Marco Besozzi

Un primer di biostatistica con la guida all'uso di Ministat 2.1

Il software Ministat 2.1 si trova sul sito www.bayes.it nell'area di download, ed è installabile su PC con sistema operativo Windows NT®, Windows 2000, Windows Xp®, WindowsVista®.

INDICE

CAPITOLO 6

FORMULE E ALGORITMI DI CALCOLO

CAPITOLO 1
DATI, INFORMAZIONE E CONOSCENZA

DOVE È LA SAPIENZA
CHE ABBIAMO SMARRITA NELLA CONOSCENZA

DOVE È LA CONOSCENZA
CHE ABBIAMO SMARRITA NELL'INFORMAZIONE

DOVE È L'INFORMAZIONE
CHE ABBIAMO SMARRITA NEI DATI

Questa frase di Mark Porat, incisa nell'ala delle Comunicazioni del Museo Scientifico dei Bambini, a Washington, ha un significato profondo.

Letta dall'alto verso il basso è la metafora della società moderna. I ritmi di vita sempre più serrati, hanno prima rotto (per sempre?) quell'armonico senso di unità tra corpo, spirito e natura che gli antichi definivano *sapienza*. Quindi hanno imposto (stanno sempre più imponendo) la necessità di suddividere l'enorme quantità di *conoscenza* globalmente disponibile, in sottoinsiemi (ecco che nasce la figura dello specialista) di dimensioni compatibili con la limitatezza della capacità di gestione della nostra mente (la corteccia cerebrale umana, ancorché sia una delle macchine più incredibilmente complesse e sofisticate presenti in natura, possiede un limite fisico finito, invalicabile). Ancora, i ritmi di vita sempre più serrati hanno imposto (stanno sempre più imponendo) la necessità di comprimere i processi di comunicazione. Il paradigma è rappresentato dalla comunicazione mediante immagini. Il processo di comunicazione aumenta in efficienza[1], e aumenta in efficacia[2]. Ma la comunicazione mediante immagini, a fronte di un aumento dell'efficienza e dell'efficacia del processo di comunicazione, comporta, ahimè, una perdita del dettaglio[3]. Inoltre il messaggio "visivo" ricevuto viene tendenzialmente percepito come compiuto: la società dell'immagine spinge quindi anche a perdere il valore aggiunto (ricchezza inestimabile) determinato dalla riflessione/analisi critica dell'individuo sul suo contenuto. La conoscenza ha ormai lasciato il posto all'*informazione*. L'informazione fine a sé stessa, l'informazione-business, è il grande fratello orwelliano già presente tra noi (ma quanti se ne sono accorti?). Ma c'è di peggio nell'aria. Il flusso informativo è talmente imponente, rispetto al tempo dedicato/dedicabile alla riflessione critica su di esso da parte del singolo individuo, da rischiare di diventare semplicemente un flusso di *dati*.

Letta dal basso verso l'alto la frase di Porat rappresenta la metafora della crescita e dello sviluppo dell'individuo. Già nell'utero materno il feto riceve in continuazione *dati* provenienti dalla madre. Fino a un certo punto solo di dati si tratta. Il battito del cuore materno che rimbomba regolarmente, il rumore dell'aria che entra ed esce dagli alveoli polmonari. Appena nato però il bambino impara

---

[1] *Per trasmettere il messaggio è necessario meno tempo, e meno tempo è necessario per memorizzarlo. In altre parole è necessario meno lavoro per inviare e per ricevere il messaggio.*

[2] *La comunicazione mediante immagini aumenta anche i risultati ottenuti (l'outcome). Basta pensare all'efficacia di uno spot, in grado, se ben concepito, di indirizzare milioni di consumatori verso un prodotto, spesso solo voluttuario.*

[3] *Come conseguenza della compressione, lo vedremo più avanti, si ha la perdita di parte del contenuto informativo del messaggio.*

subito ad integrare (operazione matematica) i dati provenienti dal mondo esterno. Lo fa perché solo deve sopravvivere. Pur totalmente imbelle, deve imparare immediatamente ad integrare i dati, ancora per lui confusi, provenienti dal mondo esterno (per ora solo la madre) al fine di avere una prima *informazione* essenziale: dove trovare il cibo. E impara subito a cercare il seno e a succhiare. Altre informazioni egli dovrà presto imparare a generare, integrandoli, dai dati che gli provengono dal mondo in cui vive. La scuola lo aiuterà poi ad integrare le informazioni al fine di generare la *conoscenza* che gli servirà per inserirsi definitivamente nella società. Non tutti ci riescono bene. Pochi privilegiati infine riescono ad effettuare l'integrazione/salto logico che separa il livello della conoscenza da quello della *sapienza*.

L'obiettivo di questo capitolo è di presentare le basi dei processi che consentono di effettuare il passaggio *dati* ⇨ *informazione* ⇨ *conoscenza*, trasformando prima i *dati* in *informazione*, e successivamente questa in *conoscenza*.

1.1. Astrarre, modellizzare e connettere

Prima di continuare vale la pena di ricordare le definizioni di *dato*, *informazione* e *conoscenza* fornite dallo Zingarelli[4].

Dato ⇨ *elemento o serie di elementi accertati e verificati che possono formare oggetto di indagini, ricerche, elaborazioni o che comunque consentono di giungere a determinate conclusioni*.

Informazione ⇨ *atto, effetto dell'informare o dell'informarsi* [essendo a sua volta] ***informare*** ⇨ *modellare secondo una forma*.

Conoscenza ⇨ *facoltà, atto, modo, effetto del conoscere* [essendo a sua volta] ⇨ ***conoscere*** *prendere possesso intellettualmente o psicologicamente, specialmente con un'attività sistematica, di qualunque aspetto di quella che è considerata realtà*.

Si sottolinea il fatto che questo capitolo rinuncia deliberatamente all'analisi, in quanto assolutamente impervia, dell'elemento al vertice della frase di Porat, la ***sapienza*** [definita come] ⇨ *il più alto grado di conoscenza delle cose* [ovvero definita come] ⇨ *sapere vasto e profondo uniti a doti morali e spirituali*.

È necessario infine, come ultimo passo, dare una definizione del termine ***intelligenza*** [definita dallo Zingarelli come] ⇨ *capacità generale che consente di adattarsi all'ambiente e che nell'essere umano si manifesta nei comportamenti e nel grado di elaborazione dei processi mentali*. Ricordando che il temine intelligenza deriva dal latino *intelligere*, per il quale lo stesso Zingarelli rimanda alla voce ***intelletto*** ⇨ *facoltà di intuire le idee, le rappresentazioni e i loro rapporti*.

La corrispondenza tra gli stati attraversati dalle rappresentazioni mentali sulla via dell'*intelligere* (*dati* ⇨ *informazione* ⇨ *conoscenza*) e le azioni intraprese dall'individuo sono le seguenti:

Rappresentazione mentale	*Deriva dall'azione di…*
Dato	*…ricerca [astrazione]*
Informazione	*…modellare secondo una forma [modellizzazione]*
Conoscenza	*…prendere possesso intellettualmente … della realtà [connettere]*

[4] *Vocabolario della lingua italiana di Nicola Zingarelli. Bologna: Zanichelli, 1994:2144pp.*

Il dato è la prima rappresentazione mentale, quella più elementare. Ma è già di per sé una grande conquista. Il dato è il risultato di un processo di astrazione. Fornire una tabella che riporta l'altezza in centimetri di 1000 soggetti appartenenti alla popolazione italiana, corrisponde a fornire un elenco di dati. Ma esprimere "l'altezza in centimetri" presuppone un processo di misura, che altro non è se non un'astrazione. Essendo che, come vedremo avanti, "la misura (dell'entità) di una grandezza fisica" consiste nell'esprimere la grandezza in modo quantitativo, dando ad essa un "*valore numerico*" che è un numero puro, ottenuto per confronto della "*(entità della) grandezza* in esame" con la "(entità di una) grandezza di riferimento ad essa omogenea, definita *unità di misura*". Un concetto astratto, che consente peraltro di dare misurabilità ai dati[5].

Dire che "l'altezza media, in un campione rappresentativo di 1000 italiani, è di 168,4 centimetri" corrisponde a dare un'informazione. Informazione che deriva dall'avere dato forma ai dati, applicando ad essi un qualche modello di distribuzione (per esempio il modello di distribuzione proposto da Gauss, la distribuzione gaussiana). È importante sottolineare che il dare forma ai dati, che di per sé è concepito come un fatto neutro, in taluni casi può portare a conseguenze/conclusioni altamente fuorvianti[6]. Questo accade quando il dare forma, dovuto al ricercatore, prende il sopravvento sulla forma, intrinseca, soggiacente, ma incognita, dei dati. In questi casi il desiderio legittimo di dare forma ai dati, può portare ad alterarne il loro contenuto informativo primigenio, ad introdurre nei dati una distorsione. Avremo modo di soffermarci abbondantemente su questo concetto/pericolo quando parleremo del disegno sperimentale.

Il terzo elemento, la conoscenza, intesa come "prendere possesso intellettualmente … della realtà [connettere]" è il più complesso. Per capirlo è opportuno ricorrere alla teoria dei sistemi, che definisce un sistema come un "insieme di parti connesse da relazioni", e nel quale "l'insieme rappresenta qualcosa di più della somma delle singole parti"[7]. In altre parole viene riconosciuto che le relazioni che interconnettono le parti sono in grado di fornire valore aggiunto al sistema. Nel campionato di calcio, nel campionato di basket, nel campionato di baseball, praticamente ogni anno esistono almeno due squadre con lo stesso numero di "altrettanto buoni" giocatori. Ma è frequente osservare come una squadra ottenga buoni risultati mentre l'altra non riesce ad ottenere risultati. La spiegazione risiede nel fatto che nel primo caso le relazioni (quelle tra i singoli giocatori, e quelle tra questi e l'allenatore) forniscono un valore aggiunto maggiore che nel secondo. Si noti che il termine relazioni non viene utilizzato per descrivere relazioni amicali, ma relazioni funzionali (anche se nel caso specifico le relazioni amicali svolgono sicuramente un ruolo importante). A parità di bontà dei giocatori, gli allenatori sono pagati proprio per massimizzare il valore aggiunto fornito dalle relazioni [funzionali] tra gli elementi/parti della squadra/sistema. Nel caso specifico dell'esempio, conoscere l'altezza degli italiani significa non solo possedere i dati (altezza di 1000 individui), non solo dare loro forma mediante un modello matematico (distribuzione gaussiana) dal quale derivare l'informazione riguardante la loro altezza media, ma anche ricordare che i sardi sono mediamente più bassi e i friulani mediamente più alti della media degli italiani. Dalla fusione delle

---

[5] *Non importa che la misura sia basata su una unità di misura "arbitraria" (come ad esempio il metro e il kilogrammo). Un dato misurabile può finalmente essere confrontato con un altro dato, essendo a questo punto il confronto relativo di due dati reso "oggettivo" dal concetto soggiacente di misurabilità, che consente di eliminare l'arbitrarietà del giudizio.*

[6] *Il concetto, qui applicato alla statistica, ha un valore assolutamente universale. Si pensi alle deviazioni introdotte dalla cultura magica medioevale, col dare forma di draghi, folletti e streghe ai dati della realtà. Ma si pensi anche a quanto accade ancora di questi tempi (per fortuna forse sempre meno) con le ideologie, in particolare con quelle con deviazioni di tipo totalitaristico o xenofobo. Gli assunti utilizzati per dare forma ai dati sono talmente forti da alterare la visione della realtà stessa a milioni di individui, inducendo comportamenti collettivi folli.*

[7] *Capezzuto D, Gianni D. Sistemi. Modellistica, comunicazione, misura e controllo. Milano: Hoepli, 1988:367pp.*

altezze dei sardi e di quelle dei friulani nel crogiolo degli italiani, deriva proprio questo valore aggiunto che né i singoli dati né l'informazione "l'altezza media, in un campione rappresentativo di 1000 italiani, è di 168,4 centimetri" sono in grado di dare. Seguendo la definizione data dalla teoria dei sistemi "l'insieme" delle altezze di 1000 italiani "rappresenta qualcosa di più della somma delle singole parti".

1.2. I principi forti che governano la conoscenza della realtà

Poiché il processo di conoscenza risulta inestricabilmente connesso con la realtà fisica nella quale esso si sviluppa, i principi che governano il mondo fisico finiscono inevitabilmente con il condizionare la conoscenza della realtà. Vediamo ora i fondamentali.

1.2.1. Il secondo principio della termodinamica

Perché facendo bollire un acquario si ottiene una zuppa di pesce, mentre raffreddando una zuppa di pesce non si ottiene un acquario? Perché un bicchiere va in mille pezzi, mentre mille pezzi non confluiscono un bicchiere? Perché una palla che rimbalza spontaneamente si ferma, mentre una palla ferma non si mette spontaneamente a rimbalzare?

Dall'osservazione che i corpi caldi spontaneamente si raffreddano, mentre non accade il contrario, cioè che i corpi freddi spontaneamente si riscaldino (sarebbe bello se i termosifoni si scaldassero spontaneamente), Clausius nel 1750 pose le basi per la comprensione scientifica di una asimmetria che caratterizza la natura del mondo in cui viviamo. Mentre tutto il lavoro può essere trasformato in calore, non accade il contrario: non è cioè possibile trasformare completamente il calore in lavoro.

Boltzmann nel 1880 definisce in termini matematici la legge nota come secondo principio della termodinamica[8], e introduce in concetto di entropia: in un sistema fisico termodinamicamente chiuso, come l'universo, l'entropia, cioè il disordine, può solo aumentare.

Se finora tutti i tentativi di contraddire il secondo principio della termodinamica sono inesorabilmente caduti sotto il vaglio della verifica sperimentale, è pure evidente che lo sviluppo di un organismo dalla cellula uovo va in senso opposto all'aumento dell'entropia.

La spiegazione risiede nel concetto di "*contraddizione su base locale*" del principio[9]. In altre parole, il secondo principio della termodinamica vale e continua a valere globalmente per un sistema termodinamicamente chiuso come l'universo, anche se sistemi termodinamicamente aperti come gli esseri viventi lo possono contraddire su base locale. Rimane un buffo mistero: quando da una

---

[8] *Atkins PW. Il secondo principio. Bologna:Zanichelli, 1992:227pp.*

[9] *Questo concetto è applicabile anche ad un altro principio forte, il principio di casualità Se si lancia la pallina nelle roulette un miliardo di volte, possiamo avere la "certezza" che cinquecento milioni di volte essa si fermerà sul rosso e cinquecento milioni di volte essa si fermerà sul nero. Tuttavia si può dare tranquillamente il fatto che, all'interno di questo miliardo di lanci, si abbia una sequenza nella quale la pallina si ferma per esempio 20 volte consecutive sul rosso. Questo evento rappresenta una contraddizione locale del principio di casualità, ma non lo inficia nella sua validità generale. Per inciso i concetti di "fortuna" e di "sfortuna" sono proprio legati a contraddizioni locali del principio di casualità. Se si pensa al fatto che per puro caso un giocatore potrebbe puntare per 20 volte sul rosso, "centrando" esattamente le 20 volte in cui esso esce consecutivamente, si avrà l'idea chiara di quello che si definirebbe un giocatore "sfacciatamente fortunato". Reciprocamente se per puro caso un giocatore puntasse per 20 volte sul nero, "centrando" esattamente le 20 volte in cui esce consecutivamente il rosso, si avrà l'idea chiara di quello che si definirebbe un giocatore "estremamente sfortunato".*

cellula uovo emerge progressivamente l'ordine di un nuovo essere vivente, visto che l'entropia, cioè il disordine globale dell'universo, può solo aumentare, chi ne paga le conseguenze? Forse nessuno, visto che l'ordine che si crea è solo una effimera bolla, che più o meno rapidamente ricade nel brodo del disordine primordiale dal quale era emersa.

La recente scoperta che sistemi rigorosamente deterministici possono evolvere rapidamente verso uno stato di caos, e che reciprocamente sistemi apparentemente caotici hanno un determinismo soggiacente, riassunta nella teoria dei sistemi caotici, ha aggiunto nuovi strumenti alla comprensione di questi fenomeni[10].

Ma ora vediamo due conseguenze del secondo principio della termodinamica sulla conoscenza della realtà, quando questa sia mediata attraverso gli strumenti matematici della probabilità e della statistica.

1.2.1.1. Dare ordine ai dati (tabulare i dati)

La prima conseguenza è che il "dare ordine" ai dati può portare ad un miglioramento del loro contenuto informativo (diminuzione dell'entropia). Questo banalmente accade quando, tabulando le altezze di 1000 individui, raccolte in modo disordinato, si provvede al loro ordinamento in ordine crescente. Dall'ordinamento emergono apparentemente "in modo spontaneo"[11] informazioni aggiuntive: l'altezza minima, l'altezza massima, le due altezze che, nelle posizioni 500 e 501 della lista ordinata di dati, corrispondono al "baricentro" dei dati (o, meglio, della loro distribuzione). Dare ordine ai dati significa dare loro vita.

1.2.1.2. Comprimere l'informazione in un dato statistico

Si considero ora l'espressione

$$2 + 2 = 4$$

e la domanda: "2 + 2 fa sempre 4" ?

Nel campo delle matematiche elementari è sicuramente vero che "2 + 2 fa sempre quattro". Ma in termini di informazione fornita questo non è vero. I due membri dell'uguaglianza sono "equivalenti" dal punto di vista dell'aritmetica, ma non lo sono dal punto di vista del contenuto informativo. Nel passaggio da *2 + 2* a *4* si ha un perdita di informazione o, se si preferisce, un aumento dell'entropia (disordine) che caratterizza l'informazione. Il compattare l'informazione porta al suo parziale degrado. È quello che accade quando al lettore viene fornito il valore della media dell'altezza di 1000 italiani (valore che chiameremo "un dato statistico" o "una statistica" del campione) in luogo dei 1000 valori delle altezze dei singoli individui. La ricchezza dell'informazione iniziale viene parzialmente persa comprimendola in un singolo dato statistico.

1.2.2. Il principio di casualità

Questo principio ci introduce all'essenza della probabilità e della statistica. Erwin Schroedinger (1887-1961) il fisico austriaco fondatore della meccanica ondulatoria, premio Nobel nel 1933 con

---

[10] *Casati G (a cura di). Il caos. Le leggi del disordine. Milano:Le Scienze, 1991:214pp.*

[11] *In realtà per generare l'ordine necessario per fare emergere le informazioni aggiuntive viene compiuto un lavoro. L'entropia negativa (aumento dell'ordine) dei dati è più che compensata dall'entropia positiva (aumento del disordine) derivante dal lavoro che la ha prodotta, e quindi in definitiva si ha un aumento dell'entropia globale dell'universo.*

Paul Dirac, diceva: "*...la ricerca in fisica ha mostrato, al di là di ogni dubbio, che l'elemento comune soggiacente alla coerenza che si osserva nella stragrande maggioranza dei fenomeni, la cui regolarità e invariabilità hanno consentito la formulazione del postulato di causalità, è il caso...*".

Reciprocamente si può dire che sembra esservi un determinismo soggiacente ai sistemi caotici, che tende a emergere in seguito al comportamento "collettivo" di eventi tra loro indipendenti. L'esempio più semplice è quello della roulette. Non è possibile sapere in anticipo, al lancio della pallina, se essa si fermerà sul colore rosso o sul colore nero (a meno che la roulette sia truccata: ma qui si assume che non lo sia). Il risultato di un singolo evento è puramente casuale. Tuttavia se si lancia la pallina diciamo un miliardo di volte, possiamo avere la "certezza" che cinquecento milioni di volte essa si fermerà sul rosso e cinquecento milioni di volte essa si fermerà sul nero.

In altri termini, quando ci si trova davanti a evento singolo quello che prevale è il *caso*; quando gli eventi sono molto numerosi, sembra esservi una *necessità* alla quale gli eventi finiscono con l'ubbidire. Questa è in estrema sintesi l'essenza della probabilità e della statistica, ed è anche, detto per inciso, la chiave di lettura che la biologia moderna ha adottato per spiegare l'emergere spontaneo della *vita/necessità* dal *caos/caso* primordiale, come magistralmente descritto da J. Monod[12].

Il sottile filo che lega la visione di Schroedinger (fisico) e di Monod (medico, anche lui premio Nobel) passa attraverso le sconvolgenti scoperte dei fisici dei primi decenni di questo secolo, le cui conseguenze sono ancora oggi oggetto di approfondimento sia teorico sia sperimentale, e le cui conseguenze pratiche sono alla base dello sviluppo delle moderne tecnologie, che stanno aprendo la strada ad innovazioni di portata epocale (i calcolatori quantici potrebbero essere ormai alla portata, mentre non à escluso che alcuni meccanismi, come quello che risiede alla base della memoria all'interno dei neuroni, siano realizzati a livello quantistico). Queste scoperte, che videro coinvolti Bohr, Einstein, lo stesso Schroedinger, e molti altri fisici, sono basate sull'osservazione che alcuni "strani" comportamenti della natura risultano descrivibili mediante leggi esprimibili solamente in termini probabilistici. L'argomento è di interesse fondamentale per comprendere fino in fondo come la probabilità e la statistica, al di là degli apparentemente astrusi apparati matematici, siano gli unici strumenti in grado di fornire una descrizione "valida" della realtà. L'unico scotto da pagare è quello di rinunciare alla "certezza", e di accettare una descrizione "probabilistica" degli eventi. Uno scotto pesante, che addirittura lo stesso Einstein inizialmente rifiutò, ma che con il passare dei decenni ha finito con l'essere progressivamente accettato.

1.2.2.1. Probabilità epistemica e probabilità non epistemica

I filosofi della scienza attribuiscono al termine probabilità due significati o valori: e parlano di *probabilità epistemica* e di *probabilità non epistemica*[13].

Nel caso di processi probabilistico nei quali la probabilità risulta epistemica, si ha a che fare con conclusioni che vengono espresse in modo probabilistico a causa della nostra ignoranza sullo stato reale del sistema in esame. Come impeccabilmente ricordato da Ghirardi nell'opera sopraccitata, il concetto di probabilità epistemica rispecchia perfettamente la posizione meccanicistica del grande matematico francese Simon de Laplace, che nel 1776 scriveva "*Lo stato attuale del sistema della natura consegue evidentemente da quello che esso era nell'istante precedente, e se noi*

[12] *Monod J. Il caso e la necessità. Saggio sulla filosofia naturale della biologia contemporanea. Milano:EST Edizioni Scientifiche e Tecniche Mondadori, 1972:163pp.*
[13] *Ghirardi GC. Un'occhiata alle carte di Dio. Gli interrogativi che la scienza moderna pone all'uomo. Milano:Il Saggiatore, 1997:420pp.*

immaginassimo un'intelligenza che ad un dato istante comprendesse tutte le relazioni fra le entità di questo universo, esso potrebbe conoscere le rispettive posizioni, i moti e le disposizioni generali di tutte quelle entità in qualunque istante del passato e del futuro… Ma l'ignoranza delle diverse cause che concorrono alla formazione degli eventi come pure la loro complessità, insieme coll'imperfezione dell'analisi, ci impediscono di conseguire la stessa certezza rispetto alla grande maggioranza dei fenomeni. Vi sono quindi cose che per noi sono incerte, cose più o meno probabili, e noi cerchiamo di rimediare all'impossibilità di conoscerle determinando i loro diversi gradi di verosimiglianza. Accade così che alla debolezza della mente umana si debba una delle più fini e ingegnose fra le teorie matematiche, la scienza del caso o della probabilità". Secondo Laplace, note le posizioni e le velocità di tutte le particelle dell'universo, e le leggi che ne governano i rapporti, sarebbe stato possibile prevederne l'evoluzione per l'eternità. La concezione di Laplace configura la probabilità nella descrizione dei processi fisici come accidentale, legata alla nostra ignoranza, ma in linea di principio eludibile.

Come ricorda Ghirardi, nel 1903 il grande matematico Jules-Henri Poincarè scriveva: "*Una causa piccolissima che sfugga alla nostra attenzione determina un effetto considerevole, che non possiamo mancare di vedere, e allora diciamo che l'effetto è dovuto al caso. Se conoscessimo esattamente le leggi della natura e la situazione dell'universo all'istante iniziale, potremmo prevedere esattamente la situazione dello stesso universo in un istante successivo. Ma se pure accadesse che le leggi naturali non avessero più alcun segreto per noi, anche in tal caso potremmo conoscere la situazione iniziale solo approssimativamente. Se questo ci permettesse di prevedere la situazione successiva con la stessa approssimazione non ci occorrerebbe di più e dovremmo dire che il fenomeno è stato previsto, che è governato da leggi. Ma non sempre è così: può accadere che piccole differenze nelle condizioni iniziali ne producano di grandissime nei fenomeni finali. Un piccolo errore nelle prime produce un errore enorme nei secondi. La previsione diviene impossibile, e si ha un fenomeno fortuito*".

L'estrema sensibilità alle condizioni iniziali descritta da Poincarè, apre la strada ai moderni concetti di "caos deterministico" e di "complessità". Laddove nella sua accezione più generale il concetto di complessità pone in crisi l'idea che in ogni caso lo studio dei sistemi complessi possa ricondursi allo studio dei loro costituenti.

Come commenta Ghirardi "*di fatto risulta relativamente facile dimostrare che esistono sistemi deterministici* [come il banale lancio di una moneta, N.d.A.] *con una tale sensibilità alle condizioni iniziali che la previsione del loro comportamento anche dopo tempi brevi richiederebbe una tale massa di informazioni (proprio perché le imprecisioni iniziali si amplificano esponenzialmente) che non potrebbero venire immagazzinate neppure in un computer che utilizzasse come chips tutte le particelle dell'universo e potesse immagazzinare un bit in ogni chip. La conclusione è che ci si è resi conto (e questo rappresenta indubbiamente una notevole conquista concettuale) che non sono rare situazioni in cui risulta di fatto impossibile prevedere il comportamento di un sistema per un periodo di tempo anche ragionevolmente breve… Il fatto che se anche tutto l'universo diventasse un calcolatore esso non risulterebbe abbastanza potente da permetterci di immagazzinare le informazioni necessarie a prevedere per più di qualche minuto l'evoluzione di un semplice sistema, non toglie nulla al fatto che secondo lo schema teorico che si è assunto soggiacere alla dinamica del processo, la necessità di ricorrere ad una descrizione probabilistica deriva dall'ignoranza circa le precise condizioni iniziali… Al contrario nelle schema quantomeccanico…l'aleatorietà degli esiti è incorporata nella struttura stesa del formalismo che, se assunto come completo, non consente neppure di pensare che, in generale, gli esiti siano, anche se in un modo a noi sconosciuto, predeterminati*".

La meccanica quantistica introduce il concetto di probabilità non epistemica, cioè di descrizione probabilistica degli eventi che non può essere attribuita ad ignoranza. In altre parole, la descrizione

probabilistica degli eventi non può essere attribuita ad una mancanza di informazioni sul sistema che, se fosse disponibile, ci consentirebbe di trasformare le asserzioni probabilistiche in asserzioni certe. I processi fisici microscopici sono fondamentalmente stocastici, e hanno una intrinseca e irriducibile aleatorietà.

Per concludere, interessantissima la presentazione del volume curato da Casati, prima citato, nella quale il curatore, tra le altre cose, dice: "*…è interessante osservare che la meccanica quantistica è una teoria intrinsecamente probabilistica. Una volta assegnato lo stato di un sistema al tempo t mediante una "funzione di stato" ψ(t) noi siamo in grado di fare affermazioni solamente sulla probabilità che eseguendo una misura su una data grandezza si ottenga un determinato valore. Tuttavia è diverso il discorso relativo alla previsione della evoluzione futura. Infatti in meccanica quantistica lo stato ψ(t) del sistema al tempo t è determinato univocamente dallo stato iniziale ψ(0). Problema: siamo in grado, date le leggi del moto e dato ψ(0), di determinare ψ(t)? Il fatto straordinariamente inaspettato è che, a differenza della meccanica classica, la risposta a questa domanda è positiva: per i sistemi quantistici è possibile, almeno in linea di principio, risolvere le equazioni del moto e predire lo stato futuro ψ(t). Pertanto il quadro che si va delineando è diametralmente opposto a quello che si aveva in precedenza: la meccanica classica è sempre stata considerata come una teoria deterministica; ora abbiamo visto che, a causa dell'insorgere del moto caotico, essa porta ad un comportamento statistico. D'altro lato la meccanica quantistica è intrinsecamente probabilistica; tuttavia, grazie al suo carattere di stabilità, risulta essere più predicibile della meccanica classica… La domanda che il lettore certamente si pone è quali implicazioni può avere lo studio del caos. In altri termini, a quale utilità pratica può portare il sapere che il comportamento della gran arte dei sistemi deterministici è in realtà così complicato da apparire completamente caotico, e che quindi essi si sottraggono alla nostra capacità di previsione? Anzitutto abbiamo imparato una lezione molto importante: leggi semplici non portano necessariamente a comportamenti semplici. Sarebbe alquanto vantaggioso se questo concetto fosse tenuto presente non solo nelle discipline scientifiche, ma anche nella vita politica ed economica. Un'altra lezione è che variazioni piccole nei parametri di un qualunque sistema non portano, necessariamente, a variazioni piccole nel "risultato", cioè nella evoluzione futura: per esempio un aumento del 5% nell'inquinamento non sempre porta a un peggioramento solo del 5% nel danno ecologico. Una delle caratteristiche dello studio dei fenomeni caotici è la enorme potenzialità di una unificazione culturale in cui tutta la "filosofia naturale" e le discipline economiche, umanistiche, politiche e sociali sono coinvolte. La natura stessa sembra usare il caos nel suo programma di evoluzione: ogni schema deterministico fallirebbe se utilizzato per la sopravvivenza delle forme di vita in condizioni ambientali in continua trasformazione; la natura, pertanto, genera una quantità enorme di forme di vita attraverso mutazioni casuali e, da questa ampia possibilità di scelta, la selezione naturale trova candidati che si adattano alle mutate condizioni ambientali… tutte le volte che nel suo faticoso ed esaltante cammino verso la comprensione dei fenomeni naturali, l'uomo si è trovato di fronte a delle limitazioni, ciò è stato l'occasione per nuovi grandi balzi in avanti che hanno comportato un rovesciamento della filosofia precedente. Mi riferisco alla osservazione del valore limite della velocità della luce (che ha portato alla teoria della relatività), alla limitazione sulla precisione delle nostre misure (che ha portato alla meccanica quantistica); la teoria del caos ci ha messo di fronte a una limitazione ancora maggiore: l'impossibilità di prevedere il futuro…*". Quest'ultima affermazione ci collega con quanto diremo fra poco a proposito dei programmi di ricerca per la conoscenza della realtà e del problema della previsione degli eventi futuri.

1.2.2.2. Scegliere un campione rappresentativo

Per inciso va detto che lo stesso rapporto che esiste tra il concetto di ordine/vita e il secondo principio della termodinamica, vale tra il concetto di fortuna (e il suo reciproco, il concetto di sfortuna) e il principio di casualità. La fortuna (e il suo reciproco, la sfortuna) rappresenta

semplicemente una contraddizione su base locale del principio di casualità, che rimane universalmente valido. Come dovrebbe essere ormai chiaro, il fatto che, avendo scelto il rosso per dieci volte consecutive, per altrettante volte esca il rosso (fortuna), ovvero il suo reciproco, cioè il fatto che avendo scelto il rosso per dieci volte consecutive, esca il nero (sfortuna) sono semplicemente conseguenze del fatto che dieci volte sono troppo poche perché il principio di casualità funzioni appieno. Questo concetto si applica al fenomeno della selezione dei campioni, che vedremo più avanti. Perché una parte (il campione) possa essere rappresentativa del tutto (la popolazione) è necessario, tra le alte cose, che gli elementi che la compongono siano tratti in modo assolutamente casuale dal tutto, e che siano "sufficientemente" numerosi. Solo in questo modo, la parte/campione risulterà essere rappresentativo del tutto/popolazione della quale si desidera conoscere le caratteristiche.

1.2.3. Il rapporto segnale/rumore

La legge del GIGO domina la teoria dell'informazione. GIGO come acronimo di Garbage In-Garbage Out. Se ci metti dentro spazzatura ne tiri fuori solo spazzatura. Se il segnale contiene molto rumore, l'informazione risulterà ineluttabilmente deteriorata. Se, diciamo, l'altezza di un individuo viene misurata con una approssimazione di 5 centimetri, che senso ha dire che il tal dei tali, alto 172 centimetri, è più basso del tal altro, che è alto 174 centimetri? Nessuna, ovviamente. Mentre se l'altezza viene misurata con un'approssimazione di 1 centimetro, l'affermazione ha (intuitivamente) più senso[14].

Camminando per strada a Milano, una domenica mattina di blocco totale del traffico a causa dell'inquinamento, per la prima volta sono riuscito a riconoscere, provenienti dal marciapiedi opposto al mio, in una della vie più congestionate della città, le voci alcune persone che chiacchieravano davanti ad un bar, e la musica che proveniva da una finestra aperta (erano "Le quattro stagioni" di Vivaldi"). In condizioni "normali" (notare il concetto di abnormità che può essere insito nell'espressione "normale"[15]) il rumore del traffico mi aveva fino ad allora impedito di sentirle.

---

[14] *Non ha nessun senso ricorrere a tecniche statistiche sofisticate se i dati di partenza sono viziati da un rumore di fondo (errore) elevato, che nessuna tecnica statistica è in grado di eliminare. Meglio quindi un disegno sperimentale e procedimenti di misura rigorosi, che consentano di ottenere dati con il minor rumore di fondo (errore) possibile. La cosa è intuitivamente vera: qualcuno ritiene che se si trasferisce un'incisione da cassetta su CD si possa migliorare il rapporto segnale/rumore? Se così fosse pensate che bello sarebbe sentire le incisioni di Toscanini su CD (ne ho appena acquistata una: posso garantire che, nonostante l'ottimo lavoro di rimasterizzazione fatto, purtroppo i limiti del rapporto segnale/rumore scadente della registrazione iniziale sono ancora completamente evidenti). Sicuramente è vero il contrario: se si trasferisce un'incisione da CD su cassetta si ha un peggioramento del rapporto segnale/rumore. Ma non solo. Non è vero che, come qualcuno ottimisticamente tende a ritenere, che gli errori si elidano: se due dati affetti da un certo grado di indeterminazione (rumore, errore) sono combinati tra di loro, il dato risultante sarà affetto da un'indeterminazione maggiore della più grande delle due (in altre parole i gradi di indeterminazione, gli errori, si sommano, cioè tendono purtroppo a propagarsi). Quindi, contrariamente a quanto si potrebbe superficialmente ritenere, tecniche statistiche sofisticate, con pesanti elaborazioni dei dati, tendono purtroppo ad espanderne l'incertezza, portando a conclusioni ancora più incerte!*

[15] *A causa di queste e di altre ambiguità e assurdità derivanti dall'uso del termine "normale", nella medicina di laboratorio, per esempio, si è sostituita l'espressione "valori normali", con riferimento ai valori "più probabili" in un soggetto sano, con l'espressione "intervalli di riferimento". Nel caso della concentrazione di un farmaco nel siero si parla di "intervallo terapeutico". Nel caso delle varie classi di lipidi (colesterolo, trigliceridi, colesterolo HDL, colesterolo LDL, trigliceridi)*

Da un punto di vista fisico, in generale, qualsiasi segnale può essere riconosciuto solamente se riesce a sopravanzare il "rumore di fondo" dell'ambiente. Se si ascoltano "Le quattro stagioni", dal "Cimento dell'armonia e dell'invenzione" di Antonio Vivaldi, in una incisione su CD, si possono distinguere agevolmente i suoni dei violini e dei violoncelli. Se si ascolta la stessa incisione su una musicassetta, si può riconoscere l'insieme degli archi, ma i suoni dei violini e quelli dei violoncelli non possono più essere distinti. La causa è insita nel fatto ben noto che il CD ha un rapporto segnale/rumore migliore della musicassetta. E questo consente al CD di avere un potere di risoluzione maggiore tra due suoni "vicini" quali quelli dei violini e dei violoncelli.

1.3. I programmi di ricerca per la conoscenza della realtà

Magia, filosofia e scienza sono le tre tappe fondamentali sulla via della conoscenza, intesa come "prendere possesso intellettualmente… della realtà", e utilizzano tre differenti metodi di ricerca o, come anche si dice, rappresentano tre "programmi di ricerca".

1.3.1. Magia, filosofia e scienza

La possibilità di basarsi su misure anziché sull'intuizione è uno dei due motori del passaggio dall'interpretazione/comprensione magica (e filosofica) all'interpretazione/comprensione scientifica della realtà. Il dare forma matematica anziché intuitiva ai dati è l'altro motore del passaggio dall'interpretazione/comprensione magica (e filosofica) all'interpretazione/comprensione scientifica della realtà.

La magia non si preoccupa di avere dati misurabili, né di dare loro una forma. Basa il suo programma di ricerca sulla sola intuizione, ignorando qualsiasi esigenza di oggettività delle fonti della conoscenza. Ma le conseguenze sono rappresentate da un mondo assolutamente irrazionale: l'uomo è in balia di dei capricciosi, di ciclopi, draghi e streghe, dell'andamento degli astri. Per di più non si riesce a trovare la pietra filosofale, e l'alchimia non porta ad alcun risultato. La magia fallisce nel suo obiettivo più ambizioso, quello di dominare, fino a poter cambiare, il mondo reale: anche se ancor oggi c'è sicuramente qualcuno che crede nell'efficacia di una danza della pioggia.

Con la filosofia l'uomo incomincia a ragionare in termini di modelli, attraverso i quali dare forma alla realtà. Modelli che prevedono anche intuizioni geniali, come quello di Democrito, che arriva all'dea di indivisibilità dei costituenti ultimi della materia. La capacità di modellizzare rappresenta un salto in avanti di importanza storica nello sviluppo della cultura umana. E alcuni risultati, come l'incommensurabilità tra misura del raggio e misura della circonferenza, oppure il modello di geometria euclidea, "scoperti" dalla scuola filosofica pitagorica, sono arrivati ai giorni nostri. Ma i pitagorici rappresentano l'unica eccezione in un mondo che continua ad ignorare il concetto di misurabilità dei dati come strumento per renderli oggettivi. Un fatto al quale ha probabilmente contribuito in modo determinante lo spregio della cultura ellenica nei confronti delle attività manuali, essendo il concetto di misura sicuramente collegato allo svolgimento di una attività manuale.

Aristotele afferma che "*i corpi leggeri* [come la fiamma, (nota dell'autore)] *si muovono verso l'altro, mentre quelli pesanti si muovono verso il basso*"[16]. Sono necessari 2000 anni prima che Galileo misuri la velocità di caduta dei gravi, con una serie di esperimenti storici. Galileo non solo modellizza, ma sottopone il modello a verifica. Esiste ovviamente il rischio di scoprire che il

---

si parla di "valore desiderabile". Nel caso di sostanze tossiche (metalli pesanti, metaboliti dello xilolo e del toluolo, eccetera) si parla di "intervalli di riferimento in soggetti esposti" e di "intervalli di riferimento in soggetti non esposti" al tossico.

[16] *Nicola U. Atlante illustrato di filosofia. Verona:Demetra, 1999:587pp.*

modello è falso. Ma Galileo accetta il rischio. Non solo, ma svolge lui stesso il lavoro manuale necessario per effettuare le misure richieste. E scopre che i gravi cadono tutti alla stessa velocità. L'insieme di dati misurabili e di una legge matematica quantitativa porta al primo modello scientifico di conoscenza della realtà. Galileo passa meritatamente alla storia come il fondatore del metodo sperimentale. Il metodo che consentirà a Newton, con la sua monumentale opera, *Philosophiae naturalis principia mathematica* di strutturare definitivamente l'evoluzione in scienza di una branca della filosofia[17].

Questi concetti possono essere così riassunti:

Dati	*Informazione*	*Conoscenza*
Non misurabili	*In forma "intuitiva", non modellizzata, ignora il criterio di oggettività*	*Magica*
Non misurabili	*In forma di "modello filosofico", basato su modelli ovvero "leggi filosofiche" qualitative, con al più deboli elementi di oggettività*	*Filosofica*
Misurabile	*In forma di "modello scientifico", basato su modelli ovvero "leggi matematiche" quantitative che governano il comportamento della natura, con forti elementi di oggettività ("misura" dei fenomeni)*	*Scientifica*

Magia e filosofia hanno in comune la caratteristica di essere stadi pre-scientifici della conoscenza. Tuttavia mentre la magia persiste solo come traccia di un meme ancestrale[18,19] in manifestazioni più o meno pittoresche che vanno dalla superstizione, agli oroscopi, alla new-age, la filosofia potrebbe mantenere probabilmente un suo ruolo. Privata del troncone della filosofia dei naturalisti, evolutasi nella scienza moderna, la filosofia potrebbe restare infatti come programma di ricerca per la conoscenza dell'etica.

Per questo motivo, nello sviluppare i prossimi argomenti, necessari per comprendere l'importanza della probabilità e della statistica come strumenti per la conoscenza della realtà, utilizzeremo solo la contrapposizione tra scienza e magia, intesa quest'ultima come paradigma della "non scienza".

1.3.2. Il problema della previsioni di eventi futuri

Durante i primi millenni della cultura dell'uomo, e fino all'avvento della *scienza* moderna (in questo caso le branche della matematica note come *probabilità* e *statistica*), la soluzione al problema della previsione di eventi futuri è stata fornita dalla *magia*. E residui dall'approccio magico alla soluzione di questo problema sono ancora oggi evidenti dalla presenza di maghi cui alcuni si rivolgono per prevedere il proprio futuro.

Si consideri un problema apparentemente banale, ma paradigmatico, come quello delle previsioni del tempo. Visto che vorrei fare un week-end al mare, ma vorrei evitare di trascorrerlo sotto la pioggia, e la domanda è: domani pioverà o no?

---

[17] *Guillen M. Le cinque equazioni che hanno cambiato il mondo. Longanesi & C, Milano, 1999, 294 pp.*
[18] *Dawkins R. Il gene egoista. Bologna:Zanichelli,1979:180pp.*
[19] *Ianneo F. Meme. Genetica e virologia di idee, credenze e mode. Roma:Castelvecchi, 1999:221pp.*

L'approccio magico per rispondere a questa domanda, procede utilizzando un ragionamento analogico non basato su dati misurabili (per una descrizione storica di incredibile vastità e profondità sull'approccio magico all'interpretazione degli eventi nelle diverse culture, si veda l'opera di James G. Fraser[20]). Il mago basa in effetti la previsione sulle sue capacità di "intuire" il tempo che farà domani.

L'approccio scientifico "deterministico", quello che pretende di avere previsioni "certe", si differenzia da quello magico per il fatto che la previsione non viene effettuata sulla base di una "intuizione", bensì sulla base di dati misurabili e di leggi/modelli in forma matematica. Si misura la temperatura, la pressione atmosferica, si elaborano modelli matematici della circolazione atmosferica, sulla base dei quali si prevede il tempo che farà domani. Tuttavia si scopre che la previsione, per quanti sforzi siano fatti in termini di numero dei dati misurati dai quali si parte, e di complessità dei modelli matematici utilizzati, non porta mai alla "certezza". E la visione scientifica deterministica, quella per intenderci della fisica di Newton, per la quale la scoperta di alcune delle leggi fondamentali che governano la natura (la legge di gravitazione) fa intendere il cosmo come un gigantesco orologio rigorosamente determinato nelle sue funzioni, deve lasciare il passo ad una visione scientifica probabilistica, e ad una interpretazione probabilistica dei problemi [in realtà, come ormai dovrebbe essere chiaro da quanto finora detto, la sviluppo dell'approccio scientifico probabilistico non è dovuto alla meteorologia, bensì alla scoperte dei fisici: in particolare alla scoperta delle leggi che governano la natura a livello degli elementi fondamentali che la costituiscono (*atomi* e *quanti*)[21]].

Ecco quindi il significato profondo della frase di E. Schroedinger precedentemente citata. La "*…coerenza che si osserva nella stragrande maggioranza dei fenomeni, la cui regolarità e invariabilità hanno consentito la formulazione del postulato di causalità...*", ha portato a Newton e i primi scienziati a vedere il cosmo come un orologio perfettamente determinato. Data l'assoluta regolarità e apparente inviolabilità delle leggi che governano il moto dei corpi celesti, è "realmente" possibile prevedere con 72 anni di anticipo il ripresentarsi della cometa di Halley, e la presenza di Plutone può essere prevista dallo studio delle perturbazioni del pianeta Urano molti anni prima che Plutone venga effettivamente osservato, esattamente nella posizione prevista. A livello *macroscopico* il principio di causalità funziona, al punto da divenire un postulato: il che implica una dichiarazione di certezza "a priori" della sua validità. Tuttavia a partire dai primi del '900, la fisica scopre che le leggi che governano il comportamento degli elementi ultimi che costituiscono la materia/energia dell'universo, possono essere descritte solamente mediante formulazioni di tipo probabilistico. A livello *microscopico* (intendendo con ciò, come detto, atomi e quanti) "*...l'elemento comune soggiacente…è il caso...*". Ad esempio l'orbitale, che descrive il moto di un elettrone attorno al nucleo di un atomo, è una funzione d'onda che fornisce la probabilità di "trovare" l'elettrone.

Il percorso che ha portato dalla spiegazione magica della realtà alla spiegazione della realtà data dalle leggi della scienza "deterministica" applicabile al mondo macroscopico, prima, e alla spiegazione della realtà data dalle leggi della scienza "probabilistica" applicabile al mondo microscopico, poi, è come detto una delle avventure più straordinarie del pensiero umano. Ed è anche stato il percorso che ha portato allo sviluppo delle tecnologie che stanno alla base del mondo in cui viviamo. La *teoria dei quanti* è rigorosamente probabilistica: ma senza di essa, ad esempio, non esisterebbe la microelettronica, e non esisterebbe il PC su cui sto scrivendo.

