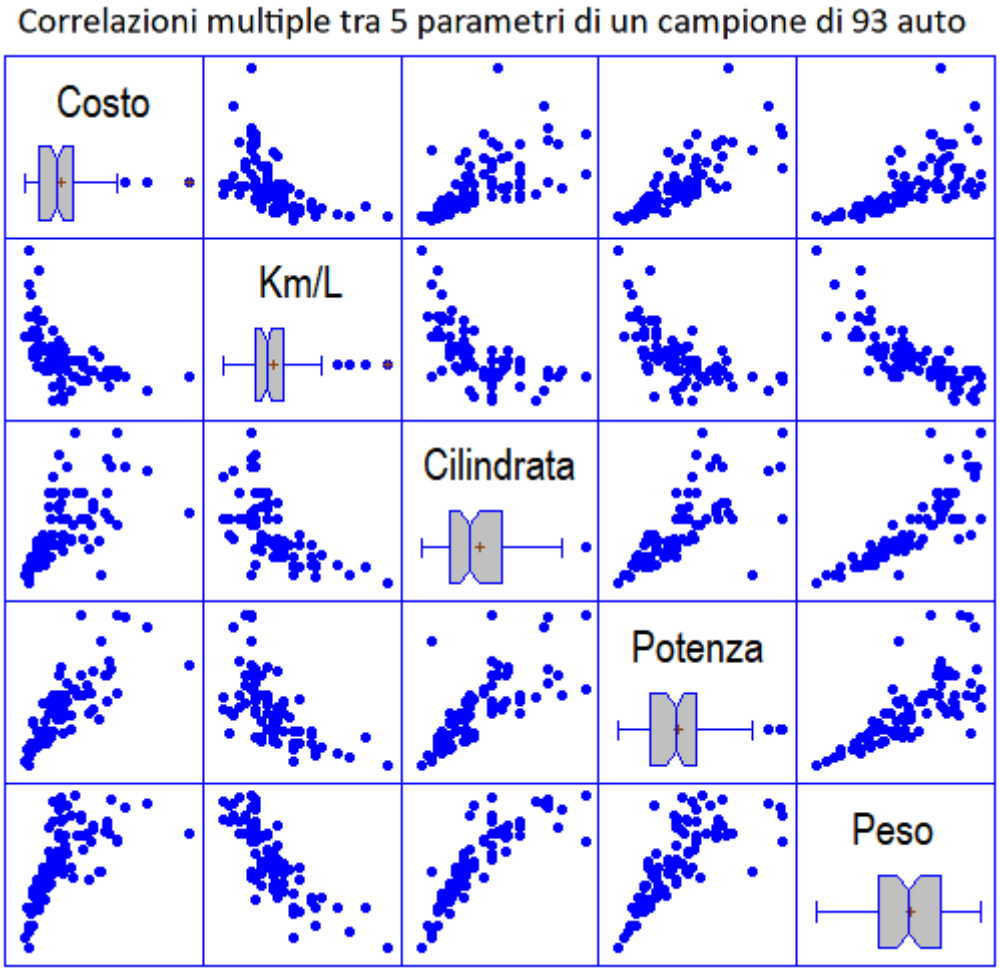
Per saggiare la significatività di r calcoliamo il rapporto t

$$t = \frac{r}{s_r} = \frac{0.945}{0.327} = 2.890$$

Il valore di r sembra grande, e anche il valore del t, ma purtroppo è solo una impressione. Con tre coppie di dati (un solo grado di libertà!) la significatività si raggiunge solo con un t pari o maggiore di 12.706 (vedi la tabella della distribuzione del t). Concludiamo pertanto che nonostante l'alto valore del coefficiente di correlazione (0.945), questo non è significativamente diverso da zero. In altre parole, non siamo autorizzati a ritenere che esista una relazione tra le due variabili, ovvero con 3 punti non si può pretendere di dimostrare niente. Ma può anche succedere il contrario: valori di r anche piccoli possono essere significativi.

Alcuni grafici con diversi valori di r:





Note:

- in ogni box-and-wisker plot il range inter-quartile (25° e 75° percentile) è indicato dal box grigio, tagliato in due dalla mediana
- il range minimo-massimo è indicato dalle due barrette del wisker; tuttavia gli outliers sono esclusi dal wisker
- la crocetta rossa vicino alla mediana rappresenta la media
- il box-and-wisker plot è stato concepito per rappresentare campioni con distribuzioni asimmetriche, e comunque non-normali, per cui la deviazione standard non è solitamente riportata nei box-and-wisker plot
- gli outliers sono rappresentati dai punti blu fuori dal wisker; sono valori bassi che distano dal 25° percentile oltre 1.5 volte il range interquartile (cioè il range tra il 25° e 75° percentile), oppure valori alti che distano altrettanto dal 75° percentile; in questi grafici troviamo solo outliers alti
- se un punto dista dal minimo o dal massimo oltre 3 volte il range interquartile è considerato outlier estremo ed è qui indicato con una crocetta sovrapposta al punto (difficile da riconoscere, vedi 'Costo' e 'Km/L')
- dalla asimmetria del box e dalla distanza tra media e mediana è possibile valutare l'asimmetria delle distribuzioni
- questi scatterplot sono molto importanti in quanto descrivono le relazioni, tuttavia la asimmetria delle distribuzioni e la forma 'a freccia' delle nuvole dei punti non consentirebbero l'applicazione dell'analisi della regressione e della correlazione. Per queste analisi occorrerebbe prima applicare delle trasformazioni di de-trendizzazione delle deviazioni standard

Correlazioni spurie e correlazioni parziali

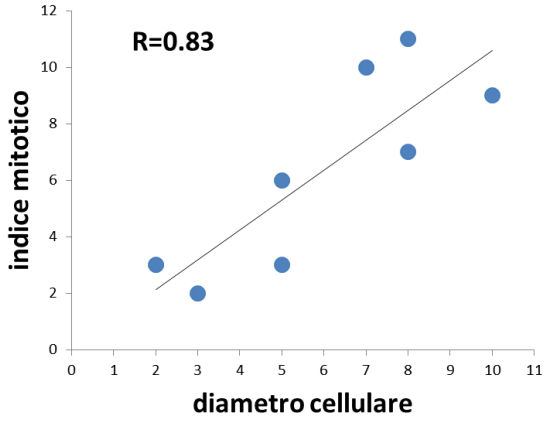
Se coltiviamo cellule in diverse capsule senza accorgerci che in alcune capsule le condizioni del pH o dei nutrienti non sono ottimali, i dati potrebbero risultare così:

capsula	indice mitotico	diametro cellulare	enzimi lisosomali	marker k	pH dato non conosciuto
1	2	3	2	5	6.1
2	3	5	2	4	6.1
3	3	2	2	2	6.2
4	6	5	7	6	6.5
5	7	8	5	8	6.6
6	10	7	11	9	7.0
7	11	8	9	7	7.1
8	9	10	9	9	7.0

A questo punto se consideriamo i dati di 2 variabili, ad es. indice mitotico e diametro cellulare...

capsula	indice mitotico	diametro cellulare
1	2	3
2	3	5
3	3	2
4	6	5
5	7	8
6	10	7
7	11	8
8	9	10

... troviamo una forte correlazione!



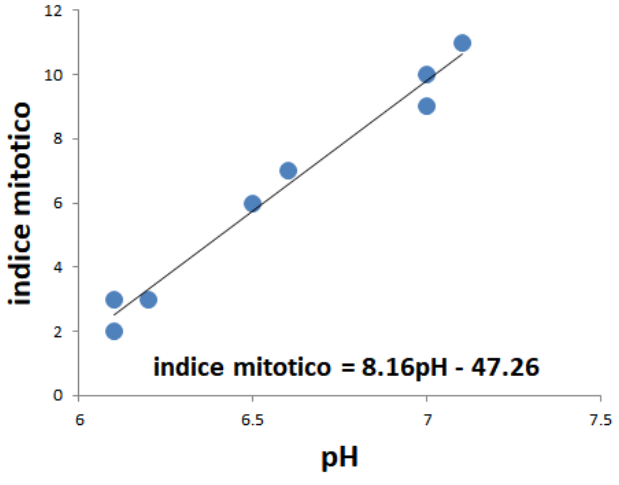
Ma questa correlazione è falsa in quanto è dovuta alle condizioni delle colture. I valori di indice mitotico e diametro cellulare sono bassi nelle capsule 1-2-3 mantenute in condizioni non buone (pH<6.5) e viceversa alti in quelle mantenute 6-7-8 in condizioni ottimali (pH~7.0). Questa forma di correlazione falsa o 'spuria' è stata da subito evidenziata da Pearson, l'autore che per

prima formulò il coefficiente di correlazione, ed è estremamente subdola in quanto difficilmente controllabile. Solo se si conosce il fattore che influenza i dati le correlazioni spurie si possono correggere calcolando le cosiddette correlazioni parziali. Correlazioni parziali significa correlazioni al netto dell'effetto di altre variabili, dette covariate. Seguendo l'esempio delle colture, se valutiamo il pH e lo consideriamo come covariata, la correlazione parziale è quella che intercorre tra indice mitotico e diametro cellulare a parità di pH, o, in altre parole, tenendo costante il pH.

Il simbolo della correlazione parziale tra due variabili x1 e x2, tenendo costante la variabile x3 (cioè azzerandone l'effetto) è:

$r_{x1\ x2,\ x3}$

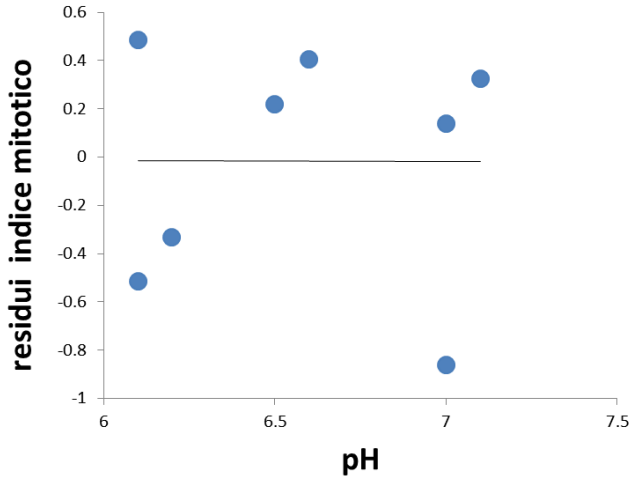
Nel nostro caso, per ottenere la correlazione parziale, calcoliamo prima la retta di regressione tra indice mitotico e pH



e mediante l'equazione di regressione i calcoliamo i residui:

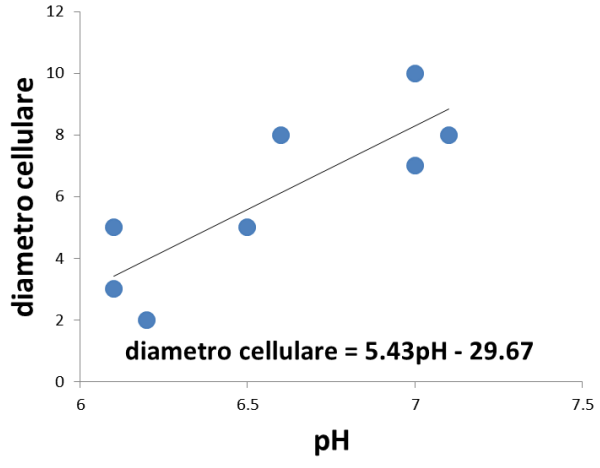
capsula	indice mitotico osservato	indice mitotico calcolato in base alla equazione di regressione in funzione del pH (sono i punti sulla retta)	residui (differenze)
1	2	2.516	-0.516
2	3	2.516	0.484
3	3	3.332	-0.332
4	6	5.78	0.22
5	7	6.596	0.404
6	10	9.86	0.14
7	11	10.676	0.324
8	9	9.86	-0.86

I residui per definizione sono del tutto indipendenti dal pH, come si vede:

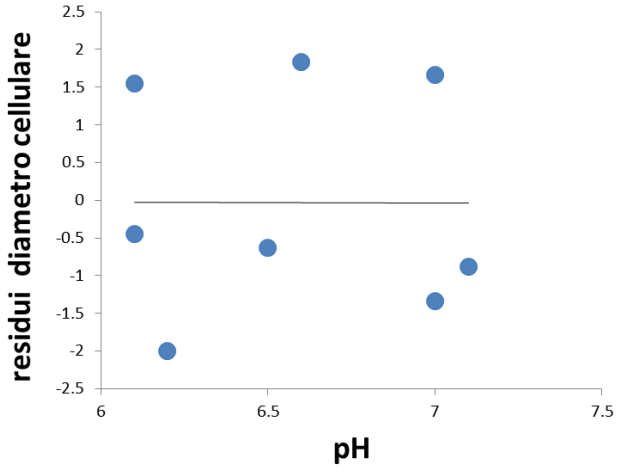


...sì, la retta è perfettamente orizzontale! Pendenza zero = nessuna relazione.

Poi facciamo la stessa cosa considerando la regressione tra diametro cellulare e pH:

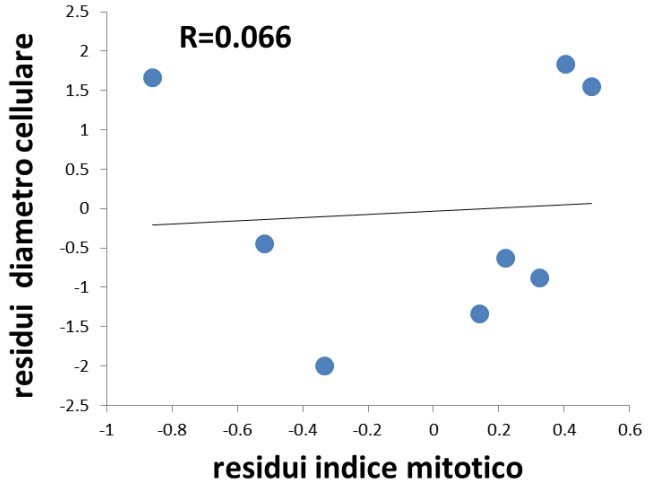


capsula	diametro cellulare osservato	diametro cellulare calcolato in base alla equazione di regressione in funzione del pH (sono i punti sulla retta)	residui (differenze)
1	3	3.453	-0.453
2	5	3.453	1.547
3	2	3.996	-1.996
4	5	5.625	-0.625
5	8	6.168	1.832
6	7	8.34	-1.34
7	8	8.883	-0.883
8	10	8.34	1.66



Anche in questo caso vediamo che la retta è perfettamente orizzontale.

I residui non sono altro che i valori delle due variabili tolto l'effetto del pH (covariata).
A questo punto, la correlazione tra le due serie di residui corrisponde alla correlazione parziale tra indice mitotico e diametro cellulare, tenendo costante il pH.



Come si vede, una volta eliminato l'effetto del pH, la correlazione tra le due variabili è estremamente bassa (r=0.066), al contrario di come sembrava all'inizio (r=0.83)!

C'è una scorciatoia che consente di calcolare la correlazione parziale tra due variabili in funzione di una terza variabile (covariata), utilizzando i tre coefficienti di correlazione tra le variabili prese a due a due:

$$r_{12,3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Ovviamente esiste una spiegazione per questa formula ma non è il caso di entrare nei dettagli in questo momento. C'è da dire che la correlazione parziale può essere calcolata anche in funzione di più covariate, così come anche la regressione può essere calcolata tra più variabili.

Esiste infine anche la possibilità di calcolare una correlazione semiparziale, cioè una correlazione ottenuta con i residui di una sola delle due variabili in esame rispetto alla covariata.

Il simbolo della correlazione semiparziale tra due variabili x1 e x2, 'parziallizando' solo x2 per la covariata x3 è:

$r_{x1\ (x2,x3)}$

e si calcola con

$$r_{1(2,3)} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{23}^2)}}$$

La correlazione semiparziale, anziché quella parziale, è lecita se siamo sicuri che la covariata ha un effetto su una sola delle due variabili da correlare.

Occorre in ogni modo stare attenti ad applicare la correlazione parziale e simiparziale solo nei casi in cui siamo sicuri che sia la covariata (x3) ad influenzare le variabili in esame (x1 e x2), e non viceversa.Se così fosse, applicando la correlazione parziale o semiparziale distruggeremo la vera correlazione tra le variabili x1 e x2. Ad ogni modo, è sempre meglio considerare bene il disegno sperimentale ed utilizzare campioni il più possibile omogenei e controllati per prevenire sin dall'inizio l'effetto di covariate.

Il chi-quadro (χ^2)

Il test del chi-quadro (χ^2) serve a saggiare la differenza tra frequenze. Generalmente si tratta di confrontare alcune frequenze osservate con le frequenze che teoricamente si sarebbero dovute osservare, le cosiddette frequenze attese.

Le ipotesi sono le solite:
H0: la differenza è dovuta al caso (campionamento, imprecisione, ecc.) e le frequenze osservate non differiscono significativamente da quelle attese.
H1: la differenza è significativa, dovuta a fattori che influiscono sulla distribuzione delle frequenze. Le frequenze osservate non provengono dal modello teorico delle frequenze attese.

Il test consiste nel semplice rapporto:

$$\chi^2 = \frac{(\text{frequenze osservate - frequenze attese})^2}{\text{frequenze attese}}$$

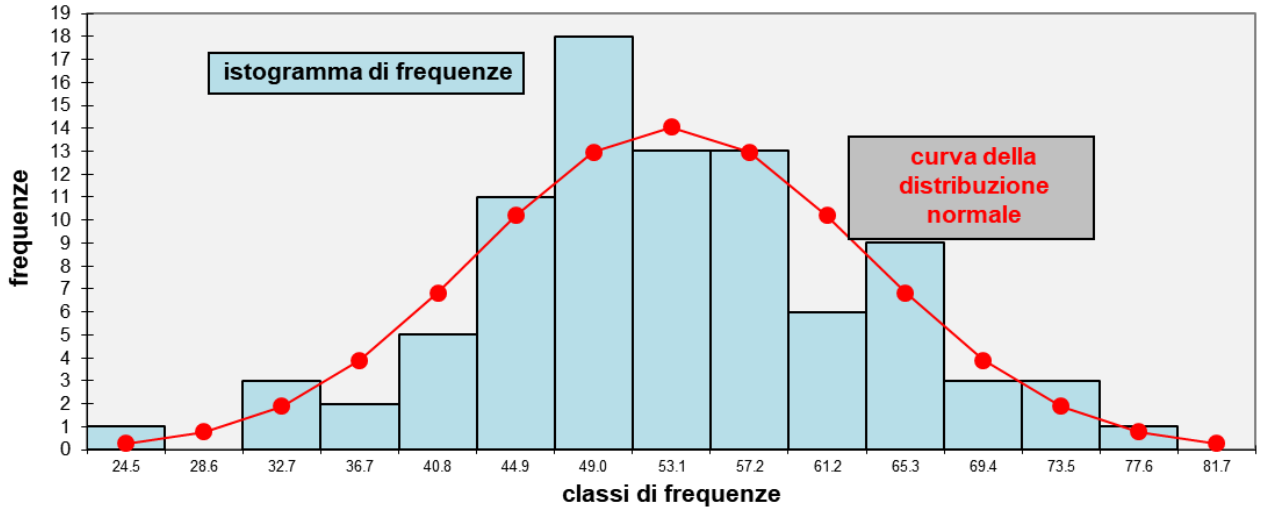
In base al valore del χ^2, la tabella o il programma indicherà la probabilità di trovare (per caso) un χ^2 tanto grande. Se la probabilità sarà inferire al 5% potremo dire che la differenza è significativa.

Notabene. Le frequenze devono essere sempre espresse come frequenze assolute. Mai utilizzare frequenze percentuali o relative.

Esistono numerose applicazioni del test del χ^2. Vediamone alcune.

Test di normalità: goodness-of-fit (bontà di adattamento)

Per testare l'adattamento dei dati alla distribuzione normale è possibile confrontare le frequenze osservate ripartite in classi (v. istogramma) con le frequenze previste dalla distribuzione normale. Il grafico riporta un istogramma sovrapposto alla curva di distribuzione normale calcolata per la stessa media e deviazione standard dei dati dell'istogramma.



Per ogni classe esiste quindi una frequenza osservata (altezza della colonna) ed una frequenza attesa (altezza del pallino della curva). La formula del χ^2 si applica a tutte le colonne sommando i risultati. I gradi di libertà sono pari al numero delle classi - 3, in quanto abbiamo 3 vincoli: la

media e deviazione standard che definiscono la distribuzione normale + una delle classi. Il valore di p ci dirà se la differenza tra le frequenze osservate e le frequenze attese per la distribuzione normale sono significative. Se non sono significative, manteniamo l'ipotesi nulla che assume che la distribuzione dei dati sia compatibile con quella normale.
Notare che le prime e ultime classi con frequenze attese inferiori a 5 vanno accorpate alle classi adiacenti. Nella figura questo è stato omesso per non complicare la dimostrazione.

Test di simmetria

Se la distribuzione è normale è necessariamente anche simmetrica attorno alla media, e quindi metà dei valori (n/2) saranno inferiori alla media e metà (n/2) saranno superiori (escludiamo che esistano valori perfettamente identici alla media, o se ve ne fossero li togliamo dal conto). In pratica, questo test corrisponde ad un confronto tra media e mediana. Si tratta in definitiva di confrontare il numero dei valori inferiori e superiori alla media contro il valore n/2.

Ad esempio, immaginiamo il seguente campione:
4, 5, 8, 3, 4, 8, 4, 5, 6, 3, 4, 2, 3, 4, 9, 5
con media=4.81 e numerosità=16.

I valori osservati inferiori alla media sono 9: 4, 3, 4, 4, 3, 4, 2, 3, 4
I valori osservati superiori alla media sono 7: 5, 8, 8, 5, 6, 9, 5

Di contro le frequenze attese inferiori e superiori alla media sono n/2: 8 e 8.

Pertanto il test consisterà nel calcolo:

$$\chi^2 = \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} = 0.25$$

Anche questa volta il risultato non è statisticamente significativo. Ma si tratta di un campione molto piccolo, per semplificare (purtroppo quando i campioni sono piccoli i test non si sbilanciano).

Questo χ^2 ha 1 solo grado di libertà in quanto le frequenze attese sono date dalla media delle due frequenze osservate. Pertanto, se la frequenza osservata dei valori inferiori alla media varia in un senso rispetto alla frequenza attesa, la frequenza dei valori superiori necessariamente varierà in senso opposto. Quindi una sola delle due frequenze osservate è libera di variare.

Tabelle 2x2

La tabella 2×2 serve a valutare l'associazione tra due caratteri/qualità/attributi o a confrontare due proporzioni. E' una sorta di correlazione tra caratteri qualitativi. La tabella 2×2 è una normale tabella a due entrate. Ogni entrata ha due modalità che devono essere:

- mutualmente esclusive (senza alcuna sovrapposizione)
- esaustive (comprendere tutti i casi che mostrano i caratteri in esame)
- indipendenti (il fatto di trovare una certa modalità in un soggetto non influisce sulla modalità presente nel soggetto successivamente campionato né in tutti gli altri).

Alcuni esempi.

Esiste una relazione tra colore degli occhi e colore dei capelli?

		Colore degli occhi	
		celeste	non-celeste
Colore dei capelli	biondi	24	6
	non-biondi	28	90

La proporzione tra promossi/bocciati è uguale in due scuole?

		Scuola	
		liceo 'A'	Liceo 'B'
Esiti scrutini	promossi o rimandati	450	640
	bocciati	110	60

L’espressione del recettore K è associata all’espressione del recettore Q?

		Recettore K	
		presente	assente
Recettore Q	presente	72	14
	assente	54	73

In tutti i casi ci si chiede se ci sia una certa relazione tra le modalità di due variabili. Se questa esiste, allora i rapporti in verticale o in orizzontale tendono a divergere.

Riprendiamo l'esempio del colore dei capelli/occhi e calcoliamo i rapporti in verticale.

		Colore degli occhi	
		celeste	non-celeste
Colore dei capelli	biondi	24/(24+28) *46%*	6/(6+90) *6%*
	non-biondi	28/(24+28) *54%*	90/(6+90) *94%*

Ma possiamo anche calcolare i rapporti in orizzontale.

		Colore degli occhi	
		celeste	non-celeste
Colore dei capelli	biondi	24/(24+6) *80%*	6/(24+6) *20%*
	non-biondi	28/(28+90) *24%*	90/(28+90) *76%*

In entrambe i modi notiamo una discrepanza tra le percentuali.

Quelle riportate sono le frequenze osservate. E le frequenze attese? Per le frequenze attese non ci si basa su una particolare distribuzione. Sono semplicemente quelle che annullano le discrepanze tra le due proporzioni. In altre parole le frequenze attese sono quelle attese per l'ipotesi nulla, che sostiene che i rapporti siano uguali, e se si vedono differenze queste sono dovute a fattori casuali. Come al solito l'ipotesi nulla è l'ipotesi dello scettico. Calcolare le frequenze attese è semplice.

Riprendiamo l'esempio del colore dei capelli/occhi. Aggiungiamo alla tabella una riga ed una colonna in cui inserire i totali, più il totale generale

		Colore degli occhi		
		celeste	non-celeste	**totali di riga**
Colore dei capelli	biondi	24	6	**30**
	non-biondi	28	90	**118**
	totali di colonna	**52**	**96**	**148**
				totale generale

Ora, per ognuna delle 4 caselle calcoliamo il prodotto del totale della sua riga × totale della sua colonna, e poi dividiamo per il totale generale:

Frequenze attese

		Colore degli occhi		
		celeste	non-celeste	totali di riga
Colore dei capelli	biondi	52×30/148= **10.5** ***(20%)***	96×30/148= **19.5** ***(20%)***	30
	non-biondi	52×118/148= **41.5** ***(80%)***	96×118/148= **76.5** ***(80%)***	118
	totali di colonna	52	96	148

Nota come le frequenze attese rendono uguali le proporzioni pur lasciando inalterati i totali di riga e di colonna ed il totale generale.

Riassumendo, abbiamo due tabelle:

Frequenze osservate

		Colore degli occhi: celeste	Colore degli occhi: non-celeste
Colore dei capelli	biondi	**24**	**6**
	non-biondi	**28**	**90**

Frequenze attese

		Colore degli occhi: celeste	Colore degli occhi: non-celeste
Colore dei capelli	biondi	**10.5**	**19.5**
	non-biondi	**41.5**	**76.5**

Ora creiamo una terza tabella in cui, per ognuna delle 4 caselle, calcoliamo il χ^2 con la formula vista all'inizio del capitolo

$$\chi^2 = \frac{(\text{frequenze osservate - frequenze attese})^2}{\text{frequenze attese}}$$

Valori di χ^2

		Colore degli occhi: celeste	Colore degli occhi: non-celeste
Colore dei capelli	biondi	**17.35**	**9.35**
	non-biondi	**4.39**	**2.38**

Il chi-quadro complessivo sarà la loro somma:

$$\chi^2 = \frac{(24-10.5)^2}{10.5} + \frac{(6-19.5)^2}{19.5} + \frac{(28-41.5)^2}{41.5} + \frac{(90-76.5)^2}{76.5} =$$

$$= \frac{13.5^2}{10.5} + \frac{13.5^2}{19.5} + \frac{13.5^2}{41.5} + \frac{13.5^2}{76.5} =$$

$$= \frac{182.25}{10.5} + \frac{182.25}{19.5} + \frac{182.25}{41.5} + \frac{182.25}{76.5} =$$

$$= 17.35 + 9.35 + 4.39 + 2.38 = 33.47$$

Confrontiamo il valore trovato (χ^2 =33.47) con il valore critico del χ^2 per probabilità del 5% con 1 grado di libertà (χ^2 =3.841). Il valore calcolato è ben superiore al valore critico. Pertanto si rigetta l'ipotesi nulla, concludendo che esiste una relazione tra colore dei capelli e colore degli occhi.
Avrete notato come la differenza tra frequenza osservata e frequenza attesa sia pari a 13.5 in tutte le 4 caselle. Infatti, dovendo rispettare i totali marginali, se sottraiamo una certa quantità ad una casella dobbiamo aggiungere la stessa quantità alla casella adiacente, sia in verticale che in orizzontale. Per questo motivo la tabella 2×2, pur sviluppando quattro quozienti, ha 1 solo grado di libertà. Mediante lo stesso ragionamento, una volta calcolata la frequenza attesa di una casella,

le frequenze attese delle altre tre sono vincolate dal rispetto dei totali marginali, per cui il modo più semplice di calcolare le frequenze attese è per differenza rispetto alla prima.

