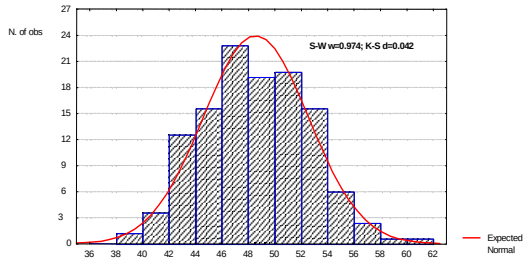
Appunti di Statistica

ALMA MATER STUDIORUM - UNIVERSITÀ DEGLI STUDI DI BOLOGNA

Scuola di *Agraria e Medicina veterinaria*

http://www.agrariaveterinaria.unibo.it



Corso di Laurea Magistrale in *Scienze e Tecnologie Alimentari*

ing. Angelo Fabbri

rev. n°102 - 25 novembre 2013

Questo documento è in fase di stesura e potrà essere modificato.

La versione più recente è disponibile all'indirizzo: ***http://tinyurl.com/2px2n3***

01. CAVEAT EMPTOR

Queste dispense, rivolte agli studenti della Facoltà di Agraria dell'Università di Bologna, non rappresentano una trattazione autonoma della materia dei corsi, ma semplicemente uno strumento per facilitare lo studio.

Nella sequenza degli argomenti e nella notazione, queste pagine ricalcano lo svolgimento delle lezioni, delle quali sono più che una trascrizione, la struttura. Lo scopo è quello di permettere agli studenti di ritrovare in "bella copia" gli argomenti illustrati a lezione. I commenti sono mantenuti volutamente stringati e gli approfondimenti (necessari ad una vera comprensione di qualsiasi materia) sono stati largamente tralasciati.

Questa versione delle dispense è a livello di bozza e sarà sottoposta a revisioni ed integrazioni. La versione più recente è disponibile alla pagina Web ***http://tinyurl.com/2px2n3***.

Ogni segnalazione di errori o imprecisioni, così come ogni critica, è ovviamente incoraggiata e ben accetta.

02. GENERALITÀ SUL CORSO

Statistica (ed elaborazione dei dati)

Docente: ANGELO FABBRI

ing. ***Angelo Fabbri*** *- INet:* {*http://www.unibo.it/docenti/angelo.fabbri* / *afabbri@agrsci.unibo.it*

🕘 Laurea Magistrale: Ottobre ÷ Dicembre;

Conoscenze e abilità da conseguire

Il corso si propone di fornire un quadro generale della statistica induttiva, con particolare attenzione agli aspetti di più diretto interesse per lo specifico corso di laurea. Segnatamente vengono sottolineati i criteri che stanno a fondamento dei metodi per il controllo statistico, utilizzabili sia con finalità legate al controllo di qualità sia alla sperimentazione di processo.

Programma/Contenuti

Richiami di statistica descrittiva. Distribuzioni di frequenza discrete. Indici di posizione e di dispersione. Numeri indice (4h). *Teoria elementare della probabilità.* Definizioni di probabilità. Distribuzioni continue di probabilità. Variabili aleatorie continue. Distribuzione normale. Intervallo di confidenza (4h). *La statistica inferenziale.* Il campionamento. Distribuzione delle somme e delle differenze campionarie. Teorema del limite centrale. (4h). Stima dei parametri della popolazione per mezzo di quelli campionari. Intervalli di confidenza per la stima della media e delle differenze. Minima ampiezza campionaria. Trattamento statistico delle misure. Significatività della differenza tra medie (8h). *Statistica multivariata.* t-test (4h), ANOVA (4h), Cluster analysis (4h).

Testi/Bibliografia

Materiale didattico distribuito durante il corso e dispense redatte dal docente;

"Statistica" di Murray R. Spiegel (Mc Graw Hill)

Metodi didattici

Lezioni frontali ed attività in laboratorio di informatica. Durante il corso vengono presi in considerazione molti esempi tratti dai settori della zootecnia, dell'agronomia, e dell'ingegneria agraria ed alimentare. Vengono inoltre impiegati alcuni pacchetti software di larga diffusione per l'organizzazione e l'analisi statistica dei dati. Le lezioni sono integrate da partecipazione a seminari specialistici e consultazione di letteratura scientifica internazionale.

Modalità di verifica dell'apprendimento

La verifica finale delle competenze di informatica e statistica si svolge attraverso un questionario, riguardante sia aspetti di carattere teorico che esercizi numerici, proposto allo studente mediante i computer del laboratorio di informatica, presso il Campus di Scienze degli Alimenti, nella sede di Cesena della Facoltà di Agraria.

E' richiesta l'iscrizione alla prova d'esame attraverso il servizio Almaesami (http://almaesami.unibo.it).

Informazioni di dettaglio sulle tecnicalità d'esame sono disponibili sulla pagina web del docente.

Strumenti a supporto della didattica

Lavagna; videoproiettore; PC; collegamento Internet; lavagna luminosa; laboratorio di informatica. Durante il corso vengono svolte esercitazioni utilizzando i programmi Winks (TexaSoft), Excel (Microsoft), Statistica (Statsoft) ed R.