Il rapporto tra l'approccio fornito dalla concezione magica e l'approccio fornito dalla concezione scientifica nei confronti del problema della previsione di eventi futuri, è illustrato qui di seguito:

---

[20] *Fraser JC. Il ramo d'oro. Torino:Boringhieri, 1973:1098 pp.*
[21] *Barrow JD. Il mondo dentro il mondo. Milano:Adelphi, 1991:491pp.*

L'approccio	*Lo strumento di previsione*	*Utilizzo di dati misurabili*	*Il livello cui si applica*	*La previsione*
Magia	Intuizione	No	Macroscopico	Apparentemente "certa", se si accettano in modo fideistico previsioni del mago. Ma si può dimostrare che anche i migliori maghi sbagliano.
Scienza	Modello matematico deterministico	Si	Macroscopico	"Certa" in riferimento ad eventi come il moto degli astri nelle orbite determinate dalla legge di gravitazione, e a *quasi tutte* le leggi che governano il mondo macroscopico.
	Modello matematico probabilistico	Si	Macroscopico	"Probabilistica" in riferimento ad eventi complessi come la circolazione dell'atmosfera (previsioni del tempo).
		Si	Microscopico	"Probabilistica" in riferimento agli eventi che caratterizzano elementi ultimi che costituiscono la materia/energia (teoria atomica e teoria dei quanti).

Per meglio illustrare questi concetti, e le loro profonde implicazioni, bisogna ora necessariamente fare un passo indietro, riconsiderando il problema del lancio della moneta, e la domanda "al prossimo lancio uscirà testa o croce?".

In merito a questa domanda, il rapporto tra concezione magica, visione deterministica e visione probabilistica è illustrato qui di seguito:

La domanda	*Lo strumento di previsione*	*La previsione*
Testa o croce?		
	La magia	Può sembrare che funzioni
	La risposta deterministica (*conclusione "certa"*)	Funziona solo se la moneta è truccata (in questo caso lanciando in un modo particolare la moneta è possibile ottenere "deterministicamente" (con certezza) un certo risultato.
	La risposta probabilistica (*conclusione "probabile"*)	Ci consente di affermare che, se il risultato di un singolo lancio è imprevedibile (legato al caso), a lungo andare metà delle volte uscirà testa e l'altra metà delle volte uscirà croce (necessità), e di esprimere quindi il risultato del lancio della moneta in termini di probabilità (la probabilità che in un dato lancio esca testa è identica alla probabilità che esca croce, ed è $p = 0,5$)

Per approfondire ulteriormente il significato insito nella rinuncia alla visione deterministica e il passaggio ad una risposta probabilistica, si consideri ora di nuovo la domanda "domani pioverà o no?".

Lo speaker che illustra le previsioni del tempo dice "domani generalmente soleggiato con possibilità di piovaschi". Si tratta di un modo colloquiale di esprimere una probabilità, diciamo (i valori effettivi di p sono ovviamente qui irrilevanti) del 90% ($p = 0,90$) che faccia bello e del 10% ($p = 0,10$) che piova. Lo speaker afferma in questo modo che, date condizioni meteorologiche quali quelle odierne, e date 100 osservazioni del tempo che ha fatto l'indomani, si è osservato che 90 volte l'indomani ha fatto bello e che 10 volte l'indomani ha piovuto.

La mia domanda diventa allora " Visto che vorrei fare un week-end al mare, ma vorrei evitare di trascorrerlo sotto la pioggia, mi piacerebbe tanto sapere: *ma domani è uno dei 90 giorni che farà bello o è uno dei 10 che giorni che pioverà?*". Ebbene, come ormai dovrebbe essere chiaro, essendo la risposta di tipo probabilistico, la mia legittima pretesa di certezza non potrà mai essere soddisfatta.

Il rapporto tra concezione magica, visione deterministica e visione probabilistica è ulteriormente sintetizzato qui di seguito:

La domanda	*Lo strumento di previsione*	*La previsione*
Domani pioverà?		
	Intuizione (*magia*)	Può sembrare che funzioni.
	Modello matematico "deterministico" (*conclusione "certa"*)	Non è possibile.
	Modello matematico "probabilistico" (*conclusione "probabile"*)	Ci consente di affermare che (ad esempio) domani c'è il 90% di probabilità che faccia bello e il 10% di *probabilità* che piova, rinunciando peraltro alla *certezza* di sapere se domani sarà uno dei 90 giorni che fa bello o piuttosto uno dei 10 giorni che pioverà.

La famosa espressione attribuita ad A. Einstein "*…non posso credere che Dio giochi ai dadi…*" si riferisce proprio a questo. Alla difficoltà che si incontra nell'accettare che buona parte delle leggi di natura siano intrinsecamente probabilistiche, sfumate (fuzzy)[22,23].

La risposta al problema sembra essere stata fornita dal matematico tedesco Kurt Gödel, che negli anni '30 ha dimostrato che qualsiasi sistema assiomatico, che parta cioè da affermazioni "certe", ma che siano anche moderatamente complicate, arriva prima o poi a delle proposizioni "indecidibili"[24]. In altre parole, anche se le assunzioni di base sono "certe", alla fine si arriva a dei passaggi per i quali la certezza non vale più, e quindi la "verità" o la "falsità" della proposizione non può più essere dimostrata. L'unica alternativa pare essere quella di accettare che la proposizione sia "vera" o "falsa" con un certo grado di probabilità. Nel caso di assoluta indecidibilità, risulta uguale la probabilità di essere vera e di essere falsa ($p = 0{,}5$). Ed è quanto accade per la moneta, e l'uscita di testa o croce al prossimo lancio[25].

---

[22] *Kosko B. Il fuzzy-pensiero, Milano: Baldini & Castoldi, 1993:365pp.*

[23] *Personalmente ritengo che la logica fuzzy sia semplicemente un modo diverso di esprimere il concetto di probabilità. Solo che risulta in effetti semanticamente più coerente con quanto si osserva nel mondo reale. Così un bicchiere pieno per tre quarti, che in termini probabilistici verrebbe descritto come "un bicchiere che ha la probabilità del 75% di essere pieno" (in effetti se su 100 bicchieri 75 sono pieni e 25 sono vuoti, mediamente ciascun bicchiere sarà pieno al 75%), in termini di logica fuzzy viene descritto semplicemente come "un bicchiere pieno al 75%" (in altri termini né pieno né vuoto). A causa del fatto che il concetto soggiacente è lo stesso, il termine* fuzzy *e il termine* probabilità *dovrebbero probabilmente essere utilizzati come sinonimi.*

[24] *Barrow JD. La luna nel pozzo cosmico. Milano:Adelphi, 1994:530pp.*

[25] *L'assioma universale aristotelico, sottostante a ogni forma di pensiero razionale, è il "principio di non contraddizione". Esso afferma che "è impossibile che la stessa cosa sia e insieme non sia". Se è dato un A, allora ogni B sarà diverso da A. A o non-A. Tertius non datur. O bianco o nero. Eppure già Zenone mette in crisi l'assioma. Raccoglie un granello di sabbia da un mucchio e chiede se il mucchio è ancora un mucchio. Zenone non trova il granello di sabbia che cambia il mucchio in un non-mucchio. Né il granello di sabbia che cambia il non mucchio, formato dal primo granello di sabbia sottratto al mucchio di partenza, in un mucchio. Il mondo reale sembra essere dominato, più che dal bianco e dal nero, dall'infinita gamma di grigi che li collegano. Il mondo è prevalentemente sfumato (fuzzy). Come dice Kosko nel suo volume* Il fuzzy-pensiero *"…La scienza rivela un mondo dai contorni sfrangiati e dalle quantità che mutano insensibilmente. Una maggior precisione nono elimina il "chiaroscuro" delle cose – semplicemente lo fissa con maggior rigore. I progressi della medicina nono hanno reso affatto più facile tracciare una linea di demarcazione fra*

Un esempio solo apparentemente banale di contraddizione logica, è fornito da uno dei tanti paradossi individuati da Bertrand Russel[26]. Si supponga che per gli abitanti dell'isola di Creta esistano solo due alternative: o affermano tutti sempre e solamente il vero, o affermano tutti sempre e solamente il falso. Si consideri ora la domanda "Un cretese dice che tutti i cretesi mentono: egli mente o non mente?". Se gli abitanti dell'isola di Creta affermano tutti sempre e solamente il vero, il cretese, che per definizione dice il vero, avrà mentito dicendo che tutti i cretesi mentono. Se gli abitanti dell'isola di Creta affermano tutti sempre e solamente il falso, il cretese, che per definizione dice il falso, avrà detto la verità. Si può uscire dalla contraddizione determinata dal paradosso solamente ammettendo che l'affermazione del cretese sia per metà vera e per metà falsa (la probabilità che sia vera è $p = 0,5$ e la probabilità che sia falsa è $p = 0,5$)[27].

1.4. Evoluzione biologica ed evoluzione culturale

Esiste l'evidenza che i sistemi biologici si siano andati strutturando in modo tale da rappresentare dei "paradigmi della conoscenza". Ci si riferisce con ciò alla capacità dei sistemi biologici di utilizzare in modo "innato" regole/leggi da applicare nel processo che consente di "dare forma ai dati".

D'altra parte esiste l'evidenza che l'evoluzione della cultura proceda attraverso la clonazione e il progressivo affinamento (di generazione in generazione) di "paradigmi della conoscenza", che forniscono la capacità di utilizzare in modo "acquisito" regole/leggi da applicare nel processo che consente di "dare forma ai dati". L'analogia tra evoluzione biologica ed evoluzione culturale, intese entrambe in senso darwiniano, cioè come risultati conseguenti alla pressione selettiva dell'ambiente, può essere così sintetizzata

---

vita e non-vita, sia alla nascita che alla morte. Anche se descrivessimo l'atmosfera della Terra molecola per molecola, nondimeno non troveremmo alcuna linea di divisione fra atmosfera e spazio. Mappe dettagliate della Terra, di Marte e della Luna non ci dicono dove finiscono le colline e cominciano le montagne".

[26] *Come riportato da J. D. Barrows nella sua opera* Il mondo dentro il mondo *precedentemente citata, Russel dice "Sulle prime pensai che sarei riuscito a venire a capo delle contraddizioni senza troppe difficoltà… Ma gradualmente divenne chiaro che le cose non stavano così… Per tutta l'ultima parte del 1901 continuai a credere che la soluzione sarebbe stata semplice ma verso la fine di quell'anno avevo ormai concluso che si trattava di un'impresa ardua… Passammo le estati del 1903 e del 1904 a Churt e a Tilford… passammo gli inverni a Londra, e in quei periodi non tentai di lavorare; ma le due estati del 1903 e del 1904 mi sono rimaste in mente come fasi di completa paralisi intellettuale. Mi era chiaro che non sarei potuto procedere senza prima avere risolto le contraddizioni, ed ero deciso a non permettere ad alcuna difficoltà di distogliermi dal completamento dei* Principia Mathematica*; ma sembrava molto probabile che avrei passato il resto della mia vita a fissare il foglio bianco, A rendere la situazione ancor più irritante era il fatto che le contraddizioni apparivano banali, e che passavo il mio tempo a pensare su questioni che non apparivano degne di seria attenzione…".*

[27] *Si lascia al lettore lo stimolo intellettuale derivante dal traslare questi concetti in campo sociologico. Per esempio alla difficoltà, dimostrata dalla storia, di individuare soluzioni "certe" ai grandi problemi sociali anche partendo da "assunti" certi (come le ideologie). Ovvero al contenuto sempre "incerto" dei discorsi degli uomini politici, che sembrano avere mutuato dal concetto di indecidibilità, in modo intuitivo ma estremamente efficace (dal loro punto di vista), il presupposto per ripresentarsi tranquillamente agli elettori anche dopo avere quasi totalmente disatteso le promesse elettorali.*

Evoluzione	Comportamento	Substrato	Principi
Biologica	Innato	Geni	Conservazione dell'informazione Applicazione della logica sfumata (fuzzy)
Culturale	Acquisito	Memi	Conservazione dell'informazione Applicazione della logica sfumata (fuzzy)

I principi basilari che assicurano l'evoluzione sono rappresentati in entrambi i casi dalla necessità di conservare l'informazione senza che essa venga degradata, e dalla applicazione della logica probabilistica, ovvero di una logica sfumata (logica "fuzzy").

1.4.1. Il principio di conservazione dell'informazione

Uno dei meccanismi alla base dei sistemi informativi "fault-tolerant", cioè dei sistemi informativi in grado di assicurare il servizio anche nel caso di crash del sistema (come la rottura improvvisa di un hard-disk) è rappresentato dal mirroring. Questa tecnica provvede ad assicurare in tempo reale una doppia copia dei dati su due diversi supporti (hard-disk). Nel caso di rottura di uno dei due, il sistema è in grado di continuare a funzionare senza interruzione, utilizzando l'hard-disk rimasto funzionante. Il sistema funziona talmente bene che l'aneddotica dei centri informativi è ormai ricca di casi nei quali ci si è accorti della rottura di uno dei due hard-disk a distanza di giorni o addirittura di settimane dalla rottura dell'altro.

Il principio del mirroring è utilizzato da tempo immemorabile nei sistemi biologici, che hanno una doppia copia del codice genetico, una per ciascuna delle due catene complementari di DNA. Gli enzimi di riparazione del DNA continuano febbrilmente, e ininterrottamente, a riparare una della due catene (danneggiata per esempio dall'attacco dei raggi UV o dei radicali liberi che alterano fisicamente o chimicamente l'informazione codificata) utilizzando come stampo l'altra catena, rimasta integra. Senza questa continua attività di riparazione le cellule accumulerebbero ben presto tali a tante alterazioni degenerative del codice genetico dall'esserne irrimediabilmente distrutte. Sia nel caso del mirroring, sia nel caso del DNA, la strategia di affidabilità nel mantenimento dell'informazione codificata passa attraverso un meccanismo di ridondanza (l'informazione viene duplicata).

1.4.2. L'applicazione della logica sfumata

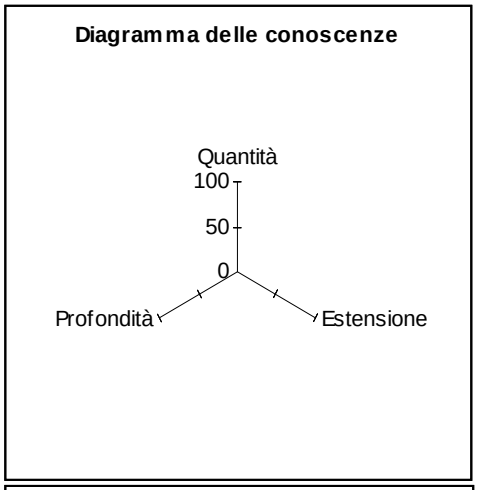
Quanto visto prima a proposito dei paradossi insiti nell'utilizzo di una logica anche semplice, come quella a due stati (vero/falso), ha anch'esso trovato una soluzione, in termini di evoluzione culturale, nell'ambito di quella che Laplace definiva "*una delle più fini e ingegnose fra le teorie matematiche…*" e cioè "*la scienza del caso o della probabilità*". Che prevede appunto l'utilizzo, nel trarre conclusioni "scientifiche", del concetto di probabilità.

Si consideri ora un falco cacciatore alla ricerca del cibo. Il falco deve effettuare una *stima* il più possibile *accurata* di dove si troverà il coniglio per avere una sufficiente *probabilità* di catturare la preda e quindi di assicurare cibo/sopravvivenza alla progenie. Il comportamento del falco, come di altri sistemi biologici, è molto specializzato. Tuttavia è ormai chiaro che la rete neurale che lo dirige con "precisione chirurgica" verso la preda arriva alla soluzione dell'incredibilmente complesso problema computazionale che gli si presenta attraverso una serie di calcoli che convergono verso la

soluzione finale (trovarsi nel posto giusto nel momento giusto) applicando leggi basate su una logica "sfumata"[28].

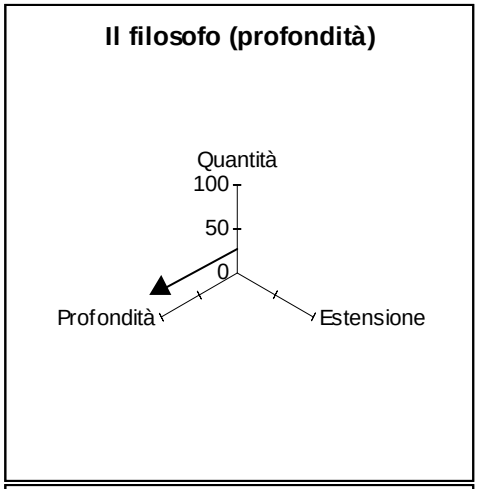
1.4.3. Geni e memi

In generale le conoscenze dell'individuo si sviluppano su tre dimensioni, profondità delle conoscenze, quantità delle conoscenze ed estensione delle conoscenze, come nel "diagramma delle conoscenze" qui riportato.



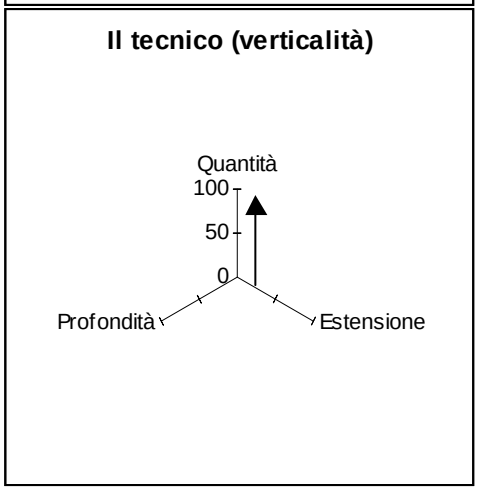
Dal punto di vista biologico:

- il substrato dell'evoluzione è rappresentato dal gene, in grado di conservare l'informazione, di accumulare mutazioni "favorevoli" sulla base della pressione selettiva dell'ambiente biologico, e di dare forma ai dati rilevati dall'ambiente utilizzando una logica "sfumata";
- lo sviluppo nella dimensione della profondità delle conoscenze corrisponde alla pulsione determinata a livello istintuale dai geni;
- lo sviluppo nella dimensione della quantità delle conoscenze corrisponde al concetto di specializzazione della specie;
- lo sviluppo nella dimensione dell'estensione delle conoscenze corrisponde al concetto di flessibilità della specie.



Dal punto di vista culturale:

- il substrato dell'evoluzione è rappresentato dal meme, in grado di conservare l'informazione, di accumulare mutazioni "favorevoli" sulla base della pressione selettiva dell'ambiente culturale, e di dare forma ai dati rilevati dall'ambiente utilizzando una logica "sfumata";
- lo sviluppo nella dimensione della profondità delle conoscenze corrisponde alla pulsione determinata a livello culturale dai memi. Il paradigma è rappresentato dal filosofo, che si muove tendenzialmente in questa sola dimensione[29];
- lo sviluppo nella dimensione della quantità delle conoscenze corrisponde al concetto di specializzazione (di verticalità) della cultura. Il paradigma è rappresentato dal tecnico, che si muove tendenzialmente in questa sola dimensione;
- lo sviluppo nella dimensione dell'estensione delle conoscenze corrisponde al concetto di flessibilità (di trasversalità) della cultura. Il paradigma è rappresentato dal politico, che si muove tendenzialmente in questa sola dimensione.

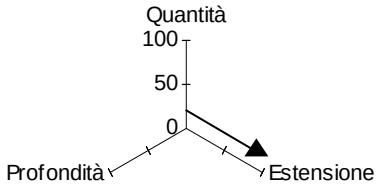


[28] Questa affermazione, qui data in termini apodittici per ovvi motivi di spazio, penso trovi tutti d'accordo di fronte alla semplice osservazione che nessuno si azzarderebbe a dire che il falco conosce il calcolo differenziale necessario a calcolare la rotta per intercettare il coniglio.
[29] In realtà il paradigma è forse rappresentato dal mistico, ma qui si è voluto utilizzare il termine filosofo in quanto più generale, intendendo il mistico come la forma estrema del filosofo.

Se si vuole, estensione e quantità sono dimensioni professionali, mentre la profondità è una dimensione esistenziale. Seguendo l'aforisma di Porat, si potrebbe dire che il sapiente integra tutte e tre le dimensioni.

Si noti che lo sforzo qui fatto di illustrare lo sviluppo della conoscenza dai dati osservati nella realtà, e di cogliere le analogie tra evoluzione biologica ed evoluzione culturale, tra geni e memi, non è gratuito, né rappresenta una divagazione fuori tema. L'obiettivo è di preparare il lettore alla consapevolezza del fatto che il modo di pensare culturale "probabilistico" e sfumato risulta determinato dalla natura soggiacente del mondo fisico e del mondo biologico, in altre parole dalla natura stessa dell'universo in cui viviamo.

Il politico (trasversalità)



CAPITOLO 2
I FONDAMENTI STATISTICI DEL PROCESSO DECISIONALE MEDICO

Come anticipato nel capitolo precedente, il falco deve effettuare una *stima* il più possibile *accurata* di dove si troverà il coniglio per avere una sufficiente *probabilità* di catturare la preda e quindi di assicurare cibo/sopravvivenza alla progenie. Mentre l'esatta definizione dei termini verrà data nei capitoli successivi, si chiede per ora di accettarli come intuitivi. Per attirare l'attenzione sul fatto che quello che nel falco è un comportamento innato, viene utilizzato in modo formale nel processo decisionale medico. Di fatto il medico deve effettuare una *stima* il più possibile *accurata* delle condizioni del paziente per avere una sufficiente *probabilità* di individuare la corretta diagnosi, e quindi di assicurare gli interventi corretti per migliorarne lo stato di salute. L'obiettivo è ora di illustrare questi concetti mediante un modello semplice ma significativo, in quanto le differenze critiche, che vedremo, rappresentano un esempio di ricorso formale all'approccio probabilistico per un processo decisionale medico[30].

2.1. Estendere i sensi del medico

Il potere disporre, per indagare le malattie, solamente dei propri sensi, ha sempre costituito per il medico un grosso limite. E' vero che anche solo basandosi su di essi il medico è potuto arrivare ad intuizioni geniali (si pensi al diabete "mellitus", dalle urine dolci come il miele)[31]. Tuttavia è chiaro che la medicina moderna si è sviluppata in modo dirompente solamente quando, sulla base dello studio "more scientifico" degli organismi biologici, si è innescato un circolo virtuoso tra questo e i suoi sviluppi tecnologici. E oggi nessun medico saprebbe rinunciare a quelle estensioni dei propri sensi che sono rappresentate dalle analisi di laboratorio, dalla diagnostica per immagini (radiografia, tomografia assiale computerizzata, risonanza magnetica nucleare, ecografia), dalle tecniche elettrofisiologiche (elettrocardiografia, elettroencefalografia, elettromiografia), dalle tecniche endoscopiche: in altre parole, oggi nessun medico saprebbe rinunciare all'informazione che da queste gli viene fornita.

2.2. Dai dati all'informazione

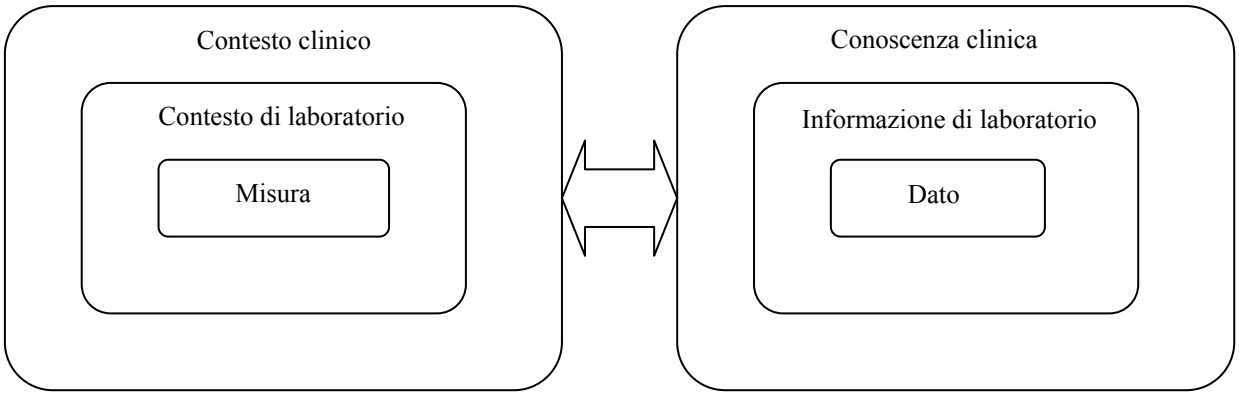
Il valore aggiunto della Medicina di Laboratorio è quindi costituito dall'informazione che le analisi di laboratorio sono in grado di fornire al medico. La trasformazione del dato analitico in informazione richiede tre passi fondamentali: (i) la *misura* della concentrazione dell'analita in questione (ci riferiremo d'ora in poi, per semplicità, alle sole situazioni nelle quali è disponibile un risultato quantitativo), (ii) la *contestualizzazione del dato* (risultato della misura) nell'ambito del quadro delle conoscenze di laboratorio (i dati sono integrati per generare l'*informazione di laboratorio*), e (iii) la *contestualizzazione dell'informazione di laboratorio* nell'ambito del quadro delle conoscenze cliniche (le informazioni sono integrate per generare la *conoscenza clinica*)[32]. Questi concetti sono riassunti nello schema sottostante:

---

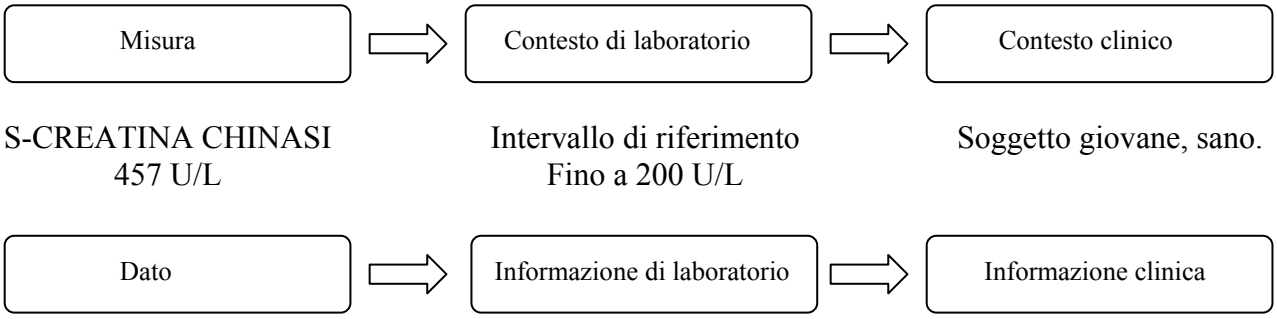
[30] *L'altro esempio di un approccio probabilistico per un processo decisionale medico è rappresentato dalla statistica bayesiana, che verrà trattata più avanti, ma senza legarla formalmente agli aspetti decisionali (decisione medica) che sono qui illustrati.*

[31] *La scoperta viene generalmente attribuita a Galeno, dal quale comunque sembra venire la prima traccia scritta del fenomeno.*

[32] *La corrispondenza tra gli stati attraversati dalle rappresentazioni mentali sulla via dell'intelligere (*dati ⇨ informazione ⇨ conoscenza*) viene qui presentata con una terminologia un poco diversa da quella introdotta all'inizio del libro, ma senza rilevanti differenze concettuali.*



Si consideri ora il seguente esempio, relativo alla determinazione della concentrazione nel siero dell'enzima creatina chinasi (CK), che abbia fornito per un determinato soggetto un valore di concentrazione pari a 457 U/L (Unità di attività enzimatica per Litro di siero):



Il dato in sé dice poco. Integrato nel contesto di laboratorio, fornisce una prima informazione: il soggetto si discosta *significativamente* (vedremo fra poco quale significato dare a quest'espressione) dal comportamento dei soggetti sani (a partire dai quali è stato determinato l'intervallo di riferimento utilizzato dal laboratorio). Ma la vera conoscenza sullo stato di salute/malattia può derivare solamente dal successivo livello d'integrazione effettuato dal medico sulla base del contesto clinico. Tutti i medici prima o poi hanno la possibilità di osservare valori di CK elevati in soggetti clinicamente sani: e una semplice indagine anamnestica può magari evidenziare il fatto che la sera prima il soggetto abbia praticato un'attività fisica alla quale, come noto, può facilmente conseguire un innalzamento della CK di durata e di intensità variabili, ma il più delle volte ancora rilevabile nelle analisi eseguite l'indomani mattina (si ricorda che anche una iniezione intramuscolare, o una semplice contusione muscolare, possono avere lo stesso effetto).
Un altro esempio potrebbe essere quello di un soggetto di mezza età che per la prima volta, inaspettatamente, presenta all'esame emocromocitometrico 60.000 leucociti, e una formula leucocitaria comprendente mieloblasti, promielociti, mielociti e metamielociti. In questo caso l'informazione fornita dal laboratorio è già da sola praticamente diagnostica di una leucemia mieloide cronica (anche se questo non esime da una serie di ulteriori indagini volte a confermare e a stadiare la malattia).

Così come tra il bianco e il nero esistono infinite sfumature di grigio, tra queste due situazioni estreme, quella della CK, nella quale l'informazione fornita dal laboratorio potrebbe essere, se astratta dal contesto clinico, in sé fuorviante, e quella dell'emocromo, nella quale l'informazione fornita dal laboratorio è praticamente diagnostica, quindi è in sé autoconsistente, esistono infinite situazioni intermedie.

Ora non è possibile conoscere a priori quale possa essere il valore informativo di una determinata analisi nel contesto clinico di uno specifico paziente. Se così fosse, sarebbero disponibili da tempo protocolli diagnostici applicabili meccanicisticamente, cosa che notoriamente non è. Questo non significa che le analisi di laboratorio debbano essere prescritte in modo acritico, visto che esse hanno comunque un costo, e che le risorse a disposizione della collettività, incluse quelle che possono essere dedicate alla sanità, sono per definizione limitate. Tuttavia è altrettanto chiaro che eventuali protocolli diagnostici (o percorsi diagnostici) devono essere considerati come comportamenti da seguire "in media", in base allo stato dell'arte, devono essere continuamente aggiornati, per seguire lo stato dell'arte, ma devono rappresentare solamente un "percorso guida" che garantisca comunque al medico i gradi di libertà necessari a permettere la loro personalizzazione (una personalizzazione ovviamente motivata) in relazione alle esigenze cliniche di ciascun specifico caso.

2.3. Diagnosi, monitoraggio e screening

Tre sono gli obiettivi con i quali il medico richiede un'analisi di laboratorio: (i) effettuare una selezione (*screening*) dei pazienti candidati ad un trattamento preventivo, (ii) fare una *diagnosi* di malattia, (iii) effettuare il *monitoraggio* di un paziente del quale la malattia è nota.

Obiettivo clinico	Risultato atteso	Esempio	Contesto di laboratorio	Note
Screening	Selezione degli individui per i quali risulta dimostrato che una diagnosi precoce consente di instaurare un trattamento preventivo efficace	Fenilchetonuria	Intervallo di riferimento	Il trattamento immediato del neonato previene l'instaurarsi di lesioni cerebrali irreversibili (oligofrenia fenilpiruvica)
		Portatori sani dell'HBsAg	Intervallo di riferimento	E' possibile la vaccinazione del partner (trasmissione per via sessuale) e dei familiari dei portatori dell'antigene Australia (soggetti HBsAg positivi)
		Ipercolesterolemia	Valore desiderabile	Il trattamento consente di ridurre l'incidenza delle lesioni cardiovascolari
Diagnosi	Conferma o esclusione dello stato di malattia di un singolo individuo	Epatite	Intervallo di riferimento	E' generalmente necessario integrare i risultati di più analisi (transaminasi, bilirubina, marcatori sierologici dell'epatite)
Monitoraggio della malattia	Valutazione dell'andamento della malattia	PSA	Differenza critica	Un aumento significativo dopo terapia radicale dell'adenocarcinoma prostatico deve indurre a ricercare le possibili recidive locali o a distanza

Monitoraggio degli effetti della terapia	Valutazione della necessità di modificare la terapia	Emoglobina glicata (HbA_{1c}) Colesterolo	Intervallo di riferimento Differenza critica	Il mantenimento dell'HbA_{1c} entro l'intervallo di riferimento è indice di buon compenso, ed è accompagnato da riduzione delle complicanze (retinopatia, nefropatia) Una riduzione significativa è indice di efficacia della terapia
Monitoraggio terapeutico dei farmaci	Mantenimento dei livelli ematici del farmaco all'interno dell'intervallo che garantisce risultati terapeutici efficaci senza o con ridotti effetti collaterali	Digossina	Intervallo terapeutico	

Come appare nella precedente tabella, il contesto di laboratorio può variare nelle differenti situazioni. Tradizionalmente esso è rappresentato dagli *intervalli di riferimento*, che esprimono i limiti (limite inferiore, limite superiore o entrambi) all'interno dei quali si collocano i risultati dell'analisi in questione nel 95% dei soggetti sani. Talora è rappresentato dai *valori desiderabili*, definiti in apposite "Consensus Conference" in relazione ad obiettivi specifici. Tale è il caso del colesterolo, per il quale è stato definito come valore desiderabile quello inferiore a 200 mg/dL, il cui raggiungimento sembrerebbe consentire una significativa riduzione delle affezioni cardiovascolari. Nel caso dei farmaci, soprattutto di quelli a ridotto indice terapeutico (quelli per i quali la dose efficace è molto vicina a quella tossica, come tipicamente avviene per i farmaci digitalici), il contesto è tipicamente rappresentato dall'*intervallo terapeutico*, vale a dire quello che garantisce effetti terapeutici efficaci senza o con ridotti effetti collaterali. Un altro tipo di contesto è quello rappresentato dalle *differenze critiche*, intese come la minima differenza significativa tra due risultati consecutivi ottenuti nel medesimo paziente. I concetti che le differenze critiche sottendono sono illustrati nei prossimi paragrafi.

2.4. Le differenze critiche

Le differenze critiche non sono di per sé un concetto nuovo. In effetti, il medico ha sempre applicato, a livello intuitivo, l'analisi comparativa di due successivi risultati di laboratorio al fine di valutare l'andamento clinico del paziente. Quello che rappresenta la vera novità è la possibilità di sostituire ad una valutazione intuitiva, una valutazione quantitativa, quindi più oggettiva, basata su un modello matematico-statistico semplice, ma in grado di fornire delle conclusioni basate su un grado di confidenza (grado di fiducia) noto. Per comprendere questo sono necessarie ancora alcune premesse.

2.4.1. La confidenza (o fiducia) statistica

La scienza è nata con il *metodo*. Tuttavia la scienza moderna è nata dopo l'abbandono della visione deterministica, e con la nascita dell'approccio *statistico/probabilistico*. Le moderne teorie scientifiche (come la meccanica quantistica) sono teorie di tipo essenzialmente statistico. Eppure, a dimostrazione della loro efficacia, basti pensare che senza di loro non esisterebbero l'energia atomica e nucleare, il laser, i semiconduttori, la microelettronica: e anche il computer sul quale sto scrivendo non esisterebbe. Tutto ciò è stato reso possibile dall'utilizzo di strumenti matematici (*probabilità e statistica*) che, rinunciando all'ambizione di fornire conclusioni certe in assoluto (dimostratasi illusoria), consentono peraltro di ottenere, riguardo gli eventi studiati, conclusioni la cui validità è comunque caratterizzata da un grado di certezza noto. La *confidenza* (o *fiducia*)

statistica fornisce una misura oggettiva del grado di affidabilità che possiamo attribuire alle nostre conclusioni. La confidenza statistica si riflette nel concetto di *significatività statistica*: risulta significativa una conclusione statistica nella quale possiamo riporre un grado sufficientemente elevato di confidenza. In genere quando la probabilità di sbagliare nel trarre dai dati una specifica conclusione risulta inferiore al 5%, la conclusione è accettata, in quanto considerata *statisticamente significativa*[33].

2.4.2. Errori e sbagli

A causa dei limiti inerenti ai sistemi (*strumenti di misura*) impiegati per rilevare i segnali provenienti dalle grandezze fisiche, ogni misura sperimentale (quindi anche quella effettuata nel laboratorio clinico) è inevitabilmente accompagnata da un qualche grado di *incertezza*. D'altra parte si può assumere che esista, ogniqualvolta viene effettuata una misura sperimentale, un valore teorico, detto *valore vero*: quello che si otterrebbe se la misura non fossa affetta da alcuna incertezza. Per effetto dell'incertezza che la caratterizza, la misura sperimentale rappresenta una *stima* più o meno approssimata del valore vero: la differenza tra una singola misura sperimentale e il suo valore vero rappresenta l'*errore*.

Ripetendo più volte una misura, e' possibile pervenire ad una migliore caratterizzazione dell'errore. In un insieme di misure ripetute, si definisce come *errore casuale* l'errore per cui le singole misure differiscono (casualmente, cioè senza nessuna regola apparente al succedersi delle misure stesse) tra loro, e come *errore sistematico* l'errore per cui l'insieme (preso globalmente) delle misure ripetute si discosta dal valore vero. L'errore, l'errore casuale e l'errore sistematico sono quindi legati all'incertezza intrinseca alle nostre conoscenze scientifiche (ai sistemi/strumenti di misura).

L'errore deve essere mantenuto distinto dallo *sbaglio* (*errore grossolano*), che e' un accidente tecnico, e che come tale si manifesta nel corso dell'applicazione delle conoscenze. Gli sbagli sono legati prevalentemente all'organizzazione e quindi ai processi di comunicazione (esempi di sbagli possono essere l'errata trascrizione di un dato numerico, l'utilizzo di un reagente scaduto, lo sbaglio nell'identificazione del campione, lo sbaglio nell'interpretazione del risultato di un test di gravidanza acquistato in farmacia ed eseguito a casa propria dalla paziente, che ha frainteso i criteri per l'interpretazione dei risultati del test). Contrariamente a quanto avviene per gli errori, gli sbagli si possono evitare operando con cura, e migliorando il sistema organizzativo. Contrariamente a quanto avviene per gli errori, non è possibile fissare un livello di tolleranza per gli sbagli, cioè definire una percentuale ammissibile di sbagli: semplicemente, gli sbagli per definizione non si devono (dovrebbero) verificare. Nel laboratorio clinico moderno il fronte degli sbagli può essere adeguatamente presidiato mediante un sistema che preveda un controllo formalmente completo del processo, cosa ottenibile per esempio applicando le norme della serie UNI EN ISO 9000 (*certificazione* secondo queste norme).

2.4.3. Variabilità analitica e variabilità biologica

A causa della presenza dell'errore, qualsiasi risultato analitico presenta un certo grado di incertezza. Chiamiamo globalmente *variabilità analitica* (CV_a) questa incertezza. Ma non basta. Il risultato analitico presenta un ulteriore grado di incertezza, legato al fatto che i valori di concentrazione degli analiti presenti nel sangue (e nei fluidi interstiziali) non sono affatto costanti, ma variano in continuo, anche nel soggetto sano, attorno al punto omeostatico. E nonostante essi rimangano mediamente costanti (un concetto squisitamente statistico), in un dato istante (per esempio al momento del prelievo di sangue) possono discostarsi dal punto omeostatico anche in modo piuttosto

[33] *Si fa notare che peraltro la conclusione potrebbe essere sbagliata!*

rilevante. Chiamiamo globalmente *variabilità biologica intraindividuale* (CV_b) l'incertezza del risultato analitico legata alle fluttuazioni fisiologiche delle concentrazioni degli analiti.
In base a queste considerazioni, risulta che la *variabilità totale* (CV_{tot}), cioè il grado di incertezza globale che accompagna un singolo risultato, è uguale alla somma (quadratica, come è necessario quando si devono sommare gli errori) della variabilità analitica e della variabilità biologica, ovvero è uguale a

$$CV_{tot} = (CV_a^2 + CV_b^2)^{1/2}$$

Com'è facile intuire, poiché qualsiasi risultato analitico risulta inevitabilmente affetto da un'incertezza pari alla somma della variabilità analitica e della variabilità biologica, l'espressione sopra riportata gioca un ruolo fondamentale nel calcolo delle differenze critiche.

Un'ulteriore ed ultima precisazione. Obiettivo del moderno laboratorio clinico non è quello di fornire risultati senza errore (obiettivo irraggiungibile), e neppure quello di ridurre sempre più l'errore analitico senza tenere conto del rapporto costi/benefici di tale continua riduzione. Obiettivo del moderno laboratorio clinico è piuttosto quello, pragmatico ma razionale ed efficace, di fornire risultati nei quali l'incertezza derivante dall'errore analitico sia sostanzialmente irrilevante rispetto a quella derivante dalla variabilità biologica intrinseca al soggetto. Se mi si consente l'analogia, che dal punto di vista della teoria dell'informazione è non solo formale bensì sostanziale, il laboratorio del 2000 sta al laboratorio degli anni '50 come il Compact Disc (CD) sta alla musicassetta. Il miglioramento del rapporto segnale/rumore del CD è tale da consentire di riconoscere, nel brano musicale, tutti i dettagli che l'orecchio potenzialmente sa discernere. Ormai il rumore è ridotto a livelli ininfluenti. Aumentare la fedeltà del CD oltre questo limite sarebbe possibile, ma sostanzialmente sarebbe uno spreco di risorse.

Attualmente si raccomanda, come obiettivo per l'errore analitico, una variabilità analitica (CV_a) inferiore alla meta' della variabilità biologica intraindividuale (CV_b), vale a dire:

$$CV_a < 0{,}5 \cdot CV_b$$

Questo fa si' che, al termine del procedimento analitico, la variabilità totale del risultato analitico sia solamente lo 11,8% (o meno) in più di quella derivante dalla sola variabilità biologica. Anche se in sé il valore dello 11,8% può apparire arbitrario, è evidente come, in tal modo, un metodo analitico finisca con l'introdurre, nel *segnale* rappresentato dalla concentrazione dell'analita, una quantità di *rumore* sostanzialmente irrilevante. Il modello delle differenze critiche qui presentato assume che la variabilità analitica sia uguale alla metà della variabilità biologica: un traguardo che è ormai stato quasi completamente raggiunto dai buoni laboratori.

2.4.4. Il modello delle differenze critiche

Il modello delle differenze critiche richiede di considerare che le misure che portano alle conclusioni (informazioni) cliniche sono affette da errore: conseguentemente le conclusioni cui si giunge sono conclusioni valide non in assoluto, ma con un livello noto di confidenza (fiducia) statistica. Come già accennato, in genere quando la probabilità di sbagliare nel trarre dai dati una specifica conclusione risulta inferiore al 5%, la conclusione viene generalmente accettata. Nel caso del modello delle differenze critiche è utilizzato il livello di significatività del 5%.

Le differenze critiche possono essere calcolate facilmente, conoscendo la variabilità analitica e la variabilità biologica che caratterizzano un dato analita, come

$$Differenza\ critica = 2{,}77 \cdot (CV_a^2 + CV_b^2)^{1/2}$$

Conoscendo il valore della variabilità analitica CV_a di un dato metodo analitico e il valore della variabilità biologica CV_b dell'analita in questione, è quindi elementare calcolare il valore della differenza critica corrispondente. Se la differenza tra due valori consecutivamente osservati in un dato soggetto eccede il valore della differenza critica, si è autorizzati a ritenere, con una probabilità del solo 5% di sbagliare (è questo il livello di confidenza statistica qui impiegato) che i due valori differiscono significativamente tra loro. Nella tabella che segue sono riportati i valori delle differenze critiche dei principali analiti.