Ma esiste una scorciatoia ancora migliore in quanto è possibile utilizzare una formula che ci dà il χ^2 in un solo passaggio. Se chiamiamo le 4 caselle con a, b, c, d, la formula immediata è:

		Colore degli occhi	
		celeste	non-celeste
Colore dei capelli	biondi	**a=24**	**b=6**
	non-biondi	**c=28**	**d=90**

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

in cui (a+b) (c+d) (a+c) (b+d) sono i 4 totali marginali ed n è il totale generale (a+b+c+d).

Nel nostro caso darà:

$$\chi^2 = \frac{(24 \cdot 90 - 6 \cdot 28)^2 \cdot 148}{30 \cdot 118 \cdot 52 \cdot 96} = 33.23$$

33.23 corrisponde al 33.47 ottenuto in base al calcolo delle frequenze attese con la differenza di qualche decimale dovuta agli arrotondamenti. Questa formula è algebricamente equivalente al procedimento indicato sopra ma non spiega la logica del test.

Tabelle mxn

E' possibile organizzare anche tabelle con più righe e più colonne (fatta salva la regola della mutua esclusività, esaustività ed indipendenza delle osservazioni). In questo caso purtroppo non esistono formule scorciatoia. Le frequenze attese si calcolano con la formula canonica: per ogni casella, la frequenza attesa è data dal totale di riga moltiplicato il totale di colonna diviso il totale generale.

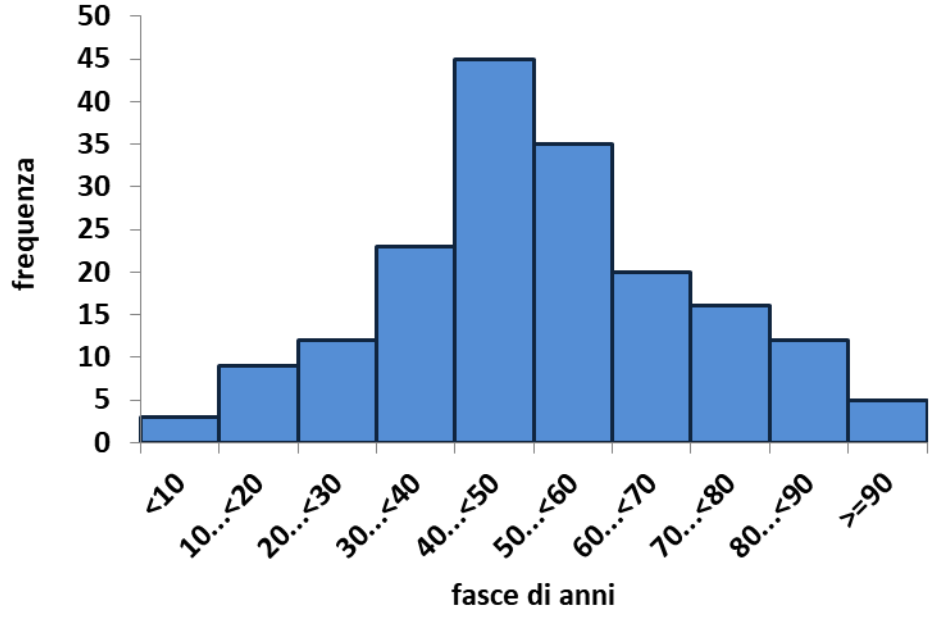
In generale le tabelle 2×2 e le altre m×n sono dette tabelle di contingenza. Per tutte le tabelle i gradi di libertà del χ^2 saranno (n° di righe-1)×(n° di colonne-1). Il motivo è lo stesso detto per la tabella 2×2: il rispetto dei totali marginali fa sì che in ogni riga ed in ogni colonna un dato sia vincolato dal valore degli altri. Per cui in una riga di n dati, solo n-1 saranno liberi di variare. Idem, in una colonna di m dati, solo m-1 saranno liberi di variare. Da cui i gradi di libertà saranno (n-1)×(m-1).

Esempio di tabella 3×3:

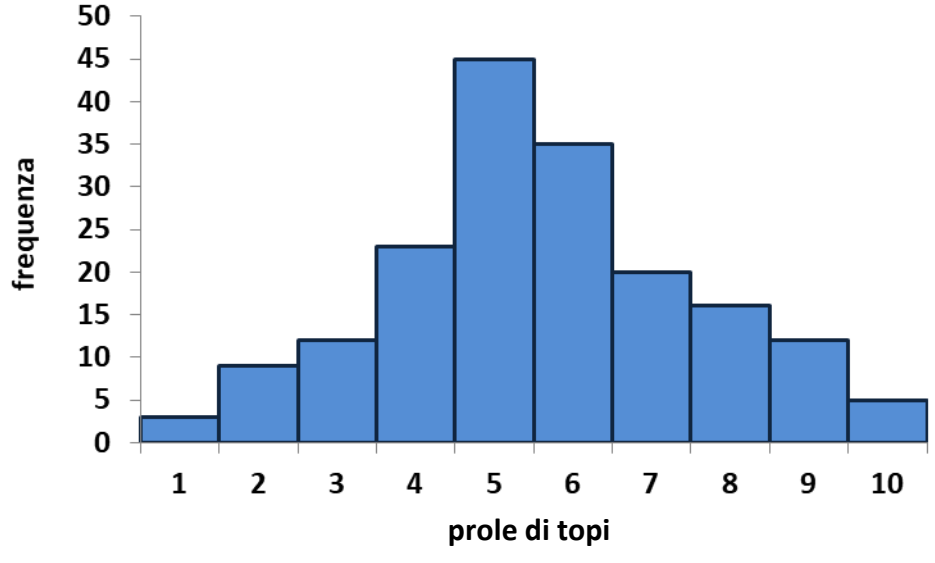
		Severità della malattia		
		lieve	moderata	grave
Quadro istologico	normale	13	6	2
	poco alterato	6	18	17
	molto alterato	2	12	24

Correzione di Yates per la continuità

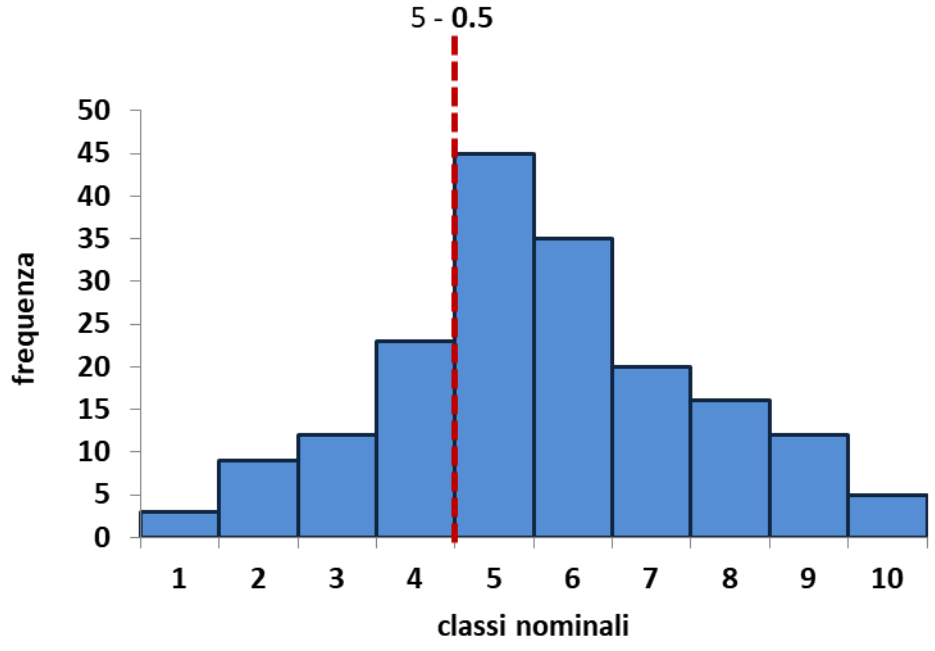
Le classi di frequenza delle variabili continue sono necessariamente riferite a intervalli di scala:



mentre le classi di frequenza delle variabili nominali espresse da conteggi hanno come riferimento valori discreti localizzati al centro di ogni classe:



Supponiamo ora di dover calcolare una certa area di quest'ultimo istogramma in base ad un determinato valore di ascissa. Ad esempio, l'area dell'istogramma dal valore 5 in su:



La vera soglia di confine tra la classe 5 e la classe precedente (classe 4) corrisponde al valore di ascissa 5 - 0.5 = 4.5. Nel test chi-quadro questa operazione è detta correzione per la continuità o correzione di Yates, ed è di fondamentale importanza quando l'area in gioco è piccola, come al livello delle code critiche della distribuzione. Questa correzione applicata alle tabelle 2×2 consiste nel sottrarre 0.5 al numeratore:

$$\chi^2_c = \sum \frac{(|\text{frequenza osservata} - \text{frequenza attesa}| - 0.5)^2}{\text{frequenza attesa}}$$

E' bene mettere una lettera 'c' a pedice del simbolo χ^2 per indicare la correzione per la continuità. Il valore del χ^2 coretto per la continuità è sempre minore di quello non corretto. La correzione per la continuità è obbligatoria quando il campione è piccolo, ma la si può comunque applicare sempre, anche quando il campione è grande, nel qual caso il suo effetto è trascurabile. E' quindi buona abitudine utilizzarla sempre. Invece la correzione per la continuità non può essere applicata alle tabelle m×n o comunque quando il χ^2 ha più di 1 grado di libertà. La formula scorciatoia del χ^2 con la correzione per la continuità per tabelle 2×2 è la seguente:

$$\chi^2_c = \frac{(|ad - bc| - n/2)^2 \, n}{(a+b)(c+d)(a+c)(b+d)}$$

Test esatto di Fisher

Quando n è più piccolo di 40 e vi sono frequenze attese minori di 5 è necessario ricorrere ad un metodo più preciso del χ^2, basato sul calcolo combinatorio, detto metodo esatto di Fisher. Tale metodo stima direttamente la probabilità di osservare la disproporzione rappresentata dalla tabella più quelle ancora più estreme. Il procedimento è basato sul seguente ragionamento: se in una tabella 2×2 aumentiamo o diminuiamo il valore di una casella (una qualsiasi delle 4), tutte le altre frequenze dovranno insieme variare per mantenere gli stessi totali marginali. Ma in un senso la disproporzione tenderà ad aumentare mentre nell'altro tenderà a ridursi, o viceversa. Supponiamo ora la tabella:

3	8
7	5

la disproporzione è **a**/**c** 3/7=**0.43** contro **b/d** 8/5=**1.60**.

Se aggiungiamo 1 alla casella **a**, per rispettare i totali marginali, siamo dobbiamo togliere 1 alle caselle **b** e **c** e ad aggiungere 1 a **d**.
La tabella con **a**+1 ecc. diventa:

3+1=4	8-1=7
7-1=6	5+1=6

la disproporzione ora è minore: 4/6=**0.666** contro 7/6=**1.166**.

Se continuiamo in questa direzione le disproporzioni tendono ad equalizzarsi, sino a raggiungere la condizione più bilanciata possibile, e cioè, con **a**+2 ecc.

3+2=5	8-2=6
7-2=5	5+2=7

la disproporzione è la minima possibile 5/5=**1** contro 6/7=**0.857**.

Ma se continuiamo a far crescere **a** i rapporti si distanziano nuovamente, questa volta in senso inverso. Con **a**+3 ecc.

3+3=6	8-3=5
7-3=4	5+3=8

la disproporzione è: 6/4=**1.50** contro 5/8=**0.63**.

ecc.

Il gioco termina quanto avremo zero in una casella, perché non si possono avere frequenze negative.

3+7=10	8-7=1
7-7=0	5+7=12

Per questa tabella la disproporzione non si può calcolare in quanto non si può dividere un numero per zero

Se invece fossimo partiti sottraendo 1 ad **a** ed a **d**, e sommando 1 a **b** e **c**, avremmo ottenuto disproporzioni sempre maggiori. Con **a**-1 ecc.

3-1=2	8+1=9
7+1=8	5-1=4

disproporzione: 2/8=**0.25** contro **2.25**

Con **a**-2 ecc.

3-2=1	8+2=10
7+2=9	5-2=3

disproporzione: 1/9=**0.111** contro 10/3=**3.333**

Con **a**-3 ecc.

3-3=0	8+3=11
7+3=10	5-3=2

disproporzione: 0/10=**0** contro 11/2=**5.5**
La disproporzione così ottenuta è la più estrema possibile.

Abbiamo così visto che in una direzione (incrementando a) la disproporzione tende a livellarsi, mentre nella direzione opposta (decrementando a) la disproporzione tende ad aumentare. Ora, il metodo di Fisher calcola la somma delle probabilità della tabella iniziale e di tutte le altre possibili nella direzione di un aumento della disproporzione. La probabilità di ogni singola tabella è data dalla la formula:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!}\frac{1}{a!b!c!d!}$$

Si tratta quindi si un calcolo molto pesante che richiede l'uso del computer, ma ha il vantaggio di ottenere direttamente la probabilità (dell'ipotesi nulla) senza passare attraverso parametri statistici. Notare che i fattoriali dei numeri interi (> 20) sono numeri giganteschi. Es. 50! = 3.0414... × 10^{64}. Per questo, nel calcolo del test esatto di Fisher i programmi si avvalgono temporaneamente della trasformazione logaritmica che trasforma un fattoriale in una semplice sommatoria, ad esempio
Log(4!) = Log(1×2×3×4) = Log(1) + Log(2) + Log(3) + Log(4)
Ultima nota: il test esatto di Fisher è a una coda, per cui nella pratica dovremmo prendere in considerazione come soglia di significatività il valore di P <0.025, anziché P<0.05.

Riassumendo

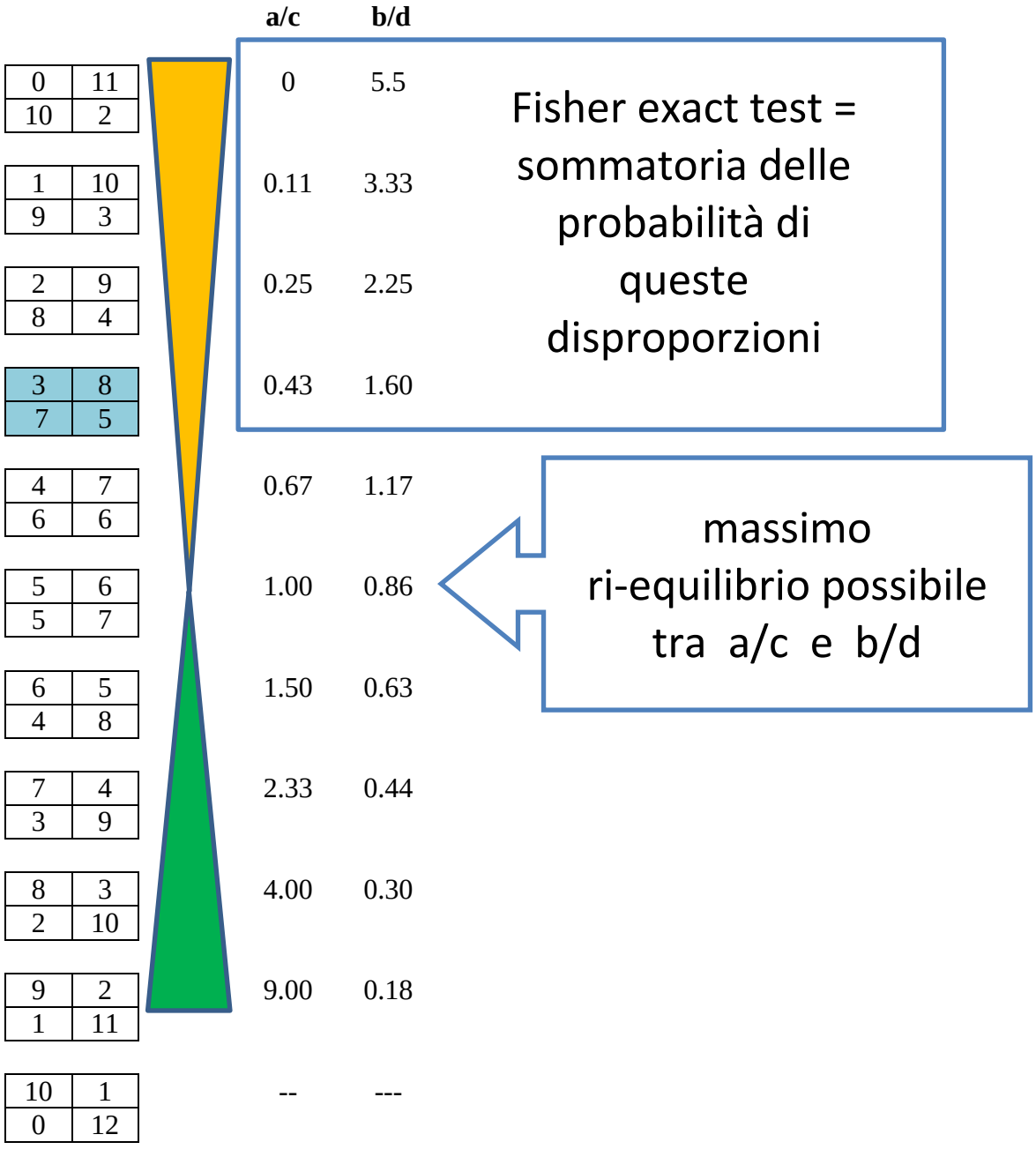


Tabella 2x2 per campioni appaiati (test di Mc Nemar)

Quando le 2 modalità dei due caratteri sono osservate nello stesso soggetto possiamo configurare diversamente la tabella. Non più nel modo visto sin qui:

		condizione B	
		B1	B2
condizione A	A1		
	A2		

ma in questo modo:

		condizione A2	
		B1	B2
condizione A1	B1		
	B2		

Supponiamo che un certo numero di volontari abbiano il compito di bere in giorni successivi una tazza di caffè o di tè prima di andare a dormire, e poi di dire se hanno potuto prendere sonno subito dopo l'assunzione di ciascuna bevanda. I dati di tale sperimentazione si possono indicare in questa tabella 2×2 per dati appaiati:

		assunzione di caffè	
		sonno facile	sonno difficile
assunzione di tè	sonno facile		
	sonno difficile		

Notare la differenza rispetto alla tabella 2×2 per dati non appaiati, che sarebbe:

		assunzione di	
		caffè	tè
sonno	facile		
	difficile		

La tabella per dati appaiati esprime l'esperienza delle due condizioni in ciascun soggetto, mentre quella per dati non appaiati esprime esperienze non associate agli stessi soggetti. Il vantaggio dei dati appaiati consiste nel fatto che in questo modo il test è più bilanciato (i più sensibili dichiareranno sonno difficile con entrambe le bevande - i meno sensibili dichiareranno sonno facile con entrambe le bevande). Il principio è simile a quello del test t per campioni appaiati.

Nella tabella per dati appaiati le caselle **a** e **d** sono indifferenti in quanto riportano gli stessi effetti delle due bevande. Il test dipende dalle altre due caselle: **c** e **b**.

		assunzione di caffè: sonno facile	assunzione di caffè: sonno difficile
assunzione di tè	sonno facile	**a**	**b**
	sonno difficile	**c**	**d**

Il χ^2 corretto per tabelle 2×2 di dati appaiati è dato da:

$$\chi_c^2 = \frac{(|b-c|-1)^2}{b+c}$$

Spiegazione per i più curiosi.
Poiché le caselle **a** e **d** sono indifferenti, si considera la differenza tra le caselle **b** e **c**. L'ipotesi nulla, che sostiene l'indifferenza, ipotizza che le frequenze attese **b** e **c** siano uguali ed assegna a ciascuna di loro il valore (**b**+**c**)/2. Queste sono le frequenze attese, per cui si può applicare la normale formula:

$$\chi_c^2 = \frac{\left(\left|b-\frac{b+c}{2}\right|-0.5\right)^2}{\frac{b+c}{2}} + \frac{\left(\left|c-\frac{b+c}{2}\right|-0.5\right)^2}{\frac{b+c}{2}} = \frac{\left(\left|\frac{b-c}{2}\right|-0.5\right)^2 + \left(\left|\frac{c-b}{2}\right|-0.5\right)^2}{\frac{b+c}{2}}$$

ma poichè $\left|\frac{b-c}{2}\right| = \left|\frac{c-b}{2}\right|$

posiamo semplificare

$$\chi_c^2 = \frac{2\left(\left|\frac{b-c}{2}\right|-0.5\right)^2}{\frac{b+c}{2}} = \frac{2\left(\left|\frac{b-c}{2}\right|-\frac{1}{2}\right)}{\frac{b+c}{2}} = \frac{2\left[\frac{1}{2}(|b-c|-1)\right]^2}{\frac{b+c}{2}} = \frac{\frac{2}{4}(|b-c|-1)^2}{\frac{1}{2}(b+c)} = \frac{(|b-c|-1)^2}{(b+c)}$$

Distribuzione binomiale

Premessa. Per la distribuzione binomiale si usano spesso le lettere **p** e **q**. Non confondiamo questa **p** con la p che solitamente è utilizzata per esprimere la probabilità dell'errore α dei test.

La distribuzione binomiale consente di calcolare la probabilità che una modalità di un evento con una certa probabilità a priori (**p**) si verifichi un determinato numero di volte (**i**) in un numero totale (**n**) di eventi. Ad esempio, la distribuzione binomiale può calcolare la probabilità che:

- su 10 figli, 7 siano maschi
  p (probabilità a priori del carattere maschio) = 0.5
  q (probabilità a priori del carattere non-maschio, complementare a p) = 1 - **p** = 0.5
  i (numero di occorrenze del carattere maschio) = 7
  n (numero totale di eventi) = 10

- su 8 lanci di dado, il due esca tre volte
  p = 1/6 = 0.16666
  q = 1-**p** = 0.83333
  i = 3
  n = 8

- sui risultati della schedina compaiano 10 pari
  p = 1/3 = 0.33333
  q = 1-**p** = 0.66666
  i = 10
  n = 13

La formula generale della distribuzione binomiale, in grado di calcolare la probabilità di un certo carattere, in una serie di **n** eventi, è:

$$P_{n,i,p} = \frac{n!}{i!(n-i)!} \cdot p^i \cdot q^{n-i}$$

Se vogliamo ottenere l'intera distribuzione binomiale dobbiamo ripetere la formula n+1 volte, mantenendo costanti **p** ed **n** e variando solo **i** (vedi l'istogramma più avanti)
Nel caso della probabilità di figli maschi su **n**=10 figli, metteremo
i = 0 per la probabilità di avere 0 figli maschi su 10 figli
i = 1 ... 1 figlio maschio su 10
i = 2 ... 2 figli maschi su 10
i = 3 ... 3 figli maschi su 10
i = 4 ... 4 figli maschi su 10
i = 5 ... 5 figli maschi su 10
i = 6 ... 6 figli maschi su 10
i = 7 ... 7 figli maschi su 10
i = 8 ... 8 figli maschi su 10
i = 9 ... 9 figli maschi su 10
i = 10 ... 10 figli maschi su 10

I valori delle 11 espressioni corrispondono esattamente ai valori che si ottengono dallo sviluppo del binomio $(\mathbf{p}+\mathbf{q})^{\mathbf{n}}$, in questo caso $(0.5+0.5)^{10}$. La distribuzione binomiale si chiama così appunto perché corrisponde allo sviluppo della potenza di un binomio. Nelle scuole medie abbiamo imparato a memoria lo sviluppo del quadrato e del cubo di un binomio. Per potenze

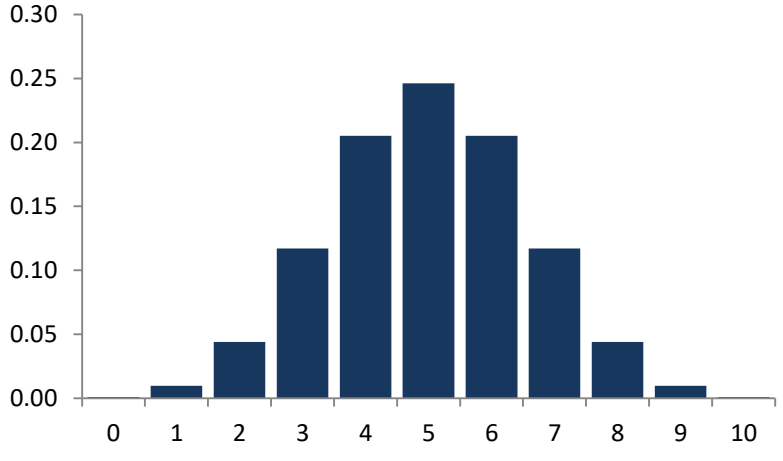
maggiori possiamo usare il triangolo di Tartaglia che dà la serie di coefficienti da applicare alla parte letterale con esponenti decrescenti (da **n** a 0) di **p**, ed esponenti crescenti (da 0 a **n**) di **q**.

Triangolo di Tartaglia (fino al 10° grado)

							1														$(p+q)^0$
						1		1													$(p+q)^1$
					1		2		1												$(p+q)^2$
				1		3		3		1											$(p+q)^3$
			1		4		6		4		1										$(p+q)^4$
		1		5		10		10		5		1									$(p+q)^5$
	1		6		15		20		15		6		1								$(p+q)^6$
		1		7		21		35		35		21		7		1					$(p+q)^7$
	1		8		28		56		70		56		28		8		1				$(p+q)^8$
1		9		36		84		126		126		84		36		9		1			$(p+q)^9$
1		10		45		120		210		252		210		120		45		10		1	$(p+q)^{10}$

Per cui $(\mathbf{p}+\mathbf{q})^{\mathbf{n}}$ sarà:

$$\mathbf{1}p^{10}q^{0} + \mathbf{10}p^{9}q^{1} + \mathbf{45}p^{8}q^{2} + \mathbf{120}p^{7}q^{3} + \mathbf{210}p^{6}q^{4} + \mathbf{252}p^{5}q^{5} + \mathbf{210}p^{4}q^{6} + \mathbf{120}p^{3}q7 + \mathbf{45}p^{2}q^{8} + \mathbf{10}p^{1}q^{9} + \mathbf{1}p^{0}q^{10}$$



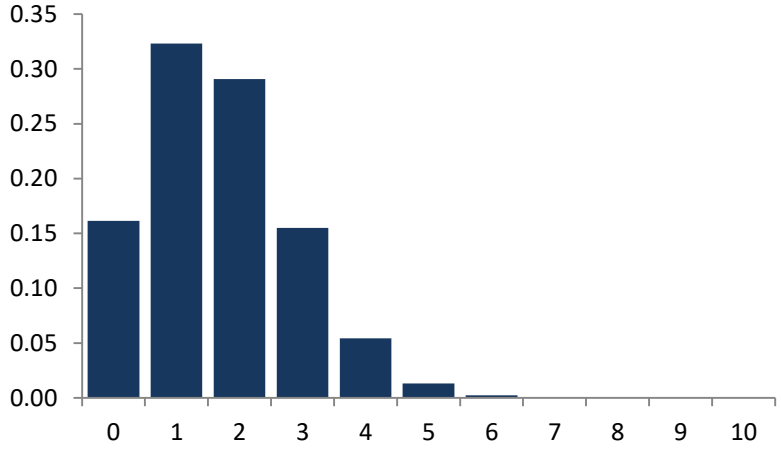
Se uno non ama Tartaglia può ottenere i coefficienti dalla formula $\frac{n!}{i!(n-i)!}$

mentre le parti letterali sono date dalla formula $p^{i} \cdot q^{n-i}$

Comunque è molto scomodo calcolare potenze di grado superiore a 10 con il triangolo di Tartaglia. E' meglio usare la formula con i fattoriali. Meglio ancora, usare un programma in grado di calcolare la distruibuzione binomiale. Excel lo fa egregiamente.

Il grafico riportato sopra mostra una distribuzione binomiale perfettamente simmetrica in quanto **p** e **q** sono uguali (entrambi 0.5).
Se invece **p** e **q** sono diversi - e quanto più sono diversi - il grafico risulta asimmetrico.

Vediamo ad es. la distribuzione binomiale dell'uscita del 2 in una serie di 10 lanci del dado



A che serve tutto ciò? Semplicemente a calcolare le frequenze attese da confrontare con quelle osservate. Se ad es. giochiamo con un dado possiamo confrontare le frequenze con cui esce il 2, in es. 50 lanci del dado, e confrontarle con le frequenze attese per la distribuzione binomiale con p=1/6 ed n=50. Poi calcoliamo il chi-quadro. Se il test è significativo vuol dire che il dado non è regolare. Potrebbe essere truccato.