Orario di ricevimento

Martedì dalle 15 alle 17 o in altri momenti previo accordo, p.e. via mail.

Programma per Guida dello studente Eramus

module of **STATISTICAL DATA ANALYSIS** - 30 hours

Angelo Fabbri PhD Eng.

Descriptive statistics. Discrete and continuous frequency distribution. The sampling theory. Statistical tests. Correlation and regression. Analysis of variance. Multivariate methods. Simple applications of statistical software.

Perchè studiamo la statistica. Esempi d'impiego delle tecniche statistiche nel settore agroalimentare

- osservazione/descrizione di insiemi di dati;
- sperimentazione (ricerca di relazioni causa/effetto) e trattamento di misurazioni di campo o di laboratorio;
- gestione della qualità (ISO9k, Vision).

- E' possibile riprendere gli argomenti principali ricorrendo ad un semplice esempio tratto dalla pratica professionale: *si vuole valutare il grado zuccherino dei frutti di un frutteto*:
  - non potendo evidentemente distruggere tutti i frutti del frutteto quanti ne dovrò considerare? E quali?
  - Come e con quale affidabilità estenderò a tutto il frutteto le misure effettuate solo su alcuni esemplari?
  - Poiché per ciascun frutto otterrò una misura differente, come dovrò comportarmi nella redazione del rapporto di prova?

    Esempio 1: Valutare la presenza di pesticida sulla superficie dei frutti di un frutteto (*stime campionarie*).

    Esempio 2: Valutare il grado zuccherino medio di una partita d'uva (*misure ripetute*).

    Esempio 3: Valutare l'accettabilità di un'acqua minerale in relazione ai limiti di legge imposti sulla concentrazione di una determinata sostanza (*verifica di ipotesi*).

    Esempio 4: Come influisce sulla resa o sulla qualità delle carni una variazione (o una serie di variazioni) nel regime alimentare di un gruppo di bovini? Ci sono differenze significative tra due diversi tipi di mangimi? (*significatività delle differenze*) Posso esprimere analiticamente le curve di crescita? In quale misura saranno affidabili a fini previsivi? (*interpolazione*).

Richiami di statistica descrittiva

materiali corso LT:

- popolazione e campione;
- statistica descrittiva/induttiva;
- distribuzioni di frequenza (frequenza relativa; piano frequenza-valori; $\Sigma Fr=1$; curve di frequenza);
- indici di posizione (valore medio = somma dei prodotti frequenza per valore; valore modale);
- indici di dispersione;
- standardizzazione.

03. L'ANALISI DEI DATI CON MICROSOFT EXCEL(*)

Ex

Generare 5'000 numeri casuali, distribuiti *normalmente* con valore medio 50 e scarto quadratico medio pari a 20.

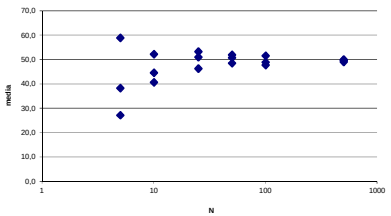
(***Nota1 MS Office 2003***) se nel menù *Strumenti* non è presente la voce *Analisi dei dati*, allora occorre attivarla con: *Strumenti → Componenti aggiuntivi → Strumenti di Analisi*.

(***Nota1 MS Office 2007***) se nel menù *Dati* non è presente la voce *Analisi dati*, allora occorre attivarla con: pulsante *MSoffice → Opzioni di Excel → Componenti aggiuntivi → Gestisci → Comp. Aggiuntivi di Excel → Vai → Strumenti di Analisi.*

(***Nota2***) per la generazione di una serie pseudocasuale è possibile inserire nel campo *Generatore* un qualsiasi numero intero positivo, p.e. le ultime 2 cifre del proprio numero di matricola.

Utilizzando le possibilità grafiche e di editing, la biblioteca di funzioni ed il modulo di analisi dei dati del programma *MS Excel*, si chiede di:

- **a)** calcolare media aritmetica e scarto quadratico medio;
- **b)** tracciare l'istogramma di frequenza assoluta e quello di frequenza relativa cumulata;
- **c)** derivare la serie standardizzata z, calcolarne media e scarto quadratico medio e confrontarli con quelli della serie di partenza;
- **d)** prima di qualsiasi valutazione numerica lo studente ipotizzi la capacità di un campione, di ampiezza pari all'1% di quella della popolazione, di rappresentare gli indici della popolazione dalla quale proviene;
- **e)** al fine di sperimentare la rappresentatività del campione, si chiede di estrarre un campione di 5 individui, calcolarne valore medio e scarto quadratico medio e confrontare i risultati con quelli della popolazione di partenza. Ripetere successivamente i calcoli su campioni di ampiezza 10, 25, 50, 100, 500. Infine tracciare un grafico dei valori campionari in funzione della numerosità campionaria;
- **f)** ripetere il punto precedente con popolazioni di ampiezza maggiore;
- **g)** quando la popolazione è molto ampia l'affidabilità del campione dipende dal rapporto tra ampiezza campionaria ed ampiezza della popolazione?
- **h)** quando la popolazione è molto ampia, come cresce l'affidabilità del campione in funzione della sua numerosità?