Analita	Variabilità biologica %	Variabilità analitica %	Differenza critica %
17-ALFA IDROSSIPROGESTERONE	15,0	7,5	46
ALANINA AMMINOTRANSFERASI (ALT,GPT)	10,0	5,0	31
ALBUMINA	2,5	1,3	8
ALDOSTERONE	15,0	7,5	46
ALFA-1-GLICOPROTEINA ACIDA	10,0	5,0	31
ALFA-AMILASI	7,5	3,8	23
ANTIGENE CARBOIDRATICO 125	10,0	5,0	31
ANTIGENE CARBOIDRATICO 15.3	7,5	3,8	23
ANTIGENE CARBOIDRATICO 19.9	15,0	7,5	46
ANTIGENE CARCINO-EMBRIONALE	12,5	6,3	39
ANTIGENE PROSTATA SPECIFICO (PSA)	12,5	6,3	39
APOLIPOPROTEINA A-I	5,0	2,5	15
APOLIPOPROTEINA B	5,0	2,5	15
APTOGLOBINA	15,0	7,5	46
APTT	5,0	2,5	15
ASPARTATO AMMINOTRANSFERASI (AST, GOT)	15,0	7,5	46
BETA-2-MICROGLOBULINA	4,0	2,0	12
BICARBONATO (Idrogenocarbonato)	2,5	1,3	8
BILIRUBINA DIRETTA	15,0	7,5	46
BILIRUBINA TOTALE	12,5	6,3	39
CALCIO TOTALE	1,5	0,8	5
CERULOPLASMINA	4,0	2,0	12
CLORURO	1,5	0,8	5
COLESTEROLO HDL	5,0	2,5	15
COLESTEROLO LDL	5,0	2,5	15
COLESTEROLO TOTALE	4,0	2,0	12
COLINESTERASI	5,0	2,5	15
COMPLEMENTO/C3	5,0	2,5	15
COMPLEMENTO/C4	7,5	3,8	23
CORTISOLO	10,0	5,0	31
CREATINA CHINASI-MB	15,0	7,5	46
CREATINCHINASI	12,5	6,3	39
CREATININA	4,0	2,0	12
EMOCROMOCITOMETRICO			
Emoglobina	1,5	0,8	5
Eritrociti	1,5	0,8	5

Leucociti	7,5	3,8	23
Piastrine	5,0	2,5	15
Ematocrito	1,5	0,8	5
EMOGASANALISI			
pH	1,5	0,8	5
pCO2	2,5	1,3	8
pO2	2,5	1,3	8
EMOGLOBINA GLICATA	4,0	2,0	12
ESTRADIOLO	12,5	6,3	39
FARMACI DIGITALICI	5,0	2,5	15
FERRITINA	7,5	3,8	23
FERRO	12,5	6,3	39
FIBRINOGENO	7,5	3,8	23
FOSFATASI ALCALINA	7,5	3,8	23
FOSFATO INORGANICO	5,0	2,5	15
FRUTTOSAMMINA	4,0	2,0	12
GAMMA-GLUTAMMIL TRANSPEPTIDASI	12,5	6,3	39
GLUCOSIO	2,5	1,3	8
GLUCOSIO-6-FOSFATO DEIDROGENASI	7,5	3,8	23
IgA	7,5	3,8	23
IgG	4,0	2,0	12
IgM	7,5	3,8	23
INSULINA	10,0	5,0	31
LATTATO DEIDROGENASI	7,5	3,8	23
LIPASI	7,5	3,8	23
LIPOPROTEINA(a)	12,5	6,3	39
LITIO	5,0	2,5	15
LUTEOTROPINA (LH)	12,5	6,3	39
MAGNESIO TOTALE	2,5	1,3	8
MICROALBUMINURIA	10,0	5,0	31
OSTEOCALCINA	10,0	5,0	31
POTASSIO	2,5	1,3	8
PROLATTINA	10,0	5,0	31
PROTEINA C REATTIVA	7,5	3,8	23
PROTEINE (ELETTROFORESI)			
Albumina	2,5	1,3	8
Alfa-1-globuline	7,5	3,8	23
Alfa-2-globuline	5,0	2,5	15
Beta-globuline	5,0	2,5	15
Gamma-globuline	5,0	2,5	15
PROTEINE TOTALI	2,5	1,3	8
RAME	4,0	2,0	12
SODIO	1,5	0,8	5
TEMPO DI PROTROMBINA	4,0	2,0	12
TEOFILLINA	10,0	5,0	31
TESTOSTERONE	10,0	5,0	31
TIREOTROPINA (TSH)	10,0	5,0	31
TIROXINA LIBERA (fT4)	7,5	3,8	23
TRANSFERRINA	4,0	2,0	12

TRIGLICERIDI	15,0	7,5	46
TRIIODOTIRONINA LIBERA (fT3)	7,5	3,8	23
URATO	7,5	3,8	23
UREA	7,5	3,8	23
VES (1 ora)	12,5	6,3	39

Si consideri il caso di un paziente al quale una prima volta sia stata rilevata una concentrazione del colesterolo nel siero pari a 243 mg/dL. Il paziente, dopo un periodo di dieta di opportuna durata, effettua una seconda determinazione del colesterolo, che risulta ora pari a 225 mg/dL. La differenza tra la prima e la seconda determinazione è uguale a 18 mg/dL. La differenza critica riportata nella tabella precedente è pari al 12%. Il 12% di 243 mg/dL è uguale a 29,16 mg/dL (arrotondato a 29 mg/dL).

Conclusione: la differenza osservata (18 mg/dL) non è significativa, in quanto risulta inferiore alla differenza critica (29 mg/dL). Per un paziente con una concentrazione iniziale di 243 mg/dL, il successivo valore avrebbe dovuto essere pari o inferiore a 214 mg/dL (243 - 29) per consentire di concludere, con un adeguato grado di confidenza, che la dieta è stata efficace nel ridurre la concentrazione del colesterolo nel siero.

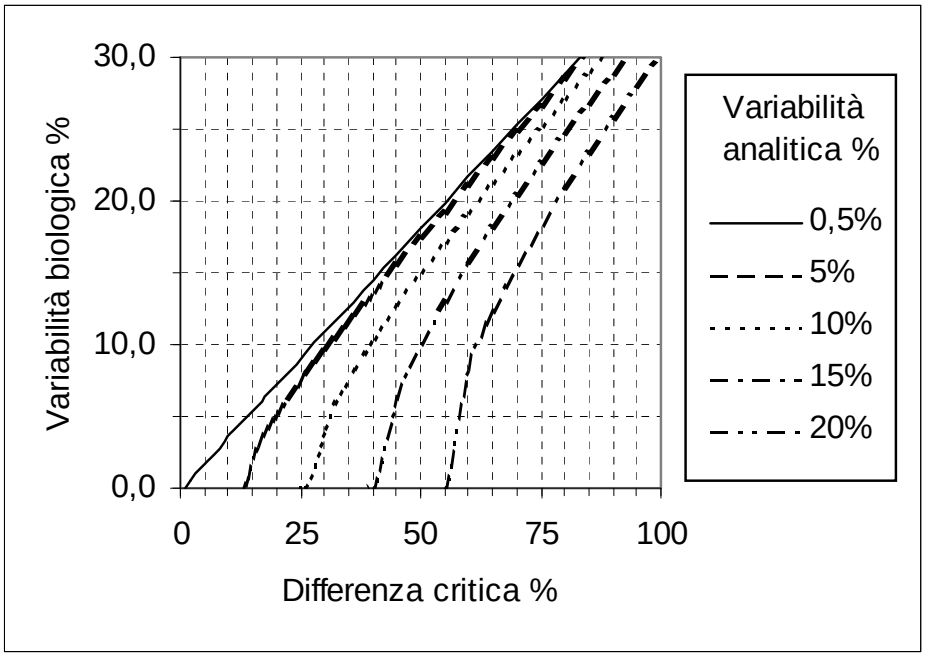
Simmetricamente, con una concentrazione iniziale di 243 mg/dL, il successivo valore avrebbe dovuto essere pari o superiore a 272 mg/dL (243 - 29) per consentire di concludere, con un adeguato grado di confidenza, che la dieta ha indotto piuttosto un aumento della concentrazione del colesterolo nel siero (come può avvenire veramente: esiste un meccanismo di retroazione negativa sulla sintesi endogena del colesterolo mediato appunto dal colesterolo di origine alimentare, e la ridotta introduzione di colesterolo con gli alimenti può pertanto determinare una attivazione della sua sintesi endogena, con un conseguente, ancorché indesiderato aumento della concentrazione del colesterolo nel siero).

Al fine di consentire un'adeguata familiarizzazione con lo strumento presentato, la tabella che segue contiene alcuni ulteriori esempi. Nella prima colonna viene riportato l'analita, nella seconda colonna il valore della differenza critica corrispondente, nella terza colonna un possibile obiettivo clinico del monitoraggio, nella quarta colonna il valore (ipotetico) riscontrato alla prima osservazione, e nella quinta il valore al quale (e oltre il quale) si deve considerare significativa la differenza alla seconda osservazione, con un rischio minimo di sbagliare (solo il 5%) accettando la conclusione che la concentrazione dell'analita si è effettivamente modificata. Alla seconda osservazione, qualsiasi valore compreso tra il primo ed il secondo (colonne (a) e (b) rispettivamente) deve essere attribuito al caso, impersonato nella fattispecie dalla variabilità analitica (ridotta) e dalla variabilità biologica (prevalente).

Un'ultima chiave di lettura della tabella. Un paziente che, nel corso di indagini multiple successive, si presentasse con valori che oscillano casualmente, ma sempre compresi tra quello della colonna (a) e quello della colonna (b), è un paziente nel quale non sta cambiando nulla. Ma non è "il laboratorio che sbaglia". Quello che si sta osservando è in sostanza la sua variabilità biologica (il solo assunto da fare è che il laboratorio che fornisce i risultati sia in grado di mantenere la propria variabilità analitica a valori almeno pari, o meglio inferiore, alla metà della variabilità biologica)

Analita	Differenza critica %	Obiettivo clinico	Valore (ipotetico) riscontrato alla prima osservazione (a)	Valore al quale (oltre il quale) considerare significativa la differenza alla seconda osservazione (b)
ALANINA AMMINOTRANSFERASI (ALT,GPT)	31	Valutare la riduzione della concentrazione dell'enzima	120 U/L	83 U/L
ANTIGENE PROSTATA SPECIFICO (PSA)	39	Valutare il possibile aumento in un soggetto con problemi specifici in atto	10,0 µg/L	13,9 µg/L
CREATININA	12	Monitorare l'eventuale innalzamento in paziente con insufficienza renale	8,5 mg/dL	9,5 mg/dL
EMOCROMOCITOMETRICO				
Emoglobina	5	Valutare la correzione dell'anemia	10,8 g/dL	11,3 g/dL
FERRO	39	Monitorare la sideremia in un donatore abituale di sangue	86 µg/L	53 µg/L
GLUCOSIO	8	Valutare il miglioramento della glicemia	180 mg/dL	166 mg/dL
MICROALBUMINURIA	31	Valutare il peggioramento della funzione renale in un paziente diabetico	40 mg/L	52 mg/L
VES (1 ora)	39	Monitorare la scomparsa della flogosi	36 mm	22 mm

Per gli analiti non riportati, è possibile calcolare le differenze critiche mediante il grafico seguente.



Le variabilità analitica può essere facilmente fornita dai laboratori. Lo stesso dicasi per la variabilità biologica, per la quale peraltro va precisato che, per molti analiti, non è ancora nota con precisione. A fronte di una perdita di dettaglio, il grafico fornisce un vantaggio: consente di visualizzare in modo conciso ma efficace gli ordini di grandezza delle grandezze in gioco. Così, anche non conoscendo con esattezza la variabilità biologica di un dato analita, si può facilmente osservare come, anche ipotizzando una variabilità analitica minima (0,5%) anche ad una variabilità biologica ridotta (del 10%) corrisponda in ogni caso una differenza critica pari al 27% circa. Queste e altre facili considerazioni in merito sono lasciate al lettore.

In conclusione il modello delle differenze critiche può a ragione aggiungersi al tradizionale contesto del laboratorio, in quanto è già in grado di fornire al clinico, con ragionevole attendibilità, un'informazione aggiuntiva rispetto ai più tradizionali strumenti rappresentati dagli *intervalli di riferimento* e dagli *intervalli terapeutici*, e ai (più recentemente introdotti) *valori desiderabili*. I suoi maggiori limiti, attualmente, sono rappresentati dal fatto che (i) la variabilità biologica non è ancora nota per tutti gli analiti e (ii) le stime della variabilità biologica sono state effettuate quasi esclusivamente su soggetti sani.

Certamente il modello dovrà ancora evolvere. Resta comunque il fatto che esso, tenuto conto della riduzione della variabilità analitica a valori trascurabili attualmente in atto, perlomeno nei laboratori di buon livello, consente di focalizzare il problema della variabilità biologica come componente fondamentale, ovviamente ineliminabile, della variabilità osservata tra differenti risultati dello stesso individuo in condizioni di salute stazionarie. Per questo è da ritenere che le differenze critiche apriranno definitivamente la strada a quello che sembra essere il contesto informativo del futuro: gli intervalli di riferimento individuali. E resta il fatto che la decisione medica, con il modello delle differenze critiche, trova un modo formale di collegarsi alla statistica e alle probabilità.

CAPITOLO 3
CONCETTI DI BASE IN BIOSTATISTICA

Il termine statistica deriva da *Status*, e venne introdotto nel XVIII secolo per indicare quella branca della scienze politiche che si occupava della descrizione delle cose dello Stato (originariamente dati economici e demografici). Da allora gli oggetti e i metodi di indagine della statistica si sono andati sempre più estendendo, e oggi la statistica è diventata una disciplina autonoma, definibile come quella "parte delle scienze matematiche che si occupa della analisi quantitativa delle osservazioni di qualsiasi fenomeno soggetto a variazione".

Nell'ambito della statistica, in particolare, la *statistica descrittiva* analizza la variabilità dei fenomeni a partire dalle informazioni che riguardano l'intera popolazione. La *statistica inferenziale*, invece, analizza fenomeni che, per ampiezza e complessità, si sottraggono all'osservazione diretta, consentendo la formulazione di ipotesi riguardanti la popolazione sulla base delle informazioni derivate da un sottoinsieme limitato della popolazione (un campione).

3.1. Probabilità

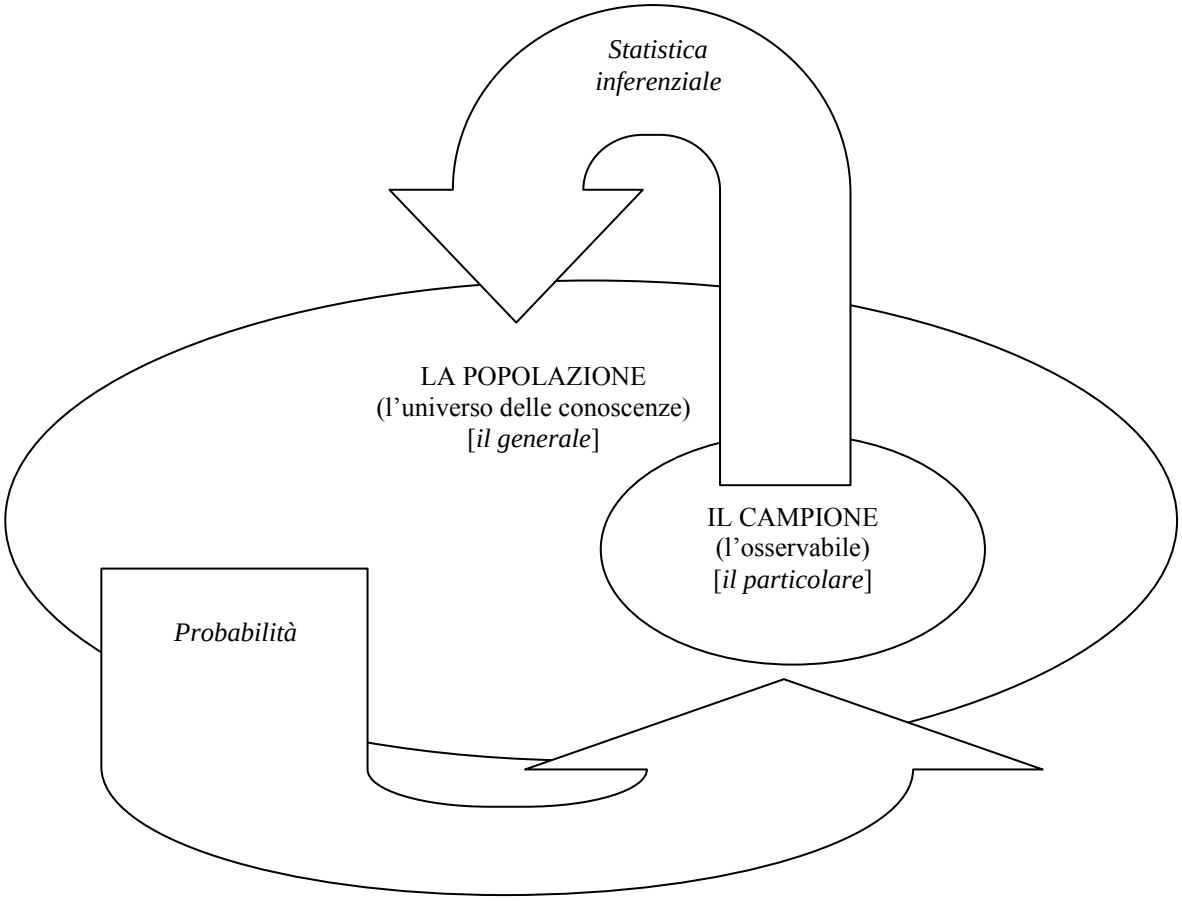
Dal punto di vista storico, la statistica è stata preceduta dagli studi riguardanti la probabilità. Gli scritti di Gerolamo Cardano (1501-1576) e di Galileo Galilei (1564-1642) rispettivamente intitolati *De ludo aleae* e *Sopra le scoperte de li dadi*, sono i primi che trattano specificamente e organicamente di problemi di probabilità. Che trova peraltro i suoi principali sviluppi nei lavori di Pascal (1623-1662) e di P. De Fermat (1601-1655). L'aspetto della probabilità che ha maggiori applicazioni in campo medico deriva dallo scritto del reverendo inglese Thomas Bayes[34], pubblicato postumo nel 1763.

Del teorema di Bayes verrà compiutamente trattato più avanti.

3.2. Statistica inferenziale

Se la probabilità consente di passare dal generale (*la popolazione*) al particolare (*il campione*), la statistica inferenziale consente di effettuare il percorso inverso, consente cioè di passare dal particolare (il campione) al generale (la popolazione).

[34] *Reverend Thomas Bayes. An assay toward solving a problem in the doctrine of chance. Philo Trans Roy Soc 1763;53:379-418.*



Le basi della statistica inferenziale sono poste da A. De Moivre (1667-1754) che pubblica l'equazione della distribuzione normale, e da Karl Friedrich Gauss (1777-1855) che ne fa conoscere l'utilizzo nell'analisi degli errori. È invece Francis Galton (1822-1911) che applica per primo il modello matematico della curva gaussiana alla descrizione di caratteristiche biologiche di soggetti impropriamente definiti come "normali". Ed è così che la curva gaussiana diventa, ancora più impropriamente, la "curva normale", e la distribuzione gaussiana la "distribuzione normale"[35].

La vera statistica, quella moderna, è molto recente. Nasce nei primi del novecento quando W. S. Gosset, con lo pseudonimo di "Student", pubblica nel 1908 i suoi risultati sulla distribuzione di *t*: le statistiche dei piccoli campioni rivoluzionano la metodologia statistica. Nel 1926 R. A. Fisher perfeziona i risultati di Gosset. Nel 1934 G. W. Snedecor pubblica "Analisis of variance and covariance" e dà il nome al test F in onore di Fisher. Nel 1938 R.A. Fisher e F. Jates pubblicano a Edimburgo le "Statistical Tables", per anni l'unico insostituibile strumento di lavoro per gli statistici, prima dell'avvento dei calcolatori, e introducono la denominazione di "rapporto tra varianze".

Uno dei problemi chiave nella statistica inferenziale è rappresentato dalla *selezione del campione*. In effetti perché l'inferenza riguardante la popolazione, inferenza effettuata a partire dal campione, sia corretta, è necessario che il campione sia rappresentativo della popolazione. La rivoluzione in questo campo è stata introdotta da R. A. Fisher. Fu lui il primo a comprendere l'importanza del campionamento casuale, in modo che dai campioni si potessero trarre conclusioni oggettive sui caratteri dell'universo di provenienza e sulla loro distribuzione originaria, e a dimostrare, con

[35] *Dell'utilizzo improprio del termine "normale" è già state detto in precedenza.*

innumerevoli lavori sperimentali, che una conclusione statistica eseguita su campioni randomizzati, è tanto più obiettiva quanto più i campioni sono stati scelti a caso.

3.2.1. La raccolta dei dati

Quando si formula una ipotesi, e si vuole verificarla sperimentalmente, il primo passo consiste nel progettare un esperimento che consenta di ottenere i dati necessari alla verifica. Una osservazione solo apparentemente banale è che l'informazione che si potrà ottenere dai dati raccolti sarà tanto migliore quanto più lo studio sperimentale sarà stato ben disegnato e ben condotto.

L'analisi statistica è solo l'ultimo di una serie di passi tra loro logicamente e operativamente concatenati. E se è vero che una cattiva analisi statistica può vanificare tutto il lavoro precedente, è altrettanto vero che nessuna analisi statistica può sopperire a difetti del disegno sperimentale o a una cattiva qualità dei dati raccolti: così se i dati raccolti sono, per esempio, affetti da errori sistematici, i risultati di qualsiasi elaborazione statistica saranno, irrimediabilmente, affetti dagli stessi errori. Per una eccellente introduzione su medicina, metodo scientifico e statistica vedere Bossi[36].

SUGGERIMENTO: una attenta definizione del disegno sperimentale e rigorose modalità di raccolta dei dati, preventivamente descritte in appositi protocolli, rappresentano un prerequisito essenziale di qualsiasi lavoro statistico.

3.2.2. Il disegno sperimentale

"...Come probabilmente ricorderete dal liceo, Procuste, il personaggio mitologico, allungava e accorciava i suoi ospiti in modo che si adattassero al letto che aveva costruito. Ma forse non sapete il resto della storia. La mattina, prima che se ne andassero, egli li misurava; per la Società Antropologica dell'Attica scrisse poi un erudito lavoro dal titolo «Sulla uniformità della statura dei viandanti»...".

Se questa è solamente una storiella divertente, attribuita ad A. S. Eddington (1882-1944), il famoso astronomo e fisico inglese[37], quella che segue è storia vera (citata da Colton[38]). Nel 1936, negli USA, il "Literary Digest" selezionò dai nomi riportati negli elenchi telefonici un campione di elettori. Si chiedeva chi avrebbe vinto le elezioni presidenziali. I candidati erano Roosevelt e Landon. Le risposte furono più di un milione, e il sondaggio predisse la vittoria di Landon. In realtà Roosevelt vinse con il più largo margine mai raggiunto in una elezione presidenziale fino al quel tempo.

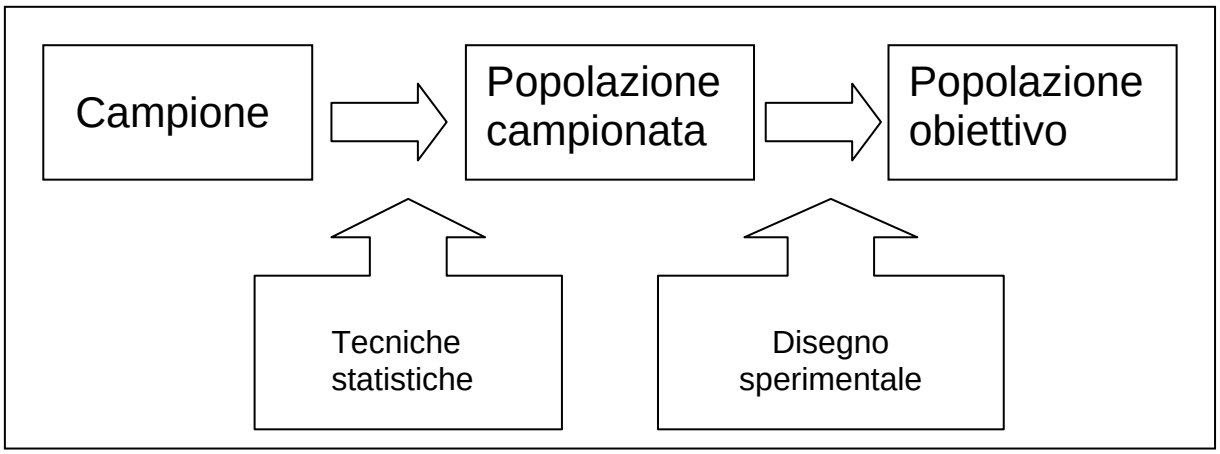
Quello che accomuna i due esempi sopra riportati è la selezione del campione. Introdotta in modo piuttosto rozzo dal ricercatore, nel caso di Procuste. Presentatasi in modo più subdolo nel caso del sondaggio del "Literary Digest". Nel qual caso il campione, pur numericamente enorme, non era rappresentativo: a causa del fatto che la maggior parte degli elettori appartenenti alle classi meno abbienti non disponeva, a quell'epoca del telefono. E, manco a farlo apposta, questi elettori erano prevalentemente orientati per Roosevelt. Per dirla in termini tecnici, l'errore fu determinato dal fatto che la popolazione campionata (elettori che disponevano del telefono) non era rappresentativa della popolazione obbiettivo (tutti gli elettori).

---

[36] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:13-34.*
[37] *Barrow JD. Il mondo dentro il mondo. Milano:Adelphi, 1991:409.*
[38] *Colton T. Statistica in medicina. Padova:Piccin, 1991:7.*

Ovviamente oggigiorno simili ingenuità non si verificano più. I sondaggi, e non solamente quelli elettorali, forniscono risultati attendibili anche utilizzando campioni numericamente assai limitati. Ma il problema della possibile mancanza di corrispondenza tra popolazione campionata e popolazione obbiettivo (il modo in cui esse differiscono è generalmente definito vizio, e le forze operanti per distinguere le due popolazioni sono generalmente indicate come fattori selettivi), risulta ancora uno degli aspetti delicati del disegno sperimentale, cioè del progetto in base al quale si effettua una ricerca scientifica.

Questi concetti sono illustrati nello schema seguente, che ricorda come, se tecniche statistiche utilizzate in modo appropriato consentono di effettuare inferenze corrette (e con un grado di confidenza noto) sulla popolazione da cui il campione origina (popolazione campionata), è solo un disegno sperimentale che può assicurare che le inferenze fatte possano essere estrapolate alla popolazione obiettivo dell'esperimento/inferenza.



Si riportano a questo proposito alcune considerazioni semplici e chiare fatte da Colton[39] in un esempio relativo ad una ricerca sull'artrite reumatoide condotta in un ospedale universitario: "...*Si supponga che un ricercatore sia interessato a determinati aspetti caratterizzanti la storia naturale dell'artrite reumatoide. Egli ha scelto come sua popolazione obbiettivo tutti i pazienti con questa malattia. Per il suo studio prende accordi con il responsabile dell'archivio dell'ospedale universitario in cui lavora perché gli venga mandato un campione di registrazioni di casi di pazienti con artrite reumatoide; per esempio, tutte le dimissioni avvenute nel periodo di un anno. L'analisi statistica e le conclusioni da essa risultanti devono essere poste in relazione con la considerazione di quali fattori selettivi e quali vizi distinguono la sua popolazione campionaria - precisamente, tutte le ammissioni di pazienti con artrite reumatoide a questo particolare ospedale universitario - dalla popolazione obbiettivo di tutti i pazienti con la malattia. Quali sono questi fattori? Balzano alla mente immediatamente due fattori selettivi. Primo l'artrite reumatoide non richiede sempre l'ospedalizzazione. Vi è una percentuale piuttosto grande di pazienti con questa malattia che non richiede ospedalizzazione, e tali pazienti non potrebbero assolutamente entrare nella ricerca qui discussa. Quindi può essere completamente erroneo trarre conclusioni sulla storia naturale dell'artrite reumatoide in tutti i pazienti, quando l'ammissione allo studio è possibile solo per i pazienti ospedalizzati. Secondo, anche tra i casi ospedalizzati il fatto che lo studio sia condotto in un ospedale universitario comporta una selezione dei casi più complessi e più seri. Le caratteristiche dei pazienti in questo ospedale possono essere molto diverse da quelle dei pazienti ricoverati in altri ospedali meno specializzati. Chiaramente fattori selettivi addizionali e possibili vizi potranno essere identificati mano a mano che vengono delineati ulteriori dettagli riguardanti le caratteristiche della ricerca...*".

[39] *Colton T. Statistica in medicina. Padova:Piccin, 1991:5.*

Risulta quindi evidente da quanto detto che le conclusioni tratte da un esperimento, intendendo per esperimento qualsiasi tentativo, legittimo ma ambizioso, di trarre da una parte (il campione) conclusioni riguardanti il tutto (la popolazione), debba essere attentamente presidiato al fine di evitare di introdurre elementi confondenti nelle conclusioni che dall'esperimento verranno tratte.

3.2.3. L'espressione dei risultati

Alla base dell'espressione di risultati vi è "la misura (dell'entità) di una grandezza fisica". Questa misura consiste nell'esprimere la grandezza in modo quantitativo, dando ad essa un "*valore numerico*" che è un numero puro, ottenuto per confronto della "*(entità della) grandezza* in esame" con la "(entità di una) grandezza di riferimento ad essa omogenea, definita *unità di misura*", essendo quindi

(entità della) grandezza / unità di misura = valore numerico

da cui si ricava

(entità della) grandezza = valore numerico · unità di misura

Da quest'ultima espressione si deduce che il "*risultato di una misura*" è dato dal prodotto di un "*numero*" per la "*unità di misura*": pertanto l'indicazione di quest'ultima non deve mai essere omessa[40].

3.2.3.1. Il sistema SI

Allo scopo di pervenire ad una immediata comprensione, in qualsiasi Paese, della espressione dei risultati di una misura, le organizzazioni internazionali e nazionali a ciò preposte hanno proceduto alla codificazione di un sistema di unità di misura delle varie grandezze, unificato nella definizione, nella nomenclatura e nella simbologia. Il sistema di base, oggi adottato, discende dal sistema metrico decimale (introdotto alla fine del XVIII secolo[41]) ed ha il nome di "Sistema Internazionale di Unità di Misura" (l'abbreviazione è SI); esso è stato sancito dalla Conferenza Generale dei Pesi e Misure (CGPM) nel 1960 e nel 1971, accettato dalla Comunità Economica Europea (CEE) nel 1980 e divenuto legale in Italia nel 1982[42].

Il sistema SI è fondato su sette grandezze e relative unità di base, indipendenti l'una dall'altra:

Grandezza di base	Simbolo	Unità di base	Simbolo
lunghezza	l	metro	m
massa	m	kilogrammo	kg
tempo	t	secondo	s
intensità di corrente elettrica	I	ampere	A
temperatura termodinamica	T	kelvin	K

[40] *Besozzi M, De Angelis G, Franzini C. Espressione dei risultati nel laboratorio di chimica clinica. Milano:Società Italiana di Biochimica Clinica, 1989:190pp.*

[41] *"…che ci sia una sola misura e un solo peso in tutto il Regno…e anche una misura uniforme per i vini, almeno nella stessa provincia" si chiedeva insistentemente nei "Cahiers de Doléances" ai tempi della rivoluzione francese.*

[42] *DPR n. 802 del 12 agosto 1982, Attuazione della direttiva (CEE) n. 80/181 relativa alle unità di misura. Suppl Ord Gazz Uff della Repubblica Italiana n. 302 del 3 novembre 1982.*

quantità di sostanza	n	mole	mol
intensità luminosa	I	candela	cd.

A questa vanno aggiunte due grandezze supplementari, che fanno pure esse parte integrante del sistema SI:

Grandezza supplementare	Simbolo	Unità supplementare	Simbolo
angolo piano	a,b,g...	radiante	rad
angolo solido	w,O	steradiante	sr

Per indicare multipli e sottomultipli delle unità sono previsti i fattori riportati nella seguente tabella:

Fattore	Nome	Simbolo
10^{3}	kilo	k
10^{6}	mega	M
10^{9}	giga	G
10^{12}	tera	T
10^{15}	peta	P
10^{18}	exa	E
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-0}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a
10^{-21}	zepto	z
10^{-24}	iocto	y

Tra le principali regole adottate dal sistema SI, si rammentano le seguenti:

- sono raccomandati i fattori che fanno variare l'unità di un fattore 1000 (kilo, mega, milli, micro, eccetera), come riportato nella precedente tabella;
- è sconsigliato l'uso dei fattori che fanno variare le unità di un fattore 10 o 100 (deca, etto, deci, centi), che pertanto nella precedente tabella non sono stati riportati ;
- dopo i simboli non si deve mettere il punto (cm e non cm., mol e non mol., eccetera): si tratta appunto di simboli, e non di abbreviazioni;
- non si devono usare i fattori da soli; il nome o il simbolo dell'unità non deve essere omesso (micrometro o μm, e non micron o μ; kilogrammo e non kilo, eccetera);
- non si devono usare unità con nomi d'uso tipo il lambda (λ) per il microlitro (μL) e il gamma (γ) per il microgrammo (μg);
- non devono essere formate unità con più di un prefisso (nanometro e non millimicrometro o, peggio ancora, millimicron, eccetera);
- i multipli e i sottomultipli dell'unità di massa (kilogrammo), che già contiene un prefisso, si formano anteponendo i prefissi al grammo (quindi μg e non nKg, eccetera).

Dalle grandezze e unità di base e supplementari SI è possibile ricavare grandezze e unità SI "derivate", di cui numerose hanno importanza in campo biomedico:

Grandezza	Nome unità	Simbolo unità
Frequenza (cicli al secondo)	hertz	Hz
Forza	newton	N
Pressione e tensione	pascal	Pa
Energia, lavoro, quantità di calore	joule	J
Potenza, flusso energetico	watt	W
Quantità di elettricità, carica elettrica	coulomb	C
Tensione elettrica, potenziale elettrico, forza elettromotrice	volt	V
Resistenza elettrica	ohm	Ω
Conduttanza	siemens	S
Capacità elettrica	farad	F
Flusso d'induzione magnetica	weber	Wb
Induzione magnetica	tesla	T
Induttanza	henry	H
Temperatura Celsius	grado Celsius	°C
Flusso luminoso	lumen	lm
Illuminamento	lux	lx
Attività (irraggiamento ionizzante)	becquerel	Bq
Dose assorbita	gray	Gy
Equivalente di dose	sievert	Sv

Della precedente tabella si fa notare in particolare che la temperatura nel sistema SI viene misurata in gradi Celsius (°C) e non in gradi centigradi come comunemente (ed erroneamente) si continua a dire.

Infine si ricordano a parte le seguenti grandezze e unità che, in quanto non SI, continuano ad essere ammesse, sono ammesse per usi particolari o, come accade per molte, non sono più ammesse:

Grandezza	Nome unità	Simbolo unità	Osservazioni
Volume	litro	l o L	ammessa
Massa	tonnellata	t	ammessa
Pressione e tensione	bar	bar	ammessa
Tempo	minuto	min	ammessa
	ora	h	ammessa
	giorno	d	ammessa
Pressione	millimetro di mercurio	mmHg	ammessa solamente per pressione sangue e fluidi biologici
Lunghezza	angström	Å	non più ammessa!
Lunghezza	micron	µ	non più ammessa!
Pressione	atmosfera (standard)	atm	non più ammessa!
Quantità di calore	caloria	cal	non più ammessa!
Potenza	cavallo vapore	CV o HP	non più ammessa!
Attività di radionuclidi	curie	Ci	non più ammessa!

Dose assorbita	rad	rad	non più ammessa!
Equivalente di dose	rem	rem	non più ammessa!
Esposizione (raggi x o γ)	röntgen	R	non più ammessa!

Dal confronto di quest'ultima tabella con quella che la precede è facile notare alcuni importanti cambiamenti intervenuti in campo biomedico:

- scomparsa dell'angström (Å);
- scomparsa del micron (μ);
- caloria (cal) sostituita dal joule (J);
- curie (Ci) sostituito dal becquerel (Bq);
- rad (rad) sostituito dal gray (Gy);
- rem (rem) sostituito dal sievert (Sv).

Per una trattazione completa del sistema SI vedere Fazio[43]. Per le applicazioni del sistema SI all'espressione dei risultati delle analisi di laboratorio vedere Besozzi[44].

3.2.3.2. Il numero di cifre significative

Per riportare i risultati con un adeguato numero di cifre significative, si rammenta innanzitutto che gli zeri dopo i numeri diversi da zero fanno parte delle cifre significative stesse mentre, nel caso di valori numerici minori di uno, gli zeri prima dei numeri diversi da zero non fanno parte delle cifre significative. Così per esempio 0,250 ha tre cifre significative, mentre 0,0025 ha due sole cifre significative.

Come regola pratica, nel caso di un numero a precisione infinita o comunque elevata, il numero delle cifre significative dovrebbe essere tale che l'ordine di grandezza del campo di variazione (differenza fra il valore massimo e quello minimo) calcolato ignorando l'eventuale presenza della virgola sia compreso tra 30 e 300. Si supponga che la misura della concentrazione dello ione calcio (Ca^{++}) nel sangue in un campione di 50 soggetti abbia fornito risultati compresi tra 2,06442 (valore minimo) e 3,05620 (valore massimo) mmol/L. La differenza tra il valore massimo e quello minimo (senza tenere conto della virgola) è pari a 99178 (305620 - 206422) mmol/L e quindi il numero delle cifre utilizzate per rappresentare i risultati è eccessivo. Arrotondando i risultati a due cifre decimali il valore minimo diventa 2,06 mmol/L, e il valore massimo diventa 3,06 mmol/L. La differenza (senza tenere conto della virgola) è pari a 100 mmol/L. Pertanto tre cifre significative sono sufficienti per rappresentare i risultati (Braga[45])

Nel caso di numeri con una precisione finita o comunque ridotta (che sono poi quelli che si incontrano nella pratica), la precisione finale con cui registrare i risultati non può essere maggiore di quanto permesso dal grado di indeterminazione (imprecisione) dal quale sono affetti i risultati. Il numero di cifre significative si deduce allora dalla deviazione standard della singola misura. Ed è un errore esprimere il valore numerico della misura con un numero di cifre superiore o inferiore a quello delle cifre significative soprattutto se non viene indicata anche la suddetta deviazione standard. Si supponga che la misura del calcio nel siero abbia fornito (media ± deviazione standard) un valore pari a 9,376 ± 0,10835 mg/dL. La deviazione standard indica 2 cifre significative, in quanto l'incertezza compare nel risultato già sulla seconda cifra da sinistra. Il modo corretto di scrivere il risultato è 9,4 ± 0,1 mg/dL.

---

[43] *Fazio M. Dizionario e manuale delle unità di misura. Bologna:Zanichelli, 1985:241pp.*

[44] *Besozzi M, De Angelis G, Franzini C. Espressione dei risultati nel laboratorio di chimica clinica. Milano:Società Italiana di Biochimica Clinica, 1989:190pp.*

[45] *Braga M. Argomenti di biometria. Quante cifre deve avere un numero per essere statisticamente significativo? Aggiornamento del medico 1990;14:XXXI-XXXVI.*

Discorso analogo deve essere fatto per il numero di cifre significative da usare nel caso di indici statistici (statistiche) come la media e la deviazione standard. Sinteticamente si può dire che la precisione con cui si deve riportare qualsiasi statistica non può essere maggiore dell'errore che si commette nel processo di stima. Quindi, nel caso più semplice, quello della media, non può essere maggiore del suo errore standard. Si supponga che, sempre nel caso precedente della determinazione dello calcio nel siero, alla media di 9,376 e alla deviazione standard di 0,10835 corrispondesse un errore standard della media pari 0,0320489 (l'errore standard è uguale alla deviazione standard divisa per la radice quadrata del numero delle osservazioni che compongono il campione). L'errore standard conferma che già la seconda cifra decimale è affetta da errore. Non avrebbe quindi senso riportare la media e la deviazione standard con più di due cifre decimali. La media deve essere scritta come uguale a 9,38 e la deviazione standard deve essere scritta come uguale a 0,11.

3.2.3.3. Alcuni segni matematici di largo uso

Si riportano qui di seguito alcuni segni matematici di largo uso, che talvolta vengono impiegati in modo improprio:

Segno	Significato
÷	intervallo da… a… (estremi inclusi)
=	uguale a
≡	identico a
≠	diverso da
≈	uguale a circa
∝	proporzionale a
∞	infinito

3.2.4. La tabulazione dei dati

La tabulazione di dati rappresenta l'approccio più elementare ma anche il più indispensabile alla successiva analisi statistica. Tabulando i dati è possibile verificare la eventuale mancanza di qualcuno di essi, ovvero la presenza di dati aberranti. Per questi ultimi si può prevedere, laddove possibile, la ripetizione della misura, al fine di stabilire se si tratti di uno sbaglio (che può avvenire per esempio nella trascrizione dei dati), se si tratti di un errore sperimentale rilevante a causa di un malfunzionamento occasionale dell'apparato di misura, o se il dato deve essere confermato (per la distinzione tra errori e sbagli vedere Baldini[46].

Sui dati in forma tabellare è anche possibile procedere ad una prima elaborazione, effettuando opportune trasformazioni dei dati (per esempio, la trasformazione in radice quadrata può rendere gaussiana una distribuzione non gaussiana), oppure combinando i risultati di due variabili in una nuova variabile (così, per esempio, nel caso di dati appaiati spesso quello che più interessa non sono i singoli valori delle coppie, ma piuttosto la differenza entro i dati di ciascuna coppia). Per ulteriori suggerimenti sulla tabulazione dei dati vedere Bossi[47] e Lantieri[48].

---

[46] *Baldini M. L'errore nella scienza. Biochim Clin 1991;15:28-38*).
[47] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:37-68*.
[48] *Lantieri PB, Risso D, Rovida S, Ravera G. Statistica medica ed elementi di informatica. Milano: McGraw-Hill, 1994:41-70*.

Il Menù Calcoli di Ministat offre una serie di interessanti e utili possibilità per la tabulazione (trasformazione ed elaborazione preliminare) dei dati.

SUGGERIMENTO: la tabulazione dei dati dovrebbe essere la prima a seguire la fase di raccolta dei dati, e dovrebbe sempre precedere la rappresentazione grafica e la ancora successiva elaborazione statistica.

Si consideri l'esempio di sei campioni di siero, analizzati con due metodi (metodo A e metodo B) per la determinazione in laboratorio dello stesso (ipotetico) analita. Si supponga che il metodo A abbia fornito sui sei campioni risultati, nell'ordine, pari rispettivamente a 30.1, 81.4, 51.1, 131.8, 8.7, 94.4, e che il metodo B abbia fornito sugli stessi sei campioni risultati, nell'ordine, pari rispettivamente a 28.4, 76.8, 48.2, 124.3, 8.2, e 89.1. Già una prima tabulazione dei dati

Metodo A	Metodo B	A - B
30,1	28,4	1,7
81,4	76,8	4,6
51,1	48,2	2,9
131,8	124,3	7,5
8,7	8,2	0,5
94,4	89,1	5,3

consente di meglio apprezzare il significato dei risultati ottenuti con i due metodi, se per esempio accanto a ciascuna coppia di valori viene riportata la differenza corrispondente. Ancora più chiare diventano le conclusioni che si possono trarre da questa semplice tabulazione dei dati se essi vengono riordinati in base alla differenza tra i risultati ottenuti con i due metodi.

Metodo A	Metodo B	A - B
8,7	8,2	0,5
30,1	28,4	1,7
51,1	48,2	2,9
81,4	76,8	4,6
94,4	89,1	5,3
131,8	124,3	7,5

Infine il riportare in una quarta colonna il valore del rapporto tra il risultato ottenuto con il metodo A e quello ottenuto con il metodo B chiarisce fa emergere un'informazione aggiuntiva.

Metodo A	Metodo B	A - B	A/B
8,7	8,2	0,5	1,06
30,1	28,4	1,7	1,06
51,1	48,2	2,9	1,06
81,4	76,8	4,6	1,06
94,4	89,1	5,3	1,06
131,8	124,3	7,5	1,06

In questo caso risulta in effetti del tutto evidente che la differenza tra i risultati ottenuti con i due metodi aumenta in valore assoluto all'aumentare della concentrazione dell'analita, mentre rimane costante il rapporto. Anche senza impiegare complicati modelli di regressione, la semplice tabulazione dei dati evidenzia la presenza di un differenza sistematica, di tipo proporzionale, tra i due metodi.

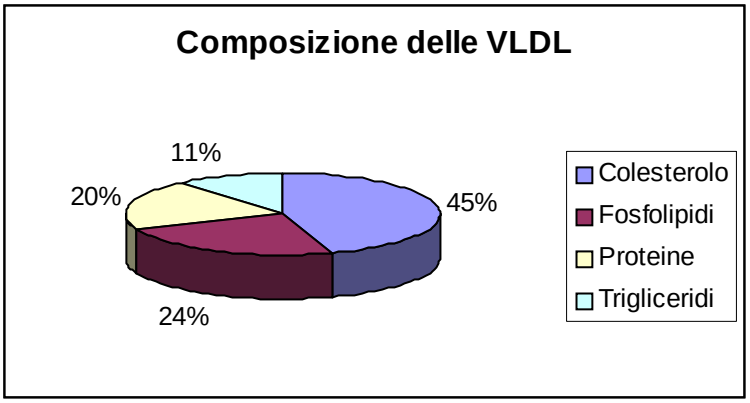
3.2.5. La rappresentazione grafica dei dati

Al pari della tabulazione, la rappresentazione grafica è un semplice ma importante strumento di analisi esplorativa dei dati: può risultare utile per la identificazione dei dati aberranti e consente di effettuare una valutazione preliminare dei risultati.
Rispetto alla presentazione tabellare, che offre una visione analitica e numericamente completa dei singoli dati raccolti, la rappresentazione grafica comporta una perdita di dettaglio (il singolo valore numerico non è più caratterizzato con esattezza): tuttavia offre il vantaggio di una visione più concisa e sintetica e, fatto molto importante, tende a fare emergere l'informazione più rilevante. Per questo motivo la rappresentazione grafica risulta utile per caratterizzare il tipo di distribuzione assunto da una variabile, come pure il tipo di relazione che intercorre tra due variabili, e addirittura può fare emergere relazioni nascoste o impreviste tra variabili. D'altra parte, proprio per lo stesso fatto di offrire una visione più concisa e sintetica e di tendere a fare emergere l'informazione più rilevante, la rappresentazione grafica è tradizionalmente usata come strumento per il riepilogo dei risultati. Per l'importanza della rappresentazione grafica come strumento per l'analisi esplorativa dei dati e per il riepilogo dei risultati vedere Bossi[49], Lantieri[50] e Campbell[51].

Il Menù Grafica di Ministat offre una serie di interessanti e utili possibilità di rappresentazione grafica dei dati.

SUGGERIMENTO: come strumento per l'analisi esplorativa dei dati, la rappresentazione grafica dovrebbe essere utilizzata dopo la raccolta e la tabulazione dei dati, e ad integrazione e supporto del giudizio derivante dalla successiva elaborazione statistica. Come strumento per il riepilogo dei risultati, può essere utilizzata al termine dell'elaborazione statistica.