La stessa cosa con il numero di figli maschi. Se il test è significativo vuol dire che c'è qualche fenomeno che interferisce sulla determinazione del sesso a livello genico o cromosomico o endocrino-embrionale.

Deviazione standard di una proporzione

Anche in questo casi, per i due termini del binomio si usano spesso le lettere **p** e **q**. Non confondiamo questa **p** con la p che rappresenta la probabilità α dei test.

La proporzione rappresenta nel campo delle frequenze ciò che è la media nel campo delle variabili continue. Infatti si usano espressioni del tipo: i fumatori sono 'in media' il ...% della popolazione, ecc. Come abbiamo visto nel paragrafo precedente, la distribuzione binomiale non è necessariamente simmetrica, anzi, quando **p** è piccolo e diverso da **q** la distribuzione si presenta fortemente asimmetrica. Ma al crescere di **n** ed al tendere di **p** a 0.5 (cioè ad essere uguale a **q**) la distribuzione binomiale tende a diventare simmetrica, e quando **n×p**>5 la distribuzione può essere considerata sufficientemente simmetrica da poter calcolare una deviazione standard da associare alla proporzione e quindi calcolare anche i limiti fiduciali della stessa proporzione. **Tutta la parte che segue è valida quando ricorre questa condizione.**

La deviazione standard di una proporzione è data dalla formula:

$$s_p = \sqrt{pq / n}$$

In questa formula possiamo utilizzare per **p** e **q** sia le frequenze relative che quelle percentuali o assolute. Invece **n** dovrà essere sempre il numero assoluto. Ovviamente a seconda di ciò che usiamo la deviazione standard sarà espressa in frequenze relative o percentuali o assolute.

Se, ad esempio, la proporzione riguarda 15 osservazioni su un totale di 70,

se usiamo le frequenze **relative**
p=15/70=0.21 q=1-p=0.79
$s_p = \sqrt{0.21 \times 0.79/70} = 0.049$
0.049 è la deviazione standard associata sia a 0.21 che a 0.79

se usiamo le **percentuali**
p =21 q=100-p=q=79
$s_{\%} = \sqrt{21 \cdot 79/70} = 4.9$
4.9 è la deviazione standard associata sia a 21 che a 79

se usiamo le frequenze **assolute**
p =15 q=70-p=55
$s_r = \sqrt{15 \cdot 55/70} = 3.4$
3.4 è la deviazione standard associata sia a 15 che a 55

I limiti fiduciali della proporzione saranno come al solito:

$$LF = p \pm t \cdot s_p$$

Considerando ora solo le frequenze relative, i limiti fiduciali della proporzione, con un livello di α=0.05 (per cui t=1.96) saranno:

$$LF = 0.21 \pm 1.96 \times 0.049$$

per cui riteniamo che la vera proporzione della popolazione sia compresa, con probabilità del 95%, tra 0.10 e 0.32.

Sempre in base alla deviazione standard della proporzione possiamo ricavare il numero di osservazioni sufficiente ad ottenere una proporzione di una certa rappresentatività. Il problema ricalca quello della grandezza del campione esaminato nel capitolo riguardante la variabilità delle medie.

Quando la distribuzione non sia approssimabile a quella normale (quando **np**<=5), i limiti fiduciali della proporzione sono asimmetrici ed il loro calcolo è piuttosto complesso, per cui occorre riferirsi ai programmi statistici (pochi in verità) che offrono questo dato.

Distribuzione di Poisson

Ci può interessare anche il solo conteggio degli oggetti o dei fenomeni, senza distinzioni di tipo. Ad esempio il numero di veicoli che passano in una certa strada, senza considerare le marche di auto, o il numero di cellule presenti in un certo tessuto, ecc. Il contenitore degli oggetti o dei fenomeni può essere un intervallo di tempo o di spazio, o di entrambi. In genere si parla di

- **frequenza** quando ci si riferisce ad intervallo di **tempo**
- **densità** quando ci si riferisce ad un intervallo di **spazio**

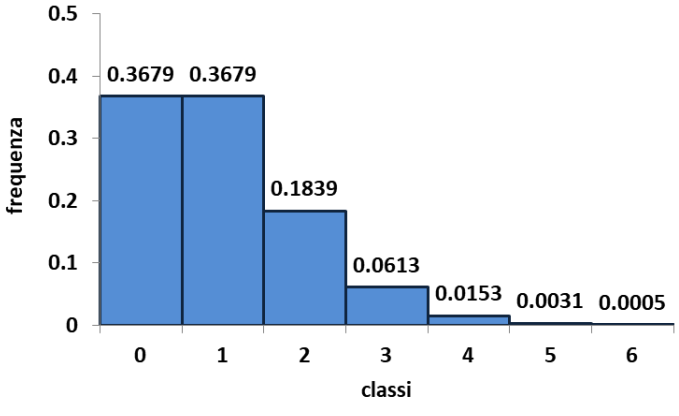
Per i nostri scopi, frequenza e densità sono equivalenti. Per semplicità parleremo di frequenza.

La frequenza di un fenomeno non è mai perfettamente costante, ma varia. E qui che entra in gioco la statistica. Un modello di variabilità è rappresentato dalla distribuzione di Poisson. Secondo tale distribuzione, la probabilità di trovare **i** oggetti in una determinata frazione di spazio o tempo è data da:

$$P_{i,m} = \frac{m^i}{i!} e^{-m}$$

in cui **e** è la base dei logaritmi naturali, circa 2.7182, ed **m** è la frequenza media attesa, in condizioni di omogeneità (numero totale di oggetti diviso per il numero di frazioni di tempo o spazio). Se ad esempio un litro di terreno di coltura contiene 1000 cellule, suddividendolo in 1000 frazioni di 1 ml, in condizioni di perfetta omogeneità, ci aspetteremo di trovare **m**=1000/1000=1 cellula in ciascuna frazione. Ma così non sarà sempre: pur se raramente, ci saranno frazioni che conterranno 0, 1, 2, 3, 4, ecc. cellule, con probabilità che molto spesso rispecchieranno la distribuzione di Poisson. Applicando la formula della distribuzione all'esempio, per valori di **i** da 0 a 6, otteniamo le seguenti probabilità:

i	P
0	0.3679
1	0.3679
2	0.1839
3	0.0613
4	0.0153
5	0.0031
6	0.0005
...	...
	totale 1



Come già detto la definizione di una distribuzione teorica è essenziale per fare previsioni e/o per valutare se esistano fattori che alterino la distribuzione. Se ad esempio le cellule esprimessero un recettore che le inducesse ad aggregare in gruppi di 4 o 5, potremmo trovare un aumento delle frazioni con 0, 4, e 5 cellule, e parallelamente una diminuzione delle frequenze delle frazioni con 1, 2, e 3 cellule. L'ipotesi che la distribuzione osservata differisca da quella attesa può essere valutata mediante test χ^2. Il test ci direbbe quindi se le cellule tendono significativamente ad aggregare.

Perché il modello di Poisson sia applicabile occorre che non si verifichi saturazione. Ad es., in un campo microscopico non possiamo trovare più di tante cellule. Come in un autobus più di tante persone non possono entrare. Se si verifica una situazione di saturazione la distribuzione di Poisson non può essere applicata.

Assortimenti

Il calcolo delle frequenze attese di eventi costituiti dall'assortimento di **r** elementi estratti da un insieme definito di **n** elementi di pende dal fatto che: (1) l'ordine degli elementi faccia o non faccia differenza e (2) gli elementi estratti siano o non siano essere ripetuti.
Queste due condizioni si combinano tra loro producendo 4 diverse tipologie:

disposizioni con elementi ripetuti
- l'ordine degli oggetti fa differenza? sì
- ci possono essere ripetizioni? sì

il loro numero è dato da: n^r
es., quante sono le possibili
... colonne di r=13 segni al totocalcio, essendo n=3 i segni (1,2,x)? 3^{13}
... parole di r=6 lettere, essendo n=21 le lettere dell'alfabeto? 21^6
... sequenze di r=10 nucleotidi, essendo n=4 i nucleotidi? 4^{10}

combinazioni con elementi ripetuti
- l'ordine degli oggetti fa differenza? no
- ci possono essere ripetizioni? sì

il loro numero è dato da: $\frac{(n + r - 1)!}{r!(n - 1)!}$
es., quante diverse compere di 4 articoli qualsiasi si possono fare in un negozio che ha un campionario di 100 articoli? - in realtà è un problema di scarsissima utilità
$\frac{(100 + 4 - 1)!}{4!(100 - 1)!}$

disposizioni con elementi non ripetuti
- l'ordine degli oggetti fa differenza? sì
- ci possono essere ripetizioni? no

il loro numero è dato da: $\frac{n!}{(n - r)!}$
es., quanti diversi ordini di arrivo di r=5 atleti si possono avere in una gara con n=20 atleti?
$\frac{20!}{(20-5)!}$

Nota: se r = n si parla di **permutazioni**, nel qual caso la formula si riduce a n!

Le permutazioni sono molto utilizzate in statistica perché se effettuate in gran numero (anche > 1000 permutazioni, quando possibile) consentono di calcolare le distribuzioni di statistiche (ad es. quella del t) rimescolando i dati, anziché usare le distribuzioni teoriche. Ad es., per 2 campioni con 5 dati ciascuno - in totale 10 dati - possiamo ottenere ben oltre tre milioni di permutazioni: (10)! = 3628800. Dividiamo quindi i 10 dati ottenuti da ciascuna permutazione in due gruppi di 5 dati - come fossero due campioni, e facciamo un test t. Ovviamente i t ottenuti con i dati rimescolati sono casuali, e quindi la loro distribuzione ci permette di stabilire quali valori di t si verificano per caso meno di 5 volte su 100, o meno di 1 volta su cento, ecc. In base a queste soglie possiamo valutare la significatività del test compiuto sui campioni reali, in un modo più consistente con la distribuzione dei dati in esame.

combinazioni con elementi non ripetuti
- l'ordine degli oggetti fa differenza? no
- ci possono essere ripetizioni? no

il loro numero è dato da: $\binom{n}{r} = \frac{n!}{r!(n - r)!}$
es., quante diverse mani di r=5 carte si possono fare giocando a poker con n=32 carte?
$\binom{n}{r} = \frac{32!}{5!(32 - 5)!}$

Probabilità, verosimiglianza e teorema di Bayes

Se A e B sono due eventi **mutualmente esclusivi**, la probabilità che si verifichi l'uno o l'altro è data dalla somma delle singole probabilità:

$p(A \text{ o } B) = p(A) + p(B)$

Due eventi mutualmente esclusivi possono essere anche **esaustivi**. In tal caso:

$p(A \text{ o } B) = p(A) + p(B) = 1$

Esempio:
- I semi di cuori e di quadri delle carte sono mutualmente esclusivi ma non esaustivi (mancano picche e fiori)
- Maschio e femmina sono mutualmente esclusivi ed esaustivi

Se un evento non influenza la probabilità che se ne verifichi un altro, allora gli eventi si dicono **indipendenti**. La probabilità che si verifichino entrambi, simultaneamente o in successione ma restando sempre indipendenti, è quindi data dal prodotto delle rispettive probabilità individuali **a priori**:

$p(A \text{ e } B) = p(A) \times p(B)$

Se invece gli eventi non sono indipendenti, nel senso che il fatto che si sia verificato l'uno influenza la probabilità che si verifichi l'altro, allora si parla di **probabilità condizionata** o **congiunta** o **a posteriori**, espressa mediante la formula:

$p(A \text{ e } B) = p(A) \times p(B/A)$

$p(B \text{ e } A) = p(B) \times p(A/B)$

ove p(A/B) sta per la probabilità di A condizionata dal fatto che si è verificato B

inversamente per p(B/A).

Nota re che p(B/A) è diverso da p(A/B). Le due espressioni non sono complementari né esiste tra loro alcuna precisa relazione definibile a priori.

Ad esempio, si sa che la probabilità di estrarre dal mazzo di 40 carte una carta rossa e che sia anche il re di quadri (K♦) si riferisce ad eventi tra loro non indipendenti, in quanto lo stesso K♦ è una carta rossa. Allora la probabilità di estrarre dal mazzo di 40 carte una carta rossa che sia anche il K♦ è data da:
$p(\text{carta rossa e } K♦) = p(\text{carta rossa}) \times p(K♦/\text{carta rossa}) = 20/40 \times 1/20 = 1/40$

Il risultato è apparentemente sciocco, in quanto sappiamo già che nel mazzo composto da 40 carte c'è un solo K♦. Ma il metodo è utile in altre situazioni di cui non abbiamo conoscenze dirette.

Esempi di eventi **indipendenti**	Esempi di eventi **dipendenti**
Regina - Carta di picche	Occhi chiari - Capelli chiari
Domenica - Giorno del mese dispari	Carta di cuori - Carta rossa
Malato di raffreddore - Calvo	Pioggia - Giornata nuvolosa

E' sempre un vantaggio potersi basare su probabilità a posteriori, anziché su probabilità a priori. Consideriamo l'ultimo esempio. Se si sa in anticipo che la giornata è di pieno sole, senza neppure una nuvola, è assai più facile prevedere che non piova. Come dire:

p(Pioggia/Nessuna nuvola)=0.

Diversamente, non sapendo affatto se ci sono o non ci sono nuvole, la probabilità che piova dipenderà unicamente dalla frequenza media di giornate piovose nell'arco dell'anno o della stagione. Tale probabilità a priori è ovviamente meno attendibile della probabilità a posteriori basata sulla presenza di nuvole in cielo.

Nel caso del mazzo di carte, del tiro di dadi, ecc. è facile calcolare le probabilità indipendenti e congiunte in quanto conosciamo perfettamente la popolazione. In altre situazioni le probabilità sono da determinare empiricamente in base alle frequenze dei vari fenomeni. In termini frequentisti, la probabilità di osservare un evento è pari alla frequenza con cui quell'evento si manifesta in una serie molto lunga, teoricamente infinita di prove.

A questo punto, entriamo nella sfera delle applicazioni concrete. Supponiamo le condizioni di essere

- affetto (M+) o non affetto (M-) da una certa malattia
- positivo (T+) o negativo (T-) ad un certo test o sintomo clinico specifico per quella malattia

Poiché malattia e test non sono indipendenti, la probabilità di essere malato e (vero) positivo al test si può esprimere come:

$p(M+ \text{ e } T+) = p(M+) \times p(T+/M+)$

Comunque lo stesso test potrebbe risultare positivo (falso positivo) anche su una piccola frazione di soggetti sani, per cui possiamo anche indicare la probabilità di essere sano e (falso) positivo al test:

$p(M- \text{ e } T+) = p(M-) \times p(T+/M-)$

Ricordiamo: p(T+/M+) e p(T+/M-) non sono uguali né complementari. Appartengono a diverse popolazioni. Quella dei malati e quella dei sani! E tuttavia la condizione di positività al test compare in entrambe le espressioni, per cui dobbiamo tener conto di tutte e due. A questo punto è legittima la vera domanda: **quale è la probabilità che un soggetto positivo al test sia effettivamente malato p(M+/T+) ?**

Un metodo assolutamente impraticabile sarebbe quello di sottoporre un gran numero di soggetti, sani e malati, allo stesso test e poi vedere quanti dei soggetti positivi sono realmente malati. Ma il metodo migliore è quello di ponderare le due formule di sopra che si riferiscono a tutte le persone positive al test. In questo modo **mescoliamo** probabilità ottenute da diverse popolazioni, quella dei malati e quella dei sani. Per non equivocare oltre sul termine di probabilità, la statistica distingue due termini di:

- **Probabilità** vera e propria quando ci si riferisce a probabilità omogenee relative ad un'unica distribuzione/popolazione, valutabili in base a modelli teorici di distribuzione o alle frequenze rilevate empiricamente in una serie estremamente lunga di prove.

- **Verosimiglianza** quando ci si riferisce a probabilità attinte da diverse distribuzioni/popolazioni, valutabili attraverso metodi di ponderazione.

In inglese le cose vanno meglio che in italiano in quanto in inglese verosimile si dice *likely* che nel linguaggio comune vuol dire probabile; e verosimiglianza si dice **likelihood**, che vuol dire anche probabilità. In italiano verosimile e verosimiglianza sono più usati come sinonimi di credibile e credibilità, più che di probabile e probabilità. Il concetto della verosimiglianza è generalmente accettato dalla statistica formale, anche se esistono posizioni più o meno favorevoli all'allargamento del suo uso a situazioni soggettive.

La soluzione al problema ci viene data dal teorema di Bayes. La verosimiglianza viene ponderata ponendo al numeratore la probabilità di essere malato e poi positivo al test essendo malato, ed al denominatore la sommatoria di tutte le probabilità di malattia e non malattia assieme alla positività per il test per le rispettive condizioni:

$$p(M+/T+) = \frac{p(M+)\cdot p(T+/M+)}{\left[p(M+)\cdot p(T+/M+)\right] + \left[p(M-)\cdot p(T+/M-)\right]}$$

$$= \frac{\text{frequenza delle persone malate che sono al tempo stesso positive al test}}{\text{somma delle frequenze di tutte le persone malate e non malate che sono al tempo stesso positive al test}}$$

in cui

p(M+)	Frequenza dei malati nella popolazione o prevalenza della malattia (beninteso della malattia M).
p(M-)	Frequenza dei sani (non malati di quella malattia M), complementare a p(M+).
p(T+/M+)	Frequenza di test positivi su malati = test veri positivi. E' detto anche *sensibilità* o *pote*nza del test.
p(T+/M-)	Frequenza di test positivi su sani = test falsi positivi. E' l'errore di I tipo. Il suo complementare (veri negativi) è detto anche *specificità* del test.

Come promemoria, ricordiamo che le probabilità condizionate riguardano la stessa distribuzione quando hanno lo stesso 'denominatore'. Pertanto:

- p(T+/M+) e p(T+/M-)
  veri positivi e falsi positivi, **non appartengono** alla stessa distribuzione,

  mentre invece

- p(T+/M+) e p(T-/M+)
  veri positivi e falsi negativi, **appartengono** alla stessa distribuzione, così come

- p(T+/M-) e p(T-/M-)
  falsi positivi e veri negativi, **appartengono** alla stessa distribuzione.

		Risultato del test		
		Positivo T+	Negativo T-	
Condizione	Sano M-	falso-positivo T+/M-	vero-negativo T-/M- *specificità*	**stessa distribuzione** ⇨complementari probabilità totale = 1
	Malato M+	vero-positivo T+/M+ *sensibilità*	falso-negativo T-/M+	**stessa distribuzione** ⇨complementari probabilità totale = 1
		⇩ non complementari: elementi di **diverse distribuzioni** verosimiglianza totale = 1	⇩ non complementari: elementi di **diverse distribuzioni** verosimiglianza totale = 1	

Nel caso in cui il significato del sintomo (lo stesso sintomo) non fosse semplicemente malato o sano, ma indicasse **n** diversi gradi o tipi della malattia, potremmo estendere la formula a tutte le diverse condizioni, mettendo al posto di M+ e M- una M_i in cui l'indice i rappresenta il grado/tipo i^{mo} della malattia

$$p(M_i / T+) = \frac{p(M_i) \cdot p(T+/M_i)}{\sum_{i=0}^{n} \left[p(M_i) \cdot p(T+/M_i) \right]}$$

in cui

$p(M_i)$	Frequenza della malattia di grado/tipo i^{mo} M_0 è la condizione di malattia di grado zero = condizione di salute
$p(T+/M_i)$	Frequenza di test positivi su soggetto con malattia di grado/tipo i^{mo} Ovviamente $p(T+/M_0)$ è la frequenza di test positivi su soggetti sani

Supponiamo ad esempio di individuare tre gradi di una certa malattia, oltre alla condizione di salute. Dovremo in tal caso conoscere i valori di:

$p(M_0)$ frequenza dei sani (per quella malattia M, ma potrebbero avere altri mali)
$p(M_1)$ frequenza di soggetti con il grado 1 della malattia M
$p(M_2)$ frequenza di soggetti con il grado 2 della malattia M
$p(M_3)$ frequenza di soggetti con il grado 3 della malattia M

$p(T+/M_0)$ frequenza del test positivo per i sani

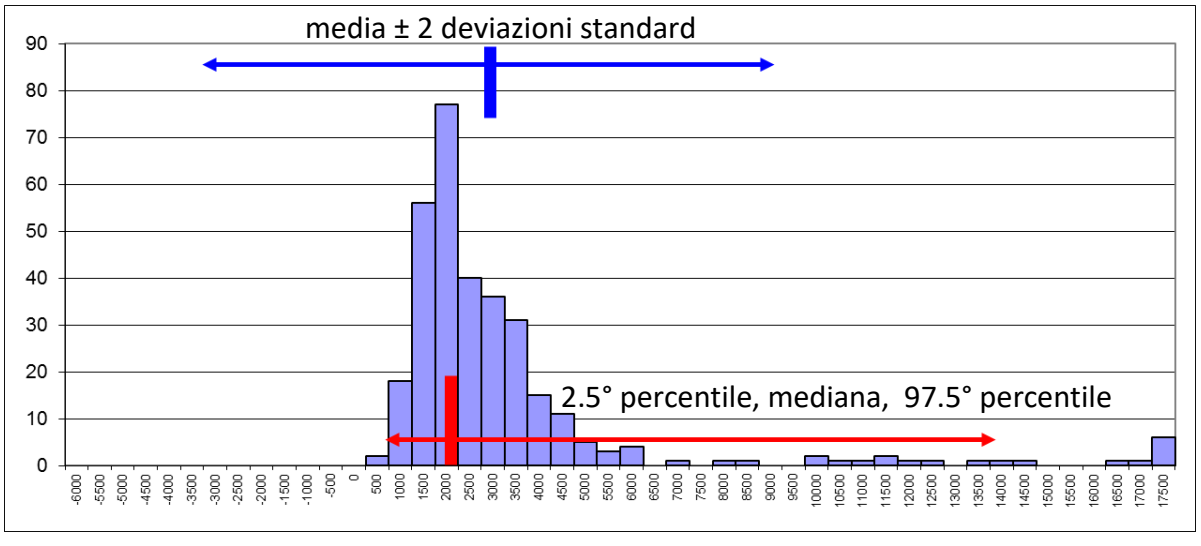
$p(T+/M_1)$ frequenza del test positivo per i soggetti con il grado 1 della malattia M
$p(T+/M_2)$ frequenza del test positivo per i soggetti con il grado 2 della malattia M
$p(T+/M_3)$ frequenza del test positivo per i soggetti con il grado 3 della malattia M

Applicheremo quindi il test 4 volte per conoscere le varie probabilità che un soggetto positivo al test sia sano M_0, o malato M_1, o malato M_2, o malato M_3.

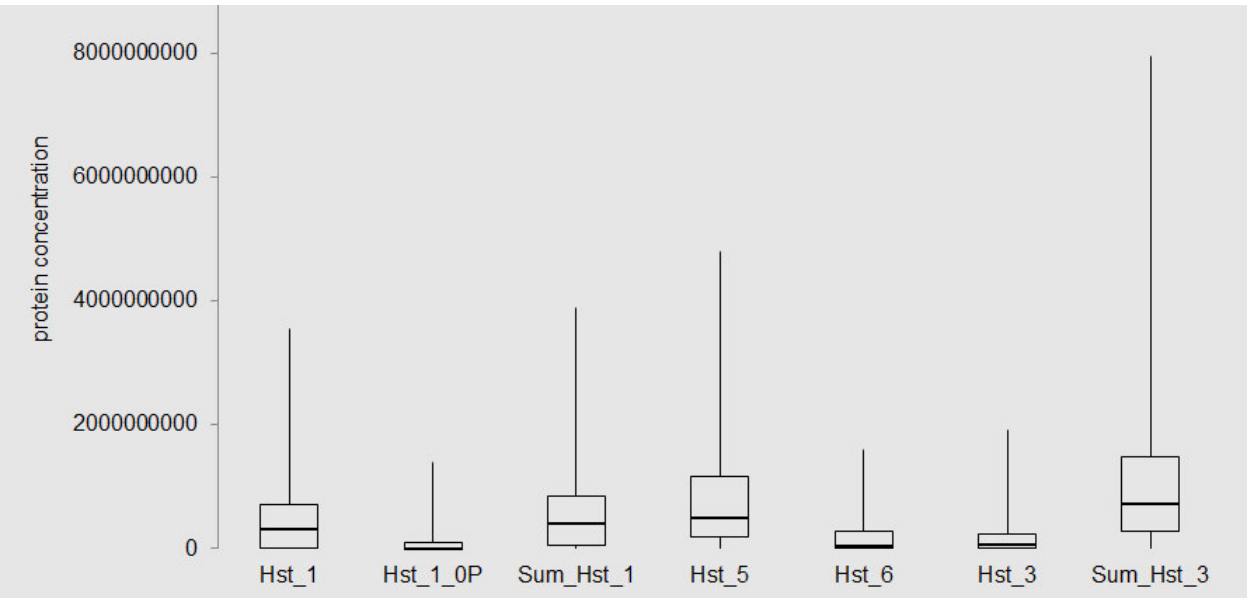
Test non-parametrici

Come abbiamo visto, il test t si può applicare a condizione che i dati siano distribuiti normalmente e i due campioni abbiano varianze uguali, anche se in realtà il test t è robusto nei confronti di distribuzioni non perfettamente normali, e la sua variante - il test di Welch - è applicabile a campioni con varianze dissimili. Conoscendo infatti la forma della distribuzione normale possiamo valutare i parametri di dispersione e con questi fare i confronti. Questa è la filosofia di tutti i test cosiddetti parametrici, in quanto media, errore standard, t, ecc. sono i parametri statistici (d'altra parte per chi lavora in laboratorio i parametri sono i parametri dell'esperimento, ovvero i trattamenti, le dosi, ecc.). Per stabilire se i dati sono distribuiti in modo normale o non-normale possiamo considerare i parametri globali di skewness e kurtosis e/o fare un Goodness-of-fit test o un test di simmetria - ne esistono diversi. Se la distribuzione è significativamente diversa da quella normale media e deviazione standard non hanno alcun significato. Al loro posto dobbiamo considerare la mediana ed i percentili della distribuzione empirica, cioè quella degli stessi dati.

Se ad esempio consideriamo l'intervallo media ± 2 deviazioni standard del peso di neonati otteniamo valori negativi (v. frecce blu nel grafico). Un peso negativo equivarrebbe a dire che i neonati più leggeri volano su come palloncini. La forma dell'istogramma dice chiaramente che la distribuzione è fortemente asimmetrica e quindi non è normale. Per abbracciare il 95% dei dati, al posto dell'intervallo media ± 2 deviazioni standard, dobbiamo quindi considerare il 2.5° ed il 97.5° percentile dei dati (v. frecce rosse nel grafico). In Excel si può usare la vecchi formula: PERCENTILE(...) oppure una delle formule più recenti e più accurate che considerano se il valore del percentile va incluso e escluso dall'intervallo, rispettivamente PERCENTILE.INC(...) PERCENTILE.EXC(...).



Nel grafico box-wisker sono rappresentati la media (la linea orizzontale più spessa), il range interquartile (il box) ed il range totale (min-max) delle concentrazioni di 7 proteine. La asimmetria dei dati rispetto alla media è evidente anche considerando il solo range interquartile.



Quando ci si accorge che i dati non sono distribuiti come richiesto dal test (parametrico), si possono fare tre cose:

1. abbandonare lo studio e dedicarsi ad altro (facile)
2. trasformare i dati in modo tale da aggiustarne la distribuzione (piuttosto difficile e non sempre possibile)
3. applicare un appropriato test non-parametrico (la scelta migliore).

Dati che non riflettono una distribuzione normale sono spesso quelli dei punteggi, votazioni, scores, ecc. utilizzati convenzionalmente da un osservatore (medico, psicologo, insegnante, giudice di gara, sommelier, ecc.) per valutare fenomeni complessi quali i comportamenti, l'intelligenza, la capacità di memoria, il rendimento a scuola, la produttività nel lavoro, la prestazione atletica, ecc. In tutti questi casi la scala non è riferita a grandezze fisiche ma a livelli qualitativi trasformati numericamente in base a qualche convenzione, ad es., i voti della scuola. In campo diagnostico queste scale sono molto accurate e talvolta persino oggetto di brevetto per cui il loro uso in forma ufficiale può essere soggetto all'acquisto di licenze.