04. TEORIA ELEMENTARE DELLA PROBABILITÀ

Il concetto di probabilità costituisce un ponte tra la l'ambito della statistica descrittiva e quello della statistica induttiva.

Definizione classica di probabilità Storicamente la probabilità di un dato **evento** è definita come il rapporto tra il numero dei casi favorevoli al suo verificarsi ed il numero totale dei casi egualmente possibili.

Formalizzazione della definizione classica di probabilità: definito un evento E, sia h il numero dei casi favorevoli al suo verificarsi, ed n il numero di tutti i casi egualmente possibili; allora la probabilità che si manifesti l'evento E si indica con $p = \Pr\{E\}$ e vale h/n.

p.e. l'evento E sia l'uscita del numero 2 sulla faccia superiore di un dado a sei facce, allora $h=1$ (numero di facce che contengono il numero 2) ed $n=6$ (numero totale di facce) e dunque la probabilità che lanciando un dado si ottenga il *2*, vale $1/6$ (circa *16%*).

p è compreso tra 0 (evento impossibile) ed 1(evento certo).

La probabilità che NON si manifesti l'evento E si indica con $q=\Pr\{\text{non E}\}$ e vale:

q = *casi sfavorevoli / casi totali* = $(n-h)/n = 1-h/n = 1-p$

Essendo $q=1-p$ risulta che $p+q=1$.

La probabilità di <u>non</u> ottenere il numero 2 vale 5/6, infatti $n=6$, $h=1$, allora $(n-h)/n = (6-1)/6 = 5/6$ che è anche uguale ad $1-1/6$.

Calcolare la probabilità che lanciando un dado si ottengano i numeri 1 o 2. $n=6$, $h=2$, dunque $p=2/6 = 1/3$.

Analisi combinatoria

Per il calcolo della probabilità associate al verificarsi di un evento occorre calcolare la quantità n, ovvero essere in grado di enumerare i casi possibili (*calcolo combinatorio*).

Permutazioni

Data una popolazione $x_1, x_2,...,x_N$ le sue permutazioni sono i gruppi diversi di N elementi che si possono formare cambiando l'ordine.

Dunque tali gruppi contengono tutti i medesimi elementi, ma in ordine differente.

P.e. le permutazioni degli elementi *a,b,c* sono: *abc, acb, bac, cab, bca, cba.*

Il numero totale di differenti permutazioni possibili risulta $N!$

$$N!=1\cdot2\cdot3\cdot4\cdot..\cdot(N-2)\cdot(N-1)\cdot N=\prod_{i=1}^{N} i$$

Per convenzione $0!=1$.

P.e. due parole composte dalle medesime lettere, ma in orine differente, sono considerate diverse. Anagrammi della parola rame [24].
Esempio con 1 elemento (*A*), 2 elementi (*A, B*), 3 (*A, B, C*), ecc., rappresentazione ad albero.

Permutazioni di N elementi presia gruppi di r

Se si considerano invece gruppi di r elementi presi dalla popolazione X di N elementi, considerati differenti se contengono elementi differenti o comunque in ordine differente, allora le permutazioni sono $N\cdot(N-1)\cdot(N-2)\cdot...\cdot(N-r+1) = N!/(N-r)!$ (Prodotto di r numeri interi decrescenti a partire da N).

P.e. parole di 5 lettere formabili con alfabeto di 25 lettere: $25\cdot24\cdot23\cdot22\cdot21$.

Combinazioni di N elementi r a r

Le combinazioni di una popolazione di N oggetti diversi, presi r alla volta, sono i gruppi di r elementi che si possono formare, considerando due gruppi diversi quando differiscono per almeno un elemento.

Dunque in questo caso due gruppi che contengano gli stessi elementi ma in ordine diverso, sono considerati uguali.

P.e.: la quantità di denaro che è possibile ottenere componedo monete di diverso valore non dipende dall'ordine nel quale le monete sono considerate.

Il numero di combinazioni di N elementi presi a gruppi di r, vale:

$$\binom{N}{r}=\frac{N!}{r!\,(N-r)!}$$

P.e. il numero dei gruppi di due lettere che è possibile formare con le lettere *abc*, vale (3 su 2) = 3!/(2!(3-2)!) = 3 [*ab, ac, cb*] mentre il numero di permutazioni vale 3·2=6 [*ab, ba, ac, ca, cb, bc*].

Definizione statistica di probabilità

La definizione classica di probabilità è inefficace, in quanto autoreferenziata, infatti definisce la *probabilità* in termini di *uguale possibilità*, ovvero un concetto derivato da quello di probabilità.Si ricorre in modo più soddisfacente ad una **definizione statistica** di probabilità (detta anche *probabilità stimata*): la probabilità che si verifichi l'evento E è il limite della frequenza relativa associata all'evento E, al tendere ad infinito del numero di osservazioni.

P.e. lanciando una moneta e rilevando la frequenza relativa associata all'evento *testa*, si genera una serie che al tendere ad infinito del numero di lanci, tende al valore 0.5.

Ex1

Empio *flip a coin* dal menu DEMO del programma *Winks*.