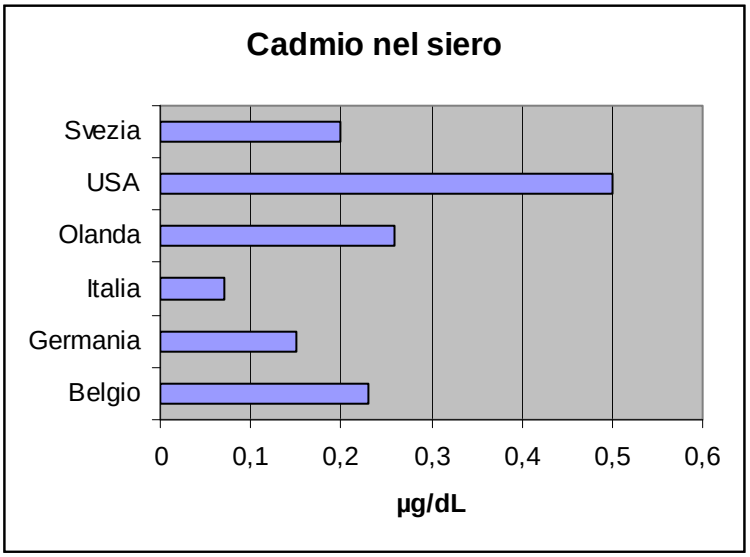
3.2.5.1. Scale nominali, scale ordinali, scale numeriche



Il tipo più semplice di scala è rappresentato dalla *scala nominale*. Corrisponde ai dati della natura più elementare, dati qualitativi che non possiedono criteri di ordinamento. Ne sono un esempio la classificazione maschio/femmina e quella sani/malati. Nei due casi precedenti si tratta di una classificazione di tipo dicotomico (due sole categorie). Ma non vi è alcuna differenza nel caso di classificazioni di tipo policotomico, cioè di dati qualitativi raggruppabili in più di due categorie (ad esempio epatite acuta virale/epatite cronica persistente/epatite cronica attiva/cirrosi epatica sono quattro categorie). La misura più intuitiva e più utilizzata, nel caso della scala nominale, è costituita dalla percentuale (o dalla proporzione o frazione). In questo caso la rappresentazione più appropriata è quella che fa uso di *grafici a torta* (*pie-chart*), come nelle figura che illustra la composizione delle VLDL (Very Light Density Lipoproteins) nel siero. Quando invece le grandezze da rappresentare sono indipendenti, e non

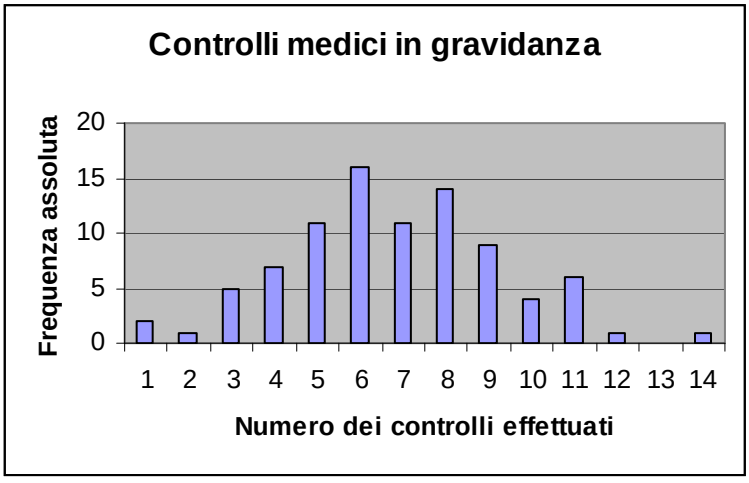
[49] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:37-68.*
[50] *Lantieri PB, Risso D, Rovida S, Ravera G. Statistica medica ed elementi di informatica. Milano: McGraw-Hill, 1994:71-101.*
[51] *Campbell MJ, Machin D. Medical statistics. A commonsense approach. Chichester: John Wiley & Sons, 1993:44-59.*

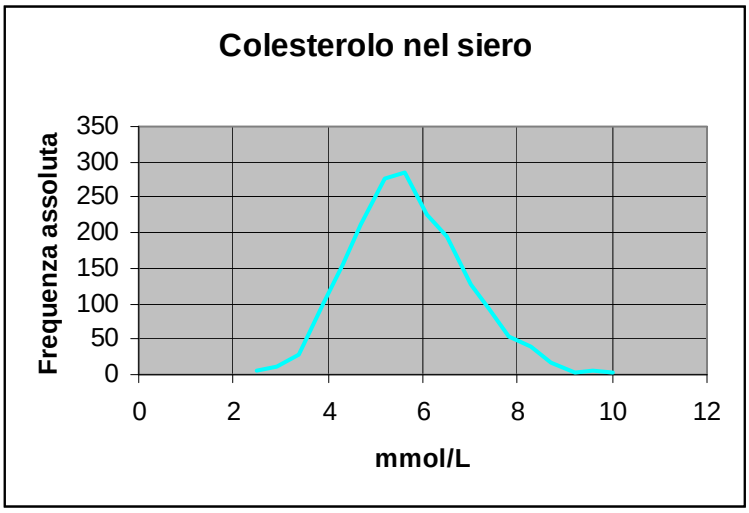
costituiscono la parte di un tutto, è consigliabile utilizzare un *grafico a barre* (*bar-chart*) con le barre staccate e disposte orizzontalmente.



Nel caso di dati qualitativi che possiedono un criterio di ordinamento si ha a che fare con una *scala ordinale*. Questo avviene per esempio nella classificazione in neonati, bambini e adulti, nella quale è intuitivo, ed automatico, applicare un ordinamento per età (neonato < bambino < adulto). Le classificazioni per ranghi ne sono un altro esempio (nelle classificazioni per ranghi il valore osservato viene trasformato nel corrispondente rango, cioè nel numero della posizione che il dato occupa nella lista ordinata dei dati). Per le scale ordinali è consigliabile utilizzare un *grafico a barre* (*bar-chart*), con le barre staccate e disposte orizzontalmente, come nella figura che illustra la concentrazione del cadmio nel siero in soggetti non esposti, in diverse nazioni.

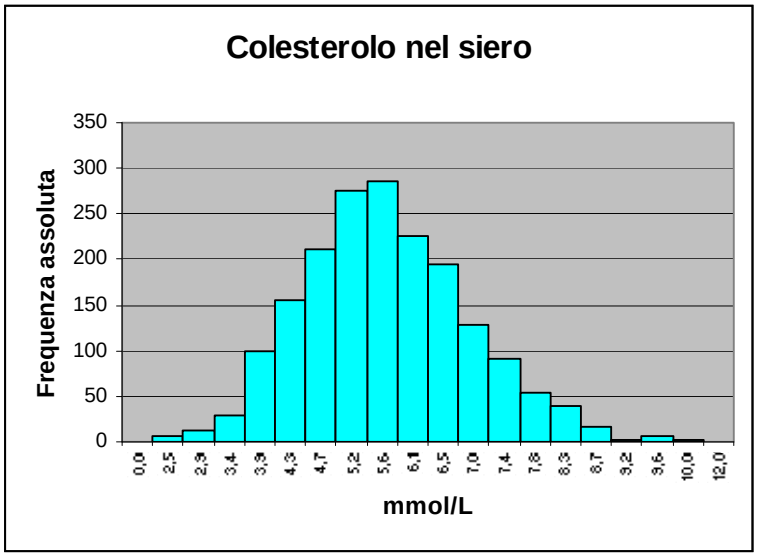
Una *scala numerica discreta* è tipicamente quella relativa a dati numerici ottenuti mediante operazioni di conteggio. Si prenda ad esempio il conteggio del decadimento di un isotopo radioattivo impiegato per un test radioimmunologico. È evidente che i colpi effettivamente contati potranno essere, ad esempio, 1822 o 1823, ma non 1822,3. La differenza minima tra due conteggi teoricamente misurabile è rappresentata da un'unità: i dati sono sì numerici, ma tra l'uno e il successivo (o il precedente) vi è un salto, essendo esclusi i valori intermedi. Si tratta di dati numerici discontinui o, meglio, discreti. Un esempio di scala numerica discreta può essere costituito dal numero di controlli medici cui si è sottoposto un gruppo di donne durante la gravidanza, come illustrato nella figura. In questo caso andrebbe utilizzato un *grafico a barre* con le barre disposte verticalmente e staccate, sempre per minimizzare l'effetto di continuità. Da evitare l'utilizzo di barre unite, come pure di un grafico a linee spezzate. I dati numerici (numero dei controlli medici effettuati) riportati in ascisse sono discreti, non esiste alcun valore compreso tra 0 e 1, tra 1 e 2, tra 2 e 3,





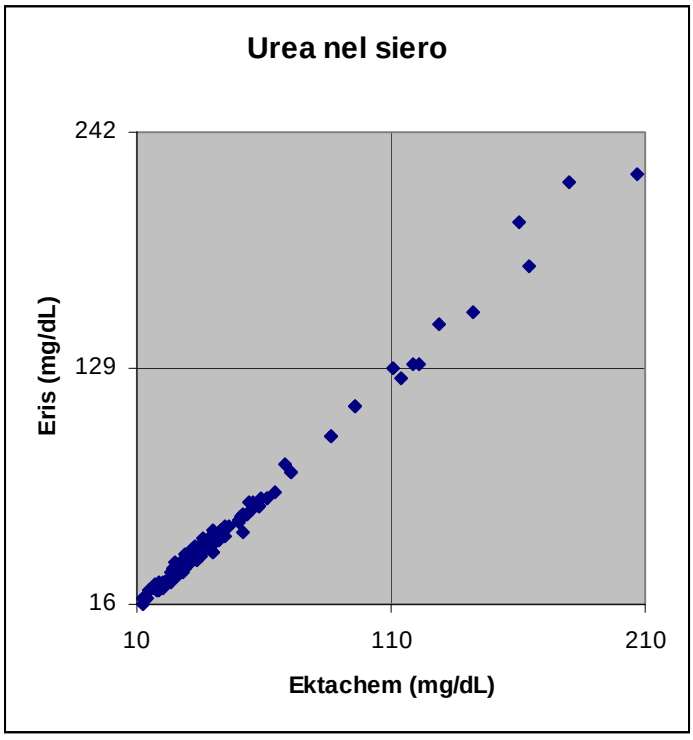
eccetera, quindi l'interpolazione mediante segmenti di retta è un errore.

La *scala numerica continua* è tipicamente quella impiegata con i dati numerici ottenuti mediante procedimenti di misura chimici e/o fisici. Un esempio può essere rappresentato dalla misura dei valori di concentrazione della creatinina nel siero. Tra un valore di 0,8 e uno di 0,9 mg/dL sono compresi infiniti possibili valori: anche se per motivi pratici, legati sia alle esigenze relative all'utilizzo clinico dei dati, sia al potere di risoluzione degli strumenti fisici impiegati, e quindi ai costi, ci si accontenta di una cifra significativa dopo la virgola. La differenza minima misurabile tra due valori di creatinina è teoricamente riducibile a piacere: i dati sono numerici, e tra l'uno e il successivo (o il precedente) sono inclusi infiniti valori intermedi. Si tratta di dati numerici continui.



Per le scale numeriche continue è appropriato di volta in volta un grafico a barre unite (*istogramma*), un *grafico a linee spezzate* (come un *poligono di frequenza*), o un *diagramma cartesiano* (nel caso in cui si debbano rappresentare contemporaneamente due variabili). Un esempio di poligono di frequenza e uno di istogramma sono illustrati nelle figure che riportano la concentrazione del colesterolo in ascisse, e in ordinate la rispettiva frequenza assoluta, osservata in 5000 soggetti apparentemente sani.

La nascita della moderna medicina di laboratorio è legata al naturale desiderio di esplorare "more scientifico" l'organismo umano, al desiderio di ottenere delle misure quantitative dei fenomeni chimici che avvengono in esso, e alla dimostrazione che alcune di queste misure hanno una insostituibile rilevanza ai fini della diagnosi e del trattamento delle malattie. Nella medicina di laboratorio sono prevalenti i dati numerici continui, e i metodi matematici e statistici per il loro trattamento. Le scale più frequentemente utilizzate sono perciò le scale numeriche continue. Seguono le scale discrete. Solo raramente sono utilizzate le scale nominali e ordinali. Nella medicina clinica è prevalente l'utilizzo delle scale nominali e ordinali.



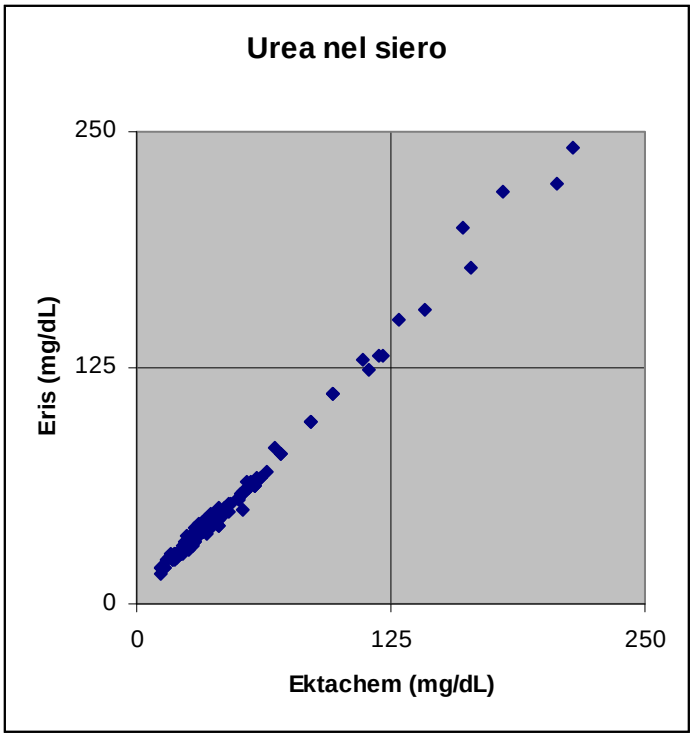
3.2.5.2. Criteri generali per la rappresentazione grafica

I grafici sono utilizzati con uno scopo ben preciso: veicolare un messaggio che sia in grado di comunicare le informazioni in modo sintetico ed efficace. Perché questo

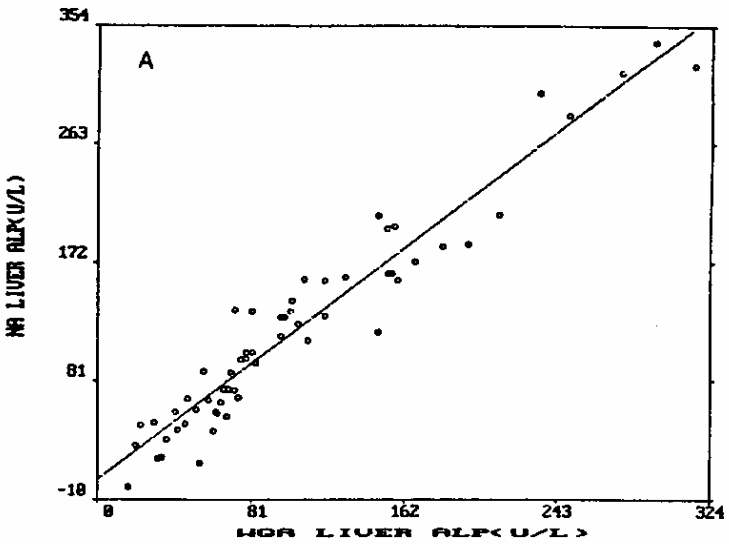
sia possibile è necessario rispettare alcuni criteri minimi, qui riassunti:

⇒ il *numero di informazioni* che si possono trasmettere con un grafico deve necessariamente essere limitato: altrimenti il messaggio non viene percepito, o viene percepito in modo confuso;

⇒ scegliere il *tipo di grafico* in funzione della natura dei dati, e quindi del genere di scala impiegata per la rappresentazione dei dati stessi (nominale, ordinale, numerica discreta e numerica continua);



⇒ le scale e quindi i *rapporti tra gli assi* devono essere scelti in modo da evitare la distorsione dei grafici e la possibilità che l'impatto visivo immediato risulti fuorviante. Nell'esempio qui riportato, relativo al confronto tra due metodi per la determinazione dell'urea nel siero, il primo diagramma cartesiano fornisce l'impressione immediata di una sostanziale equivalenza dei risultati ottenuti con due diversi analizzatori (Ektachem in ascisse e Eris in ordinate). Ciò è peraltro conseguente al fatto che le scale sono state dimensionate riportando il valore minimo e il valore massimo ottenuti con i due metodi. Dimensionando le scale in modo appropriato (in questo caso in modo che valore minimo e valore massimo siano identici sull'asse delle ascisse e sull'asse delle ordinate) risulta evidente che uno dei due metodi (quello in ordinate) presenta un errore sistematico di tipo proporzionale rispetto all'altro. Si tenga presente che l'osservazione qui fatta non è del tutto peregrina. In questa pagina viene riportato, a titolo di documentazione della serietà del problema della rappresentazione grafica, quanto comparso su una prestigiosa rivista scientifica in anni assai recenti. La distorsione del grafico, e l'utilizzo improprio delle scale nel grafico che riporta il confronto tra i risultati di due metodi per la determinazione della fosfatasi alcalina epatica, si commentano da sé;



⇒ utilizzare *scale interrotte* e non scale delle ordinate di tipo logaritmico per evitare i fenomeni di compressione di parte del grafico;

⇒ riportare sempre sugli assi il *nome della variabile* e le *unità di misura*;

⇒ di norma completare il grafico con un *titolo* che sia semplice, sintetico ed esplicativo.

Infine non dimenticare che la rappresentazione grafica dei risultati, anche se può consentire di comunicare le informazioni in modo più sintetico ed efficace, comporta inevitabilmente una perdita del dettaglio. A causa di questo può essere sconsigliabile presentare i risultati solamente in forma di grafici. Quindi sarà ogni volta necessario valutare con la massima attenzione il bilancio fra la possibilità di riassumere le informazioni in modo più sintetico ed efficace, e la necessità di fornire le stesse informazioni in modo completo (numerico). E ricorrere di volta in volta ai grafici e/o alle tabelle cercando di limitare il più possibile le rappresentazioni duplicate degli stessi dati sotto forma contemporaneamente di grafico e di tabella.

Per i criteri relativi alla corretta rappresentazione grafica dei risultati si rimanda a Bossi[52] e a Besozzi[53].

3.2.6. L'analisi statistica dei dati

Una volta effettuata la tabulazione dei dati e la loro rappresentazione grafica, si può procedere con l'analisi statistica vera e propria.

3.2.6.1. Errori e sbagli

Ripetendo più volte una misura, è possibile pervenire ad una migliore caratterizzazione dell'errore. In un insieme di misure ripetute, si definisce come *errore casuale* l'errore per cui le singole misure differiscono (casualmente, cioè senza nessuna regola apparente al succedersi delle misure stesse) tra di loro, e come *errore sistematico* l'errore per cui l'insieme (preso globalmente) delle misure ripetute si discosta dal valore vero. L'errore, l'errore casuale e l'errore sistematico sono legati all'incertezza intrinseca alle nostre conoscenze scientifiche a causa dei limiti inerenti ai sistemi (*strumenti di misura*) impiegati per rilevare i segnali provenienti dalle grandezze fisiche. Gli errori devono essere mantenuti distinti dallo *sbaglio* (*errore grossolano*), che e' un accidente tecnico, e che come tale si manifesta nel corso dell'applicazione delle conoscenze.

Nel *Vocabolario illustrato della lingua italiana* di G. Devoto e G. C. Oli si trova che "*[sbaglio]...condivide con errore...tutte le determinazioni, ma gnrl. differisce nel senso di un'attenuazione dell'importanza e della gravita'...*". A fronte di questo impiego del termine nel linguaggio comune, sta il fatto che lo sbaglio viene dallo stesso vocabolario ulteriormente definito come "*..mancanza nei confronti di un ordine corretto o di una regola...*": questa definizione si avvicina molto a quella qui data di sbaglio come accidente tecnico. Più concisamente ne *Lo Zingarelli 1995* di N. Zingarelli lo sbaglio viene definito come "*...equivoco, disattenzione, svista...*", in un senso che va ancora più chiaramente nella direzione del significato qui adottato.

Gli sbagli sono legati prevalentemente all'organizzazione e quindi ai processi di comunicazione (esempi di sbagli in chimica clinica possono essere l'errata trascrizione di un dato numerico, l'utilizzo di un reagente scaduto, lo sbaglio nell'identificazione del campione, lo sbaglio nell'interpretazione del risultato di un test di gravidanza acquistato in farmacia ed eseguito a casa propria dalla paziente, che ha frainteso i criteri per l'interpretazione dei risultati del test. Contrariamente a quanto avviene per gli errori, gli sbagli si possono evitare operando con cura, e migliorando il sistema organizzativo. Contrariamente a quanto avviene per gli errori, non e'

---

[52] *Bossi A. Guida a una corretta rappresentazione grafica dei risultati scientifici. Aggiornamento del Medico 1990;14:XXX-XXXIX.*

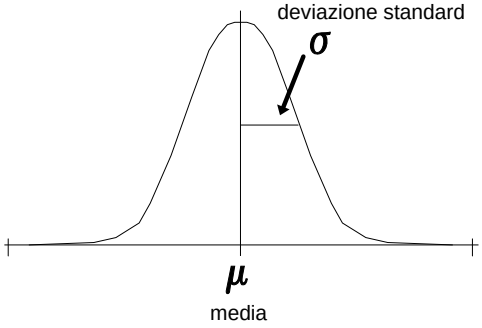
[53] *Besozzi M. Franzini C. Re G. La rappresentazione grafica dei risultati scientifici nel laboratorio Clinico. Parte I. Criteri generali per una corretta rappresentazione. Biochim Clin 1992;16:91-8[A].*

possibile fissare un livello di tolleranza per gli sbagli, cioè definire una percentuale ammissibile di sbagli: semplicemente, gli sbagli per definizione non si devono verificare[54].

4.2. La distribuzione gaussiana

Essendo π uguale a 3,1415 ed essendo e la base dei logaritmi naturali (e = 2,7183), per un dato valore della x l'equazione della distribuzione gaussiana

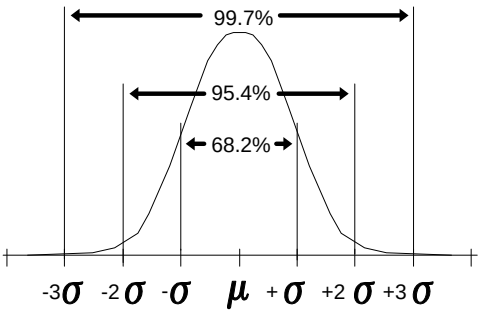
$$y = [\, 1 / (\sigma \cdot \sqrt{2\pi}) \,] \cdot e^{-(x-\mu)^2 / 2\sigma^2}$$



risulta completamente determinata dai due soli valori μ e σ.

Il valore μ (la *media della popolazione*) rappresenta la misura di posizione della distribuzione gaussiana, mentre il valore σ (la *deviazione standard della popolazione*) rappresenta la misura di dispersione della distribuzione gaussiana.

Per convenzione la distribuzione gaussiana teorica viene assunta come avere una media $\mu = 0$ e una deviazione standard $\sigma = 1$. In questo modo la superficie sottesa dalla curva gaussiana teorica ha una superficie uguale a 1, che corrisponde al valore di probabilità $p = 1$ che include tutte le osservazioni. Risulta pertanto facile calcolare il numero (espresso in percentuale) delle osservazioni che cadono all'interno di un dato intervallo (per esempio entro un numero dato di deviazioni standard rispetto alla media).



Per confrontare una distribuzione gaussiana (o ritenuta tale) con la gaussiana teorica è necessario standardizzare i risultati ottenuti in modo che la media osservata μ venga riportata a 0 e la deviazione standard osservata σ venga riportata a 1. In pratica ciò si ottiene trasformando ogni valore della x osservato nella corrispondente z (*deviata normale standardizzata*), essendo

$$z = (x - \mu) / \sigma$$

L'operazione inversa (trasformazione di una deviata normale standardizzata z nel valore x corrispondente ad un dato appartenente a una distribuzione gaussiana con media μ e deviazione standard σ risulta possibile applicando la trasformazione

$$x = \mu + \sigma \cdot z$$

Infine si ricorda che la media μ e la deviazione standard σ sono denominate globalmente "parametri" della distribuzione gaussiana. Per questo le tecniche statistiche che basano la loro validità (la validità delle conclusioni che mediante esse è possibile trarre) su assunti distribuzionali di gaussianità sono denominate *tecniche statistiche parametriche*. In contrapposizione alle tecniche

[54] *Baldini M. L'errore nella scienza. Biochim Clin 1991;15:28-38.*

che basano la loro validità su assunti distribuzionali minimi, e comunque non richiedono una distribuzione gaussiana dei dati per garantire conclusioni valide, che sono denominate *tecniche statistiche non parametriche*. Non è comunque un caso che in presenza di distribuzioni esattamente gaussiane tecniche statistiche parametriche e tecniche statistiche non-parametriche forniscono identici risultati.

I *metodi statistici parametrici* sono basati sull'assunto che i dati campionari siano estratti da una popolazione che ha una distribuzione gaussiana (esistono in realtà come vedremo poco più avanti anche *metodi statistici non-parametrici*, che possono, anzi devono essere impiegati quando i dati non sono distribuiti in modo gaussiano).

In effetti, come riconoscono Snedecor e Cochran[55] "*...è stupefacente che la distribuzione gaussiana abbia dominato sia la teoria che la pratica statistica...*". Tuttavia sono gli stessi Autori a indicare i quattro argomenti a favore dell'utilizzo della statistica parametrica:

⇒ la distribuzioni di molte variabili è approssimativamente gaussiana;
⇒ per distribuzioni non gaussiane, semplici trasformazioni matematiche (come per esempio la radice quadrata e la trasformazione logaritmica dei dati) consentono spesso di ottenere distribuzioni approssimativamente gaussiane;
⇒ la distribuzione gaussiana può essere trattata facilmente in termini matematici;
⇒ anche se la distribuzione della popolazione originaria non è gaussiana, la distribuzione delle medie campionarie tende a divenire gaussiana (teorema centrale del limite). Quest'ultimo è il singolo argomento più consistente a favore della statistica parametrica.

3.2.6.2. Misure ripetute della stessa entità

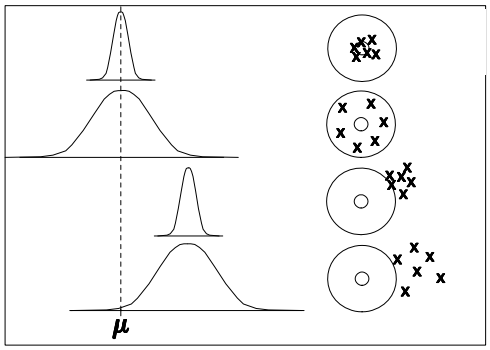
Si consideri l'operazione di misura ripetuta della concentrazione del colesterolo nel siero di uno stesso individuo. La misura viene effettuata *n* volte su un unico campione di siero, raccolto in un'unica volta.
I concetti di base sottesi sono i seguenti:

⇒ a causa dei limiti inerenti ai sistemi (*strumenti di misura*) impiegati per rilevare i segnali provenienti dalle grandezze fisiche, ogni *misura sperimentale* è inevitabilmente accompagnata da un qualche grado di *incertezza;*
⇒ si può assumere che esista, ogniqualvolta viene effettuata una misura sperimentale, un valore teorico, detto *valore vero*: quello che si otterrebbe se la misura non fossa affetta da alcuna incertezza;
⇒ per effetto della incertezza che la caratterizza, la misura sperimentale rappresenta una *stima* più o meno approssimata del valore vero;
⇒ la differenza tra una singola misura sperimentale e il suo valore vero rappresenta l'*errore* ;
⇒ ripetendo più volte una misura, è possibile pervenire ad una migliore caratterizzazione dell'errore;
⇒ in un insieme di misure ripetute, si definisce come *errore casuale* l'errore per cui le singole misure differiscono (casualmente, cioè senza nessuna regola apparente al succedersi delle misure stesse) tra di loro;
⇒ in un insieme di misure ripetute, si definisce come *errore sistematico* l'errore per cui l'insieme (preso globalmente) delle misure ripetute si discosta dal valore vero;
⇒ un insieme di misure ripetute dello stesso fenomeno può essere riassunto sotto forma di una *distribuzione di frequenze*;
⇒ nel caso degli errori di misura si assume generalmente che la distribuzione di frequenze segua un *modello gaussiano*;

[55] *Snedecor G W, Cochran WG. Statistical Methods. VII Edition. Ames:The Iowa State University Press, 1980:41.*

⇒ in una *distribuzione gaussiana* la *media aritmetica* è la *misura di posizione* della distribuzione, mentre la *deviazione standard* è la *misura di dispersione* dei dati attorno alla media;

⇒ la *precisione* è il grado di concordanza di una serie di misure ripetute. Convenzionalmente la precisione, che non ha valore numerico, viene misurata in termini di una quantità (*imprecisione*) il cui valore diminuisce all'aumentare dell'attendibilità delle misure;



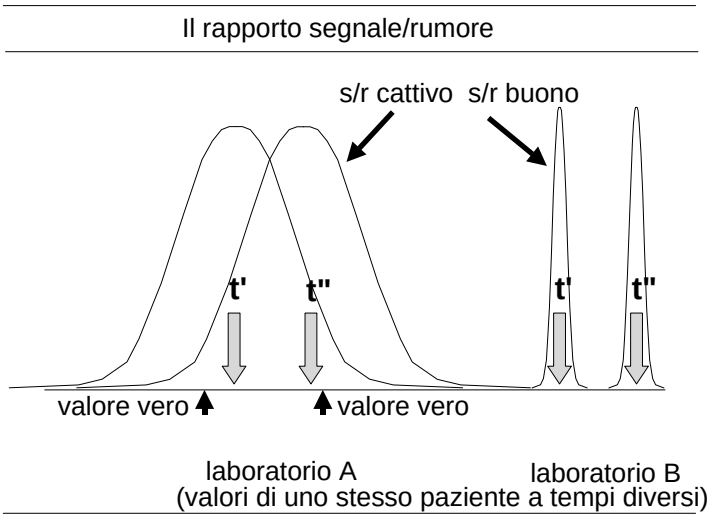
⇒ l'*imprecisione*, calcolata come *deviazione standard* (*s*), può essere espressa sia come tale (e quindi nelle unità originali) che come percentuale della media (*deviazione standard relativa*, meglio nota come *coefficiente di variazione*, *CV*). L'imprecisione rappresenta una *stima* dell'errore casuale. Tale stima è tanto più attendibile quanto più numerose sono le misure utilizzate per effettuarla (*numerosità campionaria*);

⇒ l'*accuratezza* è il grado di concordanza tra la media di una serie di misure ripetute e il valore vero. Convenzionalmente l'accuratezza viene misurata in termini di una quantità (*inaccuratezza*) il cui valore diminuisce all'aumentare dell'attendibilità delle misure;

⇒ l'*inaccuratezza*, calcolata come differenza tra la media campionaria e il valore vero, può essere espressa sia come tale (e quindi nelle unità originali) che come percentuale del valore vero. L'inaccuratezza rappresenta una *stima* dell'errore sistematico. Tale stima è tanto più attendibile quanto più numerose sono le misure utilizzate per effettuarla (numerosità campionaria).

Il fatto che imprecisione e inaccuratezza siano tra loro indipendenti è ben noto, ed è dovuto al fatto che esse sono per l'appunto le due grandezze indipendenti che concorrono a determinare la distribuzione gaussiana, ed è evidenziato nella figura.

L'indipendenza di tra imprecisione e inaccuratezza, può essere esemplificata mediante un esempio, che chiarisce anche l'importanza del concetto di rapporto segnale/rumore già illustrata in modo teorico in precedenza.



A un paziente sono effettuati due prelievi di sangue, uno al tempo t' e uno al tempo t", al fine di determinare la concentrazione di un farmaco nel siero. Quello che interessa, in questo caso, è di stabilire se la concentrazione del farmaco al tempo t" differisca significativamente da quella rilevata al tempo t', per capire se all'aumento nella posologia del farmaco stabilito dal medico corrisponda un aumento della sua concentrazione nel siero. Per questo i due campioni sono suddivisi ciascuno in due aliquote, che sono inviate contemporaneamente in due laboratori diversi (laboratorio A e laboratorio B), che ottengono risultati diversi. Le concentrazioni rilevate nel laboratorio A sono molto vicine al valore vero della concentrazione del farmaco nel siero, essendo il valore vero il valore di concentrazione che si avrebbe se un laboratorio fosse in grado di effettuare le misure della concentrazione del farmaco senza errore. Le concentrazioni rilevate nel laboratorio B sono molto lontane al valore vero della concentrazione del farmaco nel siero, essendo il valore vero il valore di concentrazione che si avrebbe se un laboratorio fosse in grado di effettuare le misure della concentrazione del farmaco senza errore. Tuttavia il laboratorio A, che appare più accurato del

laboratorio B, è più impreciso di questo. Se consideriamo come segnale la differenza tra la concentrazione del farmaco al tempo t' e quella al tempo t", il laboratorio A fornisce risultati che sono affetti da un rapporto segnale/rumore tale da renderli indistinguibili tra di loro. In effetti l'incertezza che caratterizza i risultati delle due determinazioni appare tale che le distribuzioni da cui essi sono tratti sono praticamente quasi indistinguibili. Diversa è la situazione del laboratorio B, per il quale il rapporto segnale/rumore appare buono, per cui l'incertezza che caratterizza i risultati delle due determinazioni appare contenuta, tanto che la concentrazione al tempo t" può essere agevolmente riconosciuta come diversa da quella presente al tempo t'. Questo esempio illustra anche come, nel caso del monitoraggio, la riduzione dell'imprecisione può essere un obiettivo prioritario rispetto alla riduzione dell'inaccuratezza (ammesso che questa non sia tale da comportare la possibilità che le concentrazioni assolute del farmaco, apparentemente normali, siano in realtà nell'ambito dei valori tossici).

3.2.6.3. Misure ripetute della stessa quantità

Si consideri di nuovo l'operazione di misura ripetuta della concentrazione del colesterolo nel siero di uno stesso individuo. La misura in questo caso viene effettuata una sola volta su ciascuno di *n* campioni di siero, raccolti tutti nello stesso giorno, ad orari differenti, allo stesso individuo.
I concetti di base sottesi sono sostanzialmente identici a quelli riportati al punto precedente. Va fatta eccezione per l'imprecisione, che questa volta, dato il disegno sperimentale adottato, (si assume per semplicità come nulla la variabilità intrinseca del procedimento di misura utilizzato, cioè la variabilità analitica), rappresenterà una misura della variabilità biologica del colesterolo in un singolo individuo (variabilità biologica intra-individuale).

Si consideri infine l'operazione di misura della concentrazione del colesterolo nel siero di più individui. La misura in questo caso viene effettuata una sola volta su ciascuno di *n* campioni di siero, raccolti tutti nello stesso giorno ad altrettanti individui diversi.
I concetti di base sottesi ancora una volta sono sostanzialmente identici a quelli riportati al punto precedente. Va fatta eccezione per l'imprecisione, che questa volta, dato il disegno sperimentale adottato, (si assume per semplicità come nulla la variabilità intrinseca del procedimento di misura utilizzato, cioè la variabilità analitica), rappresenterà una misura della variabilità biologica del colesterolo in individui diversi (variabilità biologica inter-individuale).

In sostanza i tre disegni sperimentali descritti individuano il semplice percorso logico seguito nel trasporre un modello concepito per descrivere l'errore di misura a situazioni che descrivono la variabilità riscontrata entro e tra gli individui.

CAPITOLO 4
METODI DI BASE IN BIOSTATISTICA

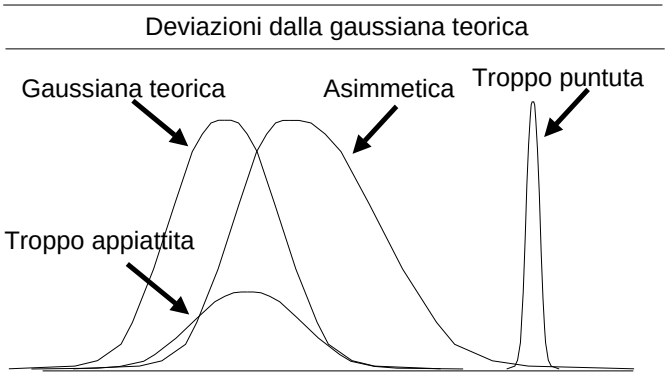
In questo capitolo sono illustrati in modo conciso i concetti alla base delle tecniche statistiche il cui utilizzo pratico mediante il programma Ministat 1.1 è presentato nel capitolo successivo (nel quale sono riportati anche i riferimento bibliografici ai testi statistici relativi), e le cui formule e algoritmi di calcolo sono ampiamente trattate nell'ultimo capitolo del libro.

4.1. Statistiche elementari parametriche

Si consideri un caso molto semplice di raccolta di dati, rappresentato dalla determinazione della concentrazione dei trigliceridi nel siero di 1000 soggetti non selezionati, a digiuno. Le domande che ci si pone sono le seguenti: (i) possibile definire dei descrittori che consentano di stabilire se le concentrazioni nei vari soggetti differiscono tra di loro "tanto" o "poco" e di darci una visione sintetica ma oggettiva dei risultati ottenuti? (ii) i descrittori ottenuti analizzando un numero limitato di casi (un campione) possono essere considerati descrittori attendibili sia del campione che dell'universo (la popolazione) da cui questo campione proviene?

La risposta alla prima domanda è "sì": in effetti la media, misura di posizione dei dati campionari nell'universo dei valori possibili, la deviazione standard, misura di dispersione dei dati attorno al loro valore medio, e la numerosità campionaria, consentono di sintetizzare in modo oggettivo l'intera informazione riguardante il campione.

La risposta alla seconda domanda è un "sì" condizionato: infatti i descrittori ottenuti dal campione rappresentano in modo attendibile il campione stesso e possono essere generalizzati alla popolazione solamente "se" gli assunti posti alla base del modello matematico che viene impiegato per calcolare tali descrittori trovano riscontro sia nel campione che nella popolazione. In caso contrario la descrizione del campione e la stima delle caratteristiche della popolazione effettuata a partire dal campione possono risultare inattendibili. Ed è esattamente quello che, nel caso delle statistiche elementari, può accadere se media e deviazione standard vengono impiegate per descrivere dei dati che non siano distribuiti in modo gaussiano, quando il modello matematico che propone media e deviazione standard come descrittori attendibili assume a fondamento della propria validità il fatto di operare su dati provenienti da una distribuzione gaussiana. Asimmetria (distribuzione con una coda troppo allungata a sinistra o a destra) e curtosi (distribuzione troppo piatta o troppo puntuta) della distribuzione osservata, forniscono una stima della deviazione della distribuzione dalla gaussiana teorica.



Si ricorda ancora una volta che la media μ e la deviazione standard σ sono i "parametri" che caratterizzano la distribuzione gaussiana. Per questo le tecniche statistiche che basano la loro validità (la validità delle conclusioni che mediante esse è possibile trarre) su assunti distribuzionali di gaussianità sono denominate in generale *tecniche statistiche parametriche*. In contrapposizione

alle tecniche che basano la loro validità su assunti distribuzionali minimi, e comunque non richiedono una distribuzione gaussiana dei dati per garantire (nei limiti in cui questo termine ha significato nella statistica inferenziale) conclusioni valide, che sono denominate in generale *tecniche statistiche non-parametriche*. Non è comunque un caso che in presenza di distribuzioni esattamente gaussiane tecniche statistiche parametriche e tecniche statistiche non-parametriche forniscono identici risultati.

Si provi ad eseguire mediante Ministat il calcolo delle Statistiche Elementari (parametriche) sul valore dei trigliceridi (campo TRIGLI) del file COLEST.DBF. Caricare il file nella tabella dati di Ministat mediante l'opzione Importa File (in formato dBase III®) del Menù File: quindi selezionare, facendo click con il mouse sul nome della colonna (variabile), la colonna TRIGLI. Come detto si tratta della determinazione della concentrazione dei trigliceridi nel siero di 1000 soggetti non selezionati, a digiuno, espressa in mg/dL.

Il calcolo delle Statistiche Elementari (parametriche) porta ad una tabella dei percentili nella quale il 2,5-esimo percentile parametrico e il 97,5-esimo percentile parametrico (rispettivamente la media meno 1,96 volte la deviazione standard e la media più 1,96 volte la deviazione standard) sono pari rispettivamente a -45,3 e a 318,3 mg/dL (i valori degli altri percentili calcolati da Ministat sono stati omessi per concisione).

		Limiti di confidenza al 90%	
Percentile	Valore	inferiore	superiore
2,5	-45,31548	-53,56551	-37,06544
97,5	318,33348	310,08344	326,58351

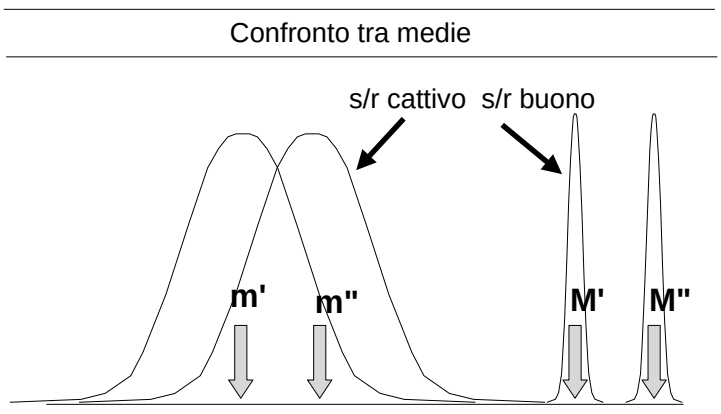
In base a questo dovremmo affermare che nel 95% dei soggetti esaminati i trigliceridi hanno una concentrazione compresa tra -45,3 e 318,3 mg/dL: un'affermazione che, a dispetto del fatto che i risultati sono stati ottenuti con un metodo statistico rigoroso, nessuno si sentirebbe di sostenere.

Questo esempio è stato utilizzato ad hoc, per spingere il lettore subito al paragrafo che illustra le statistiche elementari non parametriche, che nel caso specifico forniscono la soluzione corretta al problema.

4.2. Confronto tra medie

Si considerino due campioni con media m' e media m" rispettivamente. La domanda che ci si pone è: la media m' è significativamente diversa dalla media m"?

La risposta a questa domanda può essere assicurata sulla base di una tecnica statistica parametrica (il test t di Student) che si basa su un concetto che si rifà a quello di rapporto segnale/rumore. Molto semplicemente nel caso delle medie m' e m", sarà poco probabile che vi sia una differenza significativa, a causa del fatto che le medie sono caratterizzate da un grado di incertezza (che corrisponde all'embricarsi delle due



distribuzioni a confronto) elevato.

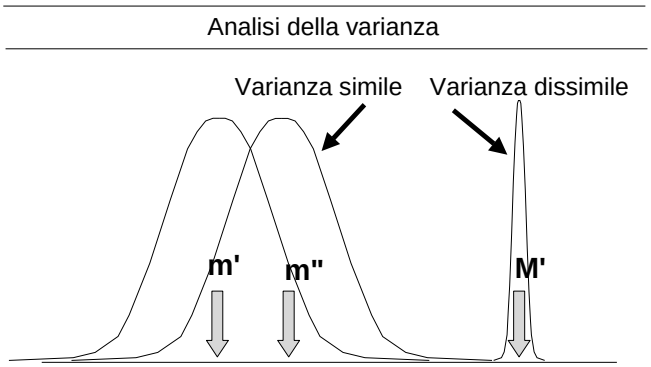
Si considerino ora due campioni con media M' e media M" rispettivamente. La domanda che ci si pone è nuovamente: la media M' è significativamente diversa dalla media M"?

Nel caso delle medie M' e M", esiste intuitivamente una differenza significativa, a causa del fatto che le medie sono caratterizzate da un grado di incertezza (che corrisponde all'embricarsi delle due distribuzioni a confronto) sostanzialmente nullo.

Nei due casi esemplificati si parla di campioni indipendenti. Un esempio potrebbe essere rappresentato dalla concentrazione del colesterolo nel siero di soggetti di sesso maschile e di sesso femminile. Tuttavia il confronto tra medie può essere utilizzato anche nel caso di dati appaiati. L'appaiamento dei dati consente di ottimizzare l'omogeneità fra i due campioni posti a confronto, che differiranno fra di loro solamente per il fattore che lo sperimentatore ha deliberatamente introdotto. Un esempio può essere quello della determinazione del colesterolo nel siero nello stesso soggetto, prima e dopo la prescrizione di una terapia. Utilizzando un numero elevato di soggetti (la numerosità campionaria può essere stabilita a priori sulla base della "forza" della conclusione cui si vuole giungere), si può arrivare a concludere se la concentrazione del colesterolo dopo la terapia sia variata significativamente.

4.3. Analisi della varianza

L'analisi della varianza generalizza il concetto del confronto tra medie, estendendolo al confronto contemporaneo tra più medie. Si considerino tre campioni con media rispettivamente m', m" e M'. Le domande che ci si pone sono: la media m' è significativamente diversa dalla media m"? E la media m' è significativamente diversa dalla media M'? E la media m" è significativamente diversa dalla media M'?



A queste tre domande si può rispondere con l'analisi della varianza[56], che segnalerà che le tre distribuzioni presentano varianze dissimili (significativamente diverse) tra di loro, in questo caso essendo evidente che è la varianza che corrisponde alla distribuzione con media M' ad essere la responsabile di tale differenza.

Se si parte dal presupposto che da una stessa popolazione sono estratti campioni con uguale varianza (e ovviamente uguale media, a meno di una differenza minima conseguente all'errore di campionamento), nel caso specifico si arriva a concludere, con un piccolo salto logico, che le medie campionarie sono significativamente diverse tra di loro. L'analisi della varianza è una tecnica generale, che si presta ad essere estesa a situazioni ancora più complesse, nelle quali peraltro alla complessità dei modelli adottati fa riscontro sempre la semplice base concettuale qui illustrata.