Gran parte dei metodi non-parametrici si basano solo sul valore ordinale dei dati, cioè sulla posizione in graduatoria, trascurando il valore reale qualora un valore reale esista. In tal modo la statistica non-parametrica si libera dei condizionamenti della distribuzione dei dati. L'analogo della media in campo non-parametrico è la mediana. Mentre la media è facilmente modificata anche da pochi valori estremamente piccoli o estremamente grandi, al contrario, la mediana è del tutto insensibile a fluttuazioni dei valori estremi. Pertanto la mediana ed i test non-parametrici sono da preferire anche quando la distribuzione complessiva dei dati non fosse significativamente diversa da quella normale ma vi fossero dati con valori estremamente piccoli o grandi che sospettiamo siano dovuti ad errori non rilevati nelle procedure o nella acquisizione dei dati, e che potrebbero inficiare media, deviazione standard, ecc.

Esiste tutta una serie di test non-parametrici (Siegel, Statistica Non-Parametrica, McGraw-Hill editore, è il testo di riferimento) che ripropongono più o meno le stesse ipotesi già viste con i test parametrici.

In definitiva, alla domanda: quando dobbiamo applicare i test non-parametrici? si risponde dicendo che la SNP deve essere applicata quando:

1) i dati non si conformano al tipo di distribuzione richiesto dalle procedure parametriche
   oppure
2) i dati si riferiscono a scale ordinali, come quelle associate a punteggi, giudizi, graduatorie, ecc.
   oppure
3) tra i dati ve ne sono alcuni molto piccoli o molto grandi da destare il sospetto che derivino da errori sperimentali ma che comunque non ci sentiamo di escludere.

Oltre ai casi raccomandati, i test non-parametrici possono essere applicati ogni volta che si voglia saggiare una ipotesi prescindendo dalla distribuzione dei dati. Certo è che se i campioni sono distribuiti normalmente e differiscono solo per la media, allora il test t è molto più efficiente dell'analogo test non-parametrico di Mann-Whitney. Tuttavia vi sono moltissime situazioni in cui la condizione di normalità è assunta passivamente solo in quanto non si può dimostrare il contrario. Ricordiamo che nei test di normalità (goodness-of-fit test ed altri) la condizione di normalità è data come ipotesi zero, e pertanto essa è da conservare sino a che la probabilità a suo favore non scende sotto il 5%. In altre parole il test esclude che la distribuzione sia normale solo quando ha un ampie evidenze, ma in mancanza di evidenze non si sente di escluderlo. Tutte le volte che consideriamo piccoli campioni, i test basati sulla skewness, kurtosis ed lo stesso Goodness-of-fit sono poco o niente efficienti. In tal caso il danno più grave non è dato tanto dai falsi-positivi (ritenere che la distribuzione sia non-normale mentre lo è) ma dai falsi-negativi (ritenere che la distribuzione sia normale mentre non lo è). Per cui, nei casi in cui la normalità sia solo presunta, non è affatto male considerare anche il responso dei test non-parametrici assieme a quelli parametrici.

In genere, la SNP converte i dati ponderali nel loro *rango:* brutto termine italiano che traduce l'inglese rank. L'inglese rank significa infatti anche posizione in graduatoria/classifica/ordine crescente, mentre l'italiano rango ha più significato di gerarchia in senso sociale/militare.

dati	41	9	84	1	67	123	81
convertiti in rango	3	2	6	1	4	7	5

Vediamo cosa succede se:

a) mettiamo 1230 al posto di 123. I ranghi non cambiano.

dati	41	9	84	1	67	**1230**	81
ranghi	3	2	6	1	4	7	5

b) mettiamo 12.3 al posto di 123. Alcuni ranghi cambiano di una posizione.

dati	41	9	84	1	67	**12.3**	81
ranghi	4	2	7	1	5	3	6

c) mettiamo 0 al posto di 123. Anche in questo caso i ranghi cambiano di poco.

dati	41	9	84	1	67	**0**	81
ranghi	4	3	7	2	5	1	6

d) eleviamo tutti i dati al quadrato. I ranghi non cambiano.

	41^2	9^2	84^2	1^2	67^2	123^2	81^2
dati	1681	81	7056	1	4489	15129	6561
ranghi	3	2	6	1	4	7	5

Questi esempi dimostrano come i ranghi siano molto robusti nei confronti di variazioni anche notevoli dei dati. Ma l'aspetto più eclatante è il fatto che i ranghi non cambiano affatto se i dati vengono trasformati in modo lineare o non-lineare (es., esponenziale o logaritmico, purché la trasformazione sia monotonica: se esistono dati negativi i quadrati diventano positivi distruggendo l'ordine dei ranghi, ed i logaritmi non si possono calcolare).

Nota. Quando esistono valori uguali, a ciascuno di essi si attribuisce la media dei ranghi che spetterebbero agli stessi valori se questi fossero diversi.

Ad esempio

dati	31	63	41	85	33	51	85	79	85	27	68
ranghi	2	6	4	**10**	3	5	**10**	8	**10**	1	7

La **somma dei ranghi** di una serie di **n** dati corrisponde alla somma dei primi n interi, data dalla formula:

$$\sum r = \frac{n(n+1)}{2}$$

Pertanto, il **rango medio** di una serie di **n** dati sarà:

$$r_{medio} = \frac{n(n+1)}{2}\frac{1}{n} = \frac{n+1}{2}$$

Di seguito sono illustrati alcuni test non-parametrici. L'esposizione è condotta in maniera di esercizio. I test non parametrici più utilizzati sono:

test non-parametrico	**analogo test parametrico**
correlazione di Kendall	correlazione (di Pearson)
correlazione di Spearman	correlazione (di Pearson)
test di Wilcoxon / Mann-Whitney per il confronto tra due campioni indipendenti o test della somma dei ranghi	test t per campioni indipendenti
test di Wilcoxon per il confronto tra due campioni appaiati o test dei segni	test t per campioni appaiati
test di Kolmogorov-Smirnov per il confronto della posizione e forma delle distribuzioni di due campioni	non esiste equivalente, sarebbe un test complessivo per la differenza tra due medie e tra due varianze
test di Kruskal-Wallis per il confronto di più gruppi	ANOVA
test di Friedman per il confronto di più trattamenti applicati agli stessi soggetti	ANOVA per misure ripetute
ecc.	

In ultimo occorre aggiungere che la SNP non comporta sempre la trasformazione dei dati in rango. Anche se non è stato detto in precedenza, il test del χ^2 può anch'esso definirsi un test non-parametrico per dati di frequenze. Se infatti consideriamo la semplice tabella 2×2, notiamo che l'ipotesi nulla non fa riferimento a nessun tipo di distribuzione in particolare, ma solo ai rapporti dei dati all'interno della stessa tabella ed al contrasto risultante tra le frequenze osservate e quelle attese.

Correlazione di Kendall

I dati della variabile Y vengono riscritti ordinati a seconda del valore crescente della variabile X. Poi si considera la successione dei valori della variabile Y. Per ogni valore di Y si conta il numero di successivi valori maggiori ($S_{maggiori}$) e il numero di successivi valori minori (S_{minori}). Se tra i successivi si incontra un valore uguale si incrementa di 0.5 sia il totale $S_{maggiori}$ sia il totale S_{minori}.
La somma di $S_{maggiori} + S_{minori}$ deve dare n(n-1)/2, altrimenti i conti sono sbagliati.
Si calcola la differenza $S_{diff} = S_{maggiori} - S_{minori}$.
Il coefficiente di correlazione non-parametrico τ (tàu di Kendall) è compreso tra +1 e -1 ed è dato da:

$$\tau = \frac{S_{diff}}{\frac{n(n-1)}{2}}$$

Si confronta il valore di S_{diff} con i valori critici riportati in tabella per verificare la significatività della correlazione.

Quando Y varia in modo del tutto irregolare rispetto a X avremo $S_{maggiori} = S_{minori,}$, per cui S_{diff} sarà **nullo**. $\tau = 0$
Quando Y cresce costantemente al crescere di X, avremo $S_{minori}=0$, per cui S_{diff} sarà **massimo** = n(n-1)/2 $\tau = 1$
Viceversa, quando Y decresce costantemente al crescere di X, avremo $S_{maggiori}=0$, per cui S_{diff} sarà **minimo =** S_{minori} = -n(n-1)/2 $\tau = -1$

Esempio:

dati iniziali

X	Y
8	11
13	14
6.6	5.9
3	4.5
9	10
11	15
5	2

X	Y	considerando la sequenza dei valori Y dall'alto verso il basso ordinata in funzione dei valori crescenti di X, calcolare per ciascun valore di Y... da quanti valori maggiori è seguito	da quanti valori minori è seguito
3	4.5	5	1
5	2	5	0
6.6	5.9	4	0
8	11	2	1
9	10	2	0
11	15	0	1
13	14	-	-
		totale $S_{maggiori}=18$	totale $S_{minori}=3$

Verifica calcolo dei ranghi e delle somme:
$n(n-1)/2=21$
$S_{maggiori} + S_{minori} = 18+3 = 21$

$S_{diff}= 18-3 = 15$

$\tau=15/21=0.71$

Correlazione di Spearman

I dati vengono trasformati nei loro ranghi, con ordinamenti separati per le due variabili X e Y, ma senza alterare l'ordine. Ai ranghi si applica la normale analisi della correlazione già vista. Notare che le due devianze sono uguali, poiché i ranghi di X e Y hanno uguali valori. Il coefficiente di correlazione non-parametrico di Spearman oscilla anch'esso tra +1 e -1, anche se non corrisponde necessariamente al τ di Kendall.

Esempio:

dati iniziali

X	Y
8	11
13	14
6.6	5.9
3	4.5
9	10
11	15
5	2

ranghi

X	Y
4	5
7	6
3	3
1	2
5	4
6	7
2	1

Codevianza = 25

$\text{Devianza}_X = 28$, $\text{devianza}_Y = 28$
(che le devianze siano uguali è dovuto al fatto che X e Y hanno gli stessi ranghi)

$$r = \frac{25}{\sqrt{28 \cdot 28}} = \frac{25}{28} = 0.89$$

Test di Wilcoxon/Mann-Whitney per 2 campioni indipendenti o test della somma dei ranghi

E' l'analogo non-parametrico del test t di Student per campioni indipendenti.
Ordinare i dati in rango, comprendendo nello stesso ordinamento i due campioni. Se i campioni non sono bilanciati, diciamo che il campione più piccolo ha numerosità **n** e quello più grande ha numerosità **m**.
La somma dei ranghi dei due campioni è

$$\Sigma_r = \frac{(n+m)(n+m+1)}{2}$$

l'ipotesi nulla di assortimento casuale dei valori nei due gruppi prevede che entrambe i gruppi abbiano lo stesso rango medio, e cioè

$$r_m = \frac{(n+m)(n+m+1)}{2}\frac{1}{(n+m)} = \frac{n+m+1}{2}$$

Moltiplicando il rango medio per la numerosità del campione più piccolo si ricava la somma dei ranghi del campione più piccolo, attesa per la condizione di non differenza tra i due campioni (m_T). T è la statistica test. L'effettiva somma dei ranghi del campione più piccolo (**T**) potrà differire da quella attesa, per cui il valore di T è già il risultato del test. Infatti quanto più la somma dei ranghi del campione più piccolo si discosta da quella attesa, tanto più diminuisce la probabilità che i due campioni siano assortimenti casuali di ranghi e la tabella o il programma ci dirà se questa supererà le soglie critiche di significatività. Per ogni livello di significatività, in tabella sono tabulati due valori estremi: uno molto piccolo ed uno molto grande. La significatività si raggiungerà se **T** sarà minore del valore più piccolo o maggiore del valore più grande in tabella. Si potrebbe considerare la somma dei ranghi del campione più grande, ma è lo stesso, in quanto le due somme sono mutualmente vincolate.

Vi è una versione differente del test nota come test U di Mann-Whitney. I risultati sono comunque gli stessi.

Per verificare che abbiamo fatto bene i conti verifichiamo che le somme dei ranghi dei due campioni diano insieme il valore totale.

Se i campioni sono sufficientemente grandi (m>8) è possibile fare a meno della tabella e sfruttare il fatto che **T** tende a distribuirsi normalmente attorno al valore atteso dall'ipotesi nulla (m_T), in cui

$$m_T = \text{numerosità} \times \text{rango medio} = n \cdot \frac{n+m+1}{2}$$

con deviazione standard pari a

$$s_T = \sqrt{\frac{n \cdot m \cdot (n+m+1)}{12}}$$

per cui si può calcolare la deviata standardizzata di T rispetto a m_T

$$z_T = \frac{T - m_T}{s_T} = \frac{T - \left(n \cdot \frac{n + m + 1}{2} \right)}{\sqrt{\frac{n \cdot m \cdot (n + m + 1)}{12}}}$$

Se z_T supera in valore assoluto il solito 1.96 i due gruppi possono considerarsi significativamente differenti. La suddetta formula andrebbe perfezionata (al numeratore) con la correzione per la continuità e (al denominatore) con la correzione della deviazione standard per la presenza di eventuali valori sovrapposti. Per questo si rimanda a manuali più completi.

Esempio:

dati iniziali

A	B
7.2	1
10	4
12	5
15	6
20	8
23	8
41	8
60	10
	11
	12
	15
	21
n=8	m=12

ranghi

A	B
5	1
9.5	2
12.5	3
14.5	4
16	7
18	7
19	7
20	9.5
	11
	12.5
	14.5
	17
T_A=114.5	T_B=95.5

verifica dei calcoli:

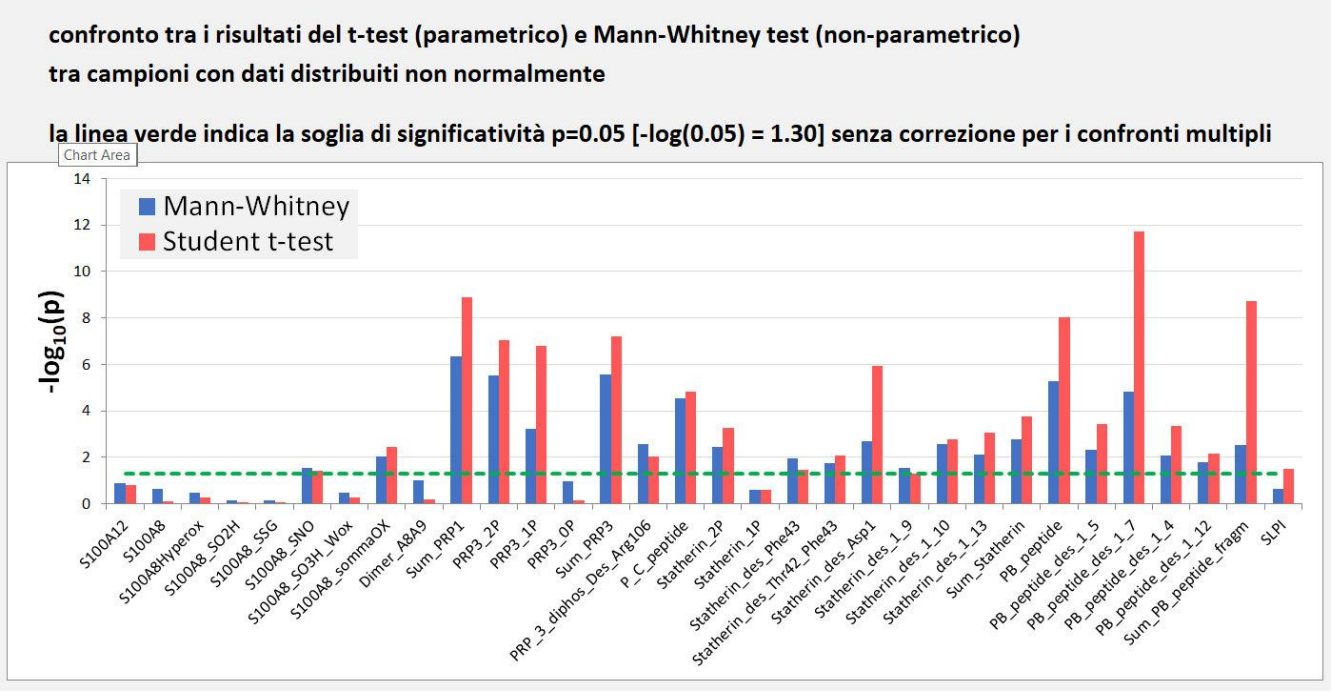
$$114.5 + 95.5 = 210$$

$$\frac{(n + m)(n + m + 1)}{2} = \frac{(8 + 12)(8 + 12 + 1)}{2} = \frac{20 \cdot 21}{2} = 210$$

verifica OK

$$z_T = \frac{114.5 - \left(8 \cdot \frac{8 + 12 + 1}{2} \right)}{\sqrt{\frac{8 \cdot 12 \cdot (8 + 12 + 1)}{12}}} = \frac{30.5}{12.96} = 2.35$$

Dal grafico è possibile apprezzare le discrepanze tra il t-test ed il test Mann-Whitney. Sull'asse Y sono riportati i valori di $\log_{10}(p)$ anziché p per esigenze di scala. Il valore $\log_{10}(p) = 2$ sta per 0.01, 3 sta per 0.001, ecc. Da notare che i dati in questione non erano distribuiti normalmente. A conferma, le oscillazioni dei risultati del test Mann-Whitney appaiono più stabili rispetto a quelle del t-test.



Test di Wilcoxon per 2 campioni appaiati o test dei segni

E' l'analogo non-parametrico del test t di Student per campioni appaiati.

Calcolare le differenze tra i due campioni.
Assegnare i ranghi alle differenze trascurando i segni (cioè i ranghi ai valori assoluti delle differenze).
Ora applicare i segni delle differenze ai ranghi.
Calcolare la somma dei ranghi segnati (**W**). W è la statistica test. E' chiaro che se i valori dei due campioni sono casualmente assortiti, le differenze si annullano tra loro e così anche la somma dei ranghi segnati tende a zero. Viceversa, se vi è una variazione coerente tra i due campioni, le differenze tendono ad essere tutte positive o tutte negative, per cui la somma totale dei ranghi segnati tende a crescere sino al limite del valore assoluto massimo

$$W_{max} = \frac{n(n+1)}{2}$$

Se n>20, si può fare a meno della tabella e sfruttare il fatto che la distribuzione di W tende a diventare normale con media

$m_W=0$

e deviazione standard

$$s_W = \sqrt{\frac{n(n+1)(2n+1)}{6}}$$

Per cui si può proporre il calcolo della deviata standardizzata di W rispetto a m_W, cioé rispetto a 0.

$$z_W = \frac{W-0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

Se z_W supera in valore assoluto il solito 1.96 la differenza tra i due gruppi può considerarsi significativa.

Esempio:

prima	dopo	differenza	valore assoluto	rango	rango segnato
3	7	4	4	6	6
5	4	-1	1	2	-2
2	9	7	7	8	8
11	16	5	5	7	7
6	5	-1	1	2	-2
7	10	3	3	5	5
4	6	2	2	4	4
2	1	-1	1	2	-2
					W=24

$$z_W = \frac{24}{\sqrt{\frac{8(8+1)(16+1)}{6}}} = 1.68$$

Test di Kolmogorov–Smirnov (K-S) per il confronto di due campioni

Il test K-S a due campioni è uno dei metodi non parametrici più utili e generali per confrontare due campioni poiché è sensibile alle differenze sia nella posizione che nella forma delle distribuzioni dei due campioni. Il test di Kolmogorov-Smirnov si basa sulla massima differenza in valore assoluto tra le funzioni cumulative. Quando tale differenza è significativa, le due distribuzioni vengono considerate diverse.

Per effettuare il test K-S occorre dapprima trasformare le frequenze assolute in frequenze relative entro ogni campione. Successivamente occorre fare un confronto tra le classi delle frequenze cumulate per trovare la deviazione o differenza massima.

Esempio:
Consideriamo le frequenze del pH dell'acqua osservate in campioni raccolti all'ingresso ed all'uscita di un depuratore. All'ingresso sono stati raccolti 10 campioni e all'uscita 12 campioni.

pH	7	6.75	6.5	6.25	6	5.75	5.5	5.25

frequenze assolute dei 10 campioni

Ingresso	0	0	0	3	6	0	1	0
Uscita	1	4	4	1	1	1	0	0

frequenze relative

Ingresso	0	0	0	0.3	0.6	0	0.1	0
Uscita	0.083	0.333	0.333	0.083	0.083	0.083	0	0

frequenze cumulate

Ingresso	0	0	0	0.3	0.9	0.9	1	1
Uscita	0.083	0.4170	0.750	0.834	0.917	1	1	1
Differenza	0.083	0.4170	**0.750**	0.534	0.017	0.1	0	0

D è la differenza massima tra le classi di frequenza cumulata, in questo caso D = 0.750

Nel caso di un test ad una coda, in cui si conosce la direzione della differenza, si deve calcolare la deviazione massima D con il segno. Per un test a due code non è importante conoscere la direzione della differenza. Lo scarto massimo è quindi calcolato in valore assoluto.

La distribuzione di D (valutata in funzione delle numerosità dei 2 campioni, 10 e 12 in questo caso) fornirà la probabilità che i due campioni provengano dalla stessa popolazione, lasciando tuttavia aperta la questione se le due popolazioni differiscano per posizione, dipersione o skewness.

Il test di Kolmogorov-Smirnov può essere modificato per fungere da Goodness-of-fit test. In questo caso i campioni sono confrontati con la distribuzione normale.

Test di Kruskal-Wallis per il confronto di più gruppi

E' l'analogo non-parametrico dell'analisi della varianza.

Assegnare i ranghi ai dati, mettendo assieme tutti i dati dei diversi gruppi (come per Wilcoxon/Mann-Whitney).
Calcolare la somma dei ranghi e poi il rango medio di ciascun gruppo r_{mgr}.
Calcolare l'esattezza dei ranghi confrontando la somma delle somme dei gruppi con N(N+1)/2, ove N è il numero totale di dati.
Calcolare il rango medio generale di tutti i dati secondo la solita formula: r_{mgen}= (N+1)/2.
Calcolare la differenza tra il rango medio di ciascun gruppo (r_{mgr}) ed il rango medio generale atteso (r_{mgen}), elevare al quadrato e moltiplicare per la numerosità del gruppo (n_{gr}). Quindi fare la sommatoria:

$$D = \sum n_{gr} (r_{mgr} - r_{m\,gen})^2$$

Il procedimento ha forti analogie col calcolo della devianza tra gruppi per l'analisi della varianza.
Normalizzare D, calcolando la statistica:

$$H = \frac{D}{\frac{N(N+1)}{12}}$$

H è la statistica test. Se i gruppi sono abbastanza numerosi (con oltre 3 gruppi e 5 soggetti per gruppo) la distribuzione di H è approssimabile a quella del χ^2 con tanti gradi di libertà quanti sono i gruppi meno 1.

Esempio:

dati iniziali.

Gruppi				
A	B	C	D	E
10	26	40	24	35
14	30	34	11	29
18	27	23	17	37
20	12	36	21	13
22	15	33	19	28
16	31	38	32	39
		41		42
				43
n_A=6	n_B=6	n_C=7	n_D=6	n_E=8
N=6+6+7+6+8=33				

ranghi

A	B	C	D	E
1	16	30	15	25
5	20	24	2	19
9	17	14	8	27
11	3	26	12	4
13	6	23	10	18
7	21	28	22	29
		31		32
				33
Σr_A=46	Σr_B=83	Σr_C=176	Σr_D=69	Σr_E=187
r_{mA}=7.7	r_{mB}=13.8	r_{mC}=25	r_{mD}=11.5	r_{mE}=23.3

verifica dei calcoli

46+83+176+69+187=561
N(N+1)/2=561

verifica OK

$r_{m\ gen}$= (N+1)/2=(33+1)/2=17

$D=6(7.7 - 17)^2 + 6(13.8 - 17)^2 + 7(25 - 17)^2 + 6(11.5 - 17)^2 + 8(23.3 - 17)^2 = 1527.4$

$$H = \frac{1527.4}{\frac{33(33+1)}{12}} = 16.34$$

La soglia del χ^2 con 3 GDL, per P = 0.05, è 7.82. Quindi il risultato del test è significativo.

Test di Friedman per il confronto di più trattamenti applicati agli stessi soggetti

E' l'analogo non-parametrico dell'analisi della varianza cosiddetta a due vie.
Esempio. Un gruppo di n soggetti, sempre gli stessi, è sottoposto a k diversi trattamenti. Ad es., gruppi di operai posti a lavorare in ambienti con diverse condizioni di luce, rumore, spazio, ecc. di cui si vuole valutare il diverso effetto sul rendimento al lavoro; oppure un gruppo di pazienti che in diverse occasioni - fatti salvi i principi di eticità – ricevono terapie differenti di cui si vuole valutare la diversa efficacia, ecc.

n = numero dei soggetti
k = numero dei trattamenti

Creare una tabella con i soggetti nelle righe ed i trattamenti nelle colonne.
Assegnare i ranghi ai dati di ciascun soggetto attraverso i diversi trattamenti (cioè riga per riga).
Calcolare il rango medio di ciascun trattamento r_{mtrat} (operando quindi colonna per colonna).
Calcolare il rango medio atteso di ciascun trattamento nell'ipotesi in cui tutti i trattamenti siano equivalenti (ipotesi di indifferenza)

$$r_{m\ atteso} = \frac{(k+1)}{2}$$ (uguale per tutti i trattamenti)

Sommare quindi, per ciascun trattamento, il quadrato della differenza tra rango medio osservato e rango medio atteso, moltiplicato per il numero di soggetti:

$$S = \sum\left[n(r_{mtrat} - r_{matteso})^2 \right]$$

Normalizzare S con il rapporto:

$$\chi_r^2 = \frac{S}{\frac{nk(k+1)}{12}}$$

χ_r^2 è la statistica test per verificare la significatività delle differenze tra i campioni. Quando i dati sono abbastanza numerosi (> 9 soggetti e 3 trattamenti, ecc.), χ_r^2 tende a distribuirsi come χ^2 con un numero di gradi di libertà pari al numero di campioni meno 1.

Nota: così come il test di Kruskal-Wallis mostra una stretta analogia con l'ANOVA parametrica ad una via (o ad un criterio di classificazione), il test di Friedman è analogo all'ANOVA parametrica a due vie, tenendo presente tuttavia che Friedman valuta solo le differenze tra i campioni (per l'effetto dei trattamenti, in questo caso), e non si esprime su possibili differenze tra i soggetti, che invece l'ANOVA a due valuta.

Esempio:

dati iniziali.

	trattamenti			
soggetti	A	B	C	D
a	21	65	39	10
b	43	87	61	6
c	32	43	83	29
d	15	76	58	4
e	17	43	34	10
f	34	49	54	29

ranghi

	trattamenti			
soggetti	A	B	C	D
a	2	4	3	1
b	2	4	3	1
c	2	3	4	1
d	2	4	3	1
e	2	4	3	1
f	2	3	4	1
	r_{mA} =12/6 =2	r_{mB} =22/6 =3.667	r_{mC} =20/6 =3.333	r_{mD} =6/6 =1

$$r_{m\,atteso} = \frac{(4+1)}{2} = 2.5$$

$$S = 6\,(2-2.5)^2 + 6\,(3.667-2.5)^2 + 6\,(3.333-2.5)^2 + 6\,(1-2.5)^2 = 27.3$$

$$\chi^2_r = \frac{164}{\frac{6 \cdot 4(4+1)}{12}} = \frac{164}{10} = 16.4$$

Statistica multivariata

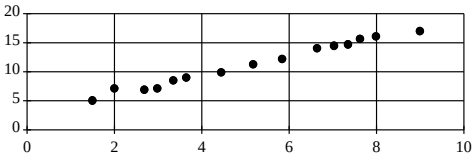
I metodi riportati in queste poche pagine rappresentano una brevissima panoramica ed i concetti di base dei metodi multivariati.

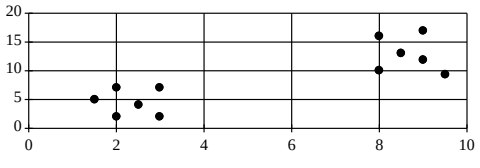
Partiamo da un argomento noto. Come sappiamo, la terra è sferica per cui è impossibile costruire una carta 2D della Terra che rispetti tutte le distanze tra i diversi punti. Tutte le carte geografiche si basano su un certo criterio di proiezione per rappresentare la superficie terrestre al meglio, a seconda delle esigenze. Ogni proiezione implica quindi una certa deformazione.