Distribuzione continua di probabilità

La definizione statistica di probabilità istituisce una analogia tra il concetto di distribuzione di frequenza relativa e quello di distribuzione di probabilità:

Probabilità stimata = Limite della Frequenza relativa per $N\to\infty$

Quando la variabile x è **continua**, e il numero di osservazioni diviene grande ($N\to\infty$), allora la distribuzione di frequenza discreta, può pensarsi come composta da un numero di classi che tende ad infinito ($k\to\infty$).

Dunque l'istogramma tende ad una curva continua (denominata ***funzione di densità di probabilità, φ(x)***) e l'area sottesa, che rappresenta una frequenza relativa, come conseguenza della definizione di probabilità stimata, tende ad esprimere una probabilità.

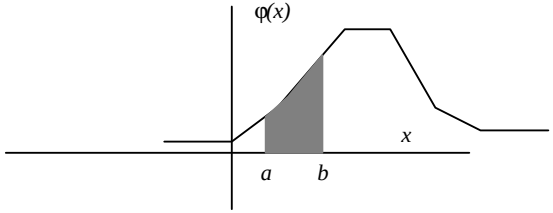
Ricordiamo che per un istogramma di frequenza relativa, l'area di ciascun rettangolo rappresenta la frequenza relativa di ciascuna classe, che l'area sottesa da tutta la curva di frequenza vale *1*. Abbiamo infatti dimostrato la seguente identità:

$$\sum_i f_{Ri} = 1$$

Invece la sola area dell'istogramma compresa tra la classe a e la classe b, rappresenta la frazione di popolazione compresa tra la classe a-esima e la classe b-esima:

$$\sum_{i=a}^{b} f_{Ri} = \sum_{i=a}^{b} \frac{f_i}{N} = \frac{1}{N}\sum_{i=a}^{b} f_i = \frac{N_{ab}}{N} = f_{Rab} \le 1$$

In analogia con le osservazioni già fatte sugli istogrammi di frequenza, sulla base della definizione di probabilità stimata, possiamo dunque concludere che l'area compresa tra le ascisse $x=a$ ed $x=b$ della curva di densità di probabilità esprime la probabilità che la variabile continua x assuma un valore compreso tra a e b = $\Pr\{a<x<b\}$.



Occorre infine ripetere come la *funzione di densità di probabilità* non definisca un valore di probabilità in ordinata, bensì attraverso l'area che sottende, ovvero attraverso la sua funzione integrale.

N finito	$N\to\infty;\ k\to\infty;\ f_{Ri}\to\varphi(x)$	$N\to\infty$
$\sum_{i=1}^{k} f_{Ri} = 1$	*La sommatoria di Simpson tende ad una funzione integrale*	$\int_{xmin}^{xmax} \varphi(x)dx = 1$
$f_{Rab} = \sum_{i=a}^{b} f_{Ri}$	$\frac{Lim}{N\to\infty} f_{Rab} = Pr(a<x<b)$	$Pr\{a<x<b\} = \int_{a}^{b} \varphi(x)dx$

05. ESERCIZI SULLA TEORIA ELEMENTARE DELLA PROBABILITÀ

Ex2

Calcolare la probabilità per gli eventi seguenti:

a) in un lancio di dado esca un numero dispari [3/6];

b) nell'estrarre una carta da un mazzo di 52 esca un asso [4/52].

Ex3

Una pallina viene estratta a caso da una scatola contenente 6 palline rosse, 4 bianche, e 5 blu. Calcolare la probabilità che la pallina estratta sia a)rossa [6/15], b)bianca [4/15], c)blu [5/15], d) non rossa [1-Pr{a}=1-6/15=3/5], e') bianca O rossa O blu [Pr{a}+Pr{b}+Pr{c}=1]; e)rossa o bianca [Pr{a}+Pr{b}=6/15+4/15].

Ex4

Una scatola contiene 8 palline bianche e 2 nere, mentre un'altra ne contiene 2 bianche e 8 nere. Le due scatole sono indistinguibili. Si estrae una pallina da una scatola scelta a caso e, senza guardarla, la si ripone nell'altra scatola. Successivamente si estrae una pallina da quest'ultima scatola. Quanto vale la probabilità che la pallina sia bianca?

06. DISTRIBUZIONI DI FREQUENZA CONTINUE E DISTRIBUZIONE NORMALE

La variabile *aleatoria* discreta

Una popolazione X, sia formata da tutti i valori $x_1, x_2,\ldots$ che possono essere assunti da una variabile discreta x, eventualmente ripetuti con una determinata frequenza.

Alcuni valori si presentano con frequenza maggiore di altri ($f_1, f_2, \ldots$), ovvero i valori di x sono distribuiti sull'intervallo di variazione di x. In tali ipotesi si dice *aleatoria* una qualsiasi quantità (x) estratta a caso dalla popolazione (X).

P.e.: diametro medio, grado zuccherino o peso di un frutto.

Ovvero a differenza del concetto di variabile (reale o intera) dell'analisi matematica, che rappresenta un valore qualsiasi appartenente ad un insieme, la *variabile aleatoria* rappresenta un insieme di valori associati ad una distribuzione di frequenza.