[56] *Si rammenta che la varianza è il quadrato della deviazione standard, che a sua volta fornisce la misura dell'ampiezza della distribuzione.*

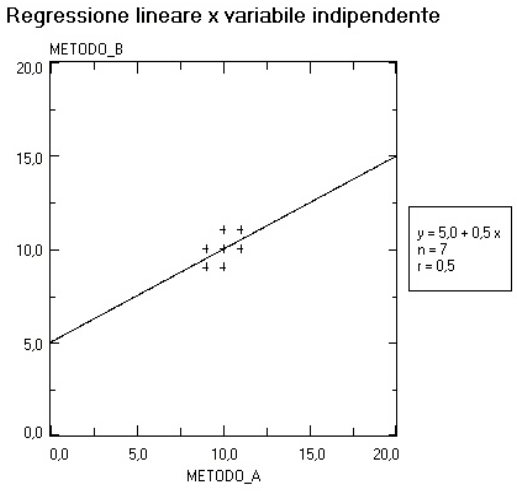
4.4. Regressione lineare

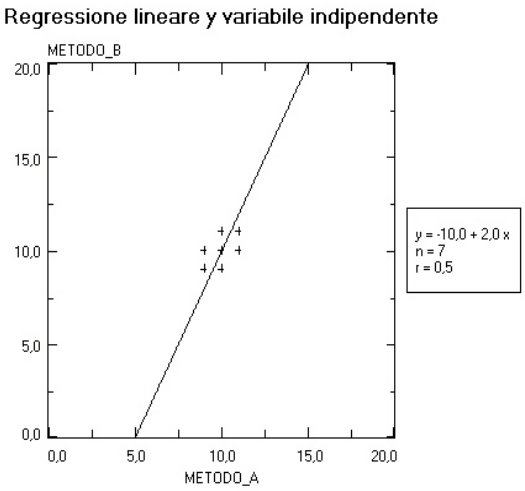
Esistono numerose occasioni nelle quali quello che interessa è ricostruire la relazione di funzione che lega due variabili, la variabile *y* (variabile dipendente, in ordinate) alla variabile indipendente (variabile *x*, in ascisse); se si ritiene che la relazione esistente fra le due variabili possa essere convenientemente descritta mediante una retta, l'equazione di tale retta può essere calcolata mediante la tecnica statistica nota come regressione lineare. Tale denominazione deriva dagli studi sull'ereditarietà condotti da F. Galton sul finire dell'800. Nel corso di questi studi vennero, fra l'altro, registrate le altezze dei componenti di più di 1000 gruppi familiari. Ponendo su un sistema di assi cartesiani in ascisse le altezze dei padri e in ordinate le altezze dei figli, si notò un fatto: sebbene in genere padri più alti avessero figli più alti (come del resto era atteso), padri che erano di 16 centimetri circa più alti della media dei padri, avevano figli che erano solamente 8 centimetri circa più alti della media dei figli. In altre parole sembrava che vi fosse un "tornare indietro", una regressione delle altezze dei figli rispetto a quelle dei padri : e il termine che descriveva il risultato di questa iniziale applicazione, finì con l'essere impiegato per indicare la tecnica statistica, ed è rimasto ancora oggi nell'uso, anche se l'attributo di "regressione" non avrebbe più alcun significato di essere.

Si consideri il confronto tra due metodi analitici per la determinazione del calcio nel siero. E la domanda: "i due metodi forniscono risultati identici?". Il confronto dà i seguenti risultati (concentrazione in milligrammi per decilitro di siero, mg/dL):

Metodo A	Metodo B
9,0	9,0
10,0	10,0
11,0	11,0
9,0	10,0
10,0	9,0
10,0	11,0
11,0	10,0

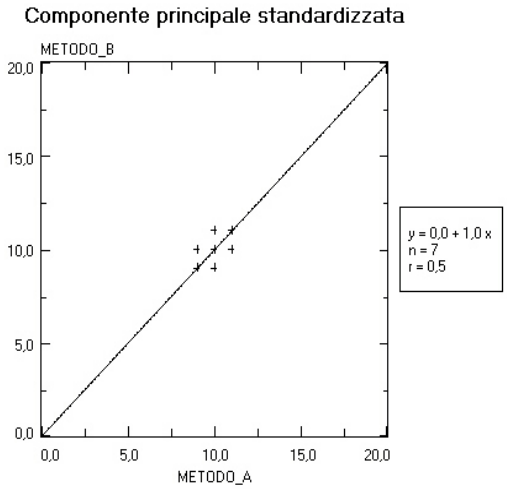
Utilizzando tre diversi modelli di regressione lineare, si ottengono i seguenti risultati:





Come è possibile che differenti modelli di regressione forniscano conclusioni così diverse? Quale modello fornisce i risultati più attendibili? Ebbene, gli assunti alla base dei modelli di regressione

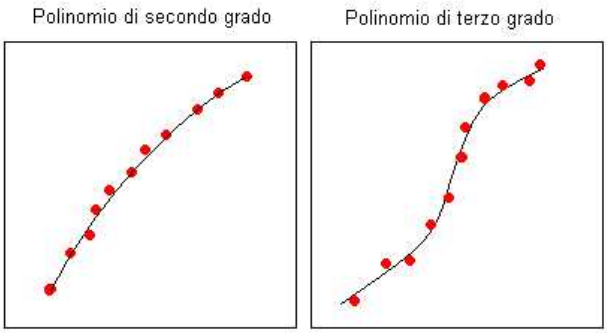
lineare utilizzati (x variabile indipendente, y variabile indipendente e componente principale standardizzata) sono identici, tranne che per il fatto che la regressione x variabile indipendente assume che la variabile in ascisse (x) sia la variabile misurata senza errore, la regressione y variabile indipendente assume che la variabile in ordinate (y) sia la variabile misurata senza errore, mentre la componente principale standardizzata assume che sia la variabile in ascisse sia la variabile in ordinate siano misurate con un errore dello stesso ordine di grandezza[57].



La risposta alle due domande è che, nel caso specifico, il rapporto tra il segnale contenuto nei dati e in rumore introdotto dagli assunti intrinseci a ciascuno dei diversi modelli di regressione è troppo basso. In effetti il fatto che r sia uguale solamente a 0,5 indica che l'informazione contenuta nei dati è troppo scarsa per essere attendibile. In questo caso la "forma" che viene dati ai dati, e quindi la conclusione che da essi viene tratta, risulta influenzata più dagli assunti del modello di regressione che dal contenuto informativo dei dati stessi!

4.5. Regressione polinomiale

La regressione polinomiale utilizza lo stesso metodo matematico della regressione lineare, ma assume che la relazione di funzione che caratterizza i dati sia meglio descritta, anziché da una retta, da un polinomio (in Ministat possono essere utilizzati rispettivamente un polinomio di secondo grado e un polinomio di terzo grado). La possibilità che la relazione di funzione che lega la variabile in ordinate (y) a quella in ascisse (x) sia meglio descritta da una curva, anziché da una retta, può essere verificata mediante un test per la linearità (incluso in Ministat) che, se significativo, consente di affermare che la relazione è presumibilmente meglio descritta da una funzione curvilinea.



In realtà in tutti i casi nei quali è necessario descrivere una relazione tra due variabili, rappresentabile su un piano cartesiano, è indispensabile che tutta una serie di ragionamenti a monte abbiano portato a concludere che le retta rappresenta un modello applicabile. Questo è vero per

[57] *Esiste anche un modello di regressione lineare, la regressione di Deming, con la quale la retta di regressione può essere calcolata attribuendo alla x e alla y l'esatto entità dell'errore di misura che le caratterizza. Si fa notare che la regressione di Deming (i) nel caso in cui l'entità dell'errore attribuito alla variabile x sia di gran lunga inferiore a quello attribuito alla variabile y, fornisce lo stesso risultato della regressione x variabile indipendente, (ii) nel caso in cui l'entità dell'errore attribuito alla variabile x sia di gran lunga superiore a quello attribuito alla variabile y, fornisce lo stesso risultato della regressione y variabile indipendente, e (iii) nel caso in cui l'entità dell'errore attribuito alla variabile x e alla variabile y sia dello stesso ordine di grandezza, fornisce lo stesso risultato della componente principale standardizzata.*

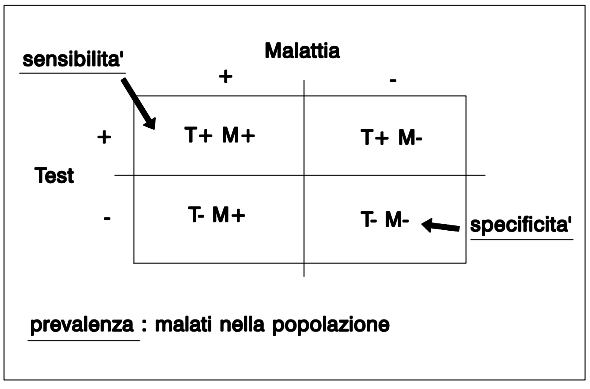
esempio nel caso della relazione tra concentrazione (in ascisse) e assorbanza (in ordinate) nel caso di reazioni fotometriche, per le quali è applicabile la legge di Lambert e Beer. In questo caso, assunta la linearità della relazione, il test di linearità può essere applicato per verificare l'ambito di intervalli di concentrazione entro i quali la relazione lineare è valida nel caso del metodo analitico in questione. In altri casi la relazione può essere per definizione assunta non lineare, come avviene per esempio nel caso di reazioni antigene-anticorpo, nelle quali la quantità di immunocomplessi che si forma (e quindi il segnale prodotto dalla reazione) è noto che varia in modo meno che proporzionale all'aumentare dell'antigene.

4.6. Regressione multipla

La regressione lineare multipla, quando applicata a due variabili, si riduce alla regressione lineare standard, cioè alla rappresentazione di una retta su un piano. Quando applicata a tre variabili, consente di rappresentare un piano in uno spazio a tre dimensioni. Inutile dire cosa rappresenta nel caso di quattro o più variabili: sta di fatto che la regressione lineare multipla può essere applicata a quattro e più variabili al fine di definire il loro grado di associazione. In generale la regressione lineare multipla consente di calcolare l'equazione di regressione multipla che lega una variabile dipendente (y) a due o più variabili indipendenti, assumendo che le variabili siano linearmente correlate tra di loro. Si fa notare che in questo caso risulta non difficile, bensì letteralmente impossibile verificare con quello che resta comunque il primo e in molti casi migliore metodo, e cioè la rappresentazione grafica, l'effettiva validità della relazione lineare nel descrivere la relazione tra le variabili. Per questo motivo un utilizzo di questa tecnica "alla cieca" espone a rischi inaccettabili nell'opinione dello scrivente.

4.7. Statistica bayesiana

Il *teorema di Bayes* consente, conoscendo la prevalenza di una malattia, e la sensibilità (positività nei malati) e la specificità (negatività nei sani) di un test per la sua diagnosi, di calcolare la probabilità di malattia in caso di test positivo (o la probabilità di assenza della malattia in caso di test negativo).



Si consideri il caso di una malattia che può essere, in un soggetto, presente (soggetto malato, indicato come M+) o assente (soggetto sano, M-). Si consideri un test diagnostico per questa malattia (può essere indifferentemente un test di laboratorio o un segno clinico), test che può essere positivo (T+) o negativo (T-). Le possibili combinazioni di test e malattia sono quattro. Si definisca come *sensibilità* (T+M+) la positività del test nei malati (una sensibilità del 100% significa che il test è positivo nel 100% dei malati, una sensibilità del 90% significa che il test è positivo nel 90% dei malati, e così via). Si definisca come *specificità* (T-M-) la negatività del test nei sani (una specificità del 100% significa che il test è negativo nel 100% dei sani, una specificità del 90% significa che il test è negativo nel 90% dei sani, e così via). Si definisca infine come *prevalenza* della malattia il numero dei soggetti malati presenti, in un dato istante, nella popolazione. Il teorema di Bayes, date queste tre grandezze (sensibilità, specificità e prevalenza) consente di calcolare il *valore predittivo del test positivo*, (come già accennato, è possibile anche calcolare il *valore predittivo del test negativo*).

Un esempio di come il teorema di Bayes in realtà possa essere applicato in modo intuitivo, è il seguente. Si consideri un test destinato a rivelare al presenza nel siero di anticorpi anti-HIV. Si

assuma che questo test abbia una sensibilità del 100% (il test, quindi, è positivo nel 100% dei malati). Si assuma che questo test abbia una specificità del 99.7% (il test, quindi, è negativo nel 99.7% dei soggetti sani). Si sa che la prevalenza della positività agli anticorpi anti-HIV è del 3 per mille (nella popolazione, su 1000 soggetti presi a caso, 3 sono positivi agli anticorpi anti-HIV). Supponendo di effettuare il test su 1000 soggetti presi a caso i risultati saranno i seguenti: 3 soggetti presenteranno positività agli anticorpi anti-HIV, in quanto la prevalenza è del 3 per mille (veri positivi); inoltre effettuando l'analisi su 1000 soggetti, a causa del fatto che la specificità del test è del 99.7% ci dovremo aspettare 3 positivi su 1000 soggetti sani (falsi positivi). Essendo in totale 6 i soggetti positivi, e 3 i veri positivi, avremo quindi un valore predittivo del test positivo pari a 3/6, cioè un valore predittivo del test positivo pari al 50% (*p* pari a 0.5).

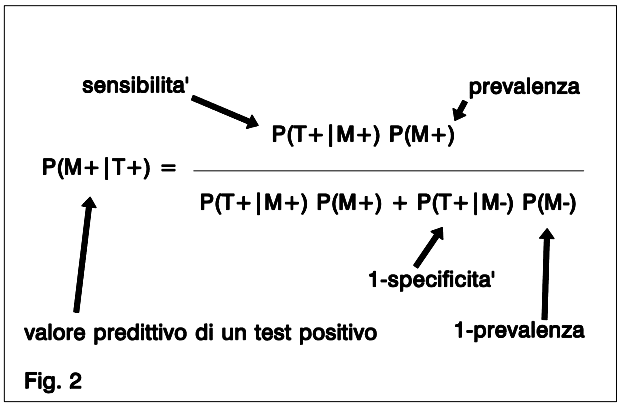


Fig. 2

L'esempio riportato si presta a tre importanti considerazioni. La prima è che il teorema di Bayes rappresenta l'unico strumento che consente di fornire una misura quantitativa, e quindi oggettiva, del valore aggiunto fornito da un test (per esempio una analisi di laboratorio). Nel caso esemplificato, la differenza tra la probabilità a posteriori (dopo avere effettuato il test, pari al 50%) e la probabilità a priori (prima di avere effettuato il test, pari al 3 per mille) rappresenta appunto il valore aggiunto che il test è in grado di fornire, in termini di informazione, alla diagnosi clinica.

La seconda considerazione parte dall'osservazione che un test con sensibilità del 100% e specificità del 100% è un test che ha per definizione un valore predittivo del test positivo del 100%. Nel caso esemplificato degli anticorpi anti-HIV le caratteristiche del test erano pressoché ideali (sensibilità del 100% e specificità del 99,7%), Nonostante questo il valore predittivo del test positivo risultava del 50% "solamente". Ecco un ottimo esempio di come piccole variazioni della specificità di un test possano influenzarne pesantemente il valore predittivo (questo è particolarmente vero in situazioni di bassa prevalenza, quale quella del 3 per mille dell'esempio). La simulazione di condizioni di diversa prevalenza possibile con Ministat consente di valutare il comportamento del valore predittivo del test positivo e del valore predittivo del test negativo in funzione della prevalenza, e quindi di familiarizzare con questi fondamentali concetti.

La terza considerazione è che ancora una volta l'analisi bayesiana è la sola che possa fornire tutti gli elementi (valore predittivo del test, veri positivi, veri negativi, falsi positivi, falsi negativi) necessari per effettuare una valutazione oggettiva del rapporto costi/benefici di una strategia diagnostica (anche se la praticabilità in sé della strategia diagnostica continuerà ovviamente a dipendere dal contesto economico e culturale, quindi sostanzialmente dai vincoli economici e dai vincoli etici che alla strategia sono attribuiti).

4.8. Statistica non parametrica

Tutti i concetti di base relativi alle statistiche elementari parametriche sopra esposti si prestano ad essere generalizzati a situazioni di maggiore complessità, e in particolare a situazioni nelle quali la struttura dei dati assume una forma tabellare. Ovviamente gli aspetti tecnici e computazionali delle statistiche parametriche diventano rapidamente più complessi, mentre la modellizzazione diventa ancora più critica. Tuttavia è questa l'essenza delle tecniche statistiche parametriche più avanzate, come il confronto tra medie, la regressione lineare, l'analisi della varianza, e le molte altre incluse in Ministat.

Analogamente rimane la possibilità, anche nelle situazioni nelle quali la struttura dei dati assume una forma tabellare, di mantenere, sulla base degli assunti distribuzionali posti alla base del modello matematico-statistico di volta in volta adottato, la distinzione tra tecniche statistiche parametriche e tecniche statistiche non-parametriche, anche se per queste ultime la modellizzazione e la complessità risultano ancora e di gran lunga maggiori.

L'opzione Statistica Non Parametrica del Menù Statistica include una serie di test statistici non-parametrici che consentono di affrontare i principali problemi statistici senza dovere ricorrere ad assunti distribuzionali di gaussianità e con la possibilità avere di una stima statistica "robusta", poco influenzata da parte di singoli dati (si richiama in particolare l'attenzione sull'utilità del test non-parametrico di Wilcoxon in sostituzione del test t di Student, e della regressione lineare non-parametrica in luogo della tradizionale regressione lineare parametrica).

4.8.1. Statistiche elementari non parametriche

Come già più volte ricordato, la media μ e la deviazione standard σ sono i "parametri" che caratterizzano la distribuzione gaussiana. Per questo le tecniche statistiche che basano la loro validità (la validità delle conclusioni che mediante esse è possibile trarre) su assunti distribuzionali di gaussianità sono denominate in generale *tecniche statistiche parametriche*. In contrapposizione alle tecniche che basano la loro validità su assunti distribuzionali minimi, e comunque non richiedono una distribuzione gaussiana dei dati per garantire (nei limiti in cui questo termine ha significato nella statistica inferenziale) conclusioni valide, che sono denominate in generale *tecniche statistiche non-parametriche*. Non è comunque un caso che in presenza di distribuzioni esattamente gaussiane tecniche statistiche parametriche e tecniche statistiche non-parametriche forniscono identici risultati.

Eseguire il calcolo delle Statistiche Elementari (parametriche) sul valore dei trigliceridi (campo TRIGLI) del file COLEST.DBF. Caricare il file nella tabella dati di Ministat mediante l'opzione Importa File (in formato dBase III®) del Menù File: quindi selezionare, facendo click con il mouse sul nome della colonna (variabile), la colonna TRIGLI. Come detto si tratta della determinazione della concentrazione dei trigliceridi nel siero di 1000 soggetti non selezionati, a digiuno, espressa in mg/dL.

Il calcolo delle Statistiche Elementari (parametriche) porta ad una tabella dei percentili nella quale il 2,5-esimo percentile parametrico e il 97,5-esimo percentile parametrico (rispettivamente la media meno 1,96 volte la deviazione standard e la media più 1,96 volte la deviazione standard) sono pari rispettivamente a -45,3 e a 318,3 mg/dL (i valori degli altri percentili calcolati da Ministat sono stati omessi per concisione).

		Limiti di confidenza al 90%		Valore non parametrico
Percentile	Valore	inferiore	superiore	comparativo
2,5	-45,31548	-53,56551	-37,06544	44,61001
97,5	318,33348	310,08344	326,58351	381,72472

Utilizzando le statistiche elementari parametriche, dovremmo affermare che nel 95% dei soggetti esaminati i trigliceridi hanno una concentrazione compresa tra -45,3 e 318,3 mg/dL: un'affermazione che, a dispetto del fatto che i risultati sono stati ottenuti con un metodo statistico rigoroso, nessuno si sentirebbe di sostenere.

Le cose vanno meglio nel caso delle statistiche non parametriche: il valore non parametrico comparativo riportato nella stessa tabella per il 2,5-esimo e per il 97,5-esimo percentile è pari rispettivamente 44,5 e 381,7 mg/dL.

Quello illustrato è, ancorché basato su dati reali, un caso limite, nel quale la contraddizione in termini biologici rappresentata da un valore di concentrazione negativo evidenzia l'errore nell'applicazione della metodologia statistica. Errore derivante dal fatto che l'assunto (accettato implicitamente) che la distribuzione della concentrazione dei trigliceridi nel siero sia di tipo gaussiano in questo caso non è corretto.

Si considerino ora i seguenti dati: 5, 6, 6, 7. La loro media è uguale a 6, e la loro mediana è anch'essa uguale a 6.

Si considerino ora invece i seguenti dati: 5, 6, 7, 54. Rispetto ai dati precedenti la loro media è cambiata grandemente, essendo ora uguale a 18, mentre la loro mediana è rimasta praticamente immodificata, essendo ora uguale a 6,5.

Questo banale esempio illustra una seconda, e in certe situazioni, formidabile caratteristica delle statistiche non-parametriche: la loro robustezza, definita come la caratteristica per cui una statistica risulta poco influenzata dalla presenza nella distribuzione di singoli dati che si discostano notevolmente dalla media degli altri. Supponendo che i valori 5, 6, 6, 7 derivino da un esperimento, che i valori 5, 6, 7, 54 derivino da una replica dello stesso esperimento, e che non vi siano ragioni oggettive per scartare il valore apparentemente "aberrante" di 54, l'impiego della mediana (tipica statistica non-parametrica), consente di ottenere in entrambi i casi conclusioni sostanzialmente simili per la misura di posizione della distribuzione.

L'opzione Statistica Non Parametrica del Menù Statistica include il calcolo delle Statistiche Elementari Non Parametriche, e consente quindi di calcolare le statistiche elementari (mediana, range, quartili della distribuzione) nel caso di distribuzioni non gaussiane ovvero nel caso in cui sia importante che la stima risulti scarsamente influenzata da singoli dati (questo è particolarmente vero per la misura di posizione, cioè la mediana, lo è molto meno o per niente per la misura di dispersione o per i percentili marginali della distribuzione).

SUGGERIMENTO: prestare sempre attenzione alla distribuzione dei dati e, nel caso in cui ci si trovi in presenza di distribuzioni non gaussiane, ricorrere all'utilizzo di test statistici non-parametrici.

4.8.2. Confronto tra mediane

Come il test t di Student (parametrico) consente di confrontare tra di loro le medie (misura di posizione parametrica), il test di Wilcoxon (non parametrico) consente di confrontare tra di loro le mediane (misura di posizione non parametrica).

Il test di Wilcoxon per dati appaiati è l'equivalente non parametrico del test *t* di Student per dati appaiati, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano.

Sebbene spesso chiamato test di Mann-Whitney, l'equivalente non parametrico del test *t* di Student per campioni indipendenti è dovuto anch'esso a Wilcoxon, come ricorda lo Snedecor: e qui si è voluto adottare l'eponimo che sembra essere storicamente più corretto.

4.8.3. Regressione lineare non parametrica

Il modello standard di regressione lineare (regressione lineare parametrica) è basato su assunti di gaussianità che possono essere disattesi. È possibile allora impiegare, sia per la regressione *x* variabile indipendente che per la regressione *y* variabile indipendente, un modello di regressione lineare non parametrico.

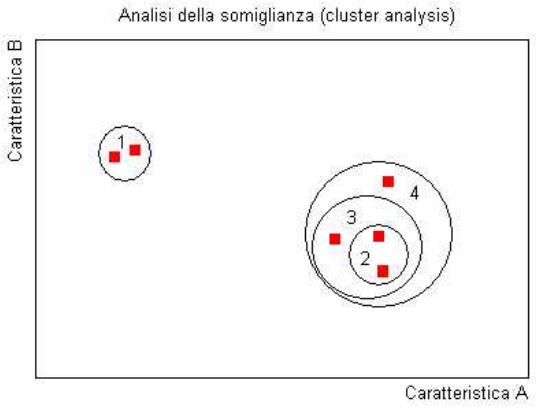
La regressione *x* variabile indipendente non è raccomandata per l'analisi statistica di dati relativi al confronto fra due metodi analitici in quanto, contrariamente a quanto assunto dal modello matematico, non solo la variabile indipendente è affetta da un errore di misura, ma per di più spesso l'errore di misura della variabile dipendente non è distribuito normalmente e la sua varianza cambia al variare del valore della x. Inoltre impiegando la regressione *x* variabile indipendente si ottengono, a seconda che sia un metodo o l'altro a essere posto in ascisse, equazioni della retta di regressione diverse, che possono portare a conclusioni contraddittorie. Come equivalente non parametrico della componente principale standardizzata viene utilizzato il modello di regressione lineare non parametrica proposto da Passing e Bablok.

4.8.4. Test chi-quadrato

Esistono numerose situazioni nelle quali la descrizione dei fenomeni che interessano può essere effettuata solamente in termini qualitativi o, nella migliore delle ipotesi, semiquantitativi: fumatore / non fumatore, sano / ammalato, HBsAg positivo / HBsAg negativo, soggetto in sovrappeso / soggetto con peso ideale / soggetto sottopeso, ecco solamente alcuni esempi presi a caso di situazioni in cui ci si imbatte quotidianamente. In casi di questo genere spesso interessa verificare la frequenza di certi fenomeni: controllare la frequenza di due differenti forme morbose in soggetti fumatori e non-fumatori, la frequenza di positività dell'antigene di superficie dell'epatite B (HBsAg) in differenti gruppi della popolazione. Il test chi-quadrato consente di verificare se le frequenze osservate nei diversi gruppi sono uguali. Per applicare il test chi-quadrato nella forma generalizzata per tabelle di contingenza, è necessario che ciascun elemento del campione in esame possa essere classificato per una caratteristica in un numero R di classi, e per una seconda caratteristica in un numero C di classi, in modo tale che i dati possano essere organizzati in una tabella di R righe per C colonne.

4.8.5. Analisi della somiglianza

Il concetto alla base dell'analisi della somiglianza (cluster analysis) è semplice: raggruppare oggetti omogenei (organismi biologici, minerali, eccetera) in insiemi (cluster), partendo dagli oggetti più simili, aggiungendo progressivamente gli oggetti più dissimili. All'inizio del processo di classificazione ad ogni oggetto corrisponde un cluster (e viceversa). In questo stadio tutti gli oggetti sono considerati dissimili tra di loro.



Al passaggio successivo i due oggetti più simili sono raggruppati in un unico cluster. Il numero dei cluster risulta quindi pari al numero di oggetti diminuito di 1. Il procedimento viene ripetuto ciclicamente, fino ad ottenere (all'ultimo passaggio) un unico cluster. Nella figura viene presentato un esempio di clusterizzazione, I due oggetti più vicini (gli oggetti sono rappresentati in un diagramma cartesiano in base a due caratteristiche, A e B) sono raggruppati nel cluster 1.

Successivamente si forma un secondo cluster (cluster 2) che raggruppa gli oggetti più vicini tra loro dopo i primi due. Un ulteriore oggetto viene raggruppato nel cluster tre, che viene in tal modo a comprendere, oltre a quest'ultimo oggetto, i due oggetti precedentemente raggruppati nel cluster 2. Infine i tre oggetti precedentemente raggruppati nel cluster 3 confluiscono nel cluste 4 insieme all'oggetto più vicini. Stabilire a quale livello di aggregazione degli oggetti fermarsi, e quindi quali conclusioni trarre, dipende in larga parte dal giudizio di merito dell'utilizzatore. Per questo la cluster analysis riveste un ruolo centrale, in statistica, limitatamente alla analisi esplorativa dei dati (nel caso specifico sembrerebbe che in definitive gli oggetti finiscano con il confluire in due diverse "famiglie", significativamente diverse tra di loro).

CAPITOLO 5
GUIDA ALL'USO DI MINISTAT

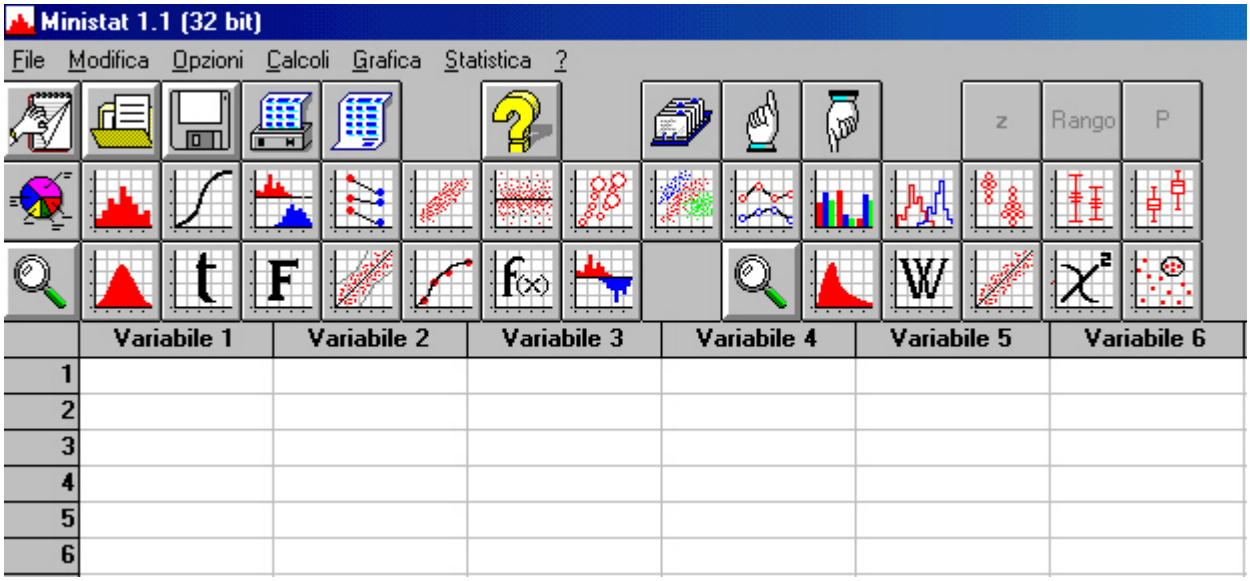
5.1. Accesso alla Guida on-line

Durante l'utilizzo di Ministat, è possibile richiamare la Guida in qualsiasi momento, selezionando Guida nel menù ? oppure facendo click sull'icona ? o ancora premendo il tasto di funzione F1.
Fare click sulle icone per una guida rapida, oppure utilizzare i collegamenti ipertestuali riportati subito dopo per accedere alla guida completa.

5.2. Caratteristiche di Ministat

5.2.1. Strutturazione dei dati

La tabella dati di Ministat, che compare nel menù principale,



costituisce la base di partenza per qualsiasi ulteriore operazione, come i calcoli, la rappresentazione grafica e l'analisi statistica.

⇒ Ogni colonna rappresenta una variabile.
⇒ Ogni cella contiene un dato (un valore di una variabile).
⇒ È possibile gestire contemporaneamente fino a 10 variabili, comprendenti ciascuna fino a un massimo di 1000 dati.

Un caso particolare è costituito da due aspetti dell'Analisi della Varianza (l'analisi della varianza a due fattori e lo studio delle componenti della variabilità), e dal Test Chi-Quadrato, nei quali la strutturazione dei dati non segue la regola colonne/variabili e righe/casi. Si rimanda alle informazioni e agli esempi forniti sotto le voci specifiche per la comprensione (in realtà molto facile) di come strutturare i dati in questi casi particolari.

5.2.2. Immissione dei dati

I dati possono essere immessi nella tabella di Ministat da tastiera, essere letti da un file in formato Ministat (con l'opzione Apri del Menù File) nel quale siano stati salvati in precedenza, o essere importati da un database esterno mediante l'opzione Importa File (in formato Access® o dBase III®) del Menù File.

⇒ Nelle celle dei dati possono essere immessi solamente caratteri di tipo numerico (inclusi i separatori virgola o punto).

⇒ I dati devono essere limitati a 11 cifre significative, sei prima della virgola e cinque dopo la virgola (vedere l'opzione Numero di Decimali del Menù Opzioni).

Si tratta di limiti solo apparentemente ristretti, se si tiene conto che le misure ottenute mediante i comuni dispositivi di misura comprendono in genere non più di due o tre cifre significative, rarissimamente quattro (per il concetto di numero di cifre significative in una misura, da non dare sempre per scontato, vedere Taylor[58].

⇒ Il separatore dei decimali è quello configurato nelle Impostazioni Internazionali del Pannello di Controllo di Windows® (il punto [.] se si è scelta l'impostazione USA, la virgola [,] se si è scelta l'impostazione italiana).

⇒ Possono essere immessi direttamente da tastiera solamente valori positivi: per i valori negativi fare doppio click sulla cella dopo avere immesso il valore.

⇒ Per correggere i valori immessi è attivo il solo tasto di backspace.

⇒ È possibile cancellare l'intero contenuto della cella anche selezionando Menù Modifica, quindi selezionando Cella e infine Vuota.

5.2.3. Salvataggio dei dati

I dati, che siano stati immessi da tastiera o importati da un file esterno, possono essere salvati solamente nel particolare formato utilizzato da Ministat.

Questo è in realtà un formato molto semplice (ASCII formato fisso), con 1001 righe di 140 caratteri ciascuna. Ogni riga comprende 10 campi di 14 caratteri, per un totale di 140 caratteri (alla mancanza di un carattere corrisponde a un carattere spazio). La prima riga contiene il nome delle variabili, le righe rimanenti (dalla 2 alla 1001) contengono i dati, esattamente così come appaiono nella tabella del menù principale di Ministat. Ecco, a titolo di esempio, come appaiono, quando siano visualizzati con un qualsiasi editor di testi ASCII, i primi 70 caratteri delle prime 6 righe del file IGA.MI1, un file dimostrativo in formato Ministat che si trova nella directory nella quale è stato installato il programma:

```
Normali        NCAH           CPH            CAH            AC
1,22           7,44           2,45           2,35           3,51
2,81           4,58           1,63           3,21           4,23
4,02           3,71           3,44           3,88           7,66
2,23           4,94           2,47           1,56           9,54
1,96           4,32           5,78           4,56           2,14
```

Il formato dati impiegato da Ministat per il salvataggio dei dati è talmente semplice che qualsiasi utente mediamente esperto è in grado di utilizzare direttamente i file di Ministat per esportare i dati verso altre applicazioni, ovvero di importare i dati dei suoi applicativi in Ministat (per questo sono sufficienti le informazioni qui fornite ed eventualmente un editor di testi ASCII). Tuttavia con Ministat viene fornito un ulteriore strumento per garantire la completa standardizzazione nel

[58] *Taylor JR Introduzione all'analisi degli errori. Bologna: Zanichelli, 1990:223pp.*

trasferimento dei dati. L'opzione Esporta File compresa nel Menù File consente infatti di esportare i dati con due modalità (file ASCII in formato fisso e file ASCII in formato delimitato) assolutamente standard e accessibili a qualsiasi altro programma (in pratica tutti i tabelloni elettronici, i database e i programmi di wordprocessing sia in ambiente Windows® che in ambiente DOS® sono forniti delle opzioni necessarie per importare dati indifferentemente da file ASCII dell'uno e dell'altro tipo).

5.2.4. Selezione dei dati da elaborare

Per selezionare i dati da elaborare è necessario evidenziare, nella tabella dati di Ministat, le/e colonna/e contenente/i i dati stessi.
Per selezionare una sola colonna (variabile) è sufficiente fare click con il tasto sinistro del mouse sulla cella contenente il nome della variabile stessa: tutta la colonna verrà evidenziata e continuerà a rimanere evidenziata finche non si deselezionerà la colonna facendo click su una cella qualsiasi. La selezione di una colonna determina l'inattivazione di tutte le voci dei menù che non risultano più pertinenti: in particolare risulteranno inattivate tutte le voci che corrispondono ad operazioni (calcoli, statistiche o elaborazioni grafiche) che possono essere eseguite solamente su più variabili.
Per selezionare più colonne (variabili) è possibile procedere in due modi:

⇒ fare click con il tasto sinistro del mouse sulla cella contenente il nome della variabile corrispondente alla prima colonna da selezionare, tenere il tasto del mouse premuto, portarsi sulla cella contenente il nome della variabile corrispondente all'ultima colonna da selezionare, e rilasciare il tasto del mouse;

⇒ fare click con il tasto sinistro del mouse sulla cella contenente il nome della variabile corrispondente alla prima colonna da selezionare, rilasciare il tasto del mouse (la colonna risulterà selezionata), premere il tasto SHIFT e, tenendolo premuto (il tasto SHIFT è quello che serve a generare le lettere maiuscole), premere il tasto FRECCIA A DESTRA tante volte quante sono quelle necessarie a selezionare tutte le colonne (variabili) aggiuntive desiderate.

La selezione di più colonne (variabili) determina l'inattivazione di tutte le voci dei menù che non risultano più pertinenti: in particolare risulteranno inattivate tutte le voci che corrispondono ad operazioni (calcoli, statistiche o elaborazioni grafiche) che possono essere eseguite solamente su una colonna (variabile). Qualsiasi sia il numero delle colonne selezionate, vengono comunque sempre inattivate tutte le voci del Menù File.
Per l'effettuazione delle rappresentazioni grafiche e dei test statistici, Ministat parte dalla prima riga (riga 1) e considera terminati i dati da elaborare al raggiungimento della prima cella vuota (eventuali dati successivi vengono ignorati). Per questo motivo l'inserimento di una cella vuota all'interno di una colonna (variabile) rappresenta un utile mezzo per limitare il numero dei dati da sottoporre ad elaborazione. L'inserimento di una riga vuota nella tabella dati determina un risultato analogo, esteso contemporaneamente a tutte le colonne (variabili) della tabella dati.
Per selezionare contemporaneamente tutte le colonne fare click con il tasto sinistro del mouse sulla cella nell'angolo in alto a sinistra.
Per selezionare una riga fare click con il tasto sinistro del mouse sulla cella contenente il numero della riga. È possibile selezionare solamente una singola riga per volta.

5.2.5. Limiti nell'elaborazione dei dati

Il primo limite nell'elaborazione dei dati è ovviamente quello derivante dal numero massimo di colonne e di righe contenute nella tabella del menù principale di Ministat: quindi 10 variabili con 1000 dati ciascuna.
Tutte la rappresentazioni grafiche e i calcoli e la maggior parte dei test statistici possono essere eseguiti sul numero massimo di dati previsti. Limiti più restrittivi possono determinarsi per alcuni particolari test statistici, che implicano l'utilizzo di risorse di memoria veramente imponenti. In particolare alcune opzioni del Menù Statistica, come la Regressione Multipla, e alcune opzioni della

Statistica Non Parametrica, come l'Analisi della Somiglianza (cluster analysis) e la Regressione Lineare Non Parametrica, comportano l'impiego di matrici le cui dimensioni aumentano in ragione geometrica all'aumentare del numero dei dati da analizzare. In questi casi la quantità di memoria disponibile sul sistema utilizzato e l'ambiente nel quale è stato sviluppato Ministat possono drasticamente limitare il numero di dati elaborabili. In ogni caso viene tentata l'elaborazione dei dati, ed eventualmente viene segnalata l'impossibilità di portarla a termine.

5.2.6. Messaggi di errore

Molti errori possono essere intercettati dal sistema operativo: in questo caso i messaggi di errore sono generati da Windows®, la cui guida fornisce l'aiuto necessario per interpretare sia le cause sia i relativi rimedi. Tra gli errori intercettati dal sistema operativo vi sono tipicamente quelli determinati dalle periferiche: stampanti non in linea, dischetti non inseriti nel drive, dischetti protetti da scrittura o danneggiati, e così via.
Gli errori non intercettati dal sistema operativo sono gestiti da Ministat, che genera i relativi messaggi di errore. Spesso questi errori hanno origine banale, come quando per esempio si cerca di calcolare la regressione lineare su due variabili che non contengono lo stesso numero di dati. Occasionalmente tuttavia questi errori possono riflettere problemi più complessi. In particolare questo può accadere quando si utilizza la funzione che consente di importare dati da database esterni nella tabella di Ministat (vedere File: Importa File). Le difficoltà sono sostanzialmente riconducibili a tre situazioni: (1) i motivi più banali e più frequenti di difficoltà sono costituiti da file danneggiati, per danni al mezzo fisico (supporto magnetico) o per danni logici dei dati (tabella di allocazione dei file [FAT], file-system, settori del disco o dischetto). Si raccomanda come prima cosa di controllare il disc(hett)o mediante una utility del sistema operativo (come SCANDISK), che consente di identificare (e il più delle volte correggere) tali danni, e che oltretutto sarebbe buona norma utilizzare a scadenza regolare per assicurare la manutenzione del disco. Come seconda cosa utilizzare una utility per escludere la presenza di virus (sempre possibile); (2) in una tabella di database (ovvero in un database, se questo è strutturato come un'unica tabella), al fine di non consentire ambiguità, non vi possono essere due campi con lo stesso nome. Alcuni tabelloni elettronici consentono di manipolare i dati e di salvarli in formato dBase III® (file con estensione .DBF) senza un congruo controllo sui nomi dei campi: Ministat non consente di importare dati da file (anche se apparentemente nel formato corretto) con nomi di campi duplicati; (3) Windows® è un sistema operativo in continua evoluzione, e questo può creare problemi di compatibilità nella gestione di applicativi che utilizzano versioni diverse dei driver del sistema operativo. In particolare l'impossibilità di importare file in formato dBase III® potrebbe derivare da conflitti di questo genere. In questo caso, dopo avere accuratamente escluso tutte le possibili cause sopra elencate, mediante un editor di testi ASCII, accedere alla directory nella quale è stato installato Ministat, aprire il file MINISTAT.INI, togliere il punto e virgola (;) dalla riga in cui si trova

```
[Installable ISAMs]
dBASE III=C:\WINDOWS\SYSTEM\xbs110.dll
;dBASE III=C:\WINDOWS\SYSTEM\xbs200.dll
```

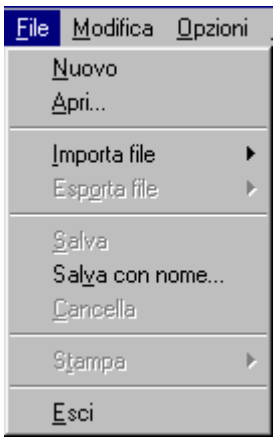
e aggiungerlo alla riga immediatamente precedente

```
[Installable ISAMs]
;dBASE III=C:\WINDOWS\SYSTEM\xbs110.dll
dBASE III=C:\WINDOWS\SYSTEM\xbs200.dll
```

In questo modo viene attivato un driver che consente di risolvere il problema.

5.3. Menù di Ministat

5.3.1. Menù File



Questo menù comprende tutte le opzioni necessarie per importare, salvare, esportare e stampare i file di dati. La selezione di una riga o di una o più colonne della tabella dati determina l'inattivazione di tutte le voci del Menù File.

⇒ Le opzioni Esporta File, Salva, Cancella e Stampa risultano attive solamente dopo che i dati della tabella sono stati salvati come file in formato Ministat.

5.3.1.1. Nuovo

Consente di svuotare completamente la tabella dati di Ministat, al fine di preparare un nuovo file in formato Ministat. Impiegare questa opzione solamente quando si desidera immettere i dati da tastiera: quando si importano i dati da un database esterno, la tabella dati di Ministat viene svuotata automaticamente. In ogni caso viene richiesta la conferma, in quanto i dati preesistenti, se non salvati su disco in un file, sono persi. Le opzioni Esporta File, Salva, Cancella e Stampa vengono disattivate.

5.3.1.2. Apri

Apre un file in formato Ministat (file con estensione .MI1), e lo carica nella tabella dati di Ministat: i dati eventualmente preesistenti nella tabella sono persi se non sono stati in precedenza salvati su disco in un file (vedere File: Salva con Nome). Una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file da aprire. In seguito alla apertura di un file in formato Ministat vengono attivate le opzioni Esporta File, Salva, Cancella e Stampa.

5.3.1.3. Importa file

Consente di importare i dati direttamente da un file/database in formato ASCII, in formato Access® 1.x oppure in formato dBase III®. Come prima cosa una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file da importare (per i file ASCII

viene assunta l'estensione .txt, per i file Access® 1.x viene assunta l'estensione .mdb, per i file dBase III® viene assunta l'estensione .dbf). Quindi la tabella dati di Ministat viene svuotata, previa richiesta di conferma: i dati della tabella non salvati su disco in un file (vedere: File: Salva con Nome) sono ovviamente persi. A questo punto le procedure differiscono.
Nel caso di file ASCII, per i quali non esiste un formato standard, il file viene caricato dal disco e presentato in una tabella, con un carattere per ciascuna cella della tabella. L'utilizzatore dovrà selezionare manualmente le colonne che contengono i caratteri che costituiscono un campo, quindi, dopo avere selezionato l'opzione Importa Dalle Colonne del Menù File, scegliere in quale colonna/variabile della tabella dati di Ministat importare i valori, selezionando prima tale colonna/variabile, e quindi selezionando l'opzione Importa Nella Colonna del Menù File. È possibile ripetere le operazioni sopra indicate, per importare altri campi, selezionando l'opzione Importa del Menù File, ovvero terminare l'importazione del file selezionando l'opzione Esci del Menù File.
Nel caso di un file/database in formato Access® 1.x, una ulteriore finestra consente di selezionare il nome di una delle tabelle che costituiscono il file/database: verranno quindi importati i dati della sola tabella selezionata.
Nel caso di un database in formato dBase III®, verranno importati tutti i dati del file selezionato, in quanto un file/database dBase III® corrisponde ad una singola tabella di database. I dati vengono importati così come sono: qualora si renda necessario, possono essere riportati al numero di cifre significative gestibile da Ministat (undici più il separatore, sei cifre per gli interi e cinque per i decimali) prima trasformandoli (mediante le opzioni Moltiplica e Dividi del Menù Calcoli) quindi arrotondandoli opportunamente (mediante l'opzione Numero di Decimali del Menù Opzioni).
Qualora si dovessero riscontrare problemi nell'importare i file, consultare la sezione dedicata ai Messaggi di Errore.