In statistica, ogni variabile rappresenta una dimensione. Quindi quando succede di dover analizzare fenomeni descritti da 3 o più variabili si pone il problema analogo di rappresentare e riassumere l'informazione su pochi assi di riferimento e grafici a due dimensioni, accettando un certo compromesso. Tecnicamente si parla di riduzione della dimensionalità. Questa è la premessa della statistica multivariata (SM). Il vantaggio è che questo consente di considerare i fenomeni (biologici, clinici, fisici, economici, sociali, psicologici ecc.) nella loro interezza, tenendo conto di molteplici caratteristiche o variabili. Per tale vocazione la SM è orientata a fornire rappresentazioni del reale più che a valutare test di ipotesi. Semmai i test servono a verificare se la rappresentazione ottenuta sia o non sia frutto del caso. Oltre tale premessa non esiste una precisa definizione di statistica multivariata. In queste poche pagine tenteremo di tenere l'orizzonte il più ampio possibile.

La possibilità di analizzare più variabili contemporaneamente non significa che sia sempre meglio utilizzare il massimo numero possibile di variabili. E' vero semmai il contrario. Se infatti alle variabili di interesse vengono aggiunte altre variabili di scarso o nessun interesse si abbassa l'efficacia dell'analisi. Esistono comunque procedure preliminari che aiutano a valutare l'importanza relativa delle diverse variabili ed a suggerire se sia bene mantenerle o escluderle dall'analisi. Questa operazione è chiamata 'feature selection'.

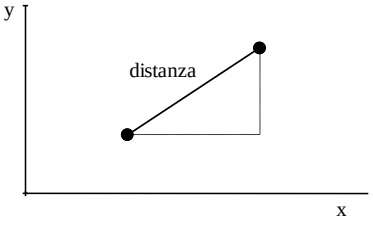
Gli esempi più semplici per introdurre le tecniche multivariate considerano solo due variabili, in modo da consentire di rappresentare e ragionare su grafici 2D. Con due sole variabili (tecniche bivariate) abbiamo già esaminato l'analisi della regressione e della correlazione che analizzano appunto la relazione tra variabili. Ma è giunto il momento di analizzare anche le relazioni tra i soggetti. Le **relazioni tra variabili** e **quelle tra soggetti** rappresentano infatti differenti prospettive della stessa struttura dei dati. Strano ma vero, normalmente quanto più è forte la relazione tra le variabili, tanto meno forte è la relazione tra i soggetti, e viceversa:





Il grafico a sinistra dimostra una forte correlazione tra le due variabili. I soggetti (punti del grafico) si piegano a questa relazione e non si intravedono gruppi (clusters) separati di soggetti. Viceversa il grafico a destra dimostra due cluster di soggetti, mentre la relazione tra le variabili è molto più debole. Possiamo dire quindi che variabili variabili poco correlate (indipendenti tra loro) mettono spesso in luce interessanti relazioni tra soggetti.

Se la relazione tra due variabili è valutabile mediante la correlazione, quella tra due soggetti può valutarsi come distanza:



Le due variabili sono riportate negli assi x e y. La distanza (euclidea) è calcolabile come l'ipotenusa di un triangolo rettangolo, i cui cateti sono le differenze tra i valori delle due variabili:

cateto orizzontale: $x_2 - x_1$
cateto verticale: $y_2 - y_1$
ipotenusa = distanza euclidea: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Quando le variabili sono più di due, dobbiamo immaginare uno spazio rappresentato da tante dimensioni quante sono le variabili. Si tratta di una rappresentazione matematica del tutto astratta, ma la distanza euclidea tra due punti **a** e **b** in uno spazio a n dimensioni (n variabili, $\mathbf{v_1, v_2, ..., v_n}$) può essere comunque calcolata come estensione della formula di sopra:

$$d_{a,b} = \sqrt{(v_{1a} - v_{1b})^2 + (v_{2a} - v_{2b})^2 + (v_{3a} - v_{3b})^2 + ... + (v_{na} - v_{nb})^2}$$

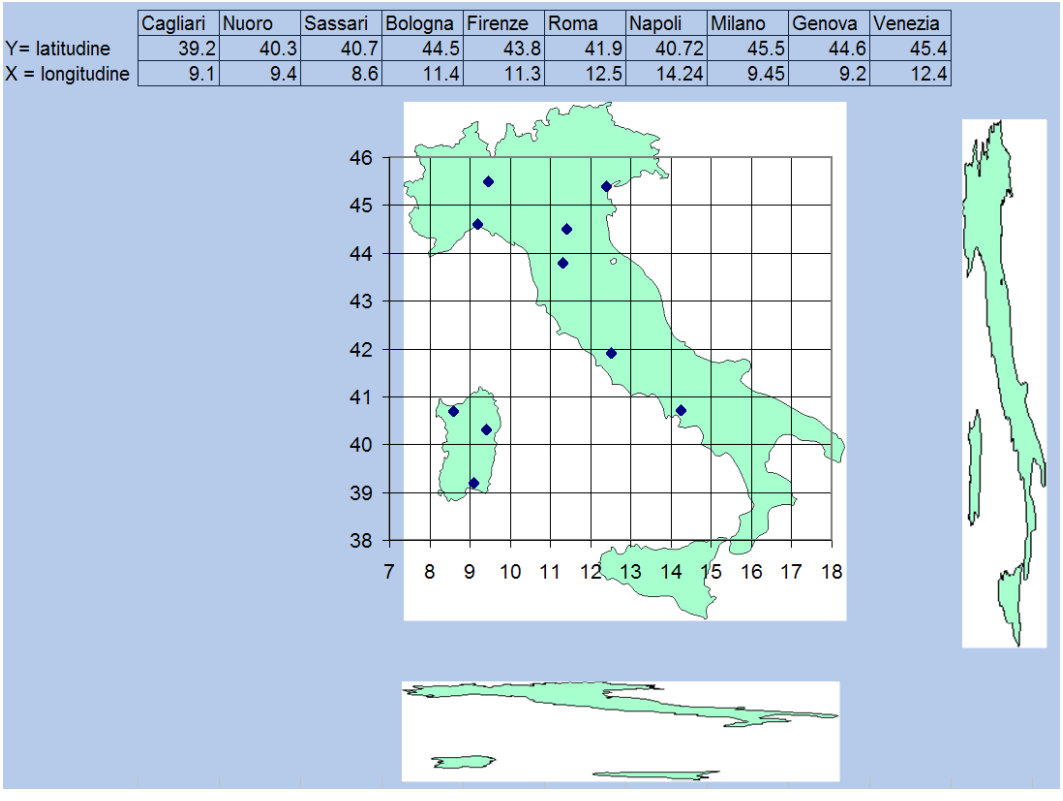
Quanto detto vale se e solo se:

1. Gli assi sono **ortogonali,** altrimenti sarebbe come applicare il teorema di Pitagora ai lati di un triangolo non rettangolo. In termini di variabili, la condizione di ortogonalità (90°) corrisponde alla condizione di assenza di correlazione. Quanto più due variabili sono correlate, tanto più il loro angolo vettoriale diventa stretto. Paradossalmente, se si applica la stessa variabile a entrambe gli assi, lo spazio vettoriale si riduce ad una retta. Con metodi matematici è possibile rendere le variabili ortogonali, in modo da poter calcolare valide distanze. In tal senso sono state proposte diverse distanze multivariate che superano alcuni limiti della distanza euclidea. Anche se non possiamo soffermarci su questo punto, citiamo la **distanza di Mahalanobis** come una delle distanze più utilizzate.

2. Gli assi hanno la **stessa metrica** (altrimenti sarebbe come applicare il teorema di Pitagora avendo un cateto misurato in centimetri e l'altro in litri) e la stessa **unità di misura** (centimetri e centimetri). Ma talvolta è difficile dire se i dati sono appropriati anche quando sono espressi con la stessa unità di misura. Consideriamo ad es. il peso in grammi di un fegato e di un linfonodo. Pur essendo entrambi espressi in grammi, i pesi del fegato avranno medie e varianze molto maggiori rispetto ai pesi di un linfonodo. In questi casi occorre standardizzare i dati.

Vediamo ad es. cosa succede in caso contrario:

	Cagliari	Nuoro	Sassari	Bologna	Firenze	Roma	Napoli	Milano	Genova	Venezia
Y= latitudine	39.2	40.3	40.7	44.5	43.8	41.9	40.72	45.5	44.6	45.4
X = longitudine	9.1	9.4	8.6	11.4	11.3	12.5	14.24	9.45	9.2	12.4



Nella mappa a destra, l'asse X è stato compresso di un fattore 1:5. Sassari appare più vicina a Napoli che a Cagliari. Invece nella mappa in basso, con l'asse Y compresso allo stesso modo, Roma e Venezia risultano vicinissime. Modificando la metrica, le distanze sono falsate.

In SM, generalmente i dati sono organizzati in una matrice con i soggetti nelle righe e le variabili nelle colonne. Ma non sempre. Nelle analisi 'omiche' (es genomica), avendo una gran quantità di variabili (es. decine di migliaia di geni) e relativamente pochi casi (al massimo qualche centinaio di pazienti o animali), le variabili (geni) sono solitamente poste nelle righe ed i casi nelle colonne, in quanto per il nostro occhio è più comodo scorrere un gran numero di righe (in verticale) che un gran numero di colonne (in orizzontale). Per lo stesso motivo, anche in Excel è disponibile un numero di righe molto maggiore rispetto al numero di colonne.

Gli esempi che seguono riguardano dati piuttosto banali di 4 variabili di 10 studenti, presi durante una esercitazione del corso di statistica quando anni fa questa materia era inserita nel corso di laurea di Medicina.

	Età	Altezza	Peso	Piede
Paolo	19	178	83	43
Fred	34	177	82	41
Pippo	20	179	82	41
Minni	19	165	57	39
Betty	19	164	52	37
Wilma	18	162	51	38
Poldo	19	170	69	41
Clara	19	163	40	33
Dino	18	169	67	45
Alice	19	170	50	36

I parametri fondamentali delle 4 variabili sono:

	Età	Altezza	Peso	Piede
n	10	10	10	10
media	20.4	169.7	63.3	39.4
s	4.812	6.395	15.535	3.534

Dai dati possiamo ottenere la matrice (simmetrica) delle correlazioni tra tutte le variabili. Notare la diagonale costituita tutta da valori 1 in quanto si tratta delle correlazioni tra le stesse variabili.

	Età	Altezza	Peso	Piede
Età	**1.00**	.46	.46	.14
Altezza	.46	**1.00**	.91	.60
Peso	.46	.91	**1.00**	.81
Piede	.14	.60	.81	**1.00**

Da notare anche che le correlazioni possono essere sempre calcolate anche tra variabili con metrica diversa, in quanto la correlazione è già di per sé è un indice normalizzato. Tuttavia se le variabili hanno differente metrica non possiamo calcolare le distanze 'euclidee' tra i soggetti. Occorre prima standardizzare le variabili. La standardizzazione non altera le correlazioni. La matrice dei dati standardizzati così risulta:

	Età norm.	Altezza norm.	Peso norm.	Piede norm.
Paolo	-0.291	+1.298	+1.268	+1.019
Fred	+2.826	+1.141	+1.204	+0.453
Pippo	-0.083	+1.454	+1.204	+0.453
Minni	-0.291	-0.735	-0.406	-0.113
Betty	-0.291	-0.891	-0.727	-0.679
Wilma	-0.499	-1.204	-0.792	-0.396
Poldo	-0.291	+0.047	+0.367	+0.453
Clara	-0.291	-1.048	-1.500	-1.811
Dino	-0.499	-0.109	+0.238	+1.585
Alice	-0.291	+0.047	-0.856	-0.962

Una volta standardizzate, le medie e deviazioni standard delle variabili sono tutte 0 ed 1:

	Età	Altezza	Peso	Piede
n	10	10	10	10
media	0	0	0	0
s	1	1	1	1

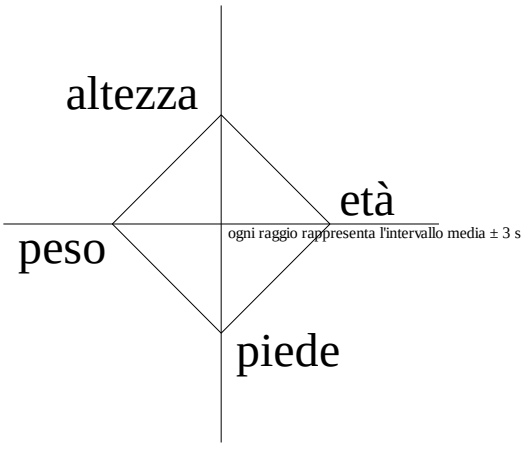
Questa sotto è la matrice simmetrica delle distanze (euclidee in questo caso) tra tutti i soggetti a due a due. Notare la diagonale costituita tutta da valori 0 in quanto si tratta delle distanze tra gli stessi soggetti.

	Paolo	Fred	Pippo	Minni	Betty	Wilma	Poldo	Clara	Dino	Alice
Paolo	**0**	10.07	0.39	8.21	11.66	12.55	2.70	21.17	3.40	10.00
Fred	10.07	**0**	8.56	16.15	18.86	21.26	11.62	26.94	14.83	17.16
Pippo	0.39	8.56	**0**	7.75	10.55	11.94	2.72	18.74	4.83	8.27
Minni	8.21	16.15	7.75	**0**	0.45	0.49	1.53	4.18	3.73	1.53
Betty	11.66	18.86	10.55	0.45	**0**	0.23	3.36	1.90	6.71	0.98
Wilma	12.55	21.26	11.94	0.49	0.23	**0**	3.67	2.57	6.18	1.93
Poldo	2.70	11.62	2.72	1.53	3.36	3.67	**0**	9.81	1.37	3.50
Clara	21.17	26.94	18.74	4.18	1.90	2.57	9.81	**0**	15.47	2.33
Dino	3.40	14.83	4.83	3.73	6.71	6.18	1.37	15.47	**0**	7.75
Alice	10.00	17.16	8.27	1.53	0.98	1.93	3.50	2.33	7.75	**0**

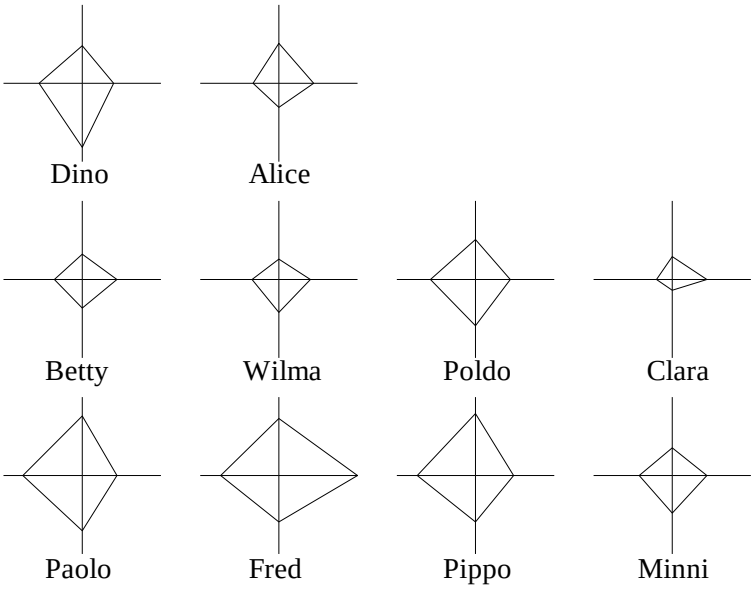
Le matrici di correlazione tra variabili e delle distanze tra soggetti sono alla base della maggior parte delle tecniche multivariate.

Sun-ray-plot

E' una semplice ma efficace rappresentazione grafica. Per ogni soggetto si costruisce una stella con tanti raggi quante sono le variabili. Il raggio rappresenta l'intervallo media ± 3 deviazioni standard, ponendo la media al centro del raggio, i valori sotto la media verso l'interno ed i valori sopra la media verso l'esterno.



Unendo i raggi si ottiene un poligono. Dall'esame delle forme e dimensioni dei poligoni è semplice fare valutazioni e confronti tra i soggetti.

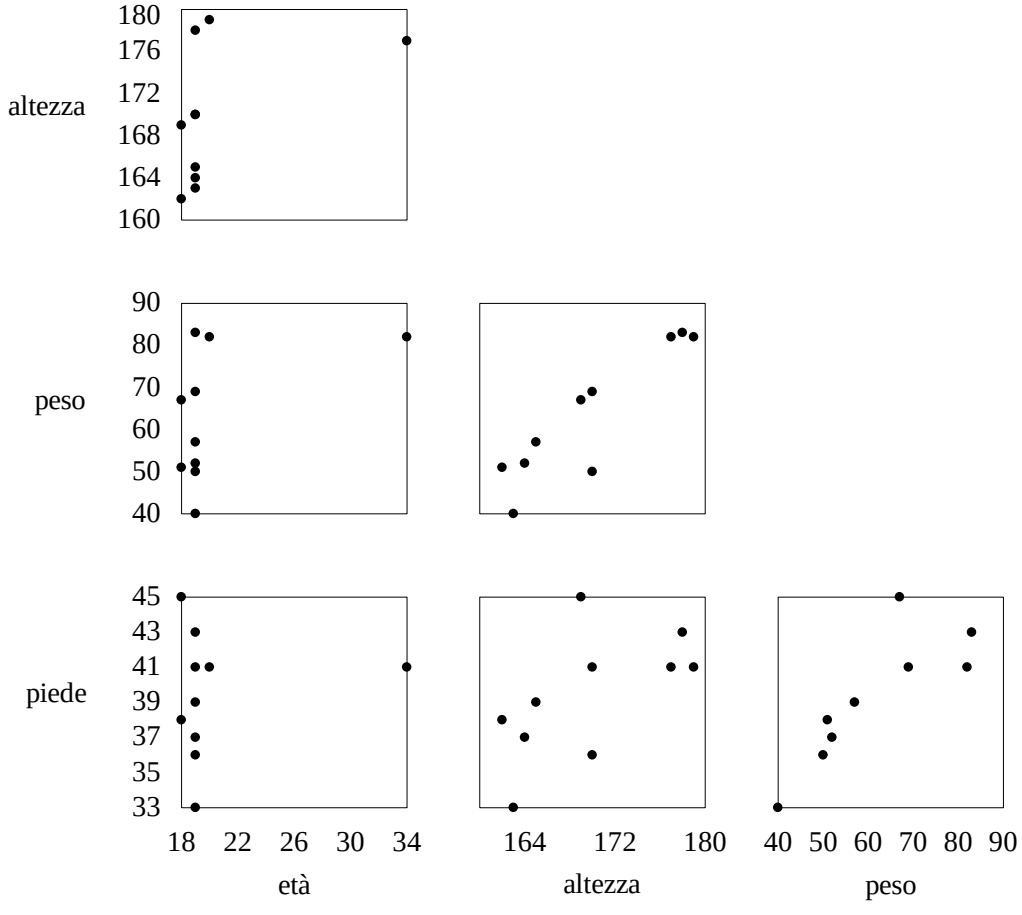


Sono evidenti:
- l'età attempata di Fred
- la taglia minuta di Clara
- il gran piede di Dino
- la somiglianza tra Pippo e Poldo
- la somiglianza tra Minni e Wilma

Per ovvie ragioni, il Sun-ray-plot va bene sino a che le variabili non sono più di una decina.

Grafici delle variabili prese a 2 a 2

E' banale ma efficace. Accompagna quindi bene la matrice delle correlazioni. Per ovvie ragioni, anche questa tecnica è utile quando il numero di variabili è limitato.



Per ragioni di spazio non abbiamo aggiunto i nomi vicino ai punti. Sarebbe stato interessante osservare la posizione di ogni soggetto nei diversi grafici. Notiamo che le variabili maggiormente correlate sono peso ed altezza (confronta con i dati della matrice di correlazione vista in precedenza).

Analisi delle componenti principali - Principal Component Analysis (PCA)

E' una classica tecnica multivariata. La PCA è applicata anche in altri campi con nomi diversi: analisi di Hotelling, analisi di Karhunen-Loève, analisi degli autovettori, ecc. La PCA ruota i punti descritti nello spazio rappresentato da n variabili in modo tale da trovare diverse proiezioni con dispersione (varianza) decrescente e tra loro ortogonali. La prima componente principale avrà quindi la massima varianza, cioè la massima informazione; poi la seconda, la terza e così via sino alla n^{ma} componente principale.

Per comprendere meglio, immaginiamo di dover riconoscere un oggetto dalla sua ombra. L'ombra è semplicemente una proiezione 2D dell'oggetto 3D. Noi possiamo ruotare diversamente l'oggetto ottenendo ombre di diversa estensione. Quasi sempre l'ombra di estensione massima è spesso quella più informativa della natura dell'oggetto. E' la cosiddetta silhouette.



Secondo lo stesso principio, sappiamo che è facile riconoscere un oggetto dal suo profilo 2D, quando l'area di proiezione è massima. Al contrario, è molto difficile se non impossibile riconoscere certi oggetti visti in proiezioni minime. Questo cerchio ad es.

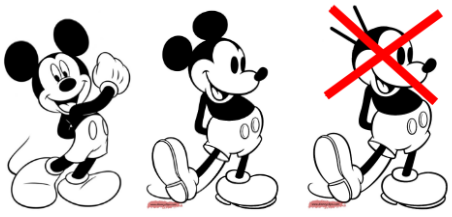
può essere è la proiezione minima di una sigaretta, di una matita, di un utensile, ecc..
Riconosciamo l'oggetto se osserviamo la sua proiezione massima (al di là dei colori)

Allo stesso modo, questo potrebbe essere la proiezione minima di una moneta, di una chiave o di una rondella:

Riconosciamo l'oggetto se osserviamo la sua proiezione massima:



Un analogo principio è da sempre utilizzato dai cartoons. Non vediamo mai le orecchie di Topolino di taglio, per quanto lo stesso Topolino sia rivolto di lato. Le orecchie sono sempre disegnate con la loro massima proiezione, per rispettare l'icona del personaggio:

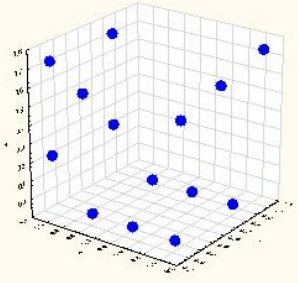


Ci sono ovviamente delle eccezioni. Ad esempio, un piccolo oggetto sotto un ombrello o sotto un sombrero. L'area di proiezione massima mostra solo l'ombrello o il sombrero, e nasconde l'oggetto.

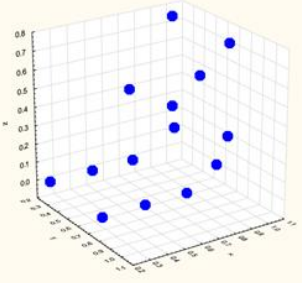
Sebbene raramente questo effetto di schermatura avviene anche con i dati. Proprio per questo metodi simili alla PCA (analisi fattoriale) consentono di ruotare i dati secondo diversi livelli (non massimi) di varianza.

Per fare un esempio più attinente al metodo della PCA, consideriamo un dado trasparente. L'immagine dei punti del dado attraverso il cubo trasparente è una proiezione 2D di un oggetto 3D. Vediamo ora come la proiezione dei punti cambia ruotando il dado.

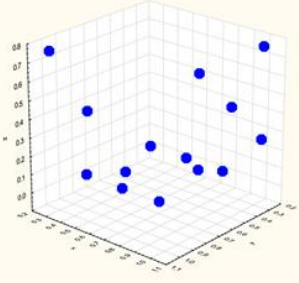
per semplicità gli esempi grafici riguardano solo 3 facce del dado



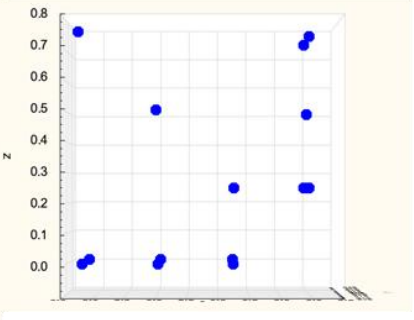
proiezione iconografica ideale



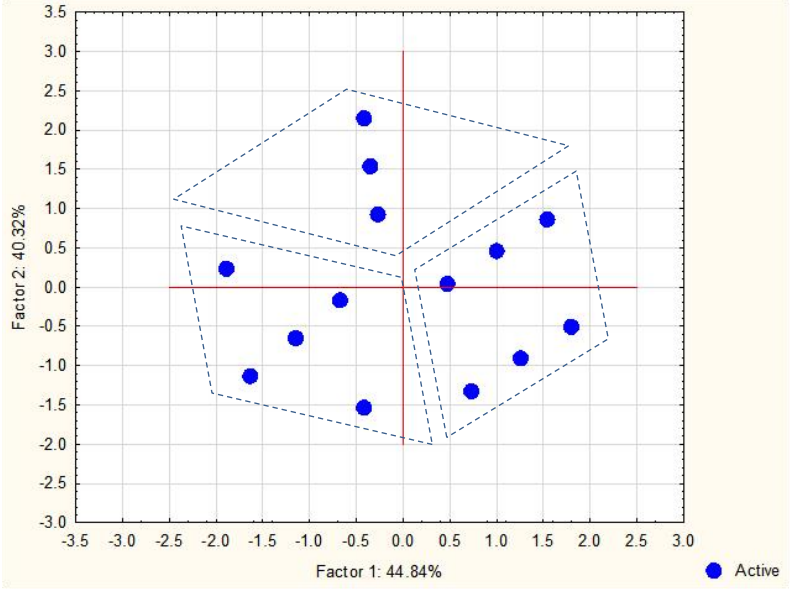
proiezione casuale



proiezione casuale



proiezione minima



proiezione massima possibile (89% della varianza totale) ottenuta dalla PCA

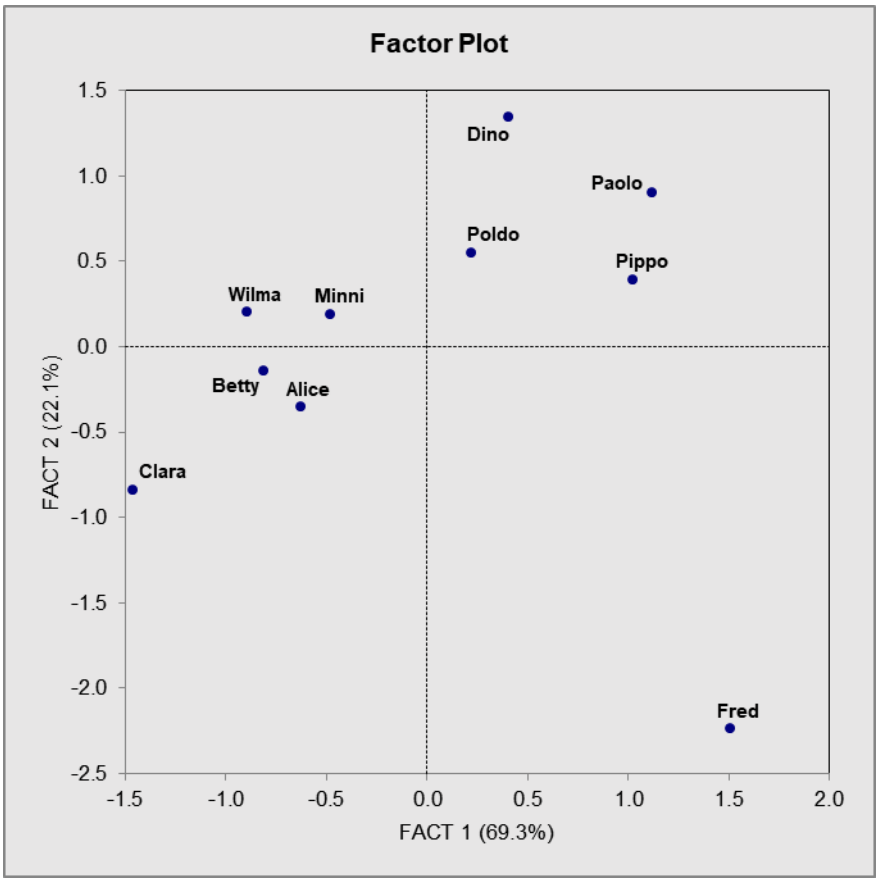
sono riconoscibili le 3 facce come nella proiezione iconografica

Nel grafico sono stati tratteggiati a mano i contorni delle tre facce del dado per apprezzare meglio la proiezione prodotta dalla PCA.