Valor medio di una variabile *aleatoria* continua

Per quanto è stato visto nel capitolo sulle distribuzioni continue di probabilità (pag.10), quando la variabile aleatoria x è continua, ed il numero di osservazioni diviene grande ($N\to\infty$) allora il numero di classi può pensarsi indefinitamente crescente ($k\to\infty$) così l'istogramma di frequenza tende ad una curva continua, detta *funzione di densità di probabilità,* indicata come $\varphi(x)$. Poichè la frequenza tende ad una probabilità, l'area sottesa dalla curva $\varphi(x)$ esprime un valore di probabilità.

Se la variabile x è continua, caratterizzata da una certa distribuzione $\varphi(x)$, come se ne calcola il valore medio?

v. discreta $N\to\infty,\ k\to\infty,\ f_{Ri}\to\varphi(x)$: ***v. continua***

la somma delle aree dei rettangoli dell'istogramma tende ad una funzione integrale

$$\overline{x} = \sum_{i=1}^{k} x_i \cdot f_{Ri} \qquad \lim_{N\to\infty} \overline{x} = \mu_x \qquad \mu_x = \int (x \cdot \varphi(x))\, dx$$

Critica della definizione di valor medio di una variabile aleatoria continua (a.a.)

Se la variabile aleatoria x può assumere tutti i valori reali compresi in un certo intervallo, ed è perciò continua, allora cade in difetto la definizione di valore medio del paragrafo precedente, poiché in qualunque intervallo finito, la variabile aleatoria vi assume infiniti valori, a ciascuno dei quali non si può far corrispondere una probabilità p_i finita e compresa tra uno e zero, come ci si rende conto pensando che altrimenti la somma delle infinite probabilità darebbe infinito, il che è assurdo.

Occorre dunque ricorrere a nuovi concetti per definire il valore medio di una variabile aleatoria continua: per ogni valore x^* di una variabile aleatoria, si può per esempio determinare la probabilità cumulativa $\Pr\{-\infty < x \le x^*\}$ che la variabile aleatoria assuma valori minori od uguali ad x^*.

Resta così definita una funzione $\Phi(x)$, detta *funzione di distribuzione*, che per ogni valore di x^* assume un valore uguale alla probabilità cumulativa $\Pr\{-\infty < x \le x^*\}$ che la variabile assuma un valore compreso nell'intervallo $]-\infty, x^*]$.

Evidentemente tale funzione è sempre positiva e monotona crescente. Il suo dominio è R, ed il suo codominio è $[0, 1]$.

$\Phi(+\infty) = 1; \qquad \Phi(-\infty) = 0$

La differenza $\Phi(x_2) - \Phi(x_1)$ fra i valori che la funzione di distribuzione assume in corrispondenza dei valori x_2, x_1 della variabile aleatoria fornisce la probabilità che i valori della variabile cadano nell'intervallo $[x_1, x_2]$, con $x_1 < x_2$.

Se la $\Phi(x)$ è continua e derivabile, allora il rapporto incrementale

$$\frac{\phi(x+h) - \phi(x)}{h}$$

per h tendente a zero, tende alla derivata della $\Phi(x)$, che viene chiamata *funzione densità di probabilità* della variabile aleatoria ed è indicata col simbolo $\varphi(x)$:

$$\frac{d\phi(x)}{dx} = \varphi(x)$$

e dunque per il teorema di Torricelli:

$$\phi(x) = \int_{-\infty}^{x} \varphi(x)dx$$

il prodotto $\varphi(x)\cdot dx$ rappresenta il differenziale di una probabilità, ovvero la probabilità che il valore della variabile aleatoria cada in un intervallo infinitesimo tra x ed $x+dx$.

Si osserva che mentre la $\Phi(x)$, che rappresenta una probabilità, è una grandezza adimensionale, la $\varphi(x)$ ha le dimensioni di un inverso di x e non è pertanto una probabilità.

Per la sua definizione la $\varphi(x)$ deve essere tale per cui

$$\int_{-\infty}^{+\infty} \varphi(x)dx = \phi(+\infty)-\phi(-\infty) = 1$$

Dunque condizioni necessarie e sufficienti affinché una funzione possa rappresentare la densità di una distribuzione di probabilità, è che tale funzione sia continua, positiva o nulla su tutto R e sottenda un'area unitaria su tutto R.

L'area sottesa dalla curva $\varphi(x)$, e l'asse x, per $a<x<b$, vale la probabilità che x sia compreso tra a e b, e si indica con $\Pr\{a<x<b\}$.

Varianza delle variabili aleatorie continue *(a.a.)*

In modo del tutto analogo a quanto visto per il valor medio di una variabile aleatoria continua, si definisce lo scarto quadratico medio della popolazione X come valor medio dei quadrati degli scarti:

$$\sigma = \int_{-\infty}^{+\infty} [(x-\mu_x)^2 \varphi(x)] \, dx$$

In tale contesto lo scarto quadratico medio di una variabile aleatoria continua viene indicato spesso semplicemente come σ^2, oppure come $\sigma^2(x)$ o σ^2_x.