5.3.1.4. Esporta file

Questa opzione risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati su disco come file in formato Ministat (vedere File: Salva con Nome). Una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file in cui salvare i dati. L'estensione (.TXT) del file viene aggiunta automaticamente. Qualsiasi altra estensione eventualmente fornita non viene considerata, e viene sostituita dal valore predefinito.
È possibile esportare i dati in due tipi di formato ASCII: formato fisso e formato delimitato. Si tratta di due formati assolutamente standard e accessibili a qualsiasi altro programma (in pratica tutti i programmi in ambiente Windows® possono importare dati indifferentemente da file ASCII dell'uno e dell'altro tipo).
I file ASCII in formato fisso contengono 1000 righe di 140 caratteri ciascuna. Ogni riga comprende 10 campi di 14 caratteri. Le righe, dalla 1 alla 1000, contengono i dati, esattamente così come appaiono nella tabella del menù principale di Ministat. Ecco, a titolo di esempio, come appaiono, quando siano visualizzati con un qualsiasi editor di testi ASCII, i primi 70 caratteri delle prime 10 righe del file IGA.MI1, quando sia stato esportato come file ASCII in formato fisso (con un nome a piacere, ed estensione .TXT):

```
1,22          7,44          2,45          2,35          3,51
2,81          4,58          1,63          3,21          4,23
4,02          3,71          3,44          3,88          7,66
2,23          4,94          2,47          1,56          9,54
2,35          3,49          1,95          1,78          11,35
1,64          3,88          4,56          2,49          6,43
2,08          4,71          7,31          3,11          5,28
1,96          4,32          5,78          4,56          2,14
```

```
1,54          4,90          3,40          5,11          4,76
```

I file ASCII in formato delimitato contengono anch'essi 1000 righe. Le righe, dalla 1 alla 1000, contengono i dati, essendo ogni dato racchiuso entro doppi apici, e i dati separati tra di loro da un virgola. Ecco, a titolo di esempio, come appaiono, quando siano visualizzate con un qualsiasi editor di testi ASCII, le prime 10 righe del file IGA.MI1, quando sia stato esportato come file ASCII in formato delimitato (con un nome a piacere, ed estensione .TXT). Notare come in questo caso il numero di caratteri per riga sia variabile:

```
"1,22","7,44","2,45","2,35","3,51","","","","","",""
"2,81","4,58","1,63","3,21","4,23","","","","","",""
"4,02","3,71","3,44","3,88","7,66","","","","","",""
"2,23","4,94","2,47","1,56","9,54","","","","","",""
"2,35","3,49","1,95","1,78","11,35","","","","","",""
"1,64","3,88","4,56","2,49","6,43","","","","","",""
"2,08","4,71","7,31","3,11","5,28","","","","","",""
"1,96","4,32","5,78","4,56","2,14","","","","","",""
"1,54","4,90","3,40","5,11","4,76","","","","","",""
"1,63","11,43","5,12","2,36","7,91","","","","","",""
```

5.3.1.5.Salva

Questa opzione risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati come file in formato Ministat (vedere File: Salva con Nome), e copia su disco il contenuto della tabella dati nel file in formato Ministat (file con estensione .MI1) in uso al momento del salvataggio.

5.3.1.6. Salva con nome

Consente di salvare su disco, in un file in formato Ministat (file con estensione .MI1), i dati della tabella. Una finestra di dialogo standard di Windows® consente di selezionare il disco, la directory e infine il nome del file in cui salvare i dati. L'estensione (.MI1) del file viene aggiunta automaticamente. Qualsiasi altra estensione eventualmente fornita non viene considerata, e viene sostituita dal valore predefinito. Il salvataggio del file in formato Ministat attiva le opzioni Esporta File, Salva, Cancella e Stampa.

5.3.1.7. Cancella

Questa opzione risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati come file in formato Ministat (vedere File: Salva con Nome). Dopo una richiesta di conferma, la tabella dati di Ministat viene vuotata e il file Ministat in uso viene cancellato dal disco: i dati sono irrimediabilmente persi.

5.3.1.8. Stampa


Questa opzione, che risulta attiva solamente se è stato aperto un file Ministat (vedere File: Apri), ovvero dopo che i dati della tabella sono stati salvati come file in formato Ministat (vedere File: Salva con Nome), consente di stampare la tabella dati di Ministat (le opzioni per la stampa dei risultati dell'elaborazione grafica e dell'elaborazione statistica sono contenuti all'interno dei menù delle specifiche elaborazioni).

⇒ È possibile stampare i dati della tabella in due parti: la prima comprendente le colonne (variabili) da 1 a 5, la seconda comprendente le colonne (variabili) da 6 a 10.

In entrambi i casi la stampa parte dalla prima riga (riga 1) e si interrompe quando viene incontrata la prima riga completamente vuota. Per questo motivo l'inserimento di una riga vuota all'interno della tabella dati di Ministat (vedere l'opzione Riga del Menù Modifica) può rappresentare un utile mezzo per limitare il numero dei dati da stampare.

5.3.1.9. Esci

Termina l'esecuzione di Ministat. Viene richiesta conferma in quanto i dati non salvati vengono persi.

5.3.2. Menù Modifica

Comprende una serie di opzioni che conferiscono una notevole efficienza nella gestione della tabella dei dati, consentendo di spostare le variabili all'interno della tabella, di ordinare i dati in ordine numerico crescente o decrescente, di aggiungere e togliere colonne, righe o anche singole celle.

5.3.2.1. Taglia colonna

Taglia la colonna (variabile) selezionata. La differenza rispetto all'opzione Copia Colonna è rappresentata dal fatto che la colonna originaria viene vuotata. I dati potranno poi essere trasferiti in un'altra colonna con l'opzione Incolla Colonna. È possibile utilizzare questa opzione anche per accorpare in un'unica tabella variabili contenute in file differenti, nel seguente modo:

⇒ aprire un file Ministat o importare un file dBase III® o Access®;

⇒ tagliare la colonna (variabile) che interessa;
⇒ aprire un nuovo file Ministat; (4) incollare la colonna (variabile) nel nuovo file;
⇒ salvare il file; aprire il primo file e tagliare una nuova colonna;
⇒ aprire il secondo file e incollare la colonna;
⇒ salvare il file;
⇒ ripetere i tre punti precedenti, tagliando via via le varie colonne dal primo file e incollandole nel secondo.

5.3.2.2. Copia colonna

Copia la colonna (variabile) selezionata. La differenza rispetto all'opzione Taglia Colonna è rappresentata dal fatto che la colonna originaria rimane immodificata. I dati potranno poi essere trasferiti in un'altra colonna con l'opzione Incolla Colonna. È possibile utilizzare questa opzione anche per accorpare in un'unica tabella variabili contenute in file differenti, nel seguente modo:
⇒ aprire un file Ministat o importare un file dBase III® o Access®;
⇒ copiare la colonna (variabile) che interessa;
⇒ aprire un nuovo file Ministat;
⇒ incollare la colonna (variabile) nel nuovo file;
⇒ salvare il file;
⇒ aprire il primo file e copiare una nuova colonna;
⇒ aprire il secondo file e incollare la colonna;
⇒ salvare il file;
⇒ ripetere i tre punti precedenti, copiando via via le varie colonne dal primo file e incollandole nel secondo.

5.3.2.3. Incolla colonna

Incolla la colonna (variabile) precedentemente tagliata (vedere Taglia colonna) o copiata (vedere Copia colonna) sulla colonna selezionata come destinazione. Eventuali dati presenti nella colonna destinazione verranno persi, sovrascritti dai dati della colonna incollata.

5.3.2.4. Annulla incolla

Dopo l'utilizzo delle opzioni Taglia Colonna e Copia Colonna, i dati della colonna (variabile) tagliata o copiata restano in attesa di una destinazione, per cui la successiva selezione di una colonna viene interpretata da Ministat come indicazione della destinazione dei dati tagliati o copiati, con il risultato di bloccare qualsiasi altra operazione. L'opzione Annulla Incolla consente di rinunciare all'operazione Incolla, riabilitando le opzioni dei menù altrimenti rese inaccessibili.

5.3.2.5. Cella

L'opzione Cella risulta attiva solamente se non sono state selezionate né una riga né una o più colonne, e agisce sulla cella selezionata dal puntatore (cella attiva).
⇒ L'opzione Vuota consente di vuotare la cella, lasciando inalterato il resto della tabella dati.
⇒ L'opzione Elimina consente di eliminare la cella: la cella immediatamente successiva viene spostata al posto della cella eliminata, e così via fino all'ultima cella della colonna.
⇒ L'opzione Inserisci consente di inserire una nuova cella, vuota, al posto della cella selezionata. La cella selezionata viene spostata di una posizione in giù, per fare posto alla nuova cella, e così via fino all'ultima cella della colonna, il cui contenuto viene ovviamente perso.

Poiché Ministat, per l'effettuazione delle rappresentazioni grafiche e dei test statistici, parte dalla prima riga (riga 1) della tabella e considera terminati i dati al raggiungimento della prima cella

vuota (eventuali dati successivi vengono ignorati), l'inserimento di una cella vuota all'interno di una colonna (variabile) rappresenta un utile mezzo per limitare il numero dei dati da elaborare per quella colonna (variabile).

5.3.2.6. Colonna

L'opzione Colonna risulta attiva solamente se è stata selezionata una colonna.
⇒ L'opzione Vuota consente di vuotare la colonna, lasciando inalterato il resto della tabella dati.
⇒ L'opzione Elimina consente di eliminare una colonna: la colonna immediatamente successiva viene spostata al posto della colonna eliminata, e così via fino all'ultima colonna.
⇒ L'opzione Inserisci consente di inserire una nuova colonna, vuota, al posto della colonna selezionata. La colonna selezionata viene spostata di una posizione a destra, per fare posto alla nuova colonna, e così via fino all'ultima colonna, il cui contenuto viene ovviamente perso.

5.3.2.7. Riga

L'opzione Riga risulta attiva solamente se è stata selezionata una riga.
⇒ L'opzione Vuota consente di vuotare la riga, lasciando inalterato il resto della tabella dati.
⇒ L'opzione Elimina consente di eliminare la riga: la riga immediatamente successiva viene spostata al posto della riga eliminata, e così via fino all'ultima riga.
⇒ L'opzione Inserisci consente di inserire una nuova riga, vuota, al posto della riga selezionata. La riga selezionata viene spostata di una posizione in giù, per fare posto alla nuova riga, e così via fino all'ultima riga, il cui contenuto viene ovviamente perso.
Poiché Ministat, per l'effettuazione delle rappresentazioni grafiche e dei test statistici, parte dalla prima riga (riga 1) e considera terminati i dati di ciascuna colonna (variabile) al raggiungimento della prima cella vuota (eventuali dati successivi vengono ignorati), l'inserimento di una riga vuota nella tabella dati rappresenta un utile mezzo per limitare contemporaneamente su tutte le colonne (variabili) il numero dei dati da elaborare.

5.3.2.8. Unisci dati

Consente di unire i dati di due colonne (variabili).
⇒ È necessario selezionare due colonne adiacenti.
I dati delle due colonne sono uniti in una terza colonna, che viene inserita automaticamente immediatamente sulla sinistra delle due colonne selezionate, e che contiene, in immediata successione e nelle sequenze originali, i dati della prima e quelli della seconda colonna. Per effetto dell'inserimento della nuova colonna il contenuto della colonna 10 viene perso.

5.3.2.9. Scambia colonne

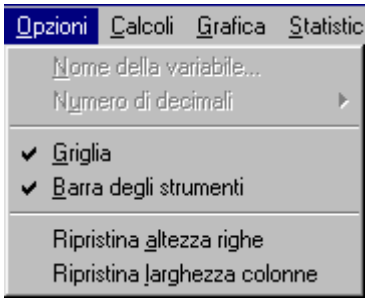
Consente di scambiare tra di loro due colonne (variabili).
⇒ È necessario selezionare due colonne adiacenti.
Al termine dell'operazione le due colonne saranno state scambiate di posizione tra di loro: quella di sinistra si troverà nella posizione di quella di destra e viceversa.

5.3.2.10. Ordina dati


Consente di ordinare In Ordine Crescente oppure In Ordine Decrescente i dati della colonna (variabile) selezionata. Al fine di mantenere l'allineamento dei dati appartenenti ad una stessa riga

(che viene quindi intesa come un record), i dati delle altre colonne (variabili) della tabella di Ministat seguono l'ordinamento della colonna selezionata. In questo modo i dati appaiati continuano a rimanere tali anche in seguito al/i processo/i di ordinamento. Prima di ordinare i dati è consigliabile numerare le righe, al fine di potere successivamente ripristinare l'ordine originario.

5.3.3. Menù Opzioni



Contiene le opzioni necessarie per gestire il nome delle variabili e il numero di decimali con cui rappresentare i dati, e le opzioni necessarie per gestire e ripristinare l'aspetto della tabella dati di Ministat.

5.3.3.1. Nome della variabile

Consente di immettere e/o correggere il nome della/e variabile/i corrispondente alla colonna (variabile) selezionata. Nel caso di dati importati da database esterni, come nome di ciascuna variabile viene assunto il nome del campo importato, eventualmente troncato a 12 caratteri. Nel caso di un nuovo file Ministat (vedere l'opzione Nuovo del Menù File), le colonne (variabili) sono denominate automaticamente con i valori predefiniti di Variabile 1 (prima colonna), Variabile 2 (seconda colonna), e così via, che possono poi ovviamente essere modificati a piacere. Il nome della variabile non può superare i 12 caratteri di lunghezza.

5.3.3.2. Numero di decimali

Consente di arrotondare al numero di decimali desiderato i valori della colonna (variabile) selezionata. Questa operazione si rende sempre necessaria quando si importano da database esterni dati rappresentati con un numero di decimali superiore a cinque (questo avviene tipicamente per campi di database che contengono valori derivati, come per esempio il rapporto tra i valori di due campi). In questi casi i dati sono importati così come sono, e devono essere ricondotti al numero di cifre significative gestibili da Ministat (undici più il separatore, sei cifre per gli interi e cinque per i decimali) arrotondandoli. L'arrotondamento comporta la perdita delle cifre in eccesso. Per questo motivo può rendersi necessario effettuare preliminarmente una riconversione delle scale. Per esempio, se viene importato il dato 0,0015437, al fine di conservare tutte le cinque cifre (ammesso che siano tutte e cinque significative) mantenendo un numero limitato di decimali, può essere opportuno moltiplicare il dato per 1000. Il Menù Calcoli contiene le opzioni Moltiplica e Dividi, che consentono di trasformare i valori delle variabili importate al fine di rappresentarli in modo congruo. Per il problema del numero delle cifre significative con cui rappresentare un dato vedere quanto riportato al capitolo 3 e anche Taylor[59].

[59] *Taylor JR Introduzione all'analisi degli errori. Bologna: Zanichelli, 1990:223pp.*

5.3.3.3. Griglia

Consente di attivare o disattivare la visualizzazione della griglia della tabella dati. Il valore predefinito comporta la visualizzazione della griglia.

5.3.3.4. Barra degli strumenti

Consente di attivare o disattivare la visualizzazione della barra degli strumenti. Il valore predefinito comporta la visualizzazione della barra degli strumenti.

5.3.3.5. Ripristina altezza righe

Posizionando il puntatore del mouse nelle celle al margine sinistro della tabella (quelle che contengono il numero delle righe), è possibile ridimensionare l'altezza della riga. Per fare questo posizionarsi esattamente sulla linea che separa due righe contigue: il puntatore del mouse cambierà di forma. Fare click sul tasto sinistro e, mantenendo la pressione sul tasto, spostarsi in su o in giù fino a conferire alla riga l'altezza desiderata, quindi rilasciare il tasto del mouse. L'opzione Ripristina Altezza Righe consente di riportare automaticamente tutte le righe all'altezza predefinita. L'opzione Ripristina Larghezza Colonne consente di riportare automaticamente alla larghezza predefinita tutte le colonne.

5.3.3.6. Ripristina larghezza colonne

Posizionando il puntatore del mouse nelle celle al margine superiore della tabella (quelle che contengono il nome della colonna (variabile)), è possibile ridimensionare la larghezza della colonna. Per fare questo posizionarsi esattamente sulla linea che separa due colonne contigue: il puntatore del mouse cambierà di forma. Fare click sul tasto sinistro e, mantenendo la pressione sul tasto, spostarsi a sinistra o a destra fino a conferire alla colonna la larghezza desiderata, quindi rilasciare il tasto del mouse. L'opzione Ripristina Larghezza Colonne consente di riportare automaticamente tutte le colonne alla larghezza predefinita. L'opzione Ripristina Altezza Righe consente di riportare automaticamente all'altezza predefinita tutte le righe.

5.3.4. Menù Calcoli

Le opzioni di questo menù consentono di effettuare una serie di utili trasformazioni numeriche sui valori delle colonne (variabili) della tabella dati. A seconda dei casi, è necessario selezionare una o due colonne. In tutti i casi i valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della/e colonna/e selezionata/e: questo consente di salvaguardare i dati originari. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi: un apposita finestra di dialogo avverte l'utilizzatore e gli consente di rinunciare eventualmente all'operazione. A causa dell'inserimento della nuova colonna, le opzioni del menu Calcoli non possono ovviamente essere effettuate oltre la colonna 9. Infine le trasformazioni richieste sono applicate solamente alle celle contenenti dati.

5.3.4.1. Somma

Consente di sommare ai dati della colonna (variabile) selezionata un numero positivo.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.2. Sottrai

Consente di sottrarre ai dati della colonna (variabile) selezionata un numero positivo.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.3. Moltiplica

Consente di moltiplicare i dati della colonna (variabile) selezionata per un numero positivo.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.4. Dividi

Consente di dividere i dati della colonna (variabile) selezionata per un numero positivo. Questa opzione risulta attiva quando viene selezionata una sola colonna. I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.5. Quadrato

Calcola il quadrato dei dati della colonna (variabile) selezionata.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.6. Radice

Calcola la radice quadrata dei dati della colonna (variabile) selezionata.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.7. Logaritmo

Calcola il logaritmo naturale (logaritmo neperiano, in base e con $e = 2{,}71828$) oppure il logaritmo decimale (logaritmo in base 10) dei dati della colonna (variabile) selezionata.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.8. Differenza

Calcola la differenza tra i dati della prima e quelli della seconda di due colonne (variabili) selezionate.
⇒ Questa opzione risulta attiva quando vengono selezionate due colonne.
I valori calcolati sono riportati in una colonna inserita ex-novo alla sinistra della prima delle due colonne selezionate. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.9. Media

Calcola la media aritmetica (somma diviso due) o geometrica (radice quadrata del prodotto) dei dati della prima e quelli della seconda di due colonne (variabili) selezionate.
⇒ Questa opzione risulta attiva quando vengono selezionate due colonne.
I valori calcolati sono riportati in una colonna inserita ex-novo alla sinistra della prima delle due colonne selezionate. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.10. Rapporto

Calcola il rapporto tra i dati della prima e quelli della seconda di due colonne (variabili) selezionate.
⇒ Questa opzione risulta attiva quando vengono selezionate due colonne.

I valori calcolati sono riportati in una colonna inserita ex-novo alla sinistra della prima delle due colonne selezionate. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.11. Cambia il segno

Cambia il segno dei dati della colonna (variabile) selezionata. Per cambiare di segno il contenuto di una singola cella (un singolo dato) fare doppio click sulla cella.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.12. Valore assoluto

Toglie il segno ai dati della colonna (variabile) selezionata.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I valori trasformati sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.13. Numera

Numera le righe, riportandone il numero di posizione.
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
Il numero di posizione di ogni riga viene riportato in una colonna inserita ex-novo alla sinistra della colonna selezionata, e denominata "Numero riga". Con l'inserimento della nuova colonna i dati della colonna 10 sono persi. Eseguire questa opzione prima di ordinare i dati. Infatti riordinandoli successivamente in base al numero di riga della colonna "Numero riga" sarà possibile ripristinare l'ordine originario .

5.3.4.14. Deviata z

Calcola la deviata normale standardizzata z dei dati della colonna selezionata. La deviata normale standardizzata z viene calcolata per ogni singolo dato come

$$z = (\text{valore del dato} - \text{media dei dati}) / \text{deviazione standard dei dati}$$

⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
Il valore della deviata normale standardizzata viene riportato in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.15. Rango

Calcola il rango, cioè il numero di posizione dei singoli dati nella lista dei dati ordinati in ordine numerico crescente. Quando due dati sono uguali, a ciascuno viene assegnato come numero di posizione la media dei numeri di posizione dei dati che risultano uguali. Così, per esempio, se il quinto e il sesto dato, in una lista ordinata, sono tutti e due uguali, e pari a 223, il rango del quinto e

del sesto dato sarà uguale, e pari a 5,5. Il procedimento viene esteso anche al caso in cui più di due dati siano uguali. Per il metodo vedere Snedecor[60].
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
Il rango viene riportato in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.4.16. Percentili

Calcola il percentile, cioè la posizione del dato all'interno della lista dei dati ordinati in ordine numerico crescente, espressa come frazione percentuale. Essendo *n* il numero dei dati, il percentile *P* (p-esimo) viene calcolato come

$$P = 100 \cdot rango / (n + 1)$$

Per il metodo vedere Snedecor[61].
⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna.
I percentili sono riportati in una colonna inserita ex-novo alla sinistra della colonna selezionata. Con l'inserimento della nuova colonna i dati della colonna 10 sono persi.

5.3.5. Menù Grafica

Le opzioni di questo menù consentono di effettuare una serie di interessanti e utili rappresentazioni grafiche dei dati, in aggiunta a quelle già disponibili all'interno delle opzioni del Menù Statistica. Per l'importanza della rappresentazione grafica vedere Bossi[62], Lantieri[63] e Campbell[64].

---

[60] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:135-148.*
[61] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:135-148.*
[62] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:37-68.*
[63] *Lantieri PB, Risso D, Rovida S, Ravera G. Statistica medica ed elementi di informatica. Milano: McGraw-Hill, 1994:71-101.*

5.3.5.1. Torta

Consente di rappresentare la distribuzione dei valori di una riga di dati, relativi ad un numero compreso tra due e dieci variabili (colonne), sotto forma di un diagramma a torta.

⇒ Questa opzione risulta attiva quando viene selezionata una sola riga.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III® del Menù File). Eliminare la prima colonna, che contiene l'età dei soggetti. Selezionare la prima riga, e osservare la ripartizione dei lipidi totale nel siero nelle quattro principali classi che li compongono, colesterolo totale, colesterolo HDL, trigliceridi e colesterolo LDL, nel soggetto in questione..

⇒ L'opzione Diagramma a Torta consente di scegliere tra una rappresentazione a due dimensioni e una a tre dimensioni del diagramma. In entrambi i casi viene esploso lo spicchio corrispondente alla variabile che compare nella prima colonna.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende non è prevista.

5.3.5.2. Istogramma

Consente di rappresentare la distribuzione dei valori di una variabile singola sotto forma di istogramma.

⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna (variabile).

⇒ Per visualizzare oltre all'istogramma anche le statistiche elementari parametriche utilizzare l'opzione Statistiche Elementari del Menù Statistica.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III® del Menù File). Selezionare la colonna COLEST, e osservare la distribuzione del colesterolo totale (in mg/dL, in ascisse) in un gruppo di soggetti non selezionati.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

---

[64] *Campbell MJ, Machin D. Medical statistics. A commonsense approach. Chichester: John Wiley & Sons, 1993:44-59.*

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.3. Cumulativa

Consente di rappresentare la distribuzione dei valori di una variabile singola sotto forma di curva cumulativa.

⇒ Questa opzione risulta attiva quando viene selezionata una sola colonna (variabile).

⇒ Per visualizzare oltre alla distribuzione cumulativa anche le statistiche elementari non parametriche utilizzare l'opzione Statistica Non Parametrica del Menù Statistica e selezionare le Statistiche Elementari.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III® del Menù File). Selezionare la colonna COLEST, e osservare la distribuzione del colesterolo totale (in mg/dL, in ascisse) in un gruppo di soggetti non selezionati. È possibile rappresentare la distribuzione cumulativa sia nella forma tradizionale che nella forma ripiegata proposta da Krouwer[65].

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.4. Confronto

Consente di rappresentare, in uno stesso grafico, la distribuzione dei valori di due variabili, sotto forma di due istogrammi.

⇒ Questa opzione risulta attiva quando vengono selezionate due colonne (variabili). La prima delle due colonne selezionate corrisponde all'istogramma riportato superiormente, la seconda corrisponde all'istogramma riportato inferiormente.

⇒ La rappresentazione viene effettuata anche se le due colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File). Selezionare le colonne COLEST e HDL, e osservare la distribuzione del colesterolo totale e del colesterolo HDL (entrambi in mg/dL) in un gruppo di soggetti non selezionati.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

[65] *Krouwer JS, Monti KL. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem 1995;33:525-7.*

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.5. Appaiamento

Consente di rappresentare i valori di due variabili appaiate (paired-dots plot). Tipicamente questo si realizza quando lo stesso fenomeno viene misurato in condizioni che differiscono tra di loro per una sola caratteristica, introdotta dal ricercatore mediante il disegno sperimentale.
⇒ Questa opzione risulta attiva quando vengono selezionate due colonne (variabili).
⇒ I valori della prima delle due colonne selezionate sono riportati sulla sinistra, quelli della seconda sulla destra.
⇒ La rappresentazione viene effettuata solamente se le due colonne (variabili) contengono lo stesso numero di dati.

Utilizzare la tabella AST del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Contiene il valore della AST (aspartato amminotransferasi) determinata su sieri freschi (colonna SUBITO), e rideterminata sugli stessi sieri dopo conservazione a +4 °C per 24 ore (colonna DOPO24ORE), al fine di stabilire se tale conservazione sia idonea a mantenere l'attività dell'enzima. Selezionare entrambe le colonne. L'andamento generale delle coppie di valori (in U/L), valutato alla luce dell'imprecisione analitica che caratterizza un metodo per la determinazione dell'AST, sembra suggerire che mediamente non vi siano differenze tra i due valori. Provare a verificare tale conclusione utilizzando il Confronto tra Medie.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.6. Dispersione

Consente di rappresentare in una sistema cartesiano una serie di punti di coordinate (x,y).
⇒ Questa opzione risulta attiva quando vengono selezionate due colonne (variabili). La prima delle due colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti.
⇒ La rappresentazione viene effettuata solamente se le due colonne (variabili) contengono lo stesso numero di dati.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File). Selezionare le colonne ETA (manca l'accento, che non può essere incluso nel nome di un campo) e COLEST, e osservare la relazione che esiste tra colesterolo totale (in mg/dL, in ordinate) ed età (in anni, in ascisse) in un gruppo di soggetti non selezionati.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.7. Altman

Consente di rappresentare in una sistema cartesiano i dati di due variabili appaiate, riportando sull'asse delle ascisse la media dei valori (cioè (valore della prima variabile + valore della seconda variabile) / 2), e sull'asse delle ordinate la loro differenze (cioè valore della prima variabile - valore della seconda variabile).

⇒ Questa opzione risulta attiva quando vengono selezionate due colonne (variabili). La prima delle due colonne selezionate corrisponde ai valori della prima variabile, la seconda corrisponde ai valori della seconda variabile.
⇒ La rappresentazione viene effettuata solamente se le due colonne (variabili) contengono lo stesso numero di dati.

Utilizzare la tabella AST del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Contiene il valore della AST (aspartato amminotransferasi) determinata su sieri freschi (colonna SUBITO), e rideterminata sugli stessi sieri dopo conservazione a +4 °C per 24 ore (colonna DOPO24ORE), al fine di stabilire se tale conservazione sia idonea a mantenere l'attività dell'enzima. Selezionare entrambe le colonne. L'andamento generale delle differenze (in U/L), sembra suggerire che mediamente le differenze non cambino al variare della concentrazione, e che globalmente la differenza media sia praticamente uguale a zero. Provare a verificare tale conclusione utilizzando il Confronto tra Medie. Il metodo è descritto da Altman[66].

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende

[66] *Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307-10.*

5.3.5.8. Associazione

Consente di rappresentare in una sistema cartesiano una serie di punti di coordinate (x,y), e vi associa il valore di una terza variabile.

⇒ Questa opzione risulta attiva quando vengono selezionate tre colonne (variabili). La prima delle tre colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti. In corrispondenza di ciascun punto viene quindi tracciato un cerchio il cui diametro risulta proporzionale al valore della terza variabile.

⇒ La rappresentazione viene effettuata solamente se le tre colonne (variabili) contengono lo stesso numero di dati.

Utilizzare il file ESEMPI.MDB: caricare la tabella COLEST_F nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Selezionare le colonne ETA (manca l'accento, che non può essere incluso nel nome di un campo), COLEST e HDL, e osservare la relazione che esiste tra colesterolo totale (in mg/dL, in ordinate), età (in anni, in ascisse), e colesterolo HDL (in mg/dL, proporzionale al diametro del cerchio) in un gruppo di soggetti non selezionati, di sesso femminile.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.9. Sovrapposizione

Consente di rappresentare in una sistema cartesiano fino a quattro serie di punti di coordinate (x,y).

⇒ Questa opzione risulta attiva quando vengono selezionate da tre a cinque colonne (variabili). La prima delle colonne selezionate corrisponde ai valori delle ascisse, le successive (fino a un massimo di quattro) corrispondono ai valori delle ordinate dei singoli punti delle diverse serie.

⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file ESEMPI.MDB: caricare la tabella COLEST_M nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Selezionare le colonne ETA (manca l'accento, che non può essere incluso nel nome di un campo), COLEST, HDL, TRIGLI e LDL, e osservare la distribuzione di colesterolo totale, colesterolo HDL, trigliceridi e colesterolo LDL (età in anni in ascisse, il resto in mg/dL in ordinate) in un gruppo di soggetti non selezionati, di sesso maschile.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche

opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.10. Linee spezzate

Consente di rappresentare in una sistema cartesiano fino a quattro serie di punti di coordinate (x,y), unendoli mediante segmenti di retta (linee spezzate).

⇒ Questa opzione risulta attiva quando vengono selezionate da due a cinque colonne (variabili). La prima delle colonne selezionate corrisponde ai valori delle ascisse, le successive (fino a un massimo di quattro) corrispondono ai valori delle ordinate dei singoli punti delle diverse serie.

⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file VARBIOL.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Contiene la concentrazione del colesterolo totale determinata ripetutamente in alcuni volontari in un periodo di 25 giorni, selezionare le colonne (variabili) da GIORNO a SOGGETTO_D, e osservare l'andamento del colesterolo totale (in mg/dL, in ordinate) nei quattro soggetti, nell'arco di tempo (in giorni, in ascisse) indicato.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.11. Barre verticali

Consente di rappresentare in una sistema cartesiano fino a quattro serie di punti di coordinate (x,y) mediante barre verticali.

⇒ Questa opzione risulta attiva quando vengono selezionate da due a cinque colonne (variabili). La prima delle colonne selezionate corrisponde ai valori delle ascisse, le successive (fino a un massimo di quattro) corrispondono ai valori delle ordinate dei singoli punti delle diverse serie.

⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Quindi numerare le righe, inserendo il numero d'ordine dei dati nella prima colonna. Selezionare infine questa colonna e le successive quattro colonne denominate Controlli, NCAH, CPH e CAH, che contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica

attiva (CAH). Osservare il diagramma a barre verticali (tutti i valori sono in g/L).

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

3.3.5.12. Istogrammi multipli

Consente di rappresentare la distribuzione dei valori di più variabili (fino a quattro contemporaneamente) sotto forma di istogrammi.
⇒ Questa opzione risulta attiva quando vengono selezionate da due a quattro colonne (variabili).
⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.
⇒ Per analizzare una a una le singole variabili utilizzare l'opzione Statistiche Elementari del Menù Statistica.

Il file BAYES.DBF contiene i risultati ottenuti in un gruppo di soggetti sani (colonna SANI) e in un gruppo di soggetti malati (colonna MALATI): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare entrambe le colonne ed effettuarne la rappresentazione.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.
⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.13. Distribuzione

Consente di rappresentare sotto forma di un diagramma a pallini (dot-plot) la distribuzione dei dati di una o più colonne (variabili). Contrariamente al diagramma Quartili e Range, in questo diagramma viene conservata la rappresentazione dei singoli dati.
⇒ Questa opzione risulta attiva quando vengono selezionate da una a quattro colonne (variabili).
⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Selezionare le colonne Controlli, NCAH, CPH e CAH: contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e

infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma quartili e range (tutti i valori sono in g/L). Il gruppo dei soggetti con epatite alcolica non cirrogena, pure presentando una distribuzione notevolmente asimmetrica verso i valori alti, è anche quello che presenta in nucleo centrale più omogeneo, e che più si discosta dal gruppo di controllo.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.14. Limiti di confidenza

Consente di rappresentare in forma grafica la media, i limiti di confidenza al 95% della media e i limiti di confidenza al 95% della distribuzione campionaria. La tacca orizzontale inferiore corrisponde al limite di confidenza inferiore della distribuzione campionaria, la tacca subito al di sotto di quella centrale corrisponde al limite di confidenza inferiore della media, la tacca centrale alla media, la tacca subito al di sopra di quella centrale corrisponde al limite di confidenza superiore della media, e infine la tacca orizzontale superiore corrisponde al limite di confidenza superiore della distribuzione campionaria.

⇒ Questa opzione risulta attiva quando vengono selezionate da una a quattro colonne (variabili).

⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Selezionare le colonne Controlli, NCAH, CPH e CAH: contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma dei limiti di confidenza (tutti i valori sono in g/L).

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.5.15. Quartili e range

Consente di rappresentare sotto forma di una scatola con i baffi (box-whisker plot) il valore minimo, il primo quartile, la mediana, il terzo quartile e il valore massimo osservati in una o più colonne (variabili). La tacca orizzontale inferiore (baffo inferiore) corrisponde al valore minimo, il margine inferiore della scatola al primo quartile, il segmento orizzontale all'interno della scatola alla

mediana, il margine superiore della scatola al terzo quartile, la tacca orizzontale superiore (baffo superiore) al valore massimo.

⇒ Questa opzione risulta attiva quando vengono selezionate da una a quattro colonne (variabili).

⇒ La rappresentazione viene effettuata anche se le colonne (variabili) contengono un diverso numero di dati.

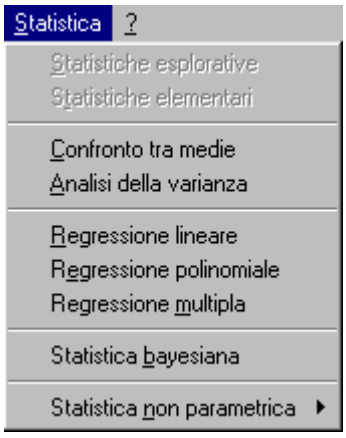
Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Selezionare le colonne Controlli, NCAH, CPH e CAH: contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), e infine in un gruppo di soggetti con epatite cronica attiva (CAH). Osservare il diagramma quartili e range (tutti i valori sono in g/L). Il gruppo dei soggetti con epatite alcolica non cirrogena, pure presentando una coda notevolmente asimmetrica verso i valori alti, è quello più omogeneo, e che più si discosta dal gruppo di controllo.

⇒ È possibile stampare il grafico mediante l'opzione Stampa del Menù File.

⇒ Per la stampa si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune e di personalizzare le legende.

5.3.6. Menù Statistica



Le opzioni di questo menù consentono di effettuare le elaborazioni statistiche desiderate. Le principali rappresentazioni grafiche (istogramma, distribuzione cumulativa, grafico dei dati e della retta e/o curva sovrapposta, eccetera) sono incluse ciascuna nell'opzione pertinente. Ulteriori interessanti e utili elaborazioni grafiche dei dati possono essere ottenute impiegando le opzioni del Menù Grafica, mentre le opzioni del Menù Calcoli consentono di effettuare sui dati eventuali trasformazioni preliminari. Tutte le opzioni del Menù Statistica presentano i risultati dell'elaborazione in una finestra di testo. È possibile selezionare il testo desiderato e copiarlo negli Appunti

5.3.6.1. Statistiche esplorative

Questa opzione consente di calcolare la media, la deviazione standard, il coefficiente di variazione percentuale (CV), l'errore standard della media e i quartili della distribuzione. Queste statistiche rappresentano il modo in cui esprimere i risultati dell'analisi elementare di dati campionari che siano distribuiti in modo gaussiano. Il metodo è riportato su tutti i testi di statistica, vedere per esempio Armitage[67].
È infine possibile valutare lo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica ricorrendo al coefficiente di asimmetria g_1 e al coefficiente di curtosi g_2, che indicano il tipo di scostamento come qui di seguito specificato:
$g_1 < 0$: asimmetria negativa, cioè coda sinistra della distribuzione eccessivamente lunga;
$g_1 > 0$: asimmetria positiva, cioè coda destra della distribuzione eccessivamente lunga;
$g_2 < 0$: platicurtosi, cioè distribuzione eccessivamente appiattita, con code troppo corte;
$g_2 > 0$: leptocurtosi, cioè distribuzione eccessivamente alta, con code troppo lunghe.
Un test per la significatività di g_1 e g_2 sufficientemente approssimato per le esigenze pratiche è il seguente: il coefficiente di asimmetria (o quello di curtosi) viene considerato significativo se il rapporto tra esso e il suo errore standard è superiore a 2,6: in questo caso si rigetta l'ipotesi che i dati siano distribuiti in modo gaussiano, e si impiegano le statistiche non parametriche. Il metodo è tratto da Snedecor[68], mentre il test approssimato per la significatività è quello suggerito da Solberg[69].
⇒ L'opzione risulta attiva quando non è stata selezionata alcuna colonna. I calcoli vengono eseguiti automaticamente su tutte le colonne contenenti dati. viene selezionata una sola colonna (variabile).

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Le colonne Controlli, NCAH, CPH e CAH contengono i valori di IgA nel siero in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), in un gruppo di soggetti con epatite cronica attiva (CAH) e infine in un gruppo di soggetti con cirrosi epatica alcolica (AC).

5.3.6.2. Statistiche elementari

Questa opzione consente di calcolare media, deviazione standard, coefficiente di variazione percentuale (CV) ed errore standard della media, i quartili della distribuzione, una tabella dei percentili più importanti, e una tabella della suddivisione dei dati in classi. È possibile rappresentare anche l'istogramma della distribuzione. Queste statistiche rappresentano il modo in cui esprimere i

[67] *Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:13-52.*
[68] *Snedecor GW, Cochran WG: Statistical methods. VII Edition. Ames:The Iowa State University Press, Ames, 1980:78-80.*
[69] *Solberg HE. The theory of reference values. Part 5. Statistical treatment of collected reference values - Determination of reference limits. J Clin Chem Clin Biochem 1983;21:749-60.*

risultati dell'analisi elementare di dati campionari che siano distribuiti in modo gaussiano. Il metodo è riportato su tutti i testi di statistica, vedere per esempio Armitage[70].
Lo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica viene valutato ricorrendo ancora al coefficiente di asimmetria g_1 e al coefficiente di curtosi g_2.
Il metodo è tratto da Snedecor[71], mentre il test approssimato per la significatività di g_1 e di g_2 è quello suggerito da Solberg[72].
⇒ L'opzione risulta attiva quando viene selezionata una sola colonna (variabile).

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Il calcolo delle Statistiche Elementari sul valore dei trigliceridi (per l'elaborazione selezionare la colonna TRIGLI) porta ad una tabella dei percentili nella quale il 2,5-esimo percentile parametrico e il 97,5-esimo percentile parametrico (rispettivamente la media meno 1,96 volte la deviazione standard e la media più 1,96 volte la deviazione standard) sono pari a -45,3 e 318,3. In base a questo dovremmo affermare che nel 95% dei soggetti esaminati i trigliceridi sono compresi tra -45,3 e 318,3 mg/dL: un'affermazione che, a dispetto del fatto che i risultati sono stati ottenuti con un metodo statistico rigoroso, nessuno si sentirebbe di sostenere. Le cose vanno meglio nel caso delle statistiche non parametriche: il valore non parametrico comparativo riportato nella stessa tabella per il 2,5-esimo e per il 97,5-esimo percentile è pari rispettivamente 44,7 e 382,0 mg/dL. Quello illustrato è, ancorché basato su dati reali, un caso limite, nel quale la contraddizione in termini biologici rappresentata da un valore di concentrazione negativo evidenzia l'errore nell'applicazione della metodologia statistica.

⇒ L'opzione Statistica Non Parametrica del Menù Statistica offre una serie di test statistici non parametrici che consentono di risolvere i principali problemi in tal senso, tra i quali il calcolo delle Statistiche Elementari con metodo non parametrico.
⇒ È possibile stampare sia i risultati dell'elaborazione statistica che l'istogramma mediante l'opzione Stampa del Menù File.
⇒ Per la stampa dell'istogramma si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire la scala delle ordinate più opportuna, di personalizzare le legende e di cambiare il numero delle classi in cui sono suddivisi i dati.
⇒ Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

[70] *Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:13-52.*
[71] *Snedecor GW, Cochran WG: Statistical methods. VII Edition. Ames:The Iowa State University Press, Ames, 1980: 78-80.*
[72] *Solberg HE. The theory of reference values. Part 5. Statistical treatment of collected reference values - Determination of reference limits. J Clin Chem Clin Biochem 1983;21:749-60.*

5.3.6.3. Confronto tra medie

Questa opzione consente di effettuare il confronto tra medie sia per dati appaiati che per campioni indipendenti. Il confronto tra medie viene effettuato mediante un test statistico parametrico: il test *t* di Student.

⇒ L'opzione risulta attiva quando vengono selezionate due colonne (variabili).

⇒ Il calcolo del test *t* di Student per dati appaiati è possibile solamente se le due colonne (variabili) contengono lo stesso numero di dati, quello per campioni indipendenti è possibile sempre.

Il confronto fra medie per dati appaiati è un caso particolare del confronto fra medie: infatti l'appaiamento dei dati consente di ottimizzare l'omogeneità fra i due campioni posti a confronto, che differiranno fra di loro solamente per il fattore che lo sperimentatore ha deliberatamente introdotto. ll test *t* di Student per dati appaiati viene in genere considerato significativo quando *p* è inferiore al 5% ($p < 0,05$). Il metodo è tratto da Snedecor[73].

Utilizzare come esempio la tabella AST del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Contiene il valore della AST (aspartato amminotransferasi) determinata su sieri freschi (colonna SUBITO), e rideterminata sugli stessi sieri dopo conservazione a +4 C per 24 ore (colonna DOPO24ORE), al fine di stabilire se tale conservazione sia idonea a mantenere l'attività dell'enzima. Selezionare entrambe le colonne. I risultati del test *t* di Student sembrano suggerire che mediamente non vi siano differenza tra i due valori.

Tuttavia è forse più frequente incontrare situazioni nelle quali i campioni sono indipendenti l'uno dall'altro: si pensi al confronto fra i valori di concentrazione di un qualsiasi analita in maschi e femmine, in adulti e in soggetti in età pediatrica, e via dicendo. Il test *t* di Student per il confronto fra le medie di campioni indipendenti è basato su un principio semplice, e tutto sommato intuitivo: a parità di valore assoluto della differenza tra le medie, a tale differenza viene data tanto maggior peso quanto minore è la dispersione dei dati campionari, e viceversa. Può quindi accadere che una ampia differenza fra le medie, alla quale si sarebbe propensi a dare importanza, risulti non significativa a causa della grande dispersione dei dati campionari, mentre per converso una piccola differenza può rivelarsi inaspettatamente significativa, qualora la dispersione dei dati campionari sia minima. Un problema tuttavia complica lievemente la situazione: il test *t* di Student ordinario tende a fornire troppo pochi risultati significativi quando il campione più grande ha la varianza (cioè la dispersione dei dati) maggiore, e troppi risultati significativi quando il campione più grande ha la varianza minore. Si consiglia perciò di utilizzare, nei casi in cui la varianza dei due campioni differisca in maniera significativa, una forma del test che preveda una opportuna correzione. Se il rapporto tra le varianze dei due campioni indica varianze significativamente diverse (test *F* con $p < 0,05$) utilizzare i risultati del test *t* per varianze non omogenee. Si fa notare che i risultati ottenuti con le due forme del test *t* di Student sono tanto più diversi quanto più differiscono tra di loro le varianze dei due campioni, mentre forniscono lo stesso identico risultato quando le varianze dei due campioni sono uguali. Infine il test *t* di Student per campioni indipendenti viene in genere considerato significativo quando *p* è inferiore al 5% ($p < 0,05$). Il metodo è tratto da Snedecor[74].

[73] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:83-89.*

[74] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:89-99.*

Il file RIBOFLAV.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene la concentrazione di riboflavina nel fegato (colonna FEGATO) e nel muscolo di bue (colonna MUSCOLO), come determinata in una apposita sperimentazione. Selezionare entrambe le colonne. Il test *F* evidenzia varianze significativamente diverse nei due casi. Il test *t* di Student per varianze non omogenee conferma una differenza nel contenuto di riboflavina dei due tessuti, la cui significatività è pero meno rilevante di quanto apparirebbe a prima vista (*p* solo di poco inferiore a 0,05).

⇒ È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
⇒ Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.4. Analisi della varianza

Questa opzione consente di effettuare l'analisi della varianza secondo diversi disegni sperimentali.
⇒ Questa opzione risulta attiva quando vengono selezionate due o più colonne (variabili), che devono contenere lo stesso numero di dati.
⇒ Se vengono selezionate due colonne, risulta abilitato il solo calcolo del rapporto tra varianze.
⇒ Se vengono selezionate più di due colonne, il calcolo del rapporto tra varianze risulta disabilitato, mentre vengono abilitati il calcolo dell'analisi della varianza a un fattore, quello dell'analisi della varianza a due fattori e quello delle componenti della variabilità.