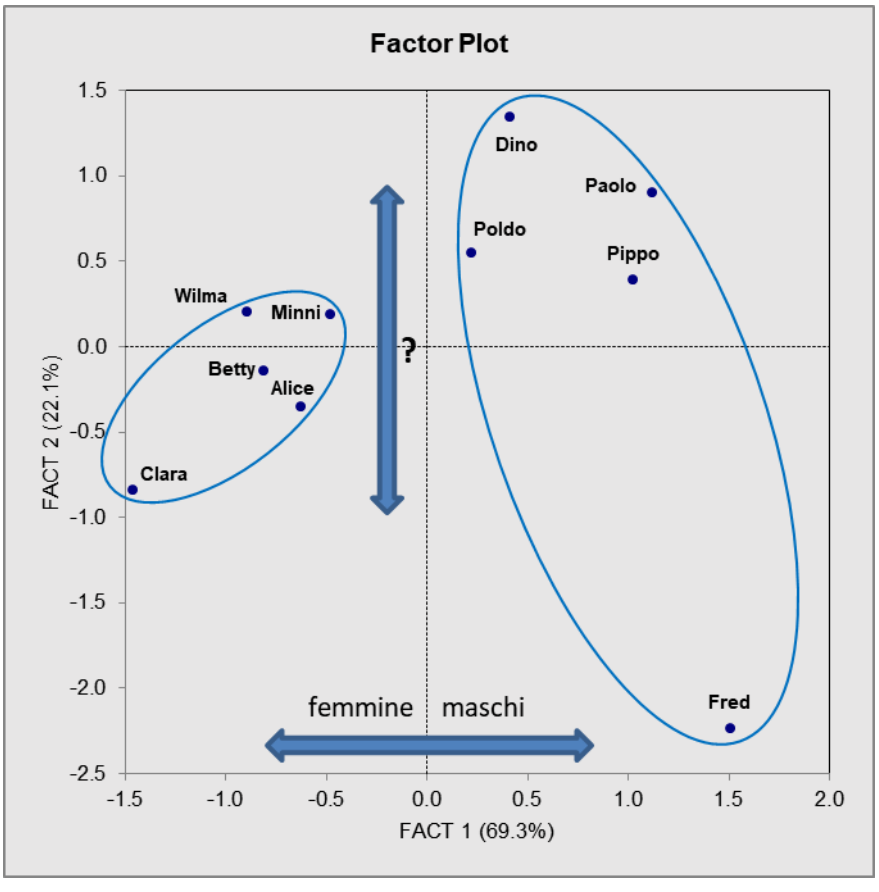
I 3 assi del dado rappresentano 3 variabili ortogonali (x, y, z). Le variabili sono dunque le coordinate di ciascun soggetto. Nel caso del dado i soggetti sono i pallini disegnati sulle facce. Nel caso di un set di dati riguardanti K (s)oggetti o casi, descritti da un numero Q di variabili, possiamo quindi concepire gli oggetti come K punti in uno spazio a Q dimensioni.
La PCA ruota lo spazio a Q dimensioni in modo da trovare inizialmente quell'asse - virtuale - su cui i punti mostrano la massima dispersione (= varianza massima). Questa è la prima componente principale. Successivamente, fissato il primo asse, si ruota lo spazio in modo da trovare un secondo asse ortogonale con dispersione massima (comunque inferiore a quella del primo asse). Questa è la seconda componente principale. Poi si continua con la terza, quarta, ecc., in teoria si possono ottenere tante componenti principali quante sono le variabili originarie, sebbene l'informazione contenuta nelle successive componenti principali sia via via minore. La maggior parte dell'informazione (60-90%) è solitamente concentrata nelle prime due o tre componenti principali. Tornando al set di dati dei 10 ragazzi, le prime due componenti principali riassumono oltre il 90% dell'intera informazione.

Componente Numero	Percentuale di Varianza	Percentuale Cumulativa
1	69.33	69.33
2	22.07	**91.40**
3	7.87	99.27
4	0.73	100.00

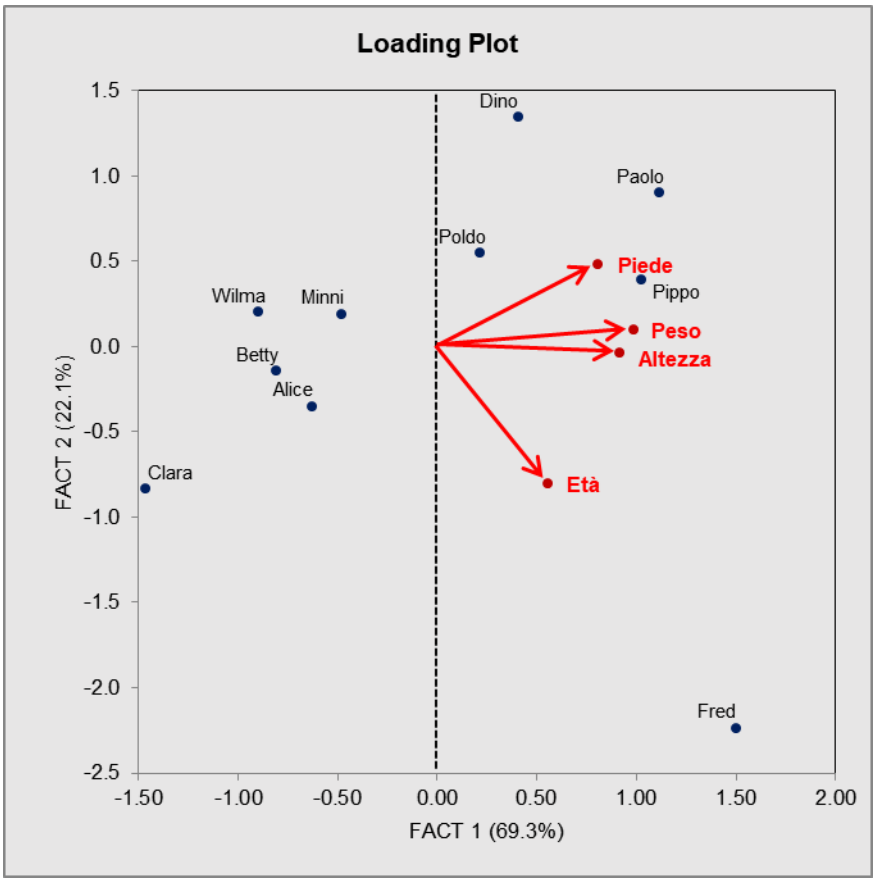
Il grafico delle prime due CP è il seguente:



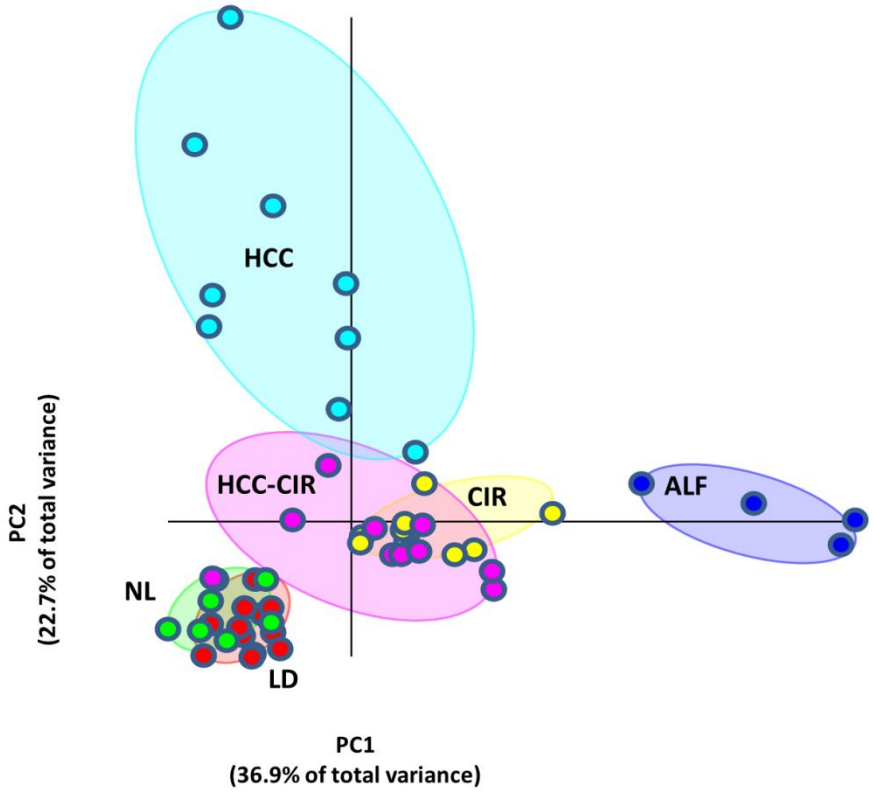
Se vogliamo dare un'interpretazione di queste prime due CP, potremo dire che la 1° CP, che rappresenta circa il 70% dell'informazione di tutti i dati, rappresenta le dimensioni dell'individuo, e quindi separa bene i maschi dalle femmine, mentre la 2° CP, che rappresenta appena il 22%, è più difficile da interpretare. Messe assieme, le prime due CP concentrano oltre il 90% dell'informazione delle quattro variabili originarie.



Se poi si è interessati a vedere come le variabili concorrono a determinare le CP, è possibile proiettare nello stesso grafico sia i soggetti che le variabili. Si vede ad es. che è il piede la variabile più allineata sull'asse obliquo che passa attraverso i maschi e le femmine e quindi è quella che concorre meglio a separare i due gruppi, anche se la prima CP è più determinata dal peso e dall'altezza. L'età non ha granché relazione con le altre tre variabili. In effetti le età degli studenti sono quasi tutte identiche, eccetto che per Fred, un po' attempato. Le frecce rappresentano solo le parti positive delle variabili, cioè i valori sopra la media, dirette verso i maschi. Le parti negative non sono tracciate per ragioni di chiarezza, e si ottengono semplicemente prolungando gli assi.



Vediamo ora una PCA di dati più seri.



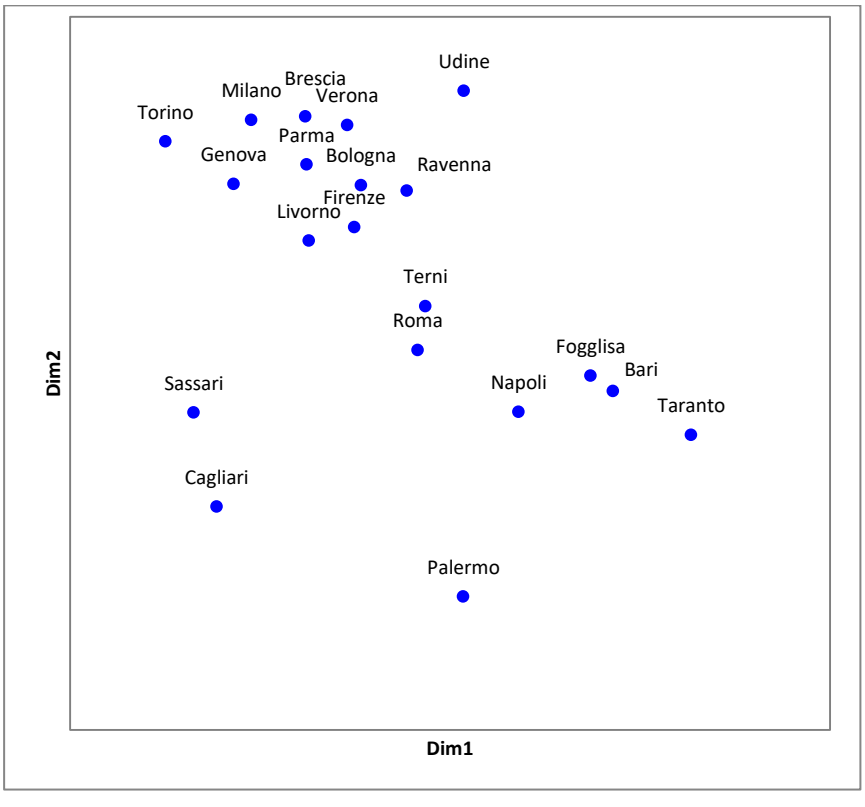
Questa PCA è stata ottenuta utilizzando circa 1000 miRNA di 55 pazienti con sei diverse condizioni del fegato: carcinoma epatocellulare associato al virus HCV e HBV (HCC), cirrosi (CIR), cirrosi associata a HCC (HCC-CIR), epatite fulminante (ALF), e fegati normali di soggetti con angiomi (NL) e di donatori (LD). I fegati normali, NL e LD, sono sovrapposti. Un'altra sovrapposizione si osserva al centro tra CIR e HCC-CIR. Invece vi è una buona separazione tra HCC, ALF e fegati normali (NL e LD). Le due componenti principali riassumono insieme circa il 60% di tutta l'informazione dei 1000 miRNA.

Multidimensional Scaling (MDS)

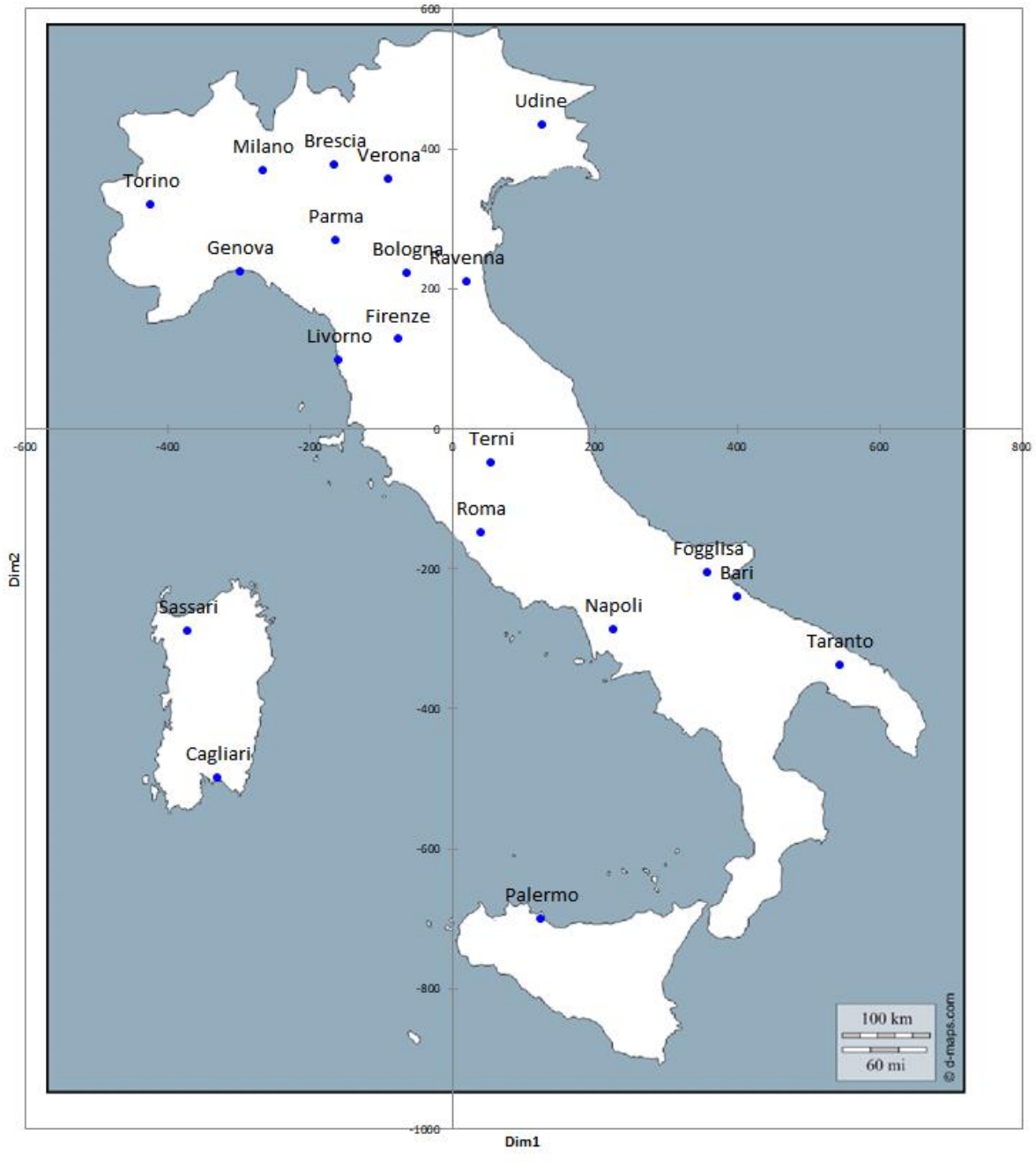
Le tecniche di multidimensional scaling (MDS) trasformano una matrice di distanze tra oggetti in un set di coordinate in modo tale che le distanze descritte da tali coordinate corrispondano il più possibile alle distanze originali. Ad esempio, mediante il MDS possiamo ricostruire la mappa delle città partendo dalle loro distanze. Qui vediamo appunto una matrice di distanze tra alcune città di Italia. Da notare che si tratta di distanze di tragitti stradali o navali e non corrispondono alle vere distanze in linea d'aria. Ad es., quella tra Firenze e Milano è molto minore dei 306 km riportati in tabella. Come vedremo il risultato del MDS sarà un po' distorto - ma solo un po' - da questo fatto.

	Milano	Roma	Cagliari	Firenze	Bologna	Genova	Torino	Udine	Sassari	Napoli	Bari	Taranto	Fogglisa	Brescia	Verona	Parma	Ravenna	Terni	Livorno	Palermo
Milano	0	600	870	306	249	147	165	397	665	819	973	1068	849	100	177	143	328	527	288	1109
Roma	600	0	510	299	385	504	660	588	436	232	461	533	324	564	522	464	358	100	322	529
Cagliari	870	510	0	676	768	724	824	1038	215	595	862	886	749	890	889	785	790	592	619	485
Firenze	306	299	676	0	95	242	398	367	510	513	678	769	548	264	230	166	127	220	96	823
Bologna	249	385	768	95	0	234	373	284	596	586	724	820	602	185	137	110	86	296	161	912
Genova	147	504	724	242	234	0	158	473	519	733	918	1007	786	201	247	141	319	447	182	990
Torino	165	660	824	398	373	158	0	562	611	890	1074	1164	943	264	337	265	458	605	340	1132
Udine	397	588	1038	367	284	473	562	0	877	728	767	871	681	298	228	334	248	488	446	1104
Sassari	665	436	215	510	596	519	611	877	0	598	866	914	736	696	705	594	634	489	437	626
Napoli	819	232	595	513	586	733	890	728	598	0	270	317	156	770	717	678	537	293	550	397
Bari	973	461	862	678	724	918	1074	767	866	270	0	104	138	902	834	830	652	478	741	566
Taranto	1068	533	886	769	820	1007	1164	871	914	317	104	0	221	1001	934	926	751	562	828	536
Fogglisa	849	324	749	548	602	786	943	681	736	156	138	221	0	784	720	706	536	342	608	521
Brescia	100	564	890	264	185	201	264	298	696	770	902	1001	784	0	79	108	250	480	278	1086
Verona	177	522	889	230	137	247	337	228	705	717	834	934	720	79	0	116	184	431	271	1049
Parma	143	464	785	166	110	141	265	334	594	678	830	926	706	108	116	0	194	386	170	982
Ravenna	328	358	790	127	86	319	458	248	634	537	652	751	536	250	184	194	0	261	218	886
Terni	527	100	592	220	296	447	605	488	489	293	478	562	342	480	431	386	261	0	266	625
Livorno	288	322	619	96	161	182	340	446	437	550	741	828	608	278	271	170	218	266	0	822
Palermo	1109	529	485	823	912	990	1132	1104	626	397	566	536	521	1086	1049	982	886	625	822	0

Di seguito vediamo la mappa ottenuta con il MDS.



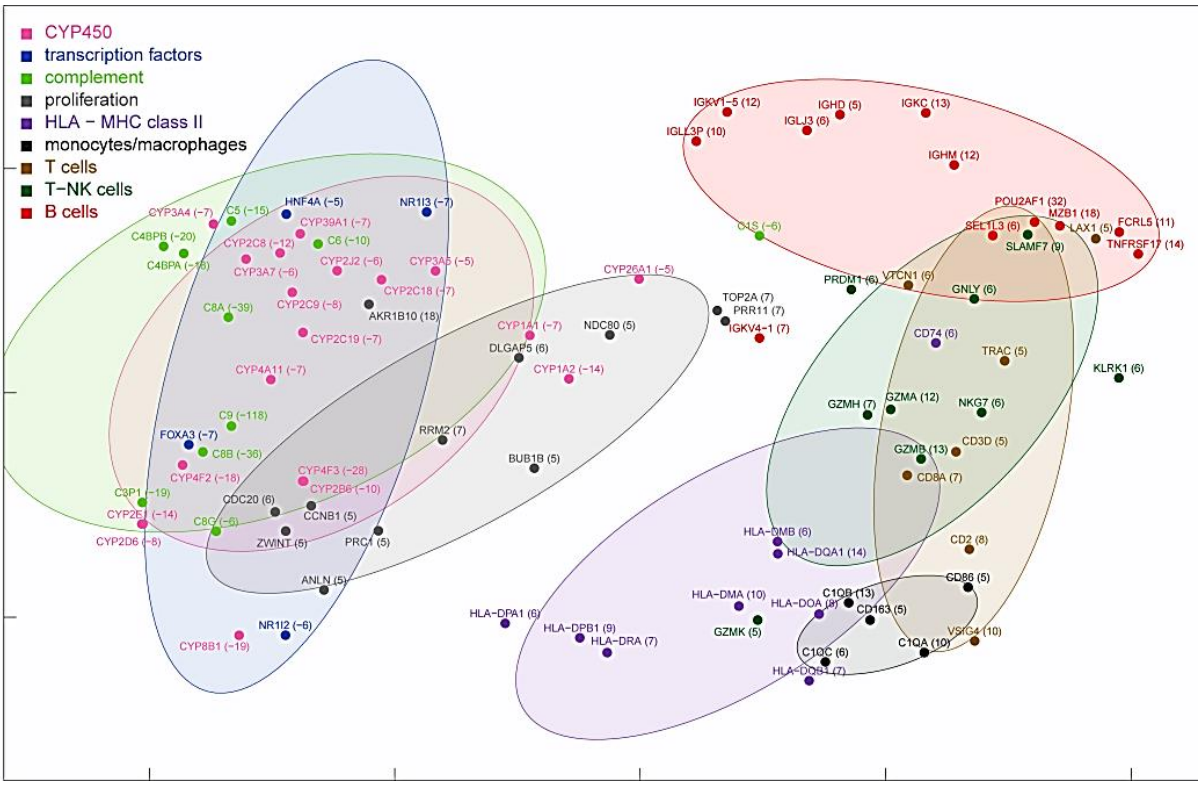
In questo secondo grafico, il grafico MDS è sovrapposto alla cartina geografica verificare la giustezza delle coordinate calcolate dal MDS.



La mappa ottenuta dal MDS è ottima ma non perfetta. Errori evidenti sono le posizioni di Bari, Taranto e Napoli che non sono sul mare. Questo è dovuto al fatto che sono state utilizzate distanze di percorsi stradatli/vanali e non distanze in linea d'aria. Comunque la ricostruzione è notevole.

Quando invece le distanze o indici di dissimilarità si riferiscono ad uno spazio a n variabili, la mappa dei prima 2 assi del MDS non potrà riprodurre interamente la struttura dei dati. Al posto della distanza possiamo considerare un indice di dissimilarità, come ad es. il complementare del coefficiente di correlazione (1-r) calcolato tra soggetti, anziché tra variabili. Questo tipo di distanza ha l'enorme vantaggio di non richiedere la standardizzazione dei dati in quanto la correlazione di per sé rappresenta una relazione standardizzata.

Il grafico successivo è un MDS ottenuto da tutte le distanze (miscorrelazioni) tra l'espressione di geni di campioni di fegato di pazienti con epatite acuta fulminante. Si riconoscono gruppi di geni (contornati da ellissi) espressi preferenzialmente da alcuni tipi cellulari o associati a determinate funzioni.



Il risultato del MDS è simile a quella della PCA, ma ha due grossi vantaggi. Uno è il fatto che con il MDS possiamo scegliere il tipo di distanze, lineari o non lineari (molto utilizzate in ecologia, psicologia e nelle discipline 'omiche': genomica, proteomica, ecc.). Il secondo vantaggio è il fatto che possiamo eliminare vizi anche gravi dovuti al campionamento. Ad esempio, se adottiamo come distanza il valore 1 - correlazione, ed individuiamo dei fattori che influenzano i dati producendo correlazioni spurie, possiamo eliminare questo effetto calcolando le correlazioni parziali.

Analisi discriminante lineare - Linear Discriminant Analysis (LDA)

La LDA, a differenza della PCA, è un metodo 'supervised' in quanto parte dalla pregressa conoscenza di gruppi o categorie di soggetti e mira a classificare cioè a stabilire a quale gruppo appartenga uno o più nuovi soggetti. Il problema è sempre quello di ruotare i dati in uno spazio a n variabili, ma questa volta non per massimizzare la varianza totale (come era nella PCA) ma per massimizzare il rapporto [varianza tra gruppi]/[varianza entro gruppi], ovviamente in campo multivariato. Una volta trovate le funzioni di rotazione che massimizzano tale rapporto con i soggetti di riferimento, si applicano le stesse funzioni ai nuovi soggetti da classificare.

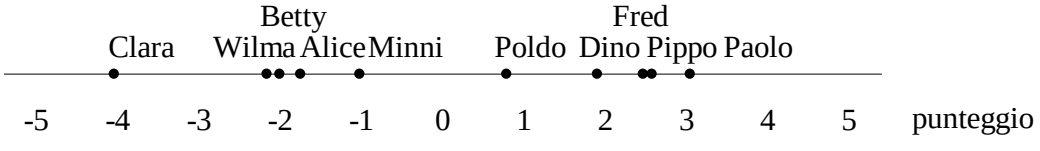
Possiamo prendere come esempio banale i due gruppi di ragazzi formati da 5 maschi e 5 femmine. Trattandosi di 2 soli gruppi, basterà solo un asse (una sola funzione) per separare i due gruppi. Vediamo i valori dei coefficienti della funzione discriminante:

costante	-36.7426
età	0.02029
altezza	0.11916
peso	0.04506
piede	0.33643

Il valore dalla funzione discriminante

$-36.7426 + (\text{età} \times 0.02029) + (\text{altezza} \times 0.11916) + (\text{peso} \times 0.04506) + (\text{piede} \times 0.33643)$

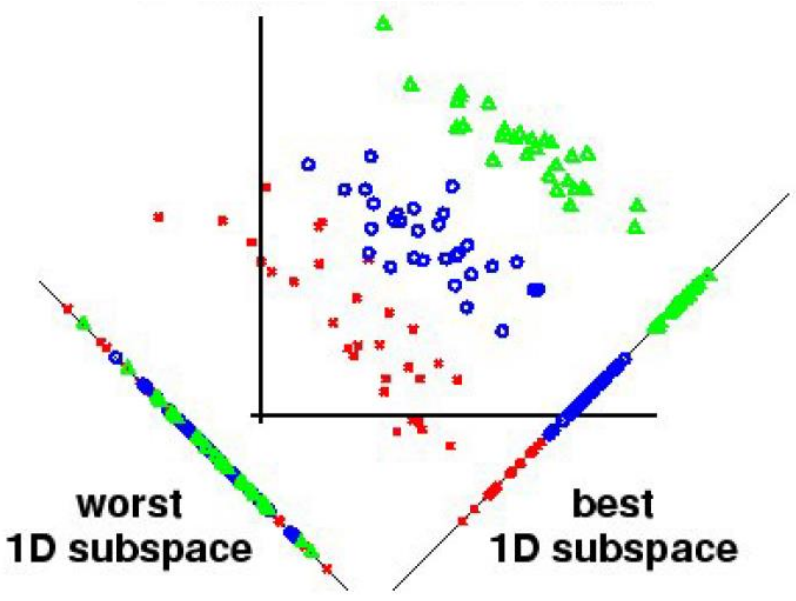
applicata ai dati di un nuovo soggetto darà un punteggio o score che assegnerà il nuovo soggetto al gruppo dei maschi o a quello delle femmine, a seconda della minor distanza dal centroide delle femmine o da quello dei maschi. Queste sono le posizioni dei 10 soggetti ottenuti in base alla funzione:



In questo caso i centroidi dei due gruppi sono:
ragazze: -2.15761
ragazzi: 2.15761
e lo zero è la soglia discriminante.