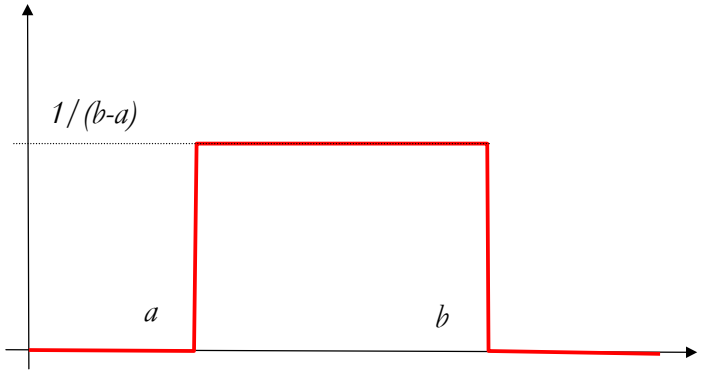
La distribuzione uniforme

Esaminiamo una distribuzione nella quale la variabile assume valori non nulli solo in un intervallo finito $[a, b]$, ed in questo intervallo tutti gli infiniti valori siano ugualmente probabili:

$\varphi(x) = 0 \quad per\ x<a$

$\varphi(x) = 0 \quad per\ x>b$

$\varphi(x) = \frac{1}{b-a} = Cost. \quad per\ x \in [a, b]$



Integrando su x si ottiene la $\Phi(x)$, nulla per $x < a$, linearmente crescente nell'intervallo $[a, b]$, costante ed uguale ad 1 per $x > b$.

É semplice rendersi conto del fatto che l'area sottesa su tutto R è unitaria ($(b-a)\cdot 1/(b-a)=1$).

Il valore medio risulta:

$$\mu_x = \int_a^b x\varphi(x)dx$$
$$= \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{2(b-a)}(b^2-a^2)$$
$$= \frac{a+b}{2}$$

La distribuzione a campana simmetrica

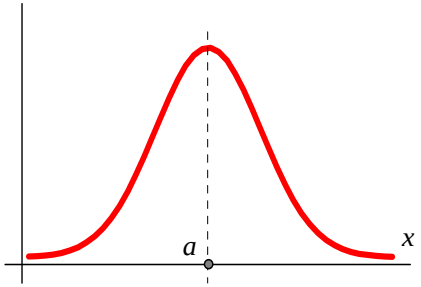
Fra le tante possibili distribuzioni di probabilità, quella a campana simmetrica ricorre frequentemente nell'analisi statistica delle misure.

Nel campo delle misure infatti gli errori più grossi sono meno frequenti (forma a campana), e gli errori per difetto sono probabili quanto quelli per eccesso (simmetria).

Una funzione continua e definita su tutto R (che giustificheremo pienamente in seguito, per mezzo del teorema del valore centrale) in grado di descrivere tale andamento, è del tipo di quella seguente:

$$\varphi(x) = e^{-[h(x-a)]^2}$$

la costante e ($e=\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x \approx 2.71818$) è la base dei logaritmi naturali.



L'area sottesa da questa curva non è unitaria, in particolare l'integrazione di questa funzione su tutto R fornisce il risultato $\sqrt{\pi}/h$, dunque per potere utilizzare tale formulazione come funzione di densità di probabilità è sufficiente dividerla per il coefficiente $\sqrt{\pi}/h$:

$$\varphi(x) = \frac{h}{\sqrt{\pi}} e^{-(x-a)^2 h^2}$$

Tale forma della funzione di distribuzione di probabilità è detta ***normale*** o *di* ***Gauss*** o ***gaussiana***.

Interpretazione della funzione di densità di probabilità normale

Cerchiamo una caratterizzazione statistica e geometrica della coppia di parametri h ed a, che definiscono univocamente una certa distribuzione di probabilità normale.

Il valor medio di una variabile aleatoria con distribuzione normale è dato da:

$$\mu_x = \int_{-\infty}^{+\infty} x \cdot \varphi(x)\, dx = \int_{-\infty}^{+\infty} \left(x \cdot \frac{h}{\sqrt{\pi}} e^{-(x-a)^2 h^2} \right) dx$$

introduciamo il cambiamento di variabile $t=h\cdot(x-a)$, ovvero $x=a+t/h$:

$$\mu_x = \int_{-\infty}^{+\infty} \left(\frac{h}{\sqrt{\pi}} \left(\frac{t}{h} + a \right) e^{-t^2} \right) \frac{dt}{h} = \frac{1}{h\sqrt{\pi}} \int_{-\infty}^{+\infty} t\, e^{-t^2} dt + \frac{a}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-t^2} dt = 0 + \frac{a\sqrt{\pi}}{\sqrt{\pi}} = a$$

risulta cioè che il parametro a altro non è se non il valor medio della variabile aleatoria.

Abbiamo così (finalmente) dimostrato analiticamente il risultato, fin qui dato per intuitivo, secondo il quale il valore medio di una distribuzione a campana simmetrica è individuato dal punto di massimo assoluto.