Il rapporto tra varianze riportato in questa opzione è identico a quello impiegato per verificare l'omogeneita delle varianza nel confronto tra medie: un test *F* con $p < 0,05$ indica varianze significativamente diverse.

Il file RIBOFLAV.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene la concentrazione di riboflavina nel fegato (colonna FEGATO) e nel muscolo (colonna MUSCOLO) di bue, come determinata in una apposita sperimentazione. Selezionare le due colonne (variabili) ed effettuare i calcoli: il test *F* evidenzia varianze significativamente diverse nei risultati ottenuti sui due tessuti.

Con le varie forme del test *t* di Student (test parametrico) e del test di Wilcoxon (test non parametrico) è possibile confrontare fra di loro due medie: l'analisi della varianza o ANOVA (da ANalysis Of VAriance) a un fattore consente di estendere il confronto a più di due medie.

Il file DIETA.DBF contiene la riduzione dei trigliceridi nel siero osservata in quattro gruppi di soggetti sottoposti a differenti diete. Selezionare le colonne da DIETA_A a DIETA_D: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. L'analisi della varianza a un fattore documenta una differenza significativa tra le riduzioni della concentrazione dei trigliceridi ottenute con le quattro diete. L'ispezione dei dati consente di attribuire alla DIETA_D la differenza significativa. Il metodo è trattato in tutti i testi di statistica; molto chiaro, fra gli altri Wonnacott[75].

Mentre l'ANOVA un fattore (detta anche "a un criterio di classificazione") consente di verificare se vi siano (in media) differenza significative fra gli elementi appartenenti alle colonne (variabili) della

[75] *Wonnacott TH, Wonnacott RJ. Introduzione alla statistica. Milano:Franco Angeli, 1980:237-54.*

tabella in cui sono stati ordinatamente raccolte le osservazioni, l'ANOVA a due fattori consente di verificare contemporaneamente se vi siano (in media) differenza significative fra gli elementi appartenenti alle righe della tabella. Che questo sia significativo dipende esclusivamente dal disegno sperimentale.

Si consideri la tabella TECNICI del file ESEMPI.MDB: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Nella tabella sono stati registrati, per cinque giorni consecutivi, i risultati del controllo di qualità di un metodo analitico eseguito in modo semi-automatico. I risultati sono stati registrati separatamente per i tre tecnici che eseguono detto metodo. Il disegno sperimentale è stato strutturato in questo modo in quanto così ci si aspetta di poter rilevare dalla variabilità tra le colonne la variabilità dovuta alla differente manualità del personale, e dalla variabilità tra le righe la variabilità tra giorni dovuta al metodo analitico stesso. Selezionare le colonne da TECNICO_A a TECNICO_C. L'ANOVA a due fattori rivela una variabilità non significativa tra i tecnici (variabilità tra le colonne), ed una variabilità significativa tra giorni (variabilità tra le righe). Per il metodo vedere ancora Wonnacott[76].

La variabilità entro duplicati consente di calcolare la variabilità presente tra misure duplicate dello stesso evento. Tipicamente viene utilizzata per determinare l'imprecisione di un metodo analitico quando si hanno a disposizione misure in doppio effettuate su campioni della routine. Ovviamente il significato dell'imprecisione misurata è condizionato in modo determinante dal disegno sperimentale adottato: così se le misure in doppio sono ottenute nella stessa serie analitica l'imprecisione misurata rappresenterà una stima dell'imprecisione entro la serie, mentre se le misure in doppio sono ottenute in giorni e quindi in serie analitiche diverse l'imprecisione misurata rappresenterà un stima dell'imprecisione tra le serie.

Il file CLORURO.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene venti coppie di valori, ottenute determinando la concentrazione del cloruro in doppio su altrettanti sieri della routine, nella stessa serie analitica. Il metodo sembra avere una buona precisione. L'esempio è tratto dal Manuale del Controllo di Qualità a cura del CROSP della Regione Lombardia[77].

L'analisi delle componenti della variabilità consente di decomporre la variabilità totale in due componenti, quella tra le colonne (variabili) e quella entro le colonne. Viene utilizzata l'analisi della varianza nella forma modificata proposta da Krouwer[78], e molto simile a quella consigliata anche dal National Comittee for Cinical Laboratory Standards[79].

La tabella POTASSIO del file ESEMPI.MDB contiene i risultati della determinazione del potassio su un singolo campione, effettuata per cinque giorni, con cinque replicati per giorno: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) Apri del Menù File. Lo scopo è di determinare l'imprecisione totale del metodo e di decomporla in due componenti: quella entro giorni e quella tra giorni. Selezionare le colonne da GIORNO_1 a GIORNO_5 ed effettuare il calcolo delle componenti della variabilità. Come atteso la variabilità tra giorni (tra le colonne) risulta superiore a quella entro giorni (entro le colonne).

---

[76] *Wonnacott TH, Wonnacott RJ. Introduzione alla statistica. Milano:Franco Angeli, 1980:254-63.*
[77] *Giunta Regionale della Lombardia, Assessorato alla Sanità. Manuale del Controllo di Qualità nel Laboratorio di Patologia Clinica. Milano: Comitato Regionale per l'Ordinamento dei Servizi di Patologia, 1978;63.*
[78] *Krouwer JS, Rabinowitz R. How to improve estimates of imprecision. Clin Chem 1984;30:2902.*
[79] *NCCLS. Proposed Standard PSEP-3. Protocol for establishing performance claims for clinical chemical methods. Replication experiment. NCCLS, 771 E. Lancaster Ave., Villanova, PA 19085.*

Interessante anche il file VARBIOL.DBF. Contiene i valori di colesterolo totale, determinati in una serie di giorni successivi in cinque volontari. Lo scopo è quello dei decomporre la variabilità biologica totale in due componenti: variabilità biologica intra-individuale (stimata dalla variabilità entro le colonne) e variabilità biologica inter-individuale (stimata dalla variabilità tra le colonne). Selezionare le colonne da SOGGETTO_A a SOGGETTO_E ed effettuare il calcolo delle componenti della variabilità. Come atteso la variabilità biologica intra-individuale è inferiore a quella tra i diversi individui (inter-individuale).

⇒ È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.

⇒ Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.5. Regressione lineare

Questa opzione consente di effettuare il calcolo dell'equazione della retta di regressione con tre diversi modelli.

⇒ Questa opzione risulta attiva quando vengono selezionate due colonne (variabili)

⇒ Le due colonne devono contenere lo stesso numero di dati.

⇒ La prima delle due colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti.

Per adattare la retta ai dati sperimentali viene impiegato il metodo dei minimi quadrati, una tecnica di approssimazione ben nota, che consente di minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta.

Il metodo dei minimi quadrati viene impiegato con tre differenti modelli: la regressione *x* variabile indipendente (regressione lineare standard), la regressione *y* variabile indipendente e la componente principale standardizzata.

A meno che non vi siano motivi particolari, che l'utilizzatore dovrà attentamente valutare, impiegare sempre per risultati della regressione lineare *x* variabile indipendente. Il modello matematico impiegato presuppone che la *x*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *y* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *x*. Tra le varie statistiche fornite, particolarmente interessanti sono la significatività della differenza del coefficiente angolare da 0 (zero) e da 1 (uno), e la significatività della differenza dell'intercetta da zero (per $p < 0{,}05$ le differenze citate possono essere considerate significative). Inoltre con il grafico della retta di regressione vengono tracciati i limiti di confidenza al 90% della regressione. Il metodo è trattato su tutti i testi di statistica, fra i quali si ricordano Armitage[80] e Snedecor[81].

Per l'interpretazione dei risultati della regressione lineare in funzione dei diversi tipi di applicazione di questa tecnica statistica nel campo della chimica clinica, vedere Davis[82].

La regressione lineare *y* variabile indipendente è semplicemente l'immagine speculare della precedente. Deve essere utilizzata solamente a scopo esplorativo dei dati. Tanto più differisce dalla

[80] *Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:151-167 e 264-266.*

[81] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:149-174.*

[82] *Davis RD, Thompson JE, Pardue HL. Characteristics of statistical parameters used to interpret least-squares results. Clin Chem 1978;24:611-20.*

regressione *x* variabile indipendente, in presenza di errori nella misura delle *x*, tanto più si dovrà considerare di esprimere i risultati della regressione in termini di componente principale standardizzata. Infatti la regressione lineare standard non è raccomandata per l'analisi statistica di dati nei quali la variabile indipendente sia affetta da un errore di misura: in questo caso, impiegando la regressione lineare standard, si ottengono, a seconda di quale variabile sia posta in ascisse, equazioni della retta di regressione diverse: e questo fatto può portare a conclusioni contraddittorie. Per i casi di questo genere (tipico il confronto tra due metodi per la determinazione dello stesso analita) si consiglia di impiegare, in alternativa alla regressione lineare standard, la componente principale standardizzata, per la quale vedere Feldman[83].

La tabella CALCIO del file ESEMPI.MDB contiene i risultati di due ipotetici metodi per la determinazione del calcio nel siero: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Sui primi tre sieri i due metodi hanno fornito identici risultati. Sui rimanenti quattro sieri hanno fornito risultati speculari. La situazione appare caratterizzata da una assoluta indecidibilità. Selezionare entrambe le colonne: i valore della prima colonna selezionata (METODO_A) saranno riportati in ascisse, quelli della seconda colonna selezionata (METODO_B) saranno riportati in ordinate. La regressione x variabile indipendente e la regressione y variabile indipendente forniscono in effetti risultati speculari (utilizzare la rappresentazione delle rispettive rette di regressione per visualizzare le rette). La componente principale standardizzata più salomonicamente fa propendere per il fatto che, con le informazioni disponibili, possiamo solo concludere che i due metodi forniscono sostanzialmente gli stessi risultati.

Dati reali, con cui sperimentare la regressione lineare, sono quelli relativi alla determinazione dell'urea (valori in mg/dL) effettuata, su una serie di sieri freschi della routine, con cinque strumenti diversi.

I dati sono contenuti nel file UREA.DBF, che può essere importato con l'opzione Importa File (in formato dBase III®) del Menù File. Da notare come in questo caso la scelta di una adeguata ampiezza della dispersione dei valori di concentrazione dell'urea consente di ottenere, con la regressione *x* variabile indipendente e con la regressione *y* variabile indipendente, risultati molto simili. In ogni caso il risultato da utilizzare dovrebbe essere quello fornito dalla componente principale standardizzata. Si fa notare infine che i risultati della componente principale standardizzata sono validi solamente quando i coefficienti angolari della regressione lineare *x* variabile indipendente e della regressione lineare *y* variabile indipendente sono entrambi positivi.

⇒ È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.

⇒ Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il valore della y corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato valore della y.

⇒ Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

[83] *Feldman U, Schneider B, Klinkers H. A multivariate approach for the biometric comparison on analytical methods in clinical chemistry. J Clin Chem Clin Biochem 1981;19:131-7.*

5.3.6.6. Regressione polinomiale

Questa opzione consente di adattare a una serie di punti, con il metodo dei minimi quadrati una retta (polinomio di primo grado) oppure, in alternativa, un polinomio di secondo grado o un polinomio di terzo grado.

⇒ Questa opzione risulta attiva quando vengono selezionate due colonne (variabili)

⇒ Le due colonne devono contenere lo stesso numero di dati.

⇒ La prima delle due colonne selezionate corrisponde ai valori delle ascisse, la seconda corrisponde ai valori delle ordinate dei singoli punti.

Per il polinomio di primo grado viene utilizzata la regressione lineare *x* variabile indipendente (regressione lineare standard): con l'aggiunta però di un test per la linearità, che consente di verificare se la retta descrive adeguatamente la relazione di funzione $y = f(x)$ che lega i valori in ordinate a quelli in ascisse come indicato in Burnett[84].

Per valori di $p < 0,05$ si può considerare che tale relazione non sia adeguatamente descritta da una retta, ma sia piuttosto curvilinea.

Il file TESTLIN.DBF contiene i valori di assorbanza (in ordinate, ASSORBANZA) ottenuti per una serie di concentrazioni note di un ipotetico analita (in ascisse, CONCENTRAZ): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare entrambe le colonne. Il test per la linearità (*p* di gran lunga inferiore a 0,05) indica che la relazione tra assorbanza e concentrazione si discosta di molto dalla linearità: la relazione viene meglio descritta da un polinomio di secondo grado.

Anche nel caso della regressione polinomiale di secondo grado il metodo dei minimi quadrati viene utilizzato per adattare la corrispondente funzione polinomiale ai dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa. Il sistema di equazioni che consente di calcolare termine noto e coefficienti del polinomio di secondo grado può essere facilmente risolto in forma matriciale, mediante la regola di Cramer (il matematico svizzero vissuto fra il 1704 e il 1752). Vedere Batschelet[85] e Spiegel[86].

La tabella APTOGLOB del file ESEMPI.MDB contiene i valori di assorbanza (in ordinate, ASSORBANZA) e di concentrazione (in ascisse, MG_DL) di una curva di calibrazione immunoturbidimetrica: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File). Selezionare entrambe le colonne: un polinomio di secondo grado descrive adeguatamente la curva.

Infine anche nel caso della regressione polinomiale di terzo grado il metodo dei minimi quadrati viene utilizzato per adattare la corrispondente funzione polinomiale ai dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa. Il sistema di equazioni che consente di calcolare termine noto e coefficienti del polinomio di terzo grado può anche in questo caso essere facilmente risolto in forma matriciale mediante la regola di Cramer (il matematico svizzero vissuto fra il 1704 e il 1752). Vedere Batschelet[87] e Scheid[88].

---

[84] *Burnett RW. Quantitative evaluation of linearity. Clin Chem 1980;26:644-6.*

[85] *Batschelet E. Introduzione alla matematica per biologi. Padova: Piccin, 1979:511-8.*

[86] *Spiegel MS. Statistica. Milano: Gruppo Editoriale Fabbri-Bompiani, 1976:221.*

[87] *Batschelet E. Introduzione alla matematica per biologi. Padova: Piccin, 1979:511-8.*

La tabella CINETICA del file ESEMPI.MDB contiene i valori di assorbanza (in ordinate, ASSORBANZA) e di tempo (in ascisse, SECONDI) registrati per una cinetica enzimatica: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File). Il polinomio di terzo grado appare l'unico ad essere in grado di descrivere adeguatamente l'andamento della cinetica nell'intervallo dei dati sperimentali.

⇒ È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.
⇒ Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).
⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il valore della y corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato valore della y.
⇒ Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.7. Regressione multipla

Questa opzione consente di definire la migliore relazione lineare che lega un numero di variabili superiore a quello che ne consente la rappresentazione in un piano cartesiano (quindi un numero di variabili superiore a due). Per il metodo vedere Snedecor[89].

⇒ Questa opzione risulta attiva quando vengono selezionate da 2 a 10 colonne (variabili)
⇒ Le due colonne devono contenere tutte lo stesso numero di dati.
⇒ L'ultima colonna selezionata deve contenere la variabile dipendente (*y*), le colonne selezionate a sinistra di questa sono assunte contenere le variabili indipendenti.

Un esempio può chiarire meglio di tutto le potenzialità di impiego di questo test statistico.

Utilizzare il file COLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Limitare a 30 i dati da analizzare: per questo è sufficiente vuotare la riga 31 (selezionare la riga 31 facendo click sulla cella contenente il numero della riga, dal Menu Modifica scegliere Riga e quindi Vuota). Selezionare le colonne COLEST, HDL, TRIGLI e LDL. Calcolare la regressione multipla. Il colesterolo LDL è la variabile dipendente (LDL, in mg/dL): le variabili indipendenti sono il colesterolo totale (COLEST, in mg/dL), il colesterolo HDL (HDL, in mg/dL) e i trigliceridi (TRIGLI, in mg/dL). Osservare i coefficienti della regressione multipla: forniscono il risultato atteso ? [Risposta: si, i coefficienti della regressione forniscono con una notevole accuratezza la formula di Friedewald, LDL = COLEST - HDL - (TRIGLI/5) con la quale è stato calcolato il colesterolo LDL].

⇒ È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.

---

[88] *Scheid F. Analisi numerica. Milano: Gruppo Editoriale Fabbri-Bompiani, Sonzogno, 1975:235-66.*
[89] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:334-64.*

⇒ Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.8. Statistica bayesiana

Questa opzione consente di calcolare, per un dato test di laboratorio, la sensibilità, la specificità, il valore predittivo del test positivo, il valore predittivo del test negativo, l'efficienza diagnostica e la curva ROC. Il valore predittivo del test positivo e il valore predittivo del test negativo possono essere ricalcolati facendo variare la prevalenza della malattia.

⇒ L'opzione risulta attiva quando vengono selezionate due colonne (variabili)

⇒ Le due colonne possono contenere un diverso numero di dati.

⇒ I risultati del test nei soggetti sani devono trovarsi nella prima delle due colonne selezionate (colonna di sinistra), i risultati del test nei soggetti malati devono trovarsi nella seconda delle colonne selezionate (colonna di destra).

Ovviamente i soggetti devono essere stati classificati nei due gruppi (sani e malati) con un criterio indipendente dal test di laboratorio di cui si vogliono determinare le caratteristiche. Tutte le grandezze citate possono essere rappresentate graficamente, sovraimposte agli istogrammi della distribuzione dei risultati del test ottenuti nel un gruppo di soggetti sani e nel gruppo di soggetti malati.

Il file BAYES.DBF contiene i risultati ottenuti in un gruppo di soggetti sani (colonna SANI) e in un gruppo di soggetti malati (colonna MALATI): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare entrambe le colonne ed effettuare i calcoli e le rappresentazioni previste dall'opzione. Si ricordi che la scelta del livello decisionale dipende dalle considerazioni di merito relative agli obbiettivi che ci si propone. Per un test con un costo contenuto, e che rivela una malattia curabile, la scelta più ovvia sarà quella di fissare il livello decisionale al massimo della sensibilità (100%), anche se ciò andrà a scapito della specificità (aumento dei falsi positivi, cioè dei falsi allarmi). Ma se i falsi positivi dovessero innescare processi diagnostici per l'esclusione della malattia complessi, molto costosi, e magari potenzialmente pericolosi, ecco magari la necessità di bilanciare meglio, anche in funzione delle risorse disponibili, sensibilità e specificità.

Per una trattazione dettagliata della statistica bayesiana illustrata da eccellenti esempi clinici delle strategie da seguire, vedere Galen e Gambino[90] e Gherardt[91]. Per una trattazione delle curve ROC vedere Robertson[92].

⇒ È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.

⇒ Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

---

[90] *Galen RS, Gambino SR. Oltre il concetto di normalità. Padova: Piccin, 1980:237pp.*

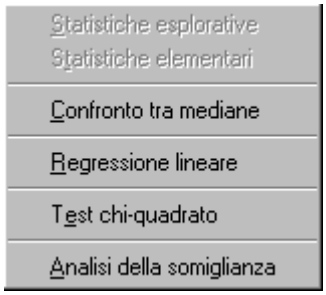
[91] *Gherardt W, Keller H. Evaluation of test data from clinical studies. Scand J Clin Lab Invest;1986:46(Supplement 181),74pp.*

[92] *Robertson EA, Zweig MH. Use of receiver operating chracteristic curves to evaluate the clinical performance of analytical systems. Clin Chem 1981;27:1569-74.*

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, e di ridefinire il valore della prevalenza della malattia (in quest'ultimo caso la sensibilità, la specificità, il valore predittivo del test positivo, il valore predittivo del test negativo, l'efficienza diagnostica e la curva ROC sono ricalcolati automaticamente).

⇒ Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.9. Statistica non parametrica



Questa opzione comprende un vero e proprio menù, e come tale viene presentata. Per un approccio generale al problema delle statistiche non parametriche vedere Maritz[93] e Siegel[94].

5.3.6.9.1. Statistiche esplorative

Questa opzione consente di calcolare la mediana, il range e i quartili della distribuzione. Queste statistiche rappresentano il modo in cui esprimere i risultati dell'analisi elementare di dati campionari che non siano distribuiti in modo gaussiano.

⇒ L'opzione risulta attiva quando non è stata selezionata alcuna colonna.

Utilizzare il file in formato Ministat IGA.MI1 (caricarlo nella tabella dati di Ministat con l'opzione Apri del Menù File). Le colonne Controlli, NCAH, CPH e CAH contengono i valori di IgA nel siero ottenuti rispettivamente in un gruppo di soggetti di controllo (Controlli), in un gruppo di soggetti con epatite alcolica non cirrogena (NCAH), in un gruppo di soggetti con epatite cronica persistente (CPH), in un gruppo di soggetti con epatite cronica attiva (CAH) e infine in un gruppo di soggetti con cirrosi epatica alcolica (AC).

5.3.6.9.2. Statistiche elementari

Questa opzione consente di calcolare la mediana, il range, i quartili della distribuzione, e una tabella dei percentili più importanti. È possibile rappresentare anche la distribuzione cumulativa dei dati, nella forma standard e nella forma di distribuzione cumulativa ripiegata proposta da Krouwer[95].

[93] *Maritz JS. Distribution-free statistical methods. London: Chapman and Hall Editor, 1984:264pp.*
[94] *Siegel S, Castellan NJ Jr. Statistica non parametrica. Milano: McGraw-Hill, 1988:477pp.*

Queste statistiche rappresentano il modo in cui esprimere i risultati dell'analisi elementare di dati campionari che non siano distribuiti in modo gaussiano. Come anche nel caso delle Statistiche Elementari determinate con metodo parametrico, sono riportati i limiti di confidenza al 90% dei percentili. Per i metodi vedere La Rocca[96].

⇒ L'opzione risulta attiva quando viene selezionata una sola colonna (variabile).

Utilizzare il file COOLEST.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Il calcolo delle Statistiche Elementari sul valore dei trigliceridi (per l'elaborazione selezionare la colonna TRIGLI) con metodo parametrico (scegliere l'opzione Statistiche Elementari dal Menù Statistica) porta ad una tabella dei percentili nella quale il 2,5-esimo percentile parametrico e il 97,5-esimo percentile parametrico (rispettivamente la media meno 1,96 volte la deviazione standard e la media più 1,96 volte la deviazione standard) sono pari a -40,3 e 318,3. In base a questo dovremmo affermare che nel 95% dei soggetti esaminati i trigliceridi sono compresi tra -40,3 e 318,3 mg/dL: un'affermazione che, e dispetto del fatto che i risultati sono stati ottenuti con un metodo statistico rigoroso, nessuno si sentirebbe di sostenere. Le cose vanno meglio nel caso delle statistiche non parametriche con metodo non parametrico, con la presente opzione (provare).

⇒ È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.

⇒ Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il percentile corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato percentile.

⇒ Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.9.3. Confronto tra mediane

Questa opzione consente di effettuare il confronto tra le mediane nel caso di campioni indipendenti e di dati appaiati.

⇒ L'opzione risulta attiva quando vengono selezionate due colonne (variabili).

⇒ Il calcolo del test di Wilcoxon per dati appaiati è possibile solamente se le due colonne (variabili) contengono lo stesso numero di dati, quello per campioni indipendenti è possibile sempre.

Il test di Wilcoxon per dati appaiati è l'equivalente non parametrico del test *t* di Student per dati appaiati, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano. La

---

[95] *Krouwer JS, Monti KL. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem 1995;33:525-7.*

[96] *La Rocca S, Milani S, Oriani G. I valori di riferimento. Principi teorici e metodologia di produzione. Biochim Clin 1989;13:168-79.*

soluzione utilizzata per il calcolo della significatività è sufficientemente accurata per n > 16. Il metodo è tratto da Snedecor[97].

Il file ARTRITER.DBF contiene i valori delle immunoglobuline, determinati in pazienti trattati con due diversi farmaci (FARMACO_A e FARMACO_B): caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Il disegno sperimentale era strutturato in modo da determinare l'appaiamento dei pazienti. Si sapeva inoltre che i valori delle immunoglobuline non sono distribuiti in modo gaussiano. Importare il file e selezionare entrambe le colonne. Calcolare innanzitutto le differenze tra le coppie di valori mediante l'opzione Differenza del Menù Calcoli. Le differenze hanno per lo più segno negativo, e indicano che i valori delle immunoglobuline rilevati dopo trattamento con il farmaco A sono tendenzialmente inferiori a quelli rilevati dopo trattamento con il farmaco B. Il test di Wilcoxon per dati appaiati conferma la significatività delle differenze legate ai due tipi di trattamento. L'esempio è tratto da Bossi[98].

Sebbene spesso chiamato test di Mann-Whitney, l'equivalente non parametrico del test *t* di Student per campioni indipendenti è dovuto anch'esso a Wilcoxon, come ricorda lo Snedecor: e qui si è voluto adottare l'eponimo che sembra essere storicamente più corretto. La soluzione utilizzata per il calcolo della significatività è sufficientemente accurata per n > 16. Il metodo è tratto da Snedecor[99].

Il file RIBOFLAV.MI1 (file in formato Ministat: caricarlo nella tabella dati con l'opzione Apri del Menù File) contiene la concentrazione di riboflavina nel fegato (colonna FEGATO) e nel muscolo (colonna MUSCOLO) di bue, come determinata in una apposita sperimentazione. Selezionare le due colonne (variabili) ed effettuare i calcoli: il test di Wilcoxon per campioni indipendenti evidenzia una differenza significativa tra i risultati ottenuti sui due tessuti.

⇒ È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.

⇒ Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.9.4. Regressione lineare

Questa opzione consente di calcolare la regressione *x* variabile indipendente, la regressione *y* variabile indipendente, e l'equivalente della componente principale standardizzata in modo non parametrico. Come equivalente non parametrico della componente principale standardizzata viene utilizzato il modello di regressione lineare non parametrica proposto da Passing e Bablok[100,101].

---

[97] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:141-3.*

[98] *Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Roma: La Nuova Italia Scientifica, 1994:341.*

[99] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:144-5.*

[100] *Passing H, Bablok W. A new biometric procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. J Clin Chem Clin Biochem 1983;21:709-20.*

[101] *Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample size. Application of linear regression procedures for method comparison studies in clinical chemistry, Part II. J Clin Chem Clin Biochem 1984;22:431-45.*

Per un articolo riassuntivo e una bibliografia abbastanza completa riguardo il problema del confronto tra metodi vedere Besozzi[102].

⇒ L'opzione risulta attiva quando vengono selezionate due colonne (variabili)

⇒ Le due colonne devono contenere lo stesso numero di dati.

La tabella CALCIO del file ESEMPI.MDB contiene i risultati di due ipotetici metodi per la determinazione del calcio nel siero: caricarla nella tabella dati di Ministat con l'opzione Importa File (in formato Access® 1.x) del Menù File. Sui primi tre sieri i due metodi hanno fornito identici risultati. Sui rimanenti quattro sieri hanno fornito risultati speculari. La situazione appare caratterizzata da una assoluta indecidibilità. Selezionare entrambe le colonne: i valore della prima colonna selezionata (METODO_A) saranno riportati in ascisse, quelli della seconda colonna selezionata (METODO_B) saranno riportati in ordinate. La regressione x variabile indipendente e la regressione y variabile indipendente forniscono in effetti risultati speculari (utilizzare la rappresentazione delle rispettive rette di regressione per visualizzare le rette). La regressione lineare di Passing e Bablok più salomonicamente fa propendere per il fatto che, con le informazioni disponibili, possiamo solo concludere che i due metodi forniscono sostanzialmente gli stessi risultati.

Dati reali, con cui sperimentare la regressione lineare, sono quelli relativi alla determinazione dell'urea (valori in mg/dL) effettuata, su una serie di sieri freschi della routine, con cinque strumenti diversi.

I dati sono contenuti nel file UREA.DBF, che può essere importato con l'opzione Importa File (in formato dBase III®) del Menù File. Da notare come in questo caso la scelta di una adeguata ampiezza della dispersione dei valori di concentrazione dell'urea consente di ottenere, con la regressione *x* variabile indipendente e con la regressione *y* variabile indipendente, risultati molto simili. In ogni caso il risultato da utilizzare dovrebbe essere quello fornito dalla regressione di Passing e Bablok. Si fa notare infine che i risultati della regressione lineare di Passing e Bablok sono validi solamente quando i coefficienti angolari della regressione lineare *x* variabile indipendente e della regressione lineare *y* variabile indipendente sono entrambi positivi.

⇒ È possibile stampare sia i risultati dell'elaborazione statistica che i grafici mediante l'opzione Stampa del Menù File.

⇒ Per la stampa dei grafici si consiglia di togliere lo sfondo grigio (scegliere Configurazione Colori per Grafica e fare click su Sfondo Grigio). Se non si dispone di una stampante a colori è anche opportuno rappresentare i dati in bianco e nero (scegliere Configurazione Colori per Grafica e fare click su Dati a Colori).

⇒ L'opzione Configurazione Scale Assi e Legende consente di definire le scale degli assi più opportune, di personalizzare le legende, di calcolare il valore della y corrispondente a un dato valore della x, e di calcolare il valore della x corrispondente a un dato valore della y.

⇒ Infine si ricorda che i risultati dell'elaborazione statistica sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

[102] *Besozzi M, Franzini C. Impiego della regressione lineare nel confronto fra metodi: nuove tecniche statistiche parametriche e non parametriche (con in appendice un programma in BASIC). Giorn It Chim Clin 1986;11:29-41*.

5.3.6.9.5. Test chi quadrato

Questa opzione consente di calcolare il test chi-quadrato nella forma generalizzata per tabelle di contingenza, che è quella impiegata da Ministat. Per questo è necessario che ciascun elemento del campione in esame possa essere classificato per una caratteristica in un numero R di classi, e per una seconda caratteristica in un numero C di classi, in modo tale che i dati possano essere organizzati in una tabella di R righe per C colonne. Il metodo è trattato in tutti i testi di statistica, vedere per esempio Ingelfinger[103].

⇒ L'opzione risulta attiva quando vengono selezionate almeno due colonne (variabili)

⇒ Le colonne devono contenere lo stesso numero di dati.

Utilizzare il file FUMO.DBF: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Selezionare le colonne VIVI e DECEDUTI (notare come non vengano importati i contenuti dei campi non numerici. Comunque la riga 1 contiene i risultati ottenuti nel gruppo dei non fumatori, la riga 2 contiene i risultati ottenuti nel gruppo di fumatori). Il file contiene i risultati di un esperimento effettuato allo scopo di verificare se il fumo di pipa determini un significativo aumento della mortalità. Per questo un gruppo di non fumatori e uno di fumatori di pipa furono seguiti per sei anni e, al termine del periodo di osservazione, venne determinato il numero di soggetti deceduti in ciascuno dei due gruppi, con i seguenti risultati: tra i non fumatori, 950 soggetti erano ancora vivi, mentre 117 erano deceduti (riga 1). Tra i fumatori di pipa, 348 erano ancora vivi, mentre 54 erano deceduti (riga 2). Il test chi-quadrato evidenzia una differenza non significativa tra la mortalità nei due gruppi. Conclusione: fumare la pipa non fa male. L'esempio, famoso per le sue conclusioni, è tratto da Snedecor[104].
Una considerazione in merito al disegno sperimentale: l'osservazione per periodi di tempo più prolungati potrebbe portare a conclusioni differenti ?.

Un'ultima osservazione: la distribuzione chi-quadrato è un'approssimazione tanto più valida quanto più grandi sono le frequenze attese. Se anche una sola frequenza attesa risulta inferiore a 1, ovvero se più del 20% delle frequenze attese risulta inferiore a 5, i risultati del test chi-quadrato, e in particolare la sua significatività, devono essere valutati con estrema prudenza.

⇒ È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.

⇒ Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®.

5.3.6.9.6. Analisi della somiglianza

Questa opzione consente di effettuare l'analisi della somiglianza (cluster analysis) di un insieme di oggetti omogenei.

---

[103] *Ingelfinger JA, Mosteller FA, Thibodeau LA, Ware JH. Biostatistica in medicina. Milano: Raffaello Cortina, 1986:195-210.*

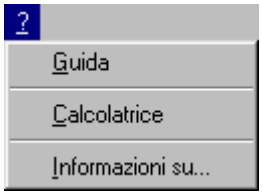
[104] *Snedecor GW, Cochran WG. Statistical methods. VII Edition. Ames: The Iowa State University Press, 1980:124.*

⇒ L'opzione risulta attiva qualsiasi sia il numero delle colonne (variabili) selezionate.
⇒ Le colonne devono contenere lo stesso numero di dati.

Il file CALCOLI.DBF contiene i dati di composizione di 10 formazioni calcolose delle vie urinarie: caricarlo nella tabella dati di Ministat con l'opzione Importa File (in formato dBase III®) del Menù File. Per ognuno dei calcoli è nota la composizione in calcio, fosfato, ossalato e magnesio. Selezionare le colonne CALCIO, FOSFATO, OSSALATO e MAGNESIO. La matrice dei cluster contiene, per ciascuno dei casi in esame, il/i cluster nei quali il caso confluisce. I cluster sono numerati in ordine progressivo di formazione, da sinistra verso destra. Ogni colonna della matrice dei cluster rappresenta in questo modo un livello crescente di aggregazione dei casi in base alla loro somiglianza. Per ogni colonna è riportata la distanza euclidea (espressa come deviata normale standardizzata z) alla quale si è formato l'ultimo cluster.
Il primo cluster, comprende i calcoli numero 6 e numero 1, che hanno una distanza euclidea di 0.94. Quindi i calcoli 1 e 6 sono i più simili tra loro in assoluto. Ma anche i calcolo 2 e 7 sono molto simili tra di loro, avendo una distanza euclidea di 0.98. Notare come all'ottavo passaggio si siano formati due cluster, il primo comprendente i calcoli 10, 5 4 e 9, il secondo comprendente i calcoli 6, 1, 2, 7, 8 e 3: perché questi due cluster confluiscano si arriva ad una distanza euclidea di 8.60. Quindi questi due gruppi di calcoli sembrano rappresentare due famiglie ben distinte per composizione.

Il metodo è tratto da Massart[105]. Per una discussione critica delle applicazioni della cluster analysis a problemi del laboratorio clinico vedere Vogt[106].
⇒ È possibile stampare i risultati dell'elaborazione statistica mediante l'opzione Stampa del Menù File.
⇒ Si ricorda che i risultati sono presentati in una finestra di testo dalla quale possono essere copiati o tagliati (opzioni Copia e Taglia del Menù Modifica) negli Appunti di Windows®. Nel caso di elaborazioni di oltre 40-50 dati, la finestra di testo potrebbe non riuscire a contenere completamente la matrice dei cluster: ciò si può evincere dal fatto che non compare il caratteristico messaggio di *fine* elaborazione. Tuttavia anche in questi casi, se l'analisi della somiglianza termina senza segnalazioni di errore, la stampa dei risultati riporterà la matrice dei cluster nella sua interezza.

5.3.7. Menù Aiuto

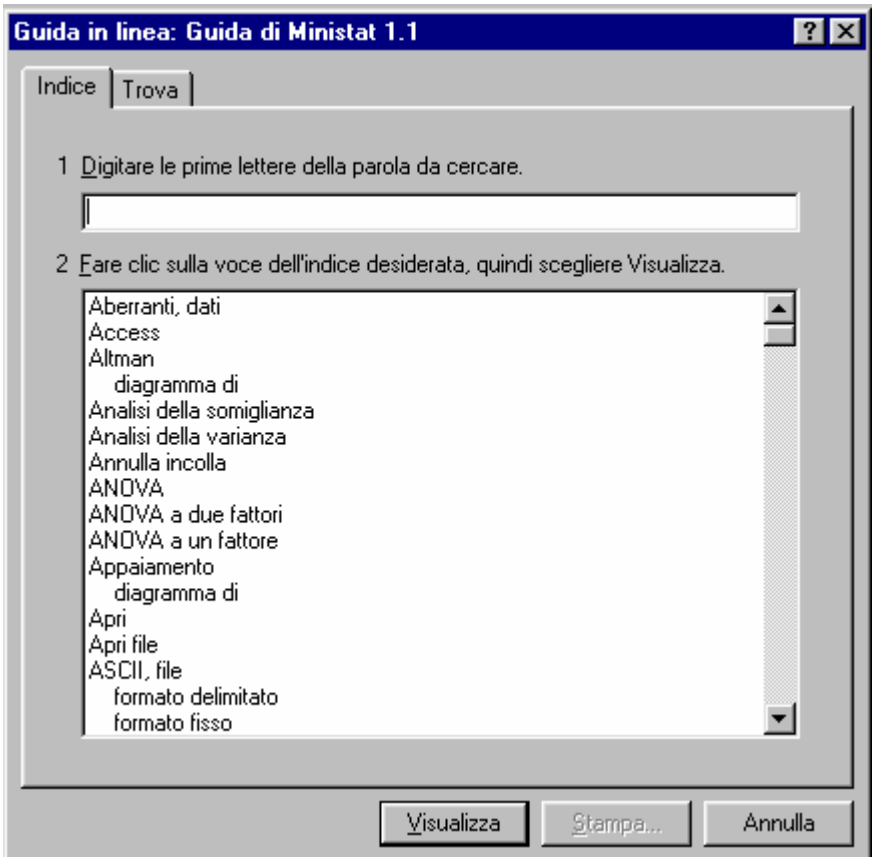


[105] *Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. Chemometrics: a textbook. Amsterdam: Elsevier Science Publishers BV, 1988:371-84.*
[106] *Vogt W, Nagel D. Cluster analysis in diagnosis. Clin Chem 1992;38:182-98.*

5.3.7.1. Guida

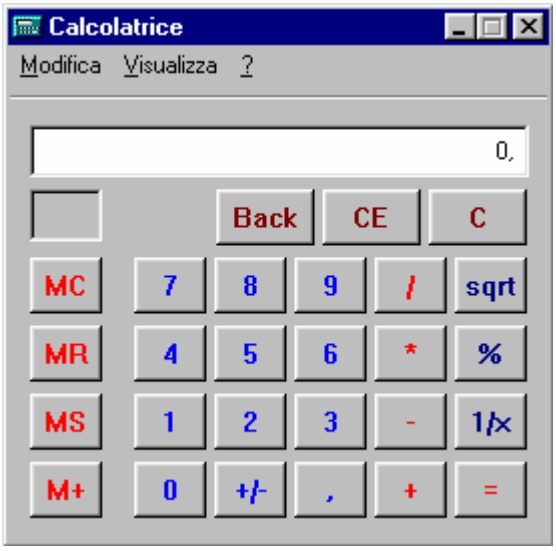
Contiene la versione elettronica on-line di questo manuale, che comprende in pratica tutto il contenuto del capitolo 5 del volume (sono quindi esclusi i quattro capitoli introduttivi e il sesto capitolo, riservato alle formule e agli algoritmi di calcolo). Si fa notare che nell'aggiornamento alla versione a 32 bit, come del resto specificato nella guida on-line, la nuova funzione che consente di tracciare diagrammi a torta è documentata solo nel testo. Il menù principale di Ministat riportato nella guida on-line e qui sotto, differisce da quello effettivo per la mancanza dell'icona corrispondente al diagramma a torta, e per una lieve differenza nella disposizione delle altre icone, che peraltro sono rimaste immodificate sia nell'aspetto che nelle funzioni.

Mediante l'opzione Cerca è possibile accedere a un Indice delle parole e delle espressioni per le quali è disponibile una ricerca rapida.



5.3.7.2. Calcolatrice

Consente di utilizzare la calcolatrice di Windows® per effettuare calcoli estemporanei.



Da notare che l'aspetto della calcolatrice dipende dalla versione di Windows® e dalle opzioni relative.

5.3.7.3. Informazioni su…

Consente di visualizzare una finestra contenente le informazioni sulla versione del programma.


Informazioni su Ministat
Ministat 1.1
Statistica per applicazioni biomediche
(C) 2000 - Versione 1.1.3216 (32 bit) by M. Besozzi
e-mail: marco.besozzi@auxologico.it
Prodotto nei Laboratori di Analisi dell'
ISTITUTO AUXOLOGICO ITALIANO
Istituto di Ricovero e Cura a Carattere Scientifico
Milano - Italia
Questo programma viene concesso in uso così-come-è alle condizioni specificate dall'ISTITUTO AUXOLOGICO ITALIANO, al quale nessun danno può essere imputato per un suo uso sia proprio che improprio.
OK

CAPITOLO 6
FORMULE E ALGORITMI DI CALCOLO

Sono riportati qui di seguito le formule e gli algoritmi più importanti, in quanto utili sia per comprendere a fondo il significato dei vari test statistici, sia per consentire ai lettori più pazienti e più intraprendenti di implementare qualcuna delle tecniche statistiche più semplici all'interno di propri applicativi, realizzabili con gli strumenti standard disponibili su PC (tipicamente i tabelloni elettronici, come Excel®).

Nell'interpretazione delle formule si tenga presente che, laddove non compaiono le parentesi a determinarle, le priorità degli operatori aritmetici sono le seguenti: prima devono essere eseguite le operazioni di sommatoria, poi quelle di elevamento a potenza e di radice, poi le moltiplicazioni e le divisioni, e infine le operazioni di somma e di sottrazione.

Per le voci bibliografiche dalle quali sono stati tratti formule e algoritmi, si rimanda alle sezioni precedenti.

6.1. Asimmetria e curtosi

Per un insieme comprendente un numero n di dati x_i, è possibile esprimere lo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica ricorrendo al coefficiente di asimmetria (g_1) e al coefficiente di curtosi (g_2), che indicano il tipo di scostamento dalla normalità come qui di seguito specificato:

$\Rightarrow$ $g_1 < 0$: asimmetria negativa, cioè coda sinistra della distribuzione eccessivamente lunga;
$\Rightarrow$ $g_1 > 0$: asimmetria positiva, cioè coda destra della distribuzione eccessivamente lunga;
$\Rightarrow$ $g_2 < 0$: platicurtosi, cioè distribuzione eccessivamente appiattita, con code troppo corte;
$\Rightarrow$ $g_2 > 0$: leptocurtosi, cioè distribuzione eccessivamente alta, con code troppo lunghe.

Essendo allora

$$m_2 = \Sigma (x_i - \bar{x})^2 / n$$
$$m_3 = \Sigma (x_i - \bar{x})^3 / n$$
$$m_4 = \Sigma (x_i - \bar{x})^4 / n$$

rispettivamente il momento di ordine secondo (m_2), il momento di ordine terzo (m_3) e il momento di ordine quarto (m_4) dalla media, i valori del coefficiente di asimmetria g_1 e del coefficiente di curtosi g_2 sono calcolati come

$$g_1 = m_3 / (m_2 \cdot \sqrt{m_2})$$
$$g_2 = m_4 / m_2^2 - 3$$

La deviazione standard s_1 del coefficiente di asimmetria e la deviazione standard s_2 del coefficiente di curtosi sono calcolate rispettivamente come

$$s_1 = \sqrt{(6 / n)}$$
$$s_2 = \sqrt{(24 / n)}$$

Il coefficiente di asimmetria g_1 viene considerato significativo se supera (in valore assoluto) di 2,6 volte la sua deviazione standard s_1, cioè se

$$|g_1 / s_1| > 2,6$$

In questo caso si rigetta l'ipotesi che l'asimmetria osservata sia compatibile con quella di una distribuzione gaussiana.

Il coefficiente di curtosi g_2 viene considerato significativo se supera (in valore assoluto) di 2,6 volte la sua deviazione standard s_2, cioè se

$$|g_2 / s_2| > 2,6$$

In questo caso si rigetta l'ipotesi che la curtosi osservata sia compatibile con quella di una distribuzione gaussiana.

Qualora si arrivi a concludere che la distribuzione campionaria osservata si discosta significativamente da quella della gaussiana teorica, sarà necessario ricorrere a tecniche statistiche che non basano la loro validità (quindi la validità delle conclusioni che mediante esse è possibile trarre) su assunti distribuzionali di gaussianità, e cioè a tecniche statistiche non-parametriche.

6.2. Test di Kolmogorov-Smirnov

In alternativa ai test di asimmetria e curtosi, è possibile verificare se la distribuzione osservata si discosta significativamente dalla distribuzione gaussiana teorica ricorrendo al test di Kolmogorov-Smirnov, che può essere calcolato procedendo come segue:

⇒ ordinare i dati in ordine numerico crescente;
⇒ calcolare la media $\bar{x}$ e la deviazione standard s dei dati;
⇒ per ciascun dato calcolare la deviata normale standardizzata (*DNS*) come

$$DNS = (x - \bar{x}) / s$$

⇒ calcolare un fattore di correzione (*FC*) come metà del rapporto fra la differenza minima misurabile per due campioni (*DMM*) e la deviazione standard calcolata (*s*), cioè

$$FC = 0,5 \cdot DMM / s$$

⇒ per ciascuno dei dati calcolare il valore della deviata normale standardizzata corretta (*DNSC*) come

$$DNSC = DNS + FC$$

⇒ per ciascuno dei dati calcolare il valore della funzione di distribuzione cumulativa normale (*FDCN*) corrispondente alla *DNSC* ;
⇒ per ciascuno dei dati calcolare il valore della distribuzione empirica (*FDE*) come rapporto fra il numero progressivo del dato (numero d'ordine nella lista ordinata dei dati) e il numero totale dei dati ;
⇒ per ciascuno dei dati calcolare il valore assoluto della differenza fra il valore della funzione di distribuzione cumulativa normale e quello della funzione di distribuzione empirica, cioè

$$|FDCN - FDE|$$

e chiamare *KS* la maggiore di tale differenze.

Essendo n il numero dei dati, un test sufficientemente approssimato per valutare la gaussianità di una distribuzione è basato sul calcolo dei valori critici del test (KS) ai livelli di probabilità del 5% e dell'1% rispettivamente come

$$KS_{0,05} = 0,886 / \sqrt{n}$$
$$KS_{0,01} = 1,031 / \sqrt{n}$$

Se il valore ottenuto supera quello previsto, al livello di probabilità prescelto, si conclude per un significativo scostamento della distribuzione dei dati trovata rispetto alla distribuzione gaussiana teorica .