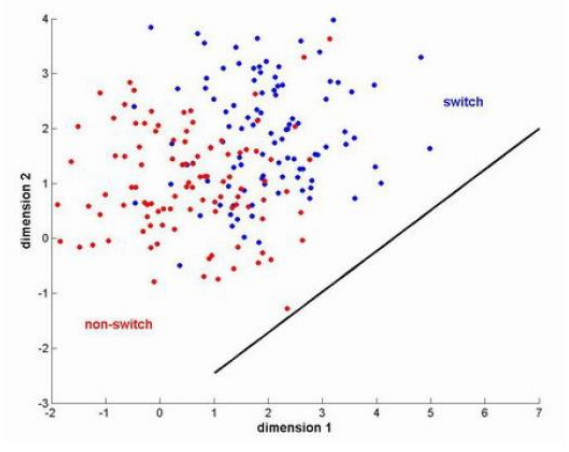
I punteggi positivi classificano i soggetti come maschi, quelli negativi come femmine. Notare che la funzione è stata trovata conoscendo il sesso dei soggetti. I dati utilizzati per calcolare la funzione discriminante costituiscono il cosiddetto set di apprendimento, o learning set. L'analisi discriminante è messa alla prova solo dopo, con l'applicazione della funzione trovata a soggetti di sesso non noto che costituiscono il cosiddetto set di applicazione, o di validazione o test set.
Per l'analisi discriminante è quindi necessario suddividere i dati in due subset per quanto possibile omogenei. Uno sarà da utilizzare come set di apprendimento e l'altro come set di validazione. Ciò che conta non è quanto bene la funzione separa i gruppi del set di apprendimento, ma quanto bene separa i gruppi del set di validazione. Nell'esempio questo non è stato fatto perché avevamo pochi dati. Chi legge può comunque provare ad applicare la funzione ai propri dati e verificare se viene classificato correttamente come maschio o femmina, purché si trovi nella fascia di età delle matricole.

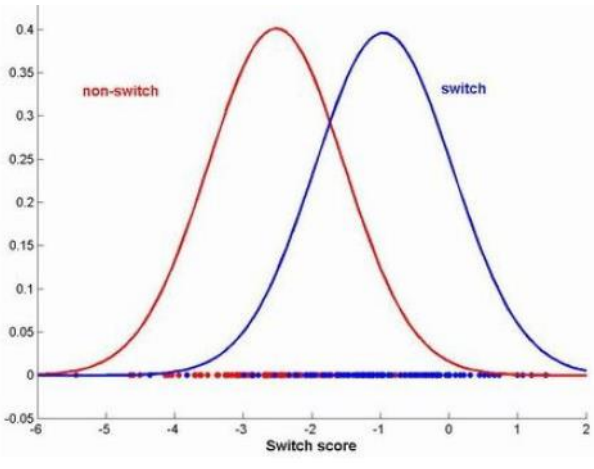
Il grafico sotto mostra 3 gruppi di oggetti decritti da due variabili.
Nessuna delle due variabili originali è in grado di discriminare i 3 gruppi. Ma se i dati vengono proiettati con un determinato angolo su di un nuovo asse i 3 gruppi sono ben discriminati. La funzione discriminante è quella che identifica la proiezione che meglio separa i gruppi.



Sempre osservando il grafico si può capire che una semplice analisi di componenti principali non sarebbe in grado di separare i gruppi. Infatti le due proiezioni mostrate (worst 1D subspace e best 1D subspace) hanno pressappoco la stessa varianza.

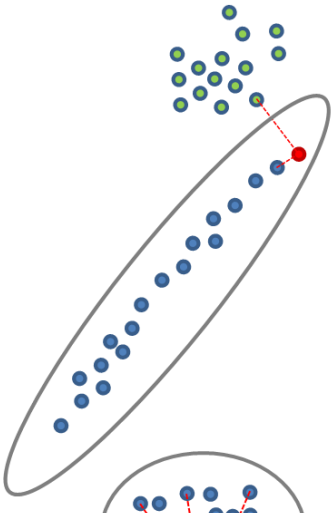
Comunque non sempre, anzi direi mai, l'analisi discriminante è in grado di discriminare al 100% i gruppi.





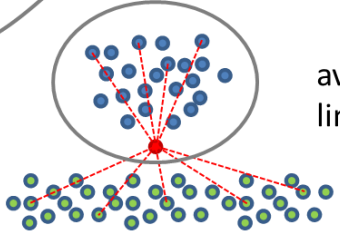
Analisi gerarchica dei gruppi - Hierarchical Cluster Analysis (HCA)

Lo scopo di questa analisi è quello di suddividere il set di soggetti oppure il set di variabili in gruppi e sottogruppi il più possibile omogenei e gerarchicamente strutturati. Tipico output della HCA è il dendrogramma, o diagramma ad albero, molto utilizzato in biologia per la costruzione di mappe filogenetiche ed in genomica per l'individuazione di geni co-espressi e co-regolati. La HCA è una tecnica 'unsupervised' in quanto prescinde da qualsiasi conoscenza a priori sul tipo di struttura dei dati. La sua è una classificazione automatica, ma tecnicamente molto critica, in quanto i risultati cambiano notevolmente a seconda di due scelte fondamentali: il tipo di distanza o dissimilarità (ad es. distanza euclidea, di Mahalanobis, di Manhattan, di Chebychev, 1-coefficiente di correlazione, ecc.) ed il criterio di clustering (tra i principali, single o nearest neighbour linkage, average linkage, complete o furthest neighbour linkage). In queste immagini vediamo come un nuovo oggetto (il pallino rosso) è assegnato ad un determinato gruppo (il gruppo dei pallini blu o verdi), a seconda dei tre diversi criteri di clustering. La distanza ovviamente è quella euclidea. L'ellisse sta ad indicare il confine del gruppo dopo l'ingresso del nuovo oggetto nel gruppo.



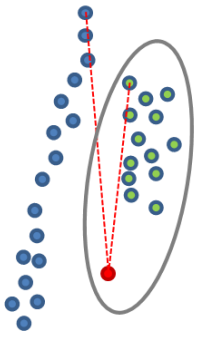
single or nearest neighbour linkage

Il single o nearest neighbour linkage assegna un oggetto al gruppo da cui dista meno, considerando il componente del gruppo più vicino.



average linkage

L'average linkage assegna un oggetto al gruppo da cui dista meno considerando la distanza media rispetto a tutti i componenti del gruppo.

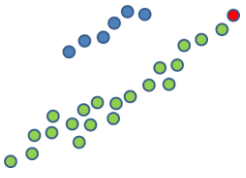


complete or furthest neighbour linkage

Il complete o furthest neighbour linkage assegna un oggetto al gruppo da cui dista meno considerando il componente del gruppo più lontano.

Partendo dalla matrice delle distanze, il primo cluster è quello costituito dai due soggetti (o variabili) più simili, secondo il criterio di clustering scelto (v. legame singolo, medio o completo). Il processo è iterativo nel senso che una volta che si individua un cluster, i suoi componenti vengono eliminati e sostituiti dal cluster stesso. Questo si ripete sino alla fusione di tutti i soggetti in un unico cluster.

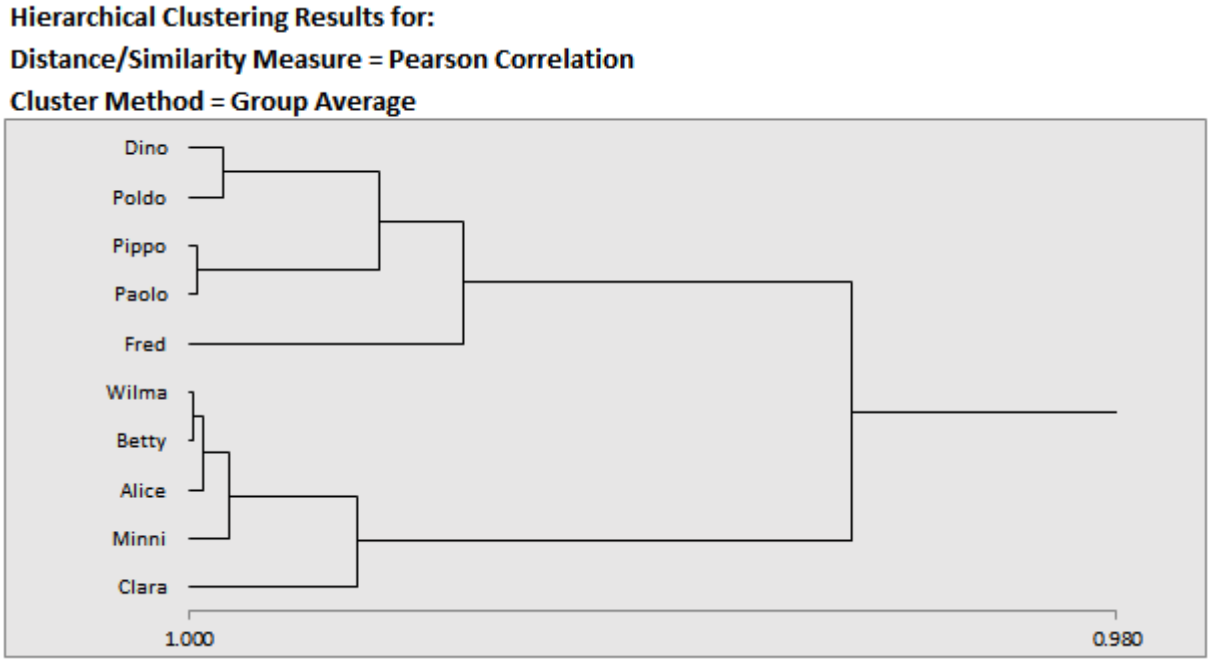
Un classico esempio di errore nell'adottare un certo criterio di clustering si verifica con l'average linkage. Questo, che sembrerebbe il criterio più equilibrato, ed è anche quello più spesso adottato, porta a gravi errori se la struttura dei gruppi è allungata. Negli esempi 2D come quelli rappresentati sopra è facile individuare la forma dei gruppi, ma nei casi concreti, in ambito multidimensionale, è difficile verificare questo. Nell'esempio riportato



è' evidente che il pallino rosso, appartiene al gruppo verde, ma se si applica il criterio dell'average linkage verrà attribuito al gruppo blu! Per questa struttura di dati, il criterio corretto è ovviamente quello del single o nearest neighbour linkage. Per questo prima di adottare un determinato criterio di clustering occorre considerare bene la struttura dei gruppi, eventualmente avvalendosi anche di altri metodi come la PCA o il MDS. Questo è il problema maggiore per questo tipo di analisi ed è responsabilità completa del ricercatore. Purtroppo, anche in lavori scientifici seri, molto spesso vengono riportati dendrogrammi (vedi oltre) senza specificare il criterio di clustering ed il tipo di distanza adottati.

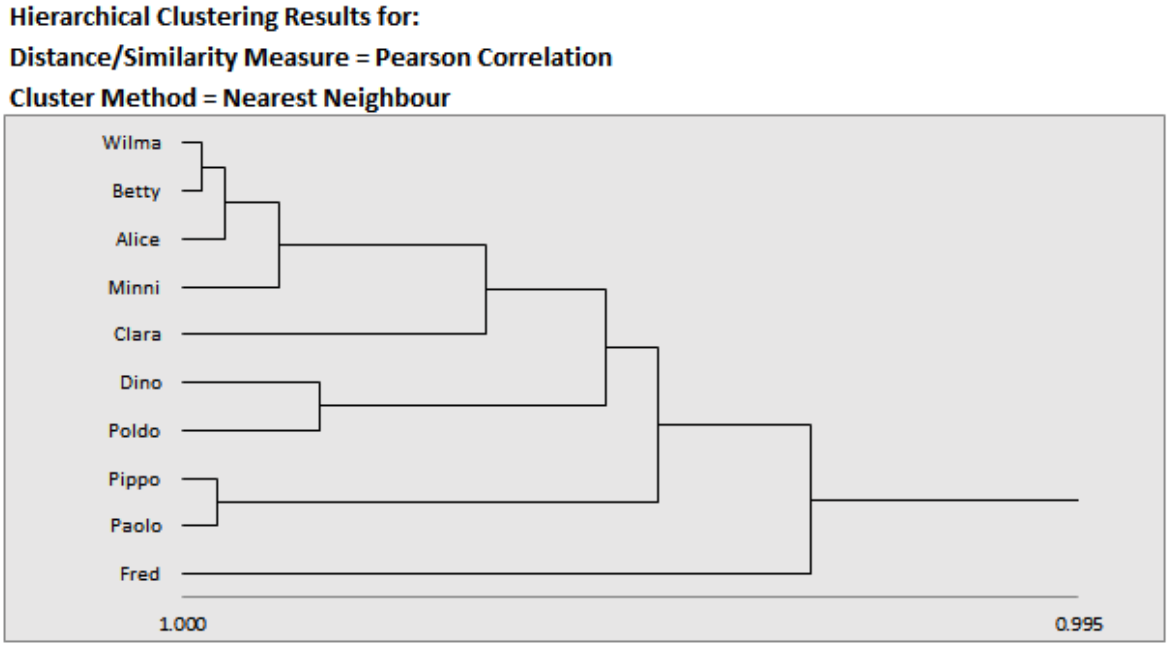
Tipicamente i gruppi sono rappresentati dal cosiddetto dendrogramma. A scopo dimostrativo osserviamo alcuni dendrogrammi ottenuti per il solito set di 10 soggetti utilizzando la distanza euclidea o il valore 1-correlazione, e diversi criteri di clustering. Si vedrà come variando il criterio di clustering e/o il tipo di distanza, varia anche la forma di aggregazione dei soggetti.

1.
Metodo: average linkage.
Distanza: 1-correlation



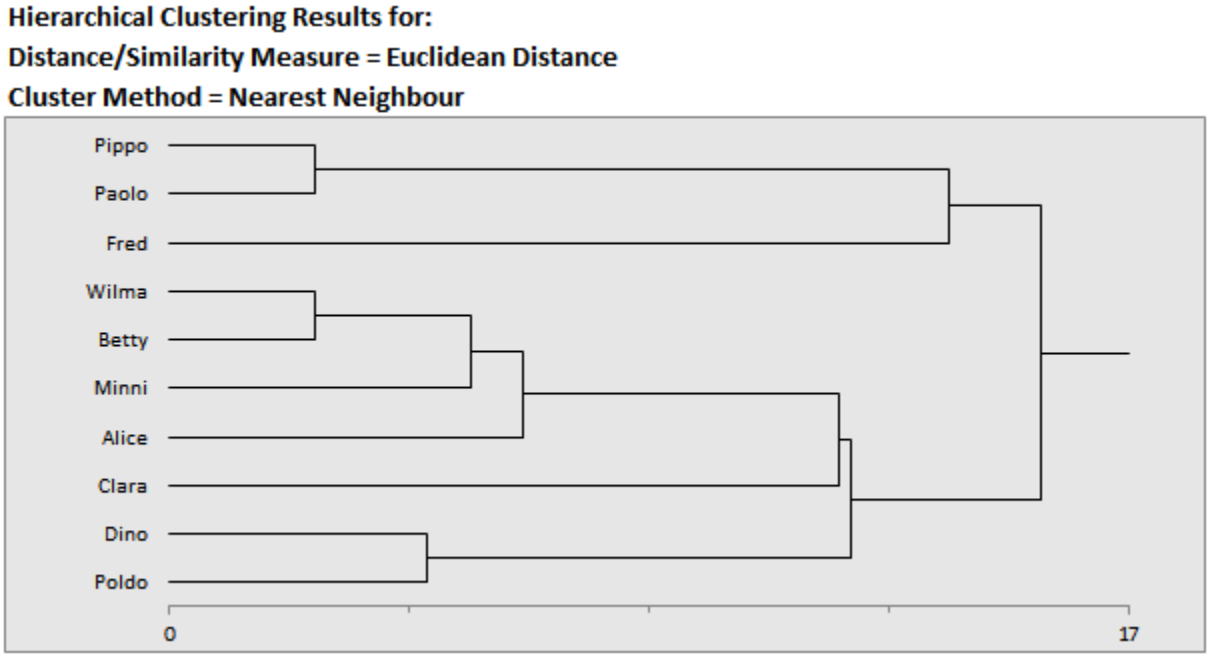
Il dendrogramma, come aveva fatto la PCA, separa molto bene maschi da femmine. Le coppie Pippo e Paolo tra i maschi e Wilma e Bettty tra le femmine sono le più omogenee. Fred appare come il soggetto più eterogeneo.

2.
Metodo: single linkage o nearest neighbour.
Distanza: 1-correlation



Non c'è una netta separazione tra maschi e femmine. Dino e Poldo sono più associati alle femmine che ai maschi.

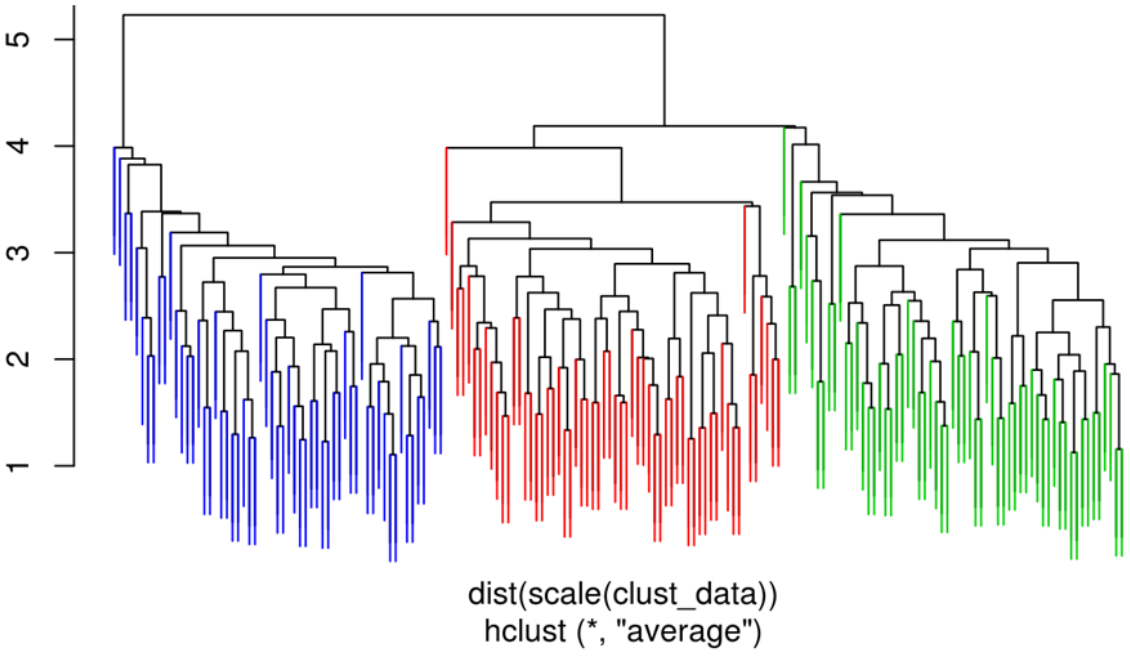
3.
Metodo: single linkage o nearest neighbour
Distanza: euclidea (sbagliato in partenza in quanto le scale delle variabili sono molto differenti e le variabili non sono state standardizzate)



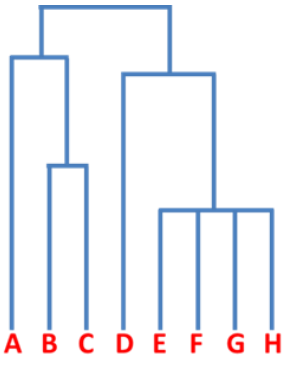
In questo modo, maschi e femmine risultano ancor più mescolati.

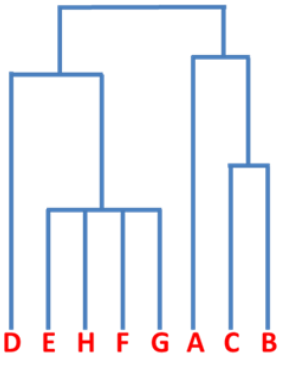
In ultimo, anche dei cluster si può valutare la significatività. O per lo meno, la significatività di alcuni cluster. Nel nostro caso, potremmo valutare se ad esempio il cluster formato da Wilma, Betty e Minni sia significativo. Oppure solo Wilma e Betty. Il test di significatività applicato all'HCA è un po' complesso ma importante, tuttavia è raramente riportato.

Altri modi di rappresentare clusters:



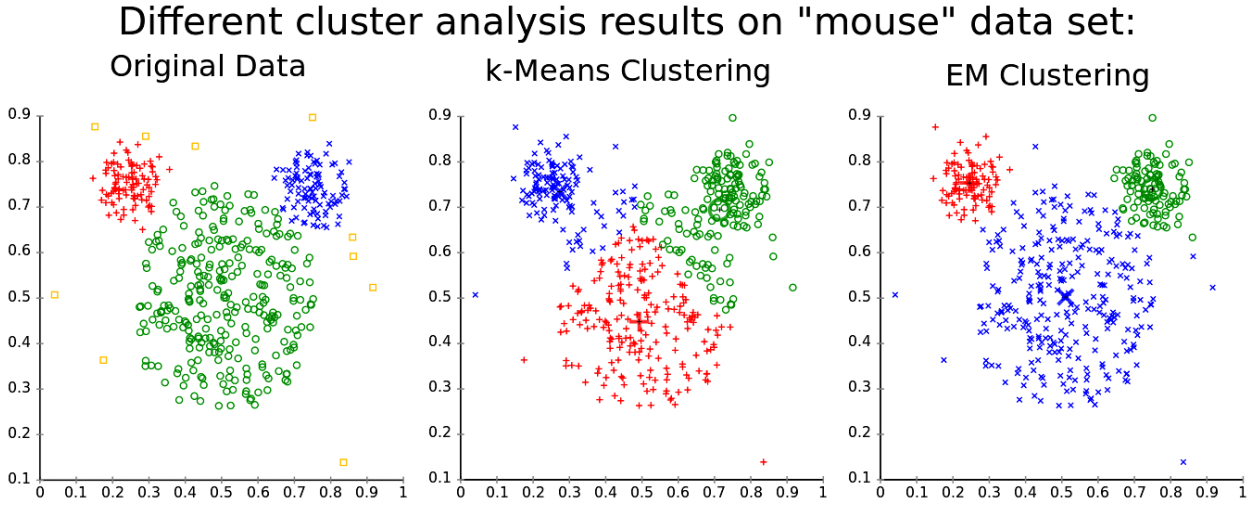
Ultima nota. Osservando questo gioco notiamo che ogni gruppo è libero di ruotare sul suo asse. Analogamente nel dendrogramma ogni cluster è libero di ruotare sulla sua radice e quindi il dendrogramma può essere rappresentato in modi diversi. I due rappresentati più sotto sono perfettamente identici. Quindi non facciamo ingannare dalla vicinanza dei rami in senso orizzontale. Ciò che conta è quanto i rami salgono per unire i diversi oggetti.





Analisi non-gerarchica dei gruppi - K-means Cluster Analysis

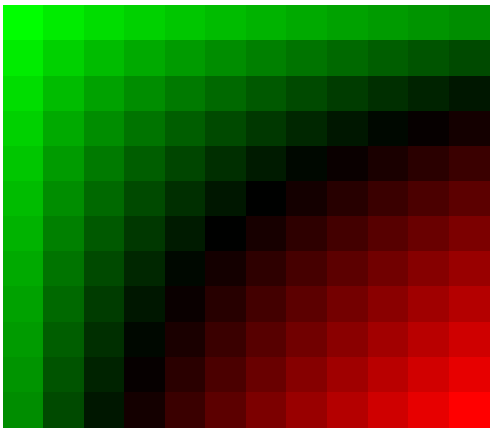
Lo scopo è quello di suddividere l'intero set di oggetti in un numero predeterminato di K gruppi. Il procedimento è iterativo e dipende dalle condizioni di partenza (solitamente affidate ad un sorting casuale) per cui produce classificazioni che possono di volta in volta essere leggermente differenti. Inoltre esiste una gran quantità di varianti di K-means clustering, basati su differenti algoritmi e con diverse proprietà. La raccomandazione è quella di fare prove preliminari su campioni noti e verificare la bontà del metodo per lo specifico tipo di dati in esame. Il grafico sotto mostra diversi risultati ottenuti con due metodi (EM sta per expectation–maximization algorithm).



Heat-map

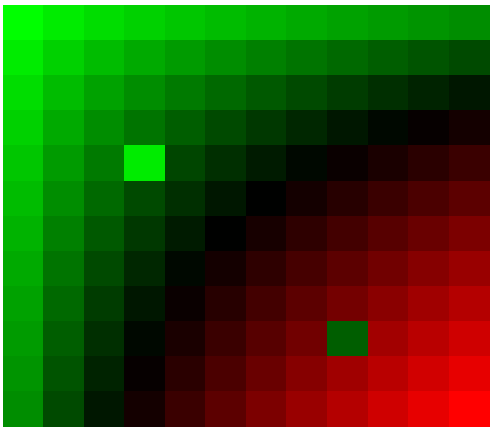
Quando occorre valutare a colpo d'occhio l'andamento di centinaia, migliaia o anche milioni di dati è utile il ricorso alla heat-map. In sostanza i grafici heat-map non sono altro che tabelle numeriche in cui i numeri sono sostituiti da colori.
Per esempio la tabellina pitagorica trasformata in heat-map appare così:

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	90	100	110	120
11	11	22	33	44	55	66	77	88	99	110	121	132
12	12	24	36	48	60	72	84	96	108	120	132	144



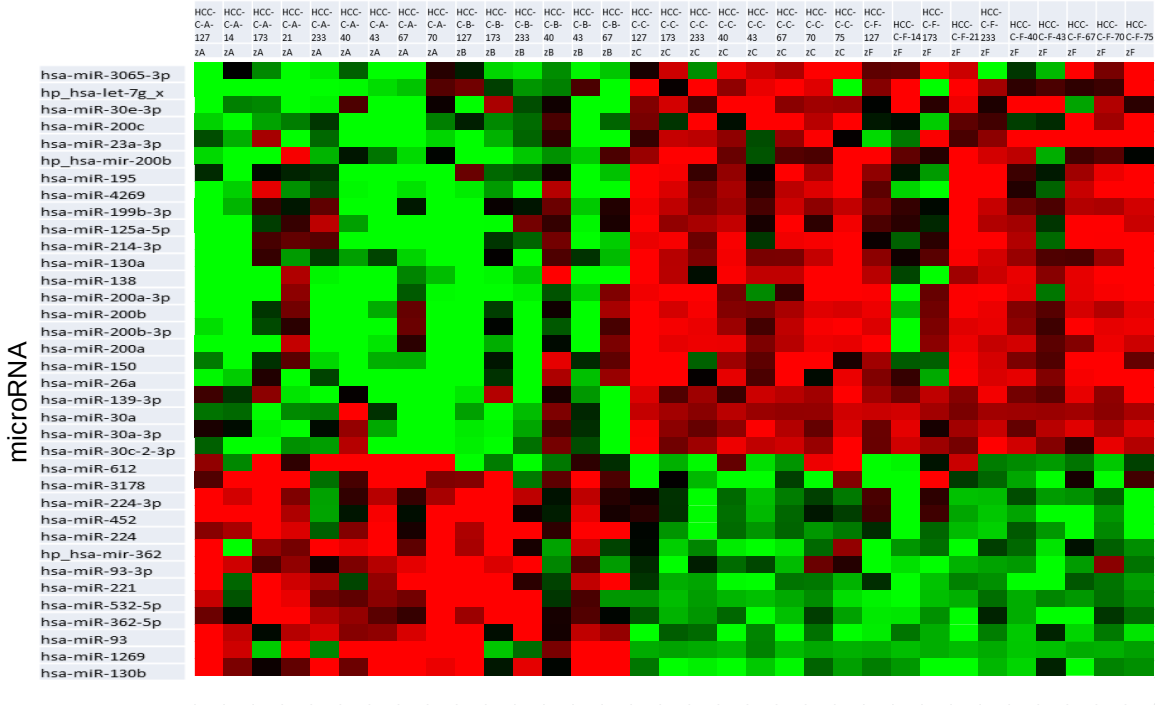
Cosa succede se alteriamo alcuni dati della tabella? Osservando la tabellina è piuttosto faticoso rendersi se e quali valori sono cambiati. La heat-map li individua all'istante.