Ricordando poi la definizione di varianza, ed operando lo stesso cambiamento di variabile si ottiene:

$$\sigma_x^2 = \int_{-\infty}^{+\infty} \left((x-\mu_x)^2 \cdot \frac{h}{\sqrt{\pi}} e^{-(x-a)^2 h^2} \right) dx = \frac{1}{h^2\sqrt{\pi}} \int_{-\infty}^{+\infty} t^2 e^{-t^2} dt = \frac{1}{2h^2}$$

e cioè il parametro h è inversamente proporzionale allo scarto quadratico medio. Allora, sostituendo μ_x ad a, ed $1/(2\sigma_x^2)$ ad h^2, l'espressione di Gauss può assumere la forma assai più significativa:

$$\varphi(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}$$

Si verifica così che per definire completamente una distribuzione gaussiana è sufficiente conoscerne i due parametri valor medio μ_x e varianza σ_x^2.

La curva presenta dunque un massimo in $x=\mu_x$, e, come è semplice verificare, due flessi, simmetrici rispetto al valor medio, in corrispondenza delle ascisse $\mu_x-\sigma_x$ e $\mu_x+\sigma_x$, distanti cioè σ_x dal valore medio.

Si può dimostrare che per la distribuzione normale risulta inoltre:
coeff. di asimmetria $\alpha_3=0$;
coeff. di curtosi $\alpha_4=3$.

Misura standardizzata della variabile aleatoria con distribuzione normale

Piccola premessa matematica

1 - Integrazione con rettangoli/trapezi. Con tale metodo è possibile ottenere un valore numerico approssimato dell'integrale definito di una funzione comunque complessa.

2 - Integrazione (definita) per sostituzione. p.e. $sin[(2x-1)/3]$, $e^{(x-1)/2}$, ecc.

3 - Integrazione definita di una funzione simmetrica. Qualsiasi siano gli estremi d'integrazione $(0,x_1;\ 0,-x_1;\ x_1,x_2;\ -x_1,x_2;\ -x_2,x_1;\ -x_1,-x_2)$, ci si può sempre ricondurre all'integrazione nella forma da 0 ad x.

Integrazione della gaussiana

Per una data distribuzione normale, noti cioè μ_x e σ_x, ha interesse conoscere con quale probabilità un determinato valore della variabile x cade all'interno di un intervallo individuato da due valori x_1 ed x_2. Per quanto è stato visto nel capitolo sulle distribuzioni continue di probabilità (pag.10), il problema è ovviamente risolto dall'integrale definito:

$$Pr\{x_1<x<x_2\} = \int_{x_1}^{x_2} \varphi(x)\,dx = \int_{x_1}^{x_2} \frac{1}{\sigma_x\sqrt{2\pi}}\, e^{-\left[\frac{(x-\mu_x)^2}{2\sigma_x^2}\right]}\,dx$$

Ad esempio ci si può chiedere quanto vale la probabilità che il peso di un frutto scelto a caso, sia compreso tra due determinati limiti. Purtroppo l'espressione di Gauss non ha una primitiva, ovvero il suo integrale non ha soluzione analitica. L'unico metodo in grado di ottenere una soluzione è quello numerico, ovvero è possibile risolvere l'integrale definito, per esempio con un metodo numerico, come quelli dei trapezi o dei rettangoli, ma non quello indefinito. Tuttavia anche l'applicazione del metodo numerico pone qualche difficoltà a livello operativo: per gli usi pratici sarebbe infatti opportuno poter disporre di tabelle precalcolate dei valori ottenuti dall'integrazione numerica, però questi sarebbero legati alle ∞^4 combinazioni di valori assumibili dai parametri μ_x, σ_x, x_1 *ed* x_2.

Si può ovviare a tale inconveniente operando un cambiamento di variabile, introducendo la sostituzione lineare (standardizzazione della variabile x):

$$z = \frac{x - \mu_x}{\sigma_x}$$

effettuando infatti la sostituzione si ottiene facilmente:

$$\int \varphi(x)\,dx = \int \left(\frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\right)\,dz$$

risulta cioè che, qualsiasi siano il valore medio di x e la sua dispersione, la variabile aleatoria z è caratterizzata da una speciale distribuzione gaussiana $\psi(z)$, con media nulla e varianza unitaria:

$$\psi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$$

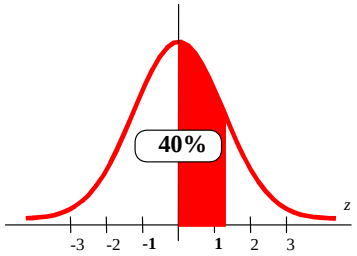
Fissato il primo estremo d'integrazione a zero e dato il fatto che tale funzione è anche simmetrica (per via del quadrato di z), si può ora affrontare il problema della determinazione del valore dell'integrale definito della gaussiana sulla base di un solo parametro (il secondo estremo d'integrazione), e quindi risulta possibile la compilazione di tabelle di uso pratico:

z^*	$p=\int_0^{z^*} \psi(z)\,dz$	$p=\int_{-z^*}^{z^*} \psi(z)\,dz$
	Appendice II – Spiegel	
0.674	0.250	**0.500**
1.000	0.341	**0.682**
1.282	0.400	**0.800**
1.645	0.450	**0.900**
1.960	0.475	**0.950**
2.576	0.495	**0.990**
3.291	0.499	**0.999**