Il test di Kolmogorov-Smirnov è qui riportato per completezza, non essendo incluso nell'attuale versione di Ministat. In realtà questo test fornisce (come atteso) gli stessi risultati dei test di asimmetria e di curtosi, con lo svantaggio di riassumere l'informazione in un'unica statistica, che non consente di separare la componente di asimmetria da quella di curtosi, e quindi non consente di identificare il contributo che ciascuna delle due fornisce alla non-gaussianità (se presente) della distribuzione osservata.

6.3. Statistiche elementari parametriche

Per una distribuzione gaussiana data, il valore μ (la *media della popolazione*) che rappresenta la misura di posizione della distribuzione gaussiana, e il valore σ (la *deviazione standard della popolazione*) che rappresenta la misura di dispersione della distribuzione gaussiana, possono essere stimati rispettivamente mediante la media campionaria $\bar{x}$ e la deviazione standard campionaria s.

Dato quindi un campione che include n dati (osservazioni) x_i, la media campionaria $\bar{x}$ e la deviazione standard campionaria s sono calcolate rispettivamente come

$$\bar{x} = \Sigma x_i / n$$
$$s = \sqrt{s^2}$$

essendo la varianza s^2 dei dati calcolata come

$$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$$

E' possibile semplificare il calcolo della varianza s^2 ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$

L'errore standard della media es viene calcolato come rapporto tra la deviazione standard s e la radice quadrata del numero n di osservazioni

$$es = s / \sqrt{n}$$

Il coefficiente di variazione CV viene calcolato come rapporto, espresso in percentuale, tra la deviazione standard s e la media $\bar{x}$, cioè come

$$CV = 100 \cdot (s / \bar{x})$$

Fatti uguale 100 il numero di dati di una distribuzione, il percentile è il valore che lascia alla sua sinistra la percentuale di dati desiderata (e corrispondentemente alla destra il complemento a 100). Così ad esempio il 2,5-esimo percentile è il valore che lascia alla sua sinistra il 2,5% dei dati osservati (e corrispondentemente lascia alla sua destra il 97,5% dei dati osservati), il 50-esimo percentile (che nel caso delle statistiche parametriche è la media campionaria $\bar{x}$) è il valore che lascia alla sua sinistra il 50% dei dati osservati (e corrispondentemente lascia alla sua destra il 50% dei dati osservati).
Per calcolare i percentili parametrici di una distribuzione ricorrere alla tabella della deviata normale standardizzata *z* riportata in tutti i testi di statistica. Alcuni percentili di frequente utilizzo sono riportati nella seguente tabella (ove $\bar{x}$ e *s* sono rispettivamente la media e la deviazione standard dei dati):

Percentile	Deviata normale standardizzata z	Valore del percentile
2,5	1,96	$\bar{x}$ - 1,96 *s*
5,0	1,645	$\bar{x}$ - 1,645 *s*
10	1,282	$\bar{x}$ - 1,282 *s*
20	0,842	$\bar{x}$ - 0,842 *s*
30	0,525	$\bar{x}$ - 0,525 *s*
40	0,253	$\bar{x}$ - 0,253 *s*
50	0	$\bar{x}$
60	0,253	$\bar{x}$ + 0,253 *s*
70	0,525	$\bar{x}$ + 0,525 *s*
80	0,842	$\bar{x}$ + 0,842 *s*
90	1,282	$\bar{x}$ + 1,282 *s*
95	1,645	$\bar{x}$ + 1,645 *s*
97,5	1,96	$\bar{x}$ + 1,96 *s*

6.4. Statistiche elementari non parametriche

Le statistiche elementari non parametriche comprendono in genere, oltre alla *mediana*, il *valore minimo* osservato, il *valore massimo* osservato, il *range* (cioè la differenza tra valore massimo e valore minimo) e spesso i *quartili*. In base al principio elementare per cui un segmento di retta può essere suddiviso in quattro parti mediante tre tacche equidistanti tra di loro e dagli estremi del segmento, i quartili sono tre. Il primo quartile è il valore che lascia alla sua sinistra il 25% dei dati (e che lascia alla sua destra il 75% dei dati). Il secondo quartile è il valore che lascia alla sua sinistra il 50% dei dati (e che lascia alla sua destra il 50% dei dati): il secondo quartile è quindi la mediana. Il terzo quartile è il valore che lascia alla sua sinistra il 75% dei dati (e che lascia alla sua destra il 25% dei dati).

Anche nel caso delle statistiche non parametriche, fatti uguale 100 il numero di dati di una distribuzione, il percentile è il valore che lascia alla sua sinistra la percentuale di dati desiderata (e corrispondentemente alla destra il complemento a 100). Così ad esempio il 2,5-esimo percentile è il valore che lascia alla sua sinistra il 2,5% dei dati osservati (e corrispondentemente lascia alla sua destra il 97,5% dei dati osservati), il 50-esimo percentile (che nel caso delle statistiche non parametriche è la mediana) è il valore che lascia alla sua sinistra il 50% dei dati osservati (e corrispondentemente lascia alla sua destra il 50% dei dati osservati). Si noti che il primo quartile corrisponde al 25-esimo percentile non parametrico, il secondo quartile corrisponde al 50-esimo

percentile non parametrico (cioè alla mediana), il terzo quartile corrisponde al 75-esimo percentile non parametrico.

6.4.1. Calcolo del percentile corrispondente a un dato

Per calcolare il percentile corrispondente a un dato, percentile che indica la posizione del dato all'interno della distribuzione osservata:

⇒ calcolare il rango del dato. Il rango è il numero di posizione dei singoli dati nella lista dei dati ordinati in ordine numerico crescente. Quando due dati sono uguali, a ciascuno viene assegnato come numero di posizione la media dei numeri di posizione dei dati che risultano uguali. Così, per esempio, se il quinto e il sesto dato, in una lista ordinata, sono tutti e due uguali, e pari a 223, il rango del quinto e del sesto dato sarà uguale, e pari a 5,5. Il procedimento viene esteso anche al caso in cui più di due dati siano uguali;

⇒ essendo n il numero dei dati, il percentile P (p-esimo) non parametrico, che indica la posizione del dato all'interno della distribuzione osservata, viene calcolato come

$$P = 100 \cdot rango / (n + 1)$$

6.4.2. Calcolo del dato corrispondente a un percentile

In un insieme di n dati, il P-esimo percentile non parametrico corrisponde al dato il cui numero C (numero progressivo, non necessariamente intero, nella serie dei dati ordinati in ordine numerico crescente) è

$$C = P \cdot (n + 1) / 100$$

Qualora C non sia un intero (accade quasi sempre) viene effettuata una interpolazione lineare fra il valore corrispondente all'intero che immediatamente precede C e il valore corrispondente all'intero che immediatamente segue C.

6.5. Rapporto tra varianze

Il test F applicato al rapporto tra due varianze campionarie, consente di verificare se le varianze di due campioni sono tra di loro omogenee (cioè sostanzialmente uguali).

Indicati con x_1 i dati del primo campione, con n_1 il loro numero e con $\bar{x}_1$ la loro media, ed indicati con x_2 i dati del secondo campione, con n_2 il loro numero e con $\bar{x}_2$ la loro media, la varianza del primo campione viene calcolato come

$$s_1^2 = \Sigma (x_1 - \bar{x}_1)^2 / (n_1 - 1)$$

con gradi di libertà pari a

$$n_1 - 1$$

la varianza del secondo campione viene calcolata come

$$s_2^2 = \Sigma (x_2 - \bar{x}_2)^2 / (n_2 - 1)$$

con gradi di libertà pari a

$$n_2 - 1$$

E' possibile semplificare il calcolo della varianza s_1^2 e della varianza s_2^2 ricordando che

$$\Sigma (x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - (\Sigma x_1)^2 / n_1$$
$$\Sigma (x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - (\Sigma x_2)^2 / n_2$$

Il valore del test *F* di omogeneità fra le varianze viene calcolato allora come rapporto fra la varianza maggiore e la varianza minore, cioè

$$F = s_1^2 / s_2^2 \qquad se \;\; s_1^2 > s_2^2$$

ovvero

$$F = s_2^2 / s_1^2 \qquad se \;\; s_2^2 > s_1^2$$

Il valore di *p* corrispondente alla statistica *F* rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude che le varianze sono significativamente diverse. Varianze diverse stanno ad indicare che i due campioni sono stati tratti da popolazioni diverse, aventi medie diverse. In questo senso il rapporto tra varianze corrisponde a un confronto tra medie.

6.6. Test t di Student per dati appaiati

Nel confronto tra medie mediante il test *t* di Student per dati appaiati la differenza fra le coppie di osservazioni diventa la variabile in esame. Dato un numero *n* di dati x_i , y_i sia d_i la differenza (presa con il segno) fra il valore del primo e il valore del secondo elemento della coppia, ovvero

$$d_i = x_i - y_i$$

Allora la differenza media $\bar{d}$ e la varianza s_d^2 delle differenza sono calcolate rispettivamente come

$$\bar{d} = \Sigma d_i / n$$
$$s_d^2 = \Sigma (d_i - \bar{d})^2 / (n - 1)$$

Si ricorda che anche in questo caso è possibile semplificare il calcolo della varianza s_d^2 ricordando che

$$\Sigma (d_i - \bar{d})^2 = \Sigma d_i^2 - (\Sigma d_i)^2 / n$$

Il valore del test t di Student è calcolato come rapporto fra la differenza media osservata e il suo errore standard, cioè

$$t = \bar{d} / \sqrt{(s_d^2 / n)}$$

con $n - 1$ gradi di libertà.

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza.

6.7. Test t di Student per campioni indipendenti

Il test t di Student per il confronto fra le medie di campioni indipendenti è basato su un principio semplice e intuitivo: a parità di valore assoluto della differenza fra le medie, a tale differenza viene data tanto maggior peso quanto minore è la dispersione dei dati campionari (cioè in definitiva quanto meno le due distribuzioni campionarie si sovrappongono), e viceversa.

Dati allora due campioni, il primo comprendente n_1 osservazioni x_1, aventi media $\bar{x}_1$, e il secondo comprendente n_2 osservazioni x_2, aventi media $\bar{x}_2$, il valore t viene calcolato come

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2 / n_1 + s^2 / n_2)}$$

con gradi di libertà pari a

$$n_1 + n_2 - 2$$

Al numeratore compare la differenza fra le medie, mentre al denominatore compare l'errore standard di tale differenza, la grandezza che, come si è precedentemente accennato, consente di dare il peso maggiore o minore alla differenza fra le medie.
La varianza s^2 è la varianza combinata dei due campioni, e viene calcolata come rapporto fra la somma delle devianza dei due campioni e la somma dei loro gradi di libertà, cioè come

$$s^2 = (\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2) / (n_1 + n_2 - 2)$$

E' possibile semplificare il calcolo della varianza s_1^2 e della varianza s_2^2 ricordando che

$$\Sigma (x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - (\Sigma x_1)^2 / n_1$$
$$\Sigma (x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - (\Sigma x_2)^2 / n_2$$

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

Un problema tuttavia complica lievemente la situazione: il test *t* di Student ordinario tende a fornire troppo pochi risultati significativi quando il campione più grande ha la varianza (cioè la dispersione dei dati) maggiore, e troppi risultati significativi quando il campione più grande ha la varianza minore. E' perciò necessario utilizzare, nei casi in cui la varianza dei due campioni differisca in maniera significativa, una forma del test che preveda una opportuna correzione. Per fare questo si calcola il rapporto tra varianze.

La varianza del primo campione viene calcolata come

$$s_1^2 = \Sigma (x_1 - \bar{x}_1)^2 / (n_1 - 1)$$

con gradi di libertà pari a

$$n_1 - 1$$

la varianza del secondo campione viene calcolata come

$$s_2^2 = \Sigma (x_2 - \bar{x}_2)^2 / (n_2 - 1)$$

con gradi di libertà pari a

$$n_2 - 1$$

Il valore del test *F* di omogeneità fra le varianze viene calcolato allora come rapporto fra la varianza maggiore e la varianza minore, cioè

$$F = s_1^2 / s_2^2 \qquad se \;\; s_1^2 > s_2^2$$

ovvero

$$F = s_2^2 / s_1^2 \qquad se \;\; s_2^2 > s_1^2$$

Il valore di *p* corrispondente alla statistica *F* rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze.

Nel caso di varianze significativamente diverse (varianze non omogenee) si calcola la statistica

$$t' = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s_1^2 / n_1 + s_2^2 / n_2)}$$

con gradi di libertà uguali a

$$(v_1 + v_2)^2 / (v_1^2 / (n_1 - 1) + v_2^2 / (n_2-1))$$

essendo rispettivamente

$$v_1 = s_1^2 / n_1$$
$$v_2 = s_2^2 / n_2$$

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

6.8. Test t di Student per una media teorica

Dato un campione che include *n* dati (osservazioni) x_i, la media campionaria $\bar{x}$ e varianza campionaria s^2 sono calcolate rispettivamente come

$$\bar{x} = \Sigma x_i / n$$
$$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$$

E' possibile semplificare il calcolo della varianza s^2 ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$

Essendo allora μ la media teorica attesa, il test *t* nella forma

$$t = (\bar{x} - \mu) / \sqrt{(s^2 / n)}$$

consente di verificare se la media osservata $\bar{x}$ differisce dalla media teorica μ.

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

Come è possibile vedere più avanti, nella parte dedicata al calcolo della regressione lineare, questa forma del test t di Student è molto importante in quanto consente di effettuare i test di significatività sui valori dell'intercetta *a* e del coefficiente angolare *b* dell'equazione della retta di regressione *x* variabile indipendente calcolata con il metodo dei minimi quadrati.

6.9. Test di Wilcoxon per dati appaiati

Il test di Wilcoxon per dati appaiati è l'equivalente non parametrico del test t di Student per dati appaiati, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano.

Per i calcoli si procede in questo modo :

⇒ determinare le differenze (con il segno) fra le *n* coppie di valori;

⇒ stabilire, per ciascuna differenza, il numero di posizione nella lista delle differenze ordinate (questa volta ignorando il segno) in ordine numerico crescente: la più piccola differenza osservata avrà numero di posizione 1, e via dicendo;

⇒ quando due o più differenze sono uguali, assegnare a ciascuna di esse la media dei numeri di posizione che esse dovrebbero avere; così, per esempio, se la quinta e la sesta differenza sono uguali, assegnare come numero di posizione nella lista il valore 5.5 a entrambe;

⇒ riassegnare il segno ai numeri di posizione nella lista (se la differenza avente quel numero di posizione era negativa, assegnare il segno meno, se era positiva assegnare il segno più);

⇒ calcolare il totale per i numeri di posizione con segno negativo e per i numeri di posizione con segno positivo, e chiamare *T* il più piccolo di questi due totali;

⇒ calcolare la deviata normale standardizzata *Z* come

$$Z = (\mu - T - 0{,}5) / s$$

essendo

$$\mu = n \cdot (n + 1) / 4$$
$$s = \sqrt{((2n + 1) \cdot \mu / 6)}$$

La statistica *Z* così calcolata corrisponde a sottoporre al test la mediana delle differenze.

Il valore di *p* corrispondente alla statistica *Z* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le mediane. Questa soluzione è sufficientemente accurata per n > 16.

6.10. Test di Wilcoxon per campioni indipendenti

Sebbene spesso chiamato test di Mann-Whitney, l'equivalente non parametrico del test t di Student per dati appaiati è dovuto anch'esso a Wilcoxon. Rappresenta l'equivalente non parametrico del test t di Student per campioni indipendenti, e va utilizzato in luogo di questo quando i dati non siano distribuiti in modo gaussiano.

Per i calcoli si procede in questo modo:

⇒ mettere i dati dei due campioni in una singola lista, badando ad etichettarli in modo che poi possano essere successivamente di nuovo distinti;

⇒ stabilire, per ciascun dato, il numero di posizione nella lista ordinata in ordine numerico crescente: il dato più piccolo avrà numero di posizione 1, e via dicendo;

⇒ quando due o più dati sono uguali, assegnare a ciascuno di essi la media dei numeri di posizione che esse dovrebbero avere; così, per esempio, se i dati dal primo al sesto sono uguali, assegnare come numero di posizione nella lista a tutti il valore 3,5 (media dei numeri da 1 a 6);

⇒ se i campioni hanno lo stesso numero di dati, calcolare il totale per i numeri di posizione del primo e per i numeri di posizione del secondo campione, e chiamare *T* il più piccolo di questi due totali;

⇒ se i due campioni hanno diverso numero di dati chiamare T_1 il totale per il campione che ha il numero minore di dati, diciamo n_1 , e quindi, essendo n_2 il numero di dati del secondo campione, calcolare

$$T_2 = n_1 \cdot (n_1 + n_2 + 1) - T_1$$

⇒ chiamare T il più piccolo fra i valori T_1 e T_2

⇒ calcolare la deviata normale standardizzata Z come

$$Z = (| \mu - T | - 0{,}5) / s$$

essendo

$$\mu = n_1 \cdot (n_1 + n_2 + 1) / 2$$
$$s = \sqrt{(n_2 \cdot \mu / 6)}$$

La statistica Z così calcolata corrisponde a sottoporre al test la mediana delle differenze.

Il valore di p corrispondente alla statistica Z rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza mediana osservata. Questa soluzione è sufficientemente accurata per n > 16.

6.11. Analisi della varianza a un fattore

Con le varie forme del test t di Student e del test di Wilcoxon è possibile confrontare fra di loro due medie: l'analisi della varianza o ANOVA (da ANalysis Of VAriance) consente di estendere il confronto a più di due medie.

Sia i (con $i = 1,2,3,....r$) un generico campione, sia j (con $j = 1,2,3,....n$) un generico replicato, e quindi $x_{i,j}$ un generico valore della tabella

Campione	Replicato					Media
	j=1	j=2	j=3	. . . . .	j=n	$\bar{x}_i$
i=1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	. . . . .	$x_{1,n}$	$\bar{x}_1$
i=2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	. . . . .	$x_{2,n}$	$\bar{x}_2$
. . . .	. . . .	. . . .	. . . .	. . . . .	. . . .	. . . .
i=r	$x_{r,1}$	$x_{r,2}$	$x_{r,3}$	. . . . .	$x_{r,n}$	$\bar{x}_r$

nella quale le $r \cdot n$ osservazioni corrispondenti a r campioni, per ciascuno dei quali sono disponibili n dati, sono riportate in modo ordinato.

Siano ancora $\bar{x}_i$ la media di un generico campione, e $\bar{x}_g$ la media generale di tutte le $r \cdot n$ osservazioni. Allora la variabilità totale (S_t) osservata viene calcolata come

$$S_t = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (x_{i,j} - \bar{x}_g)^2$$

con $r \cdot n - 1$ gradi di libertà .

La variabilità S_s spiegata dalle differenza fra le medie $\bar{x}_i$ delle righe viene calcolata come

$$S_s = n \cdot \sum_{i=1}^{i=r} (\bar{x}_i - \bar{x}_g)^2$$

con $r - 1$ gradi di libertà, mentre la variabilità casuale, non spiegata, detta anche "residua" (S_n), viene calcolata come

$$S_n = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (x_{i,j} - \bar{x}_i)^2$$

con $r \cdot (n - 1)$ gradi di libertà, tenendo presente che, per semplicità, essa può essere calcolata anche per differenza come

$$S_n = S_t - S_s$$

La varianza spiegata (V_s) e la varianza non spiegata (V_n), calcolate rispettivamente come

$$V_s = S_s / (r - 1)$$
$$V_n = S_n / (r \cdot (n - 1))$$

vengono allora impiegate per calcolare finalmente il rapporto fra varianze F

$$F = V_s / V_n$$

con $r - 1$ gradi di libertà al numeratore e $r \cdot (n - 1)$ gradi di libertà al denominatore.

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_i$ delle r righe.

6.12. Analisi della varianza a due fattori

Se l'analisi della varianza a 1 fattore (detta anche "a un criterio di classificazione") consente di verificare se vi siano(in media) differenze significative fra gli elementi appartenenti alle righe della tabella in cui sono stati ordinatamente raccolte le nostre osservazioni, l'ANOVA a 2 fattori consente di verificare contemporaneamente se vi siano (in media) differenze significative fra gli elementi appartenenti alle colonne della tabella, che si presenta così

Riga	Colonna					Media
	j=1	j=2	j=3	. . . . .	j=c	$\bar{x}_{i.}$
i=1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	. . . . .	$x_{1,c}$	$\bar{x}_{1.}$
i=2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	. . . . .	$x_{2,c}$	$\bar{x}_{2.}$
. . . .	. . . .	. . . .	. . . .	. . . . .	. . . .	. . . .
i=r	$x_{r,1}$	$x_{r,2}$	$x_{r,3}$	. . . . .	$x_{r,c}$	$\bar{x}_{r.}$
Media $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.3}$	. . . . .	$\bar{x}_{.c}$	

Ciascuno dei dati $x_{i,j}$ risulta pertanto assegnato in base a una caratteristica a una delle i ($i = 1,2,3,....r$) righe e per un'altra caratteristica a una delle j ($j = 1,2,3,....c$) colonne.

Siano allora $\bar{x}_{i.}$ la media di una generica riga, $\bar{x}_{.j}$ la media di una generica colonna, e $\bar{x}_g$ la media generale degli $r \cdot c$ dati. È facile notare che, rispetto alla tabella dell'analisi della varianza a 1 fattore, la novità consiste nell'avere introdotto un elemento di classificazione a livello delle colonne, e quindi le corrispondenti medie $\bar{x}_{.j}$.

La variabilità totale (S_t) osservata viene calcolata come

$$S_t = \sum_{i=1}^{i=r} \sum_{j=1}^{j=c} (x_{i,j} - \bar{x}_g)^2$$

con $r \cdot c - 1$ gradi di libertà.

Tuttavia, essendo stato introdotto un nuovo criterio di classificazione dei dati, essa verrà scomposta non più in due, bensì in tre componenti. La prima di esse, la variabilità S_r spiegata dalle differenze fra le medie $\bar{x}_{i.}$, cioè dalle differenze fra le medie delle righe, viene calcolata come

$$S_r = c \cdot \sum_{i=1}^{i=r} (\bar{x}_{i.} - \bar{x}_g)^2$$

con $r - 1$ gradi di libertà.

La variabilità S_c spiegata dalle differenze fra le medie $\bar{x}_{.j}$, cioè dalle differenze fra le medie delle colonne, viene calcolata come

$$S_c = r \cdot \sum_{j=1}^{j=c} (\bar{x}_{.j} - \bar{x}_g)^2$$

con $c - 1$ gradi di libertà.

Infine la variabilità casuale, non spiegata (S_n), detta anche "residua", viene calcolata come

$$S_n = \sum_{i=1}^{i=r} \sum_{j=1}^{j=c} (x_{i,j} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_g)^2$$

con $(r - 1) \cdot (c - 1)$ gradi di libertà, tenendo presente che può più semplicemente essere calcolata per differenza come

$$S_n = S_t - S_r - S_c$$

La varianza spiegata dalle differenze fra le medie $\bar{x}_{i.}$ delle righe (V_r), quella spiegata dalle differenze fra le medie $x_{.j}$ delle colonne (V_c) e la varianza non spiegata (V_n) sono allora calcolate rispettivamente come

$$V_r = S_r / (r - 1)$$
$$V_c = S_c / (c - 1)$$
$$V_n = S_n / ((r - 1) \cdot (c - 1))$$

Il rapporto fra varianze F

$$F = V_r / V_n$$

con $r - 1$ gradi di libertà al numeratore e $(r - 1) \cdot (c - 1)$ gradi di libertà al denominatore viene allora impiegato per verificare l'esistenza di una differenza significativa fra le medie delle righe ($\bar{x}_{i.}$).

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_i$ delle r righe.

Il rapporto fra varianze

$$F = V_c / V_n$$

con $c - 1$ gradi di libertà al numeratore e $(r - 1) \cdot (c - 1)$ gradi di libertà al denominatore viene impiegato per verificare l'esistenza di una differenza significativa fra le medie delle colonne ($\bar{x}_{.j}$).

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_{.j}$ delle c colonne.

6.13. Regressione lineare parametrica

Per adattare la retta ai dati sperimentali viene impiegato il metodo dei minimi quadrati, una tecnica di approssimazione ben nota, che consente di minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta.

6.13.1. Regressione lineare x variabile indipendente

Il modello matematico impiegato presuppone che la x, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la y sia distribuito normalmente e con una varianza che non cambia al variare del valore della x.

Sia allora n il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione x variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{yx} = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma (x_i - \bar{x})2$$

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2)}$$

E' possibile semplificare i calcoli ricordando che

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

La varianza residua attorno alla regressione viene calcolata come

$$s_0^2 = (\Sigma(y_i - \bar{y})^2 - s_1^2) / (n - 2)$$

essendo

$$s_1^2 = (\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}))^2 / \Sigma(x_i - \bar{x})^2$$

Infine l'errore standard della stima s_{yx} e le deviazioni standard del coefficiente angolare (s_b) e dell'intercetta (s_a), che forniscono una misura rispettivamente della dispersione dei dati attorno alla retta calcolata, e del grado di incertezza connesso con i valori ottenuti di a_{yx} e di b_{yx}, sono calcolati come

$$s_{yx} = \sqrt{(s_0^2)}$$
$$s_b = s_{yx} \cdot \sqrt{(1 / \Sigma(x_i - \bar{x})^2)}$$
$$s_a = s_b \cdot \sqrt{(\Sigma x_i^2 / n)}$$

Si consideri che la retta di regressione campionaria

$$y = a_{yx} + b_{yx} \cdot x$$

rappresenta la migliore stima possibile della retta di regressione della popolazione

$$y = \alpha + \beta \cdot x$$

Si consideri che il test t di Student per una media teorica nella forma già vista

$$t = (\bar{x} - \mu) / \sqrt{(s^2 / n)}$$

può essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = a_{yx}$$
$$\mu = \alpha$$
$$\sqrt{(s^2 / n)} = s_a$$

assumendo quindi la forma

$$t = (a_{yx} - \alpha) / s_a$$

Questo consente di sottoporre a test la differenza dell'intercetta *a* rispetto a un valore atteso (per esempio rispetto a 0, cioè all'intercetta di una retta passante per l'origine). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza dell'intercetta rispetto al valore atteso.

Si consideri che il test t di Student per una media teorica può anche essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = b_{yx}$$
$$\mu = \beta$$
$$\sqrt{(s^2 / n)} = s_b$$

assumendo quindi la forma

$$t = (b_{yx} - \beta) / s_b$$

Questo consente di sottoporre a test la differenza del coefficiente angolare *b* rispetto a un valore atteso (per esempio rispetto a 0, cioè al coefficiente angolare di una retta orizzontale, oppure rispetto a 1, cioè al coefficiente angolare di una retta a 45 gradi). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza del coefficiente angolare rispetto al valore atteso.

6.13.2. Regressione lineare y variabile indipendente

Il modello matematico impiegato presuppone che la *y*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *x* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *y*. Si noti che in questo caso inizialmente la *y* (variabile indipendente) viene posta in ascisse e la *x* (variabile dipendente) viene posta in ordinate.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{xy} = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma(y_i - \bar{y})^2$$

$$a_{xy} = \bar{x} - b_{xy} \cdot \bar{y}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico a quello calcolato mediante la regressione *x* variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione *x* variabile indipendente, si esplicita l'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} + 1 / b_{xy} \cdot x$$

Quindi l'intercetta *a* e il coefficiente angolare *b* che consentono di rappresentare la regressione *y* variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione *x* variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

6.13.3. Componente principale standardizzata

Il modello matematico impiegato presuppone tanto la *x* quanto la *y* siano affette da un errore di misura equivalente.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i, sia b_{yx} il coefficiente angolare dell'equazione della retta di regressione *x* variabile indipendente, e sia b_{xy} il coefficiente angolare dell'equazione della retta di regressione *y* variabile indipendente.

Il coefficiente angolare b_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è allora uguale a

$$b_{cps} = \sqrt{(b_{yx} \cdot b_{xy})}$$

cioè alla media geometrica tra il coefficiente angolare b_{yx} della regressione *x* variabile indipendente e il coefficiente angolare b_{xy} della regressione *y* variabile indipendente, cioè

$$b_{xy} = \sqrt{(\Sigma(y_i - \bar{y})^2 / \Sigma(x_i - \bar{x})^2)}$$

mentre l'intercetta a_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è uguale a

$$a_{cps} = \bar{y} - b_{cps} \cdot \bar{x}$$

Infine il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico sia a quello calcolato mediante la regressione *x* variabile indipendente sia a quello calcolato mediante la regressione *y* variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

6.14. Regressione lineare non parametrica

Per adattare la retta ai dati sperimentali viene impiegato un metodo che non fa ricorso ad assunti preliminari riguardo la distribuzione dei dati e degli errori ad essi associati.

6.14.1. Regressione lineare x variabile indipendente

Essendo *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (y_i - y_j) / (x_i - x_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Allora il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione *x* variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima i dati, vengono calcolati nel seguente modo:

⇒ si calcolano i coefficienti angolari delle $n \cdot (n - 1) / 2$ rette che passano per tutte le coppie possibili di punti;

⇒ si ordinano gli $n \cdot (n - 1) / 2$ coefficienti angolari così calcolati in ordine numerico crescente;

⇒ si calcola il coefficiente angolare b_{yx} dell'equazione della retta di regressione come mediana degli $N = n \cdot (n - 1) / 2$ valori di cui al punto precedente, cioè come

$$b_{yx} = b_{((N + 1)/2)} \quad \text{se } N \text{ è dispari}$$
$$b_{yx} = (b_{(N/2)} + b_{(N/2+1)}) / 2 \quad \text{se } N \text{ è pari}$$

⇒ si calcolano allora gli *n* valori possibili per l'intercetta *a* come

$$a_i = y_i - b_{yx} \cdot x_i$$

⇒ si ordinano gli *n* valori di intercetta così calcolati in ordine numerico crescente;

⇒ si calcola l'intercetta a_{yx} dell'equazione della retta di regressione come mediana dei valori di cui al punto precedente, cioè come

$$a_{yx} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{yx} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

6.14.2. Regressione lineare y variabile indipendente

Si noti che in questo caso inizialmente la *y* (variabile indipendente) viene posta in ascisse e la *x* (variabile dipendente) viene posta in ordinate.

Essendo allora *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (x_i - x_j) / (y_i - y_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Allora il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima i dati vengono calcolati nel seguente modo:

⇒ si calcolano i coefficienti angolari delle $n \cdot (n - 1) / 2$ rette che passano per tutte le coppie possibili di punti;

⇒ si ordinano gli $n \cdot (n - 1) / 2$ coefficienti angolari così calcolati in ordine numerico crescente;

⇒ si calcola il coefficiente angolare b_{xy} dell'equazione della retta di regressione come mediana degli $N = n \cdot (n - 1) / 2$ valori di cui al punto precedente, cioè come

$$b_{xy} = b_{((N + 1)/2)} \quad \text{se } N \text{ è dispari}$$
$$b_{xy} = (b_{(N/2)} + b_{(N/2+1)}) / 2 \quad \text{se } N \text{ è pari}$$

⇒ si calcolano allora gli *n* valori possibili per l'intercetta *a* come

$$a_i = x_i - b_{xy} \cdot y_i$$

⇒ si ordinano gli *n* valori di intercetta così calcolati in ordine numerico crescente;

⇒ si calcola l'intercetta a_{xy} dell'equazione della retta di regressione come mediana dei valori di cui al punto precedente, cioè come

$$a_{xy} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{xy} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione *x* variabile indipendente, si esplicita l'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} + 1 / b_{xy} \cdot y$$

Quindi l'intercetta *a* e il coefficiente angolare *b* che consentono di rappresentare la regressione *y* variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione *x* variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

6.14.3. Regressione lineare di Passing e Bablok

Essendo *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (y_i - y_j) / (x_i - x_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Nel caso in cui sia

$$x_i = x_j \text{ e } y_i = y_j$$

il valore del coefficiente angolare non è definito, e quindi viene scartato dai calcoli successivi; ugualmente vengono scartati tutti i valori del coefficiente angolare uguali a -1.

I rimanenti *N* valori vengono allora ordinati in ordine numerico crescente; essendo *K* il numero dei valori del coefficiente angolare inferiori a -1, il coefficiente angolare b_{pb} della retta di regressione calcolata con il metodo di Passing e Bablok

$$y = a_{pb} + b_{pb} \cdot x$$

viene calcolato come

$$b_{pb} = b_{((N + 1)/2 + K)} \quad \text{se } N \text{ è dispari}$$
$$b_{pb} = (b_{(N/2+K)} + b_{(N/2+1+K)}) / 2 \quad \text{se } N \text{ è pari}$$

mentre l'intercetta a_{pb} viene calcolata come mediana degli *n* valori

$$a_i = y_i + b_{pb} \cdot x_i$$

e cioè, nella lista dei valori di *a* così calcolati e ordinati in ordine numerico crescente, come

$$a_{pb} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{pb} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

6.15. Regressione polinomiale di secondo grado

Il metodo dei minimi quadrati consente di adattare un funzione $y = f(x)$ ad una serie di dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa.

In questo caso il metodo dei minimi quadrati viene impiegato per adattare ai dati un polinomio di secondo grado, nella forma

$$y = a + b \cdot x + c \cdot x^2$$

Essendo n il numero dei dati aventi coordinate (x_i,y_i), il termina noto a, e i coefficienti b e c del polinomio di secondo grado possono essere determinati risolvendo contemporaneamente le equazioni

$$a \cdot n + b \cdot \Sigma x_i + c \cdot \Sigma x_i^2 = \Sigma y_i$$
$$a \cdot \Sigma x_i + b \cdot \Sigma x_i^2 + c \cdot \Sigma x_i^3 = \Sigma\, x_i \cdot y_i$$
$$a \cdot \Sigma x_i^2 + b \cdot \Sigma x_i^3 + c \cdot \Sigma x_i^4 = \Sigma\, x_i^2 \cdot y_i$$

Il sistema può essere facilmente risolto, utilizzando qualsiasi programma che includa la soluzione dei sistemi di equazioni in forma matriciale (per esempio Excel®).

6.16. Regressione polinomiale di terzo grado

Il metodo dei minimi quadrati consente di adattare una funzione $y = f(x)$ ad una serie di dati sperimentali in modo che risulti minimizzata la somma dei quadrati delle differenze residue fra i dati sperimentali e la funzione stessa.

In questo caso il metodo dei minimi quadrati viene impiegato per adattare ai dati un polinomio di secondo grado, nella forma

$$y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$

Essendo n il numero dei dati aventi coordinate (x_i,y_i), il termina noto a, e i coefficienti b, c e d del polinomio di terzo grado possono essere determinati risolvendo contemporaneamente le equazioni

$$a \cdot n + b \cdot \Sigma x_i + c \cdot \Sigma x_i^2 + d \cdot \Sigma x_i^3 = \Sigma y_i$$
$$a \cdot \Sigma x_i + b \cdot \Sigma x_i^2 + c \cdot \Sigma x_i^{3\,+}\, d \cdot \Sigma x_i^4 = \Sigma x_i \cdot y_i$$
$$a \cdot \Sigma x_i^2 + b \cdot \Sigma x_i^3 + c \cdot \Sigma x_i^{4\,+}\, d \cdot \Sigma x_i^5 = \Sigma x_i^2 \cdot y_i$$
$$a \cdot \Sigma x_i^3 + b \cdot \Sigma x_i^4 + c \cdot \Sigma x_i^{5\,+}\, d \cdot \Sigma x_i^6 = \Sigma x_i^3 \cdot y_i$$

Il sistema può essere facilmente risolto, utilizzando qualsiasi programma che includa la soluzione dei sistemi di equazioni in forma matriciale (per esempio Excel®).

6.17. Test chi-quadrato per tabelle di contingenza

Per applicare il test chi-quadrato (χ^2) nella forma generalizzata per tabelle di contingenza, che è quella qui impiegata, è necessario che ciascun elemento del campione in esame possa essere classificato per una caratteristica in un numero R di classi, e per una seconda caratteristica in un numero C di classi, in modo tale che i dati possano essere organizzati in una tabella di R righe per C colonne, comprendente quindi un totale di $N = R \cdot C$ celle.

Essendo allora f la frequenza osservata in una data cella e F la frequenza attesa per la stessa cella, il test χ^2 viene calcolato come somma dei rapporti $(F - f)^2 / F$ per tutte le celle della tabella, cioè come

$$\chi^2 = \Sigma (f - F)^2 / F$$

Il valore di *f* per ciascuna cella è noto, essendo come detto *f* la frequenza osservata. Il valore di *F*, cioè la frequenza attesa, non è noto, ma può essere specificato qualora si operi alla luce di una ben definita ipotesi riguardante i dati. Nel caso particolare dell'ipotesi "non vi è differenza fra le frequenze osservate", cioè di quella che gli statistici chiamano la "ipotesi nulla", e che viene qui impiegata, le frequenze attese *F* possono essere stimate, e assumono ciascuna un valore pari al prodotto del totale della riga per il totale della colonna cui la cella appartiene diviso per il totale *n* dei casi osservato, ovvero

$$F = (\text{totale della riga}) \cdot (\text{totale della colonna}) / n$$

Per illustrare le modalità di sviluppo dei calcoli si consideri il caso più semplice, quello della seguente tabella 2 x 2

$$\begin{matrix} f_1 & f_2 \\ f_3 & f_4 \end{matrix}$$

in cui le frequenze osservate per ciascuna delle quattro celle sono indicate rispettivamente con f_1, f_2, f_3 e f_4.

Indicando allora:

$\Rightarrow$ il totale della riga 1 con R_1 ($R_1 = f_1 + f_2$);
$\Rightarrow$ il totale della riga 2 con R_2 ($R_2 = f_3 + f_4$);
$\Rightarrow$ il totale della colonna 1 con C_1 ($C_1 = f_1 + f_3$);
$\Rightarrow$ il totale della colonna 2 con C_2 ($C_2 = f_2 + f_4$);
$\Rightarrow$ il totale dei casi con n ($n = f_1 + f_2 + f_3 + f_4$);

i quattro valori di *F* attesi

$$\begin{matrix} F_1 & F_2 \\ F_3 & F_4 \end{matrix}$$

corrispondenti ai valori di *f* osservati saranno per definizione uguali rispettivamente a

$$\begin{matrix} F_1 = C_1 \cdot R_1 / n & \qquad F_2 = C_2 \cdot R_1 / n \\ F_3 = C_1 \cdot R_2 / n & \qquad F_4 = C_2 \cdot R_2 / n \end{matrix}$$

essendo ancora ovviamente $F_1 + F_2 + F_3 + F_4 = n$, mentre il valore di χ^2 sarà pari a

$$\chi^2 = (f_1 - F_1)^2 / F_1 + (f_2 - F_2)^2 / F_2 + (f_3 - F_3)^2 / F_3 + (f_4 - F_4)^2 / F_4$$

con $(R - 1) \cdot (C - 1)$ gradi di libertà.

E' quindi facile estendere i calcoli dalla tabella 2 x 2 a tabelle di qualsiasi estensione.

Il valore di *p* corrispondente alla statistica χ^2 rappresenta la probabilità di osservare per caso una differenza tra frequenze osservate e frequenze attese della grandezza di quella effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una differenza significativa di incidenza nei diversi gruppi del fattore in esame.

6.18. Analisi della somiglianza (cluster analysis)

I problemi di classificazione rivestono un ruolo centrale nella scienza. E classificare gli oggetti in base a criteri oggettivi, basati su quantità misurabili, è lo scopo fondamentale della cluster analysis.

Si consideri il seguente esempio, relativo a 10 calcoli delle vie urinarie, analizzati per la loro composizione in calcio, fosfato, ossalato e magnesio (dati simulati e valori espressi in unità di misura arbitrarie).

CALCOLO	CALCIO	FOSFATO	OSSALATO	MAGNESIO
1	99	81	69	61
2	78	65	53	43
3	81	66	38	54
4	45	23	19	16
5	44	18	24	19
6	102	83	72	66
7	83	68	49	45
8	74	71	41	57
9	38	19	22	14
10	48	14	21	12

All'inizio del processo di classificazione (clustering) ad ogni oggetto corrisponde un cluster (e viceversa). In questo stadio tutti gli oggetti sono considerati dissimili (diversi) tra di loro. Al passaggio successivo i due oggetti più simili sono raggruppati in un unico cluster. Il numero dei cluster risulta quindi pari al numero di oggetti - 1. Il procedimento viene ripetuto ciclicamente, fino ad ottenere (all'ultimo passaggio) un unico cluster.

Stabilire a quale livello di aggregazione degli oggetti fermarsi, e quindi quali conclusioni trarre, dipende esclusivamente dai giudizio di merito dell'utilizzatore. Per questo la cluster analysis riveste un ruolo centrale, in statistica, limitatamente alla analisi esplorativa dei dati (in effetti il livello a cui fermarsi trarre le conclusioni a questo punto non è più quantitativo, e quindi non è più oggettivo).

La prima cosa che si fa nell'analisi della somiglianza è quella di calcolare le distanze euclidee. La distanza può essere calcolata in vari modi: quello qui seguito prevede di calcolare la distanza tra tutte le possibili coppie di punti. Nel caso di due variabili (come per esempio sarebbe stato se si fossero avute, per ciascun calcolo, le sole misure di calcio e fosfato) mediante il teorema di Pitagora. Che peraltro può essere esteso dal piano cartesiano, bidimensionale, ad uno spazio tridimensionale (nel caso di tre misure sullo stesso campione) e a uno spazio n-dimensionale (nel caso di n misure sullo stesso campione).

Nel caso dell'esempio illustrato la matrice delle distanze è la seguente:

Caso	1	2	3	4	5	6	7	8	9	10
1	0,00	4,72	5,56	13,50	13,21	0,94	4,51	5,44	14,05	13,87
2	4,72	0,00	2,79	8,85	8,60	5,63	0,98	2,81	9,39	9,28
3	5,56	2,79	0,00	8,85	8,74	6,32	2,12	1,20	9,58	9,44
4	13,50	8,85	8,85	0,00	1,03	14,38	9,12	9,14	1,12	1,21
5	13,21	8,60	8,74	1,03	0,00	14,07	8,95	8,99	1,08	1,27
6	0,94	5,63	6,32	14,38	14,07	0,00	5,41	6,17	14,92	14,75
7	4,51	0,98	2,12	9,12	8,95	5,41	0,00	2,39	9,75	9,58
8	5,44	2,81	1,20	9,14	8,99	6,17	2,39	0,00	9,79	9,81
9	14,05	9,39	9,58	1,12	1,08	14,92	9,75	9,79	0,00	1,41
10	13,87	9,28	9,44	1,21	1,27	14,75	9,58	9,81	1,41	0,00

La matrice delle distanze è formata da un numero di righe e di colonne uguale, e pari al numero di casi in esame (i casi sono numerati in ordine progressivo: il caso 1 corrisponde alla prima riga di dati, il caso 2 alla seconda, e così via). La matrice contiene le distanze tra tutte le possibili coppie di casi. Notare come essa sia simmetrica rispetto alla diagonale. Le distanze sono espresse come deviata normale standardizzata (*z*). Si noti che la matrice è simmetrica, e che la diagonale assume valori uguali a zero (la distanza euclidea tra un calcolo e sé stesso è ovviamente nulla).

La corrispondente matrice dei cluster appare così

Caso	Cluster								
10	0	0	0	0	0	6	6	6	9
5	0	0	3	4	4	6	6	6	9
4	0	0	3	4	4	6	6	6	9
9	0	0	0	4	4	6	6	6	9
6	1	1	1	1	1	1	1	8	9
1	1	1	1	1	1	1	1	8	9
2	0	2	2	2	2	2	7	8	9
7	0	2	2	2	2	2	7	8	9
8	0	0	0	0	5	5	7	8	9
3	0	0	0	0	5	5	7	8	9
	0.94	0.98	1.03	1.08	1.20	1.21	2.12	4.51	8.60
	Distanza euclidea								

La matrice dei cluster contiene, per ciascuno dei casi in esame (righe) il/i cluster nei quali il caso confluisce. I cluster sono numerati in ordine progressivo di formazione, da sinistra verso destra. Ogni colonna rappresenta in questo modo un livello crescente di aggregazione dei casi in base alla loro somiglianza. Per ogni colonna è riportata la distanza (standardizzata) alla quale si è formato l'ultimo cluster.
Il primo cluster, comprende gli oggetti numero 6 e numero 1, che hanno una distanza euclidea di 0.94 (controllare anche la matrice delle distanze). Quindi gli oggetti 1 e 6 sono i più simili tra loro in assoluto. Ma anche gli oggetti2 e 7 sono molto simili tra di loro, avendo una distanza euclidea di 0.98. Notare come all'ottavo passaggio si siano formati due cluster, il primo comprendente gli oggetti 10, 5 4 e 9, il secondo comprendente gli oggetti 6, 1, 2, 7, 8 e 3: perché questi due cluster confluiscano si arriva ad una distanza euclidea di 8.60. Quindi questi due gruppi di oggetti sembrano rappresentare due famiglie ben distinte per composizione.

APPENDICE

Condizioni per l'utilizzo del software

Il software Ministat 2.1 può essere utilizzato alle seguenti condizioni:

⇒ il software è di proprietà dell'Autore;

⇒ il software viene concesso in uso "*così come è*" dall'Autore al quale nessun danno può essere imputato per un suo uso sia proprio sia improprio;

⇒ il software non può essere in alcun modo disassemblato, copiato, rimasterizzato, nemmeno in parte, e non può essere distribuito in alcuna forma a terzi.