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	2	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	20	100	110	120
11	11	22	33	44	55	66	77	88	99	110	121	132
12	12	24	36	48	60	72	84	96	108	120	132	144



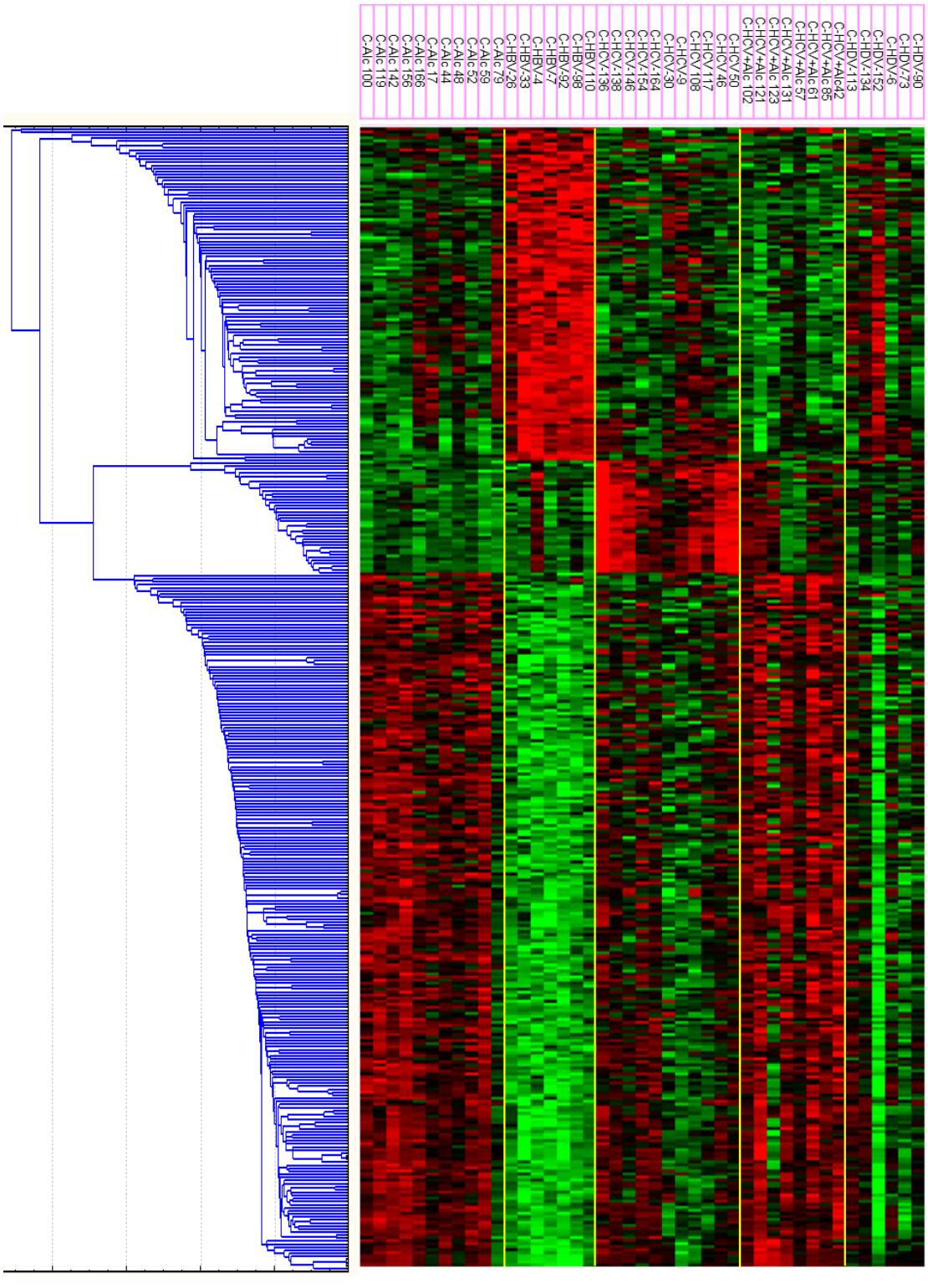
Nel campo della bioinformatica l'heat-map è estremamente utile in quanto consente di controllare a colpo d'occhio l'espressione di un numero impressionante geni di un gran numero di soggetti.

Questa ad es. è una heat-map che rappresenta l'espressione di 36 geni (si tratta in effetti di microRNA) di 33 soggetti.



La heat-map mostra chiaramente la presenza di due gruppi distinti di soggetti (a destra con HCC e a sinistra con cirrosi) ed allo stesso tempo due gruppi di microRNA (in alto ed in basso) upregolati (rossi) e downregolati (verdi).

Perché la heat-map sia efficace occorre però che i dati siano disposti in modo da avere gruppi di variabili e di soggetti il più possibile omogenei. Questo può essere fatto in due modi: (a) disponendo i dati secondo l'ordinamento ottenuto da una cluster analysis oppure (b) in base alla conoscenza di classi di dati note, ad es. parlando di geni, raggruppando assieme i geni dell'immunità innata, dell'immunità acquisita, del ciclo cellulare, ecc., ed i soggetti con determinate patologie. In questo grafico i geni (righe) sono stati ordinati in base alla sequenza del dendrogramma, mentre i casi (colonne) sono stati ordinati per tipo di patologia (cirrosi alcoolica, cirrosi virale, cirrosi associata a HCC, HCC con diverse eziologie virali).

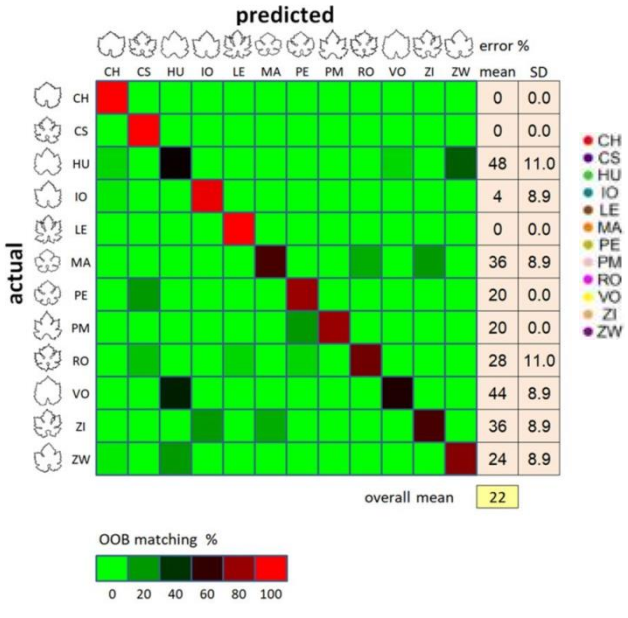


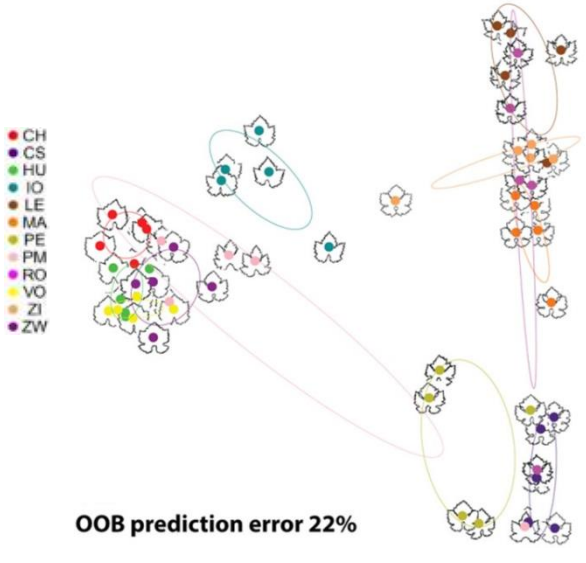
Random Forest

Quando la distribuzione dei dati è fortemente irregolare, non-normale l'analisi discriminante o anche sue varianti più flessibili si rivelano inefficaci. Per fare predizioni un metodo estremamente interessante è il Random Forest. Rispetto ad altri metodi multivariati, Random Forest

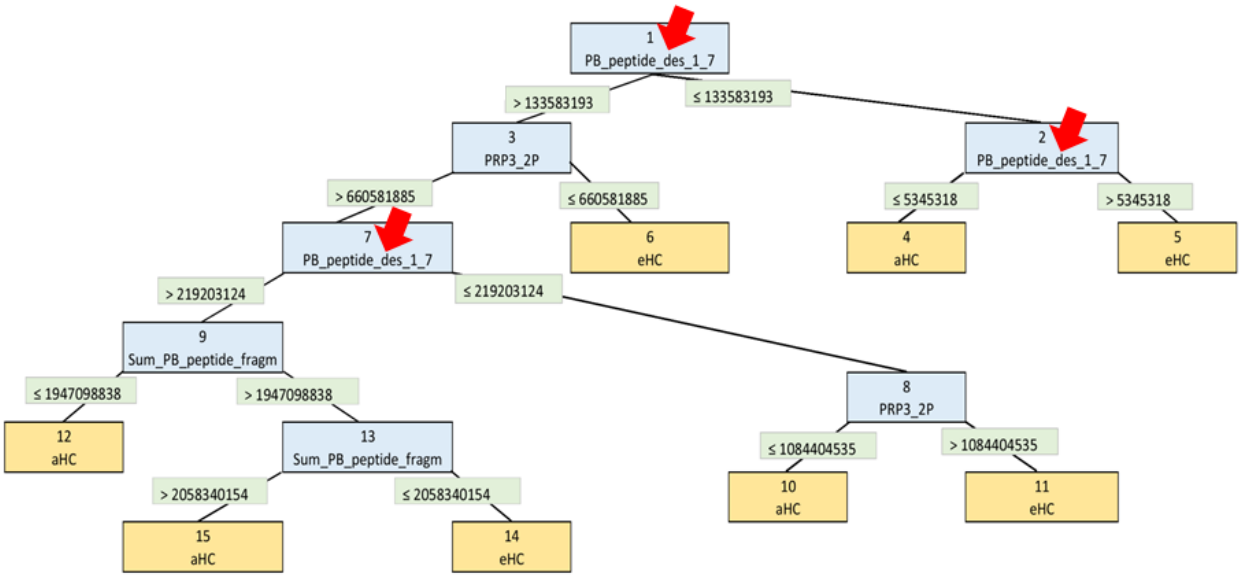
- non è condizionato dalla distribuzione dei dati, né da differenze tra le varianze dei gruppi, né da relazioni non lineari (cioè, i dati non richiedono variazioni di scala, trasformazioni lineari o non lineari o in rango o altre modifiche)
- può essere applicato sia a variabili categoriche sia a variabili continue
- ha un basso rischio di overfitting (adattamento solo apparente dei dati ad un modello)
- non richiede l'uso altri campioni per la cross-validazione dei risultati, in quanto ogni albero viene costruito omettendo di volta in volta su un piccolo numero di campioni selezionati casualmente, e poi testato su quei campioni non utilizzati per costruire l'albero (questo metodo è detto OOB, out-of-bag, 'fuori-sacco') e con questi calcolare il tasso di misclassificazione (OOB prediction error)
- consente di calcolare l'importanza relativa di ciascuna variabile intesa come la diminuzione di accuratezza della classificazione che si ottiene escludendo quella variabile dal calcolo (ma attenzione a quanto detto alla fine di questo capitolo)
- consente di ottenere una rappresentazione grafica, dimensionalmente ridotta dei risultati mediante multidimensional scaling o clustering gerarchico, usando la matrice delle distanze tra le osservazioni, prese a due a due (le distanze tra le osservazioni si ottengono come i valori complementari alle somiglianze tra le stesse osservazioni, queste ultime calcolate come le frequenze normalizzate degli alberi che contengono le due osservazioni nello stesso nodo finale)

L'immagine a sinistra mostra la heat-map dei valori di predizione di alcune varietà di vite ottenute mediante Random Forest applicato a parametri della forma della foglia. I colori, dal verde al rosso, indicano il tasso di predizione (dal verde chiaro = 0%, al rosso=100%). Una perfetta classificazione risulterebbe come una diagonale tutta rossa ed il resto della griglia verde chiaro. Le colonne marginali indicano l'errore di predizione per ciascuna varietà. L'immagine a destra rappresenta il multidimensional scaling ottenuto dalla matrice delle distanze tra le varietà.





Ma ci sono anche aspetti negativi. Random Forest è spesso paragonato ad un black box, nel senso che da esso escono risultati anche buoni o ottimi ma non è possibile ottenere una formula o funzione in grado di riprodurre quei risultati. La classificazione/predizione di nuovi campioni si può fare ma solo inserendo i nuovi dati nella 'pentola' dell'algoritmo che ha già 'cucinato' il primo set di dati (set di apprendimento). In effetti una formula o funzione non c'è proprio, in quanto Random Forest non utilizza funzioni ma la logica degli alberi decisionali, migliaia di alberi decisionali (da cui il termine Random Forest) su tanti piccoli subset di dati selezionati a caso. L'immagine sotto mostra uno specifico albero in cui sono presenti otto soggetti appartenenti a due gruppi (in giallo) e tre variabili (in celeste), estratti a caso tra i tanti analizzati (varie decine di soggetti e decine di variabili). E' importante osservare come una stessa variabile, ad es. quella indicata dalle frecce rosse, sia spezzata in tre punti per separare i gruppi.



Per quanto il procedimento possa sembrare inconcludente, sommando i risultati di tutti gli alberi si ottengono ottime classificazioni/previsioni validate internamente dai campioni OOB, e volendo validate anche da set di dati esterni. Inoltre Random Forest è in grado di valutare quanto una variabile sia importante per la classificazione/previsione, calcolando il calo di accuratezza della classificazione escludendo quella variabile dall'algoritmo. Ma nasce un problema: è possibile che una variabile importante per la classificazione possa non mostrare differenze significative se confrontiamo i gruppi con i normali test a ipotesi (es., t-test o Mann-Whitney o altro). Questo è un autentico paradosso e di questo purtroppo non c'è traccia nella letteratura statistica. Per capire e superare il problema, supponiamo di avere questi due campioni:

gruppo A	2	3	4	5	3	4	6	1	5	4	3	4	13	15	13	13	10	12	14	12	11
gruppo B	5	6	7	8	7	6	5	9	11	8	4	6	7	8	9	10	7	9	8	8	9

	gruppo A
Mean	7.5
Median	5.0
Std Dev.	4.7
25th Percentile	3.5
75th Percentile	12.5
Minimum	1.0
Maximum	15.0

	gruppo B
Mean	7.5
Median	8.0
Std Dev.	1.8
25th Percentile	6.0
75th Percentile	9.0
Minimum	4.0
Maximum	11.0

Le medie dei due gruppi non variano mentre variano altri parametri (deviazioni standard, mediane, ecc.)
Se ora facciamo un t-test o un test Mann-Whitney...

2-tailed t-Test

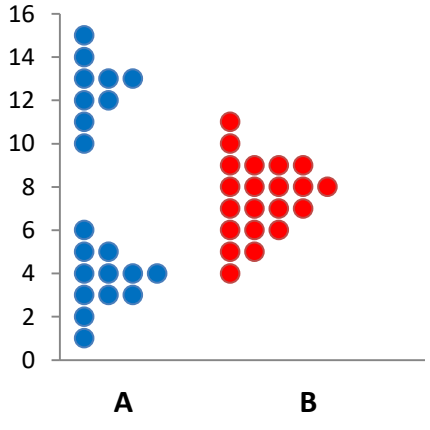
Ho. Diff	Mean Diff.	SE Diff.	T	DF	P
0.000	0.000	1.098	0.000	25.399	**1.000**

Sample variances differ at the specified alpha of 0.0500
so the individual variances are used in the t-Test.

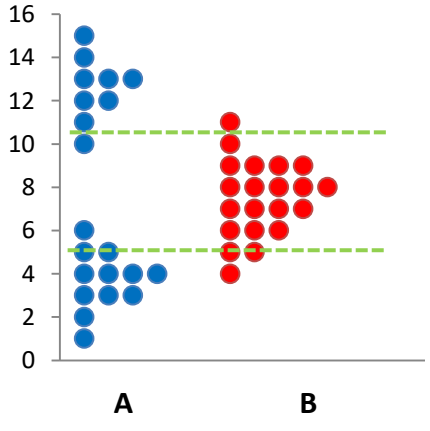
2-tailed Mann-Whitney test

	U	DF 1	DF 2	P
	243.500	21	21	**0.584**

...nessuno dei due test risulta significativo, ma se applichiamo Random Forest potrà risultare che questa variabile sia molto importante per la classificazione dei due campioni. Infatti se osserviamo il grafico di sotto notiamo che il gruppo A ha in genere valori più piccoli e più grandi rispetto al gruppo B.



In questa situazione, una buona separazione dei due gruppi si può ottenere stabilendo due soglie:



Questa operazione è normale negli alberi decisionali. In tal modo Random Forest ottiene una classificazione efficiente con basse probabilità di errore e quindi la variabile in questione potrà risultare importante per la classificazione. Tuttavia i confronti con i test a ipotesi non funzionano

allo stesso modo, e possono risultare negativi. Questo contrasto è piuttosto grave in quanto la ricerca di potenziali marker o 'signature' passa di norma attraverso singole variabili i cui valori superano (o scendono sotto) una certa soglia. Le soglie possono essere anche più di una ma sempre associate monotonicamente (positivamente o negativamente) al grado di criticità (es., livello normale, di attenzione, di allarme) in campo clinico, ambientale, ecc. Per questo l'importanza delle variabili nel procedimento di classificazione operato da Random Forest non può essere utilizzata per l'individuazione di potenziali marker. Questo vale anche per altri metodi (ad es., Support Vector Machine) che appartengono a quella branca dell'intelligenza artificiale detta Machine Learning. Questi metodi in genere utilizzano algoritmi e modelli che hanno un elevato grado di adattamento e apprendimento. Tuttavia la loro notevole flessibilità ed efficienza passa attraverso una forte segmentazione delle scale che non consente di attribuire precisi significati alle variabili d'origine considerate singolarmente. In realtà, la natura e soprattutto i fenomeni biologici sono sistemi molto complessi e probabilmente i metodi di Machine Learning hanno ragione di operare in quel modo. Ma d'altra parte, per gli attuali usi diagnostici, c'è necessità di marker associati con relazioni semplici alle condizioni critiche.

Standardizzazione e centraggio

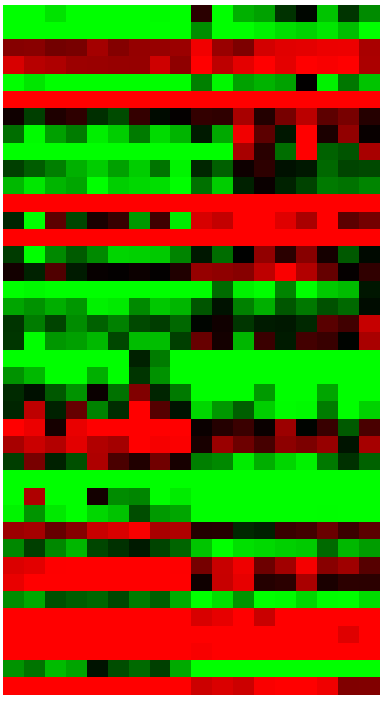
La standardizzazione, come abbiamo già visto, trasforma i dati in modo tale che la loro media diventi 0 e la deviazione standard diventi 1. Questa trasformazione può essere necessaria quando le unità di misura (es. altezza in cm, età in anni, peso in grammi, ecc.) o le scale o anche semplicemente le varianze (es. pH del plasma, pH delle urine, ecc.) sono diverse. Il centraggio invece consiste semplicemente nel sottrarre a tutti i dati la loro media. In questo modo le medie diventano uguali (zero) mentre le deviazioni standard restano invariate. Normalmente la standardizzazione o il centraggio sono applicati alle variabili, per attribuire loro un peso relativamente omogeneo e poterle confrontare. Ma le stesse trasformazioni possono anche essere applicate ai casi facendo però attenzione a ciò che questo comporta. Se ad es. standardizziamo/centriamo i casi possiamo esplorare alcune relazioni tra soggetti ma al contempo distruggiamo l'informazione contenuta nelle variabili. Possiamo anche standardizzare/centrare sia le variabili che i casi. In quest'ultimo caso, molto poco frequente, si parla di matrice bi-standardizzata o bi-centrata.

Esempio di dati centrati e dati standardizzati:

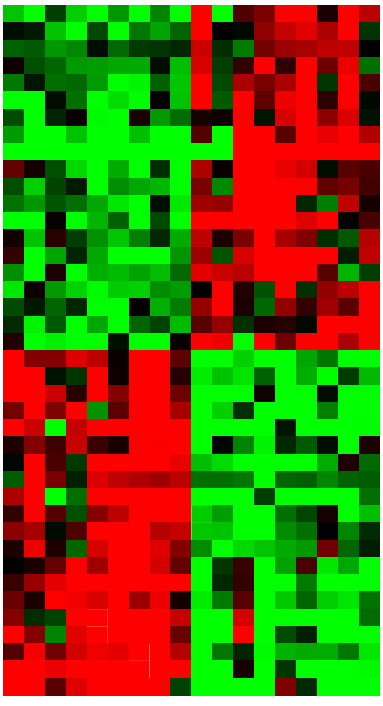
dati originali	2	4	3	1	2	6	media =	3	deviazione standard = 1.789
dati centrati x-media	-1	1	0	-2	-1	3	media =	0.00	deviazione standard = 1.789
dati standardizzati x-media/deviazione standard	-0.56	0.56	0.00	-1.12	-0.56	1.68	media =	0.00	deviazione standard = 1.000

Normalmente le variabili sono disposte nelle colonne ed i casi nelle righe. Ma nelle scienze 'omiche' dato il gran numero di variabili (migliaia di geni, centinaia di migliaia di SNP, ecc.) queste sono disposte nelle righe ed i casi (che sono molto meno) nelle colonne. E' solo questione di praticità, così come un testo occupa un numero virtualmente illimitato di righe, facili da scorrere, piuttosto che di colonne.

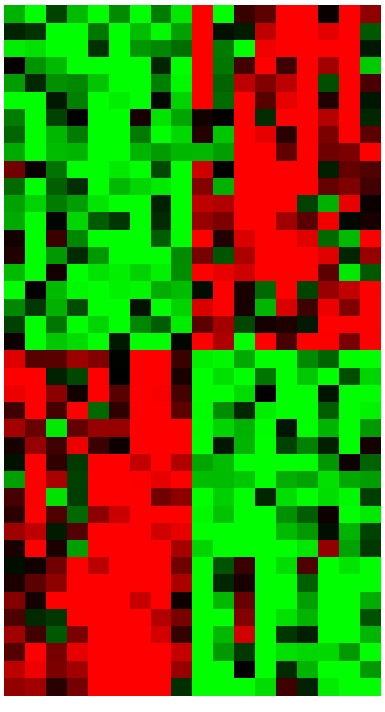
Vediamo in pratica gli effetti di queste trasformazioni su una heat-map che rappresenta 40 geni (righe) e 18 casi (9 pazienti e 9 controlli, colonne)



Valori originali
Questa prima heat-map, che rappresenta i valori originali, ci dice quali geni sono più espressi (righe prevalentemente rosse) e quali meno espressi (righe prevalentemente verdi) in assoluto. Le differenze tra i geni sono tanto forti che quasi non si notano le differenze tra i gruppi di soggetti. In altre parole, se una riga è tutta prevalentemente rossa, o verde, non apprezziamo differenze tra le colonne.



Righe (geni) centrate
Questa seconda heat-map centra le medie dei geni e quindi elimina le relative differenze tra geni. Non ci sono più righe tutte rosse o tutte verdi. Ma le deviazioni standard restano invariate, per cui è possibile apprezzare quali geni variano maggiormente relativamente alla media (righe con l'intera gamma dei colori, dal verde pieno al rosso pieno, vedi ad esempio la nona riga) e quali variano poco (righe con colori che variano poco, es tra il rosso scuro ed il verde scuro, vedi la terza riga).



Righe (geni) standardizzate
Quest'ultima heat-map non solo pareggia le medie ma anche le deviazioni standard dei geni. Tutte le righe hanno quindi la stessa gamma di colori (dal verde pieno al rosso pieno, anche la terza riga!). Non possiamo più fare alcun confronto tra i geni, che sono del tutto uniformati, ma in compenso possiamo valutare molto bene differenze tra i soggetti e quindi tra i gruppi.

Nota. Nelle heat-map mostrate il range dei colori, che va dal verde al rosso saturo, corrisponde all'intervallo tra il 20° e l'80° percentile di tutti i dati. Ma sia il tipo dei colori sia il range per le tonalità possono variare. Comunque il range dei valori adottato deve sempre accompagnare il grafico.

Software statistico

- SPSS: uno dei classici è più completi software statistici, estremamente documentato.

- XLStat (Excel addin): completo ma costoso e molto invadente (inserisce automaticamente comandi ActiveX nello spreadsheet, fogli nascosti e macro).

- R: è in assoluto il migliore per problemi di una certa complessità e soprattutto per applicazioni multivariate. Ha ottime qualità grafiche. E' del tutto gratuito (http://www.R-project.org) ed ha un'ottima documentazione. Richiede solo un piccolo impegno iniziale. Sotto Windows il migliore ambiente d'uso e di sviluppo è il programma RStudio: (https://www.rstudio.com/products/rstudio/download2/#download).
  Ma diffidate degli addin che dicono di portare R su Excel (es. RExcel). Sono molto limitati.
  Come documentazione posso consigliare:
  - F. Frascati - Formulario di Statistica con R. 2008 - https://cran.r-project.org/doc/contrib/Frascati-FormularioStatisticaR.pdf
  - M. Dell'Omodarme - Esercitazioni di statistica biomedica - alcune note su R. 2012 - https://cran.r-project.org/doc/contrib/DellOmodarme-esercitazioni-R.pdf
  - J. Oksanen - Multivariate Analysis of Ecological Communities in R: vegan tutorial. 2015 - cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

- Statistica for Windows: ottimo e completissimo.

- Statgraphics (Centurion): buono e piuttosto completo. Soffre un po' della rigidità del vecchio programma che girava sotto DOS.

- GPower: il migliore che io conosca per valutare l'errore β e determinare la grandezza del campione appropriata.

- GraphPad - Prism: uno dei più utilizzati nei laboratori.

- RealStats 2010 (Excel addin) (www.real-statistics.com): ottimo e gratuito. Oltre alla ampia gamma di applicazioni, è molto positivo il fatto che i risultati siano dinamicamente associati ai dati di input, per cui modificando i dati i risultati si aggiornano automaticamente. In Excel questa capacità la troviamo nelle funzioni ma non negli strumenti di analisi statistiche di Excel, che producono solo dei reports statici. RealStats offre due tipi di uso: (1) quello delle analisi da menù a tendina, che producono risultati molto ben documentati ma che per questo stesso motivo riempiono numerose celle del foglio Excel, oppure (2) quello di funzioni compattissime, che consentono di fare test anche molto complessi con il risultato finale dentro un'unica cella di Excel. La lista di queste funzioni è disponibile solo nel sito. Il vantaggio di questo secondo metodo è quello di applicare test complessi a numerose serie di dati, mediante copia e incolla, ottenendo i risultati sulle stesse righe/colonne dei dati. Unica pecca: RealStats non ha grafica.

- StatiXL (Excel addin): ottimo, quasi gratuito.

- Excel: alla base di tutto.

Alcuni dei grafici in copertina sono tratti da questi articoli:

Diaz G, Engle RE, Tice A, Melis M, Montenegro S, Rodriguez-Canales J, Hanson J, Emmert-Buck MR, Bock KW, Moore IN, Zamboni F, Govindarajan S, Kleiner DE, Farci P. Molecular signature and mechanisms of hepatitis D virus-associated hepatocellular carcinoma. **Mol Cancer Res.** 2018 Sep;16(9):1406-1419. doi: 10.1158/1541-7786.MCR-18-0012. Epub 2018 Jun 1. PMID: 29858376

Diaz G, Zamboni F, Tice A, Farci P. Integrated ordination of miRNA and mRNA expression profiles. **BMC Genomics.** 2015 Oct 12;16:767. doi: 10.1186/s12864-015-1971-9. PMID: 26459852

Farci P, Wollenberg K, Diaz G, Engle RE, Lai ME, Klenerman P, Purcell RH, Pybus OG, Alter* HJ. Profibrogenic chemokines and viral evolution predict rapid progression of hepatitis C to cirrhosis. **Proc Natl Acad Sci U S A.** 2012 Sep 4;109(36):14562-7. doi: 10.1073/pnas.1210592109. Epub 2012 Jul 24. PMID: 22829669

Diaz G, Melis M, Tice A, Kleiner DE, Mishra L, Zamboni F, Farci P.
Identification of microRNAs specifically expressed in hepatitis C virus-associated hepatocellular carcinoma. **Int J Cancer.** 2013 Aug 15;133(4):816-24. doi: 10.1002/ijc.28075. Epub 2013 Mar 16. PMID: 23390000

Nissim O, Melis M, Diaz G, Kleiner DE, Tice A, Fantola G, Zamboni F, Mishra L, Farci P. Liver regeneration signature in hepatitis B virus (HBV)-associated acute liver failure identified by gene expression profiling. **PLoS One.** 2012;7(11):e49611. doi: 0.1371/journal.pone.0049611. Epub 2012 Nov 21. PMID: 23185381

* HJ Alter è stato premio Nobel per la Medicina 2020