P.e. la tabella dice immediatamente che:

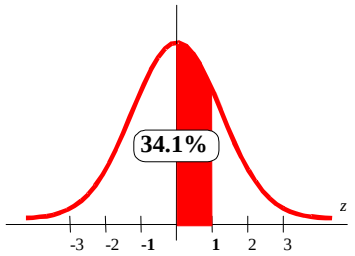
- Con il 40% di probabilità i valori di z risultano compresi nell'intervallo [0, +1.282]

$$Pr\{0<z<+1.282\} = \int_{0}^{+1.282} \psi(z)\,dz \cong 40\%$$



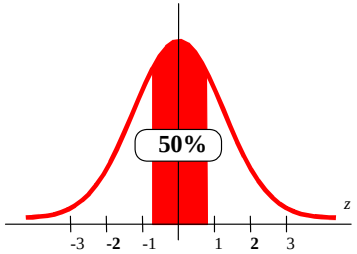
- con il 34.1% di probabilità i valori di z risultano compresi nell'intervallo [0, +1]

$$Pr\{0<z<+1\} = \int_{0}^{+1} \psi(z)dz \cong 34.1\%$$



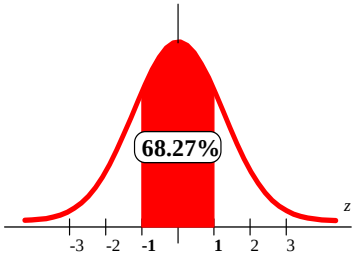
- con il 50% di probabilità i valori di z risultano compresi nell'intervallo [-0.674, +0.674]

$$Pr\{-0.674<z<0.674\} = \int_{-0.674}^{+0.674} \psi(z)\,dz \cong 50\%$$



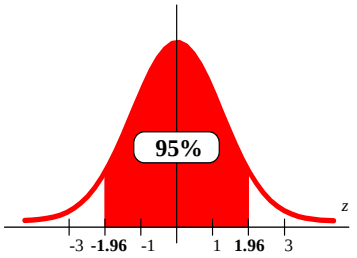
- con circa il 68% di probabilità i valori di z risultano compresi nell'intervallo [-1, +1]

$$Pr\{-1<z<1\} = \int_{-1}^{+1} \psi(z)dz \cong 68\%$$



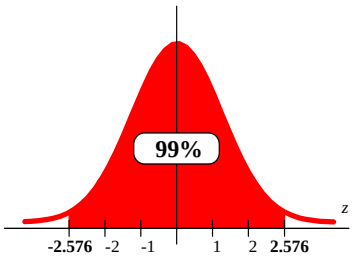
- con il 95% di probabilità i valori di z risultano compresi nell'intervallo [-1.96, +1.96]

$$Pr\{-1.96<z<1.96\} = \int_{-1.96}^{+1.96} \psi(z)dz \cong 95\%$$



- con il 99% di probabilità i valori di z risultano compresi nell'intervallo [-2.576, +2.576]

$$Pr\{-2.576<z<2.576\} = \int_{-2.576}^{+2.576} \psi(z)dz \cong 99\%$$



- Ricordando che noi vogliamo mettere in relazione un valore di probabilità con un intervallo di valori $[x_1, x_2]$, e non $[z_1, z_2]$, noti i parametri μ_x e σ_x, basta calcolare l'integrale definito della gaussiana in forma standard, tra gli estremi d'integrazione z_1 e z_2, corrispondenti ai valori standardizzati di x_1 ed x_2.

Area sottesa dalla curva di distribuzione di probabilità Normale, tra i punti di ascissa 0 e z

z	0	1	2	3	4	5	6	7	8	9
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
∞	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

07. ESERCIZI SULLA DISTRIBUZIONE NORMALE

Ex5

Calcolare il seguente integrale definito:

$$\int_{x_1=0}^{x_2=2} e^{(x-2)/3}\,dx$$

si effettua la sostituzione $z=(x-2)/3$, allora la relazione che lega il differenziale di x a quello di z è:

$$x=3z+2 \rightarrow dx=\frac{\partial(3z+2)}{\partial z}dz=3\,dz$$

allora sostituendo tutte le espressioni in x con le corrispondenti in z, otteniamo:

$$\int_{z_1=(0-2)/3}^{z_2=(2-2)/3} 3\,e^z\,dz = [3\,e^z]_{2/3}^{0} = (e^0 - e^{-2/3})$$

Ex6

Calcolare l'ascissa del punto di massimo e dei punti di flesso della gaussiana e della gaussiana in forma standard.

Ex7

All'esame di matematica la media dei voti è 22 e lo scarto quadratico medio 5. Determinare i valori standard dei voti 11, 17, 18, 21, 22, 27, 30, 32.

[z=(voto-22)/5]

Ex8

Trovare l'area sotto la curva normale standardizzata per ciascuno dei casi seguenti (tracciare i grafici):

$$\int_{z_1}^{z_2} \psi(z)\, dz = \int_{z_1}^{z_2} \left(\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \right) dz$$

a) estremi $z_1 = 0$ e $z_2 = 1.2$ [Pr(0<z<1.2) = 0.3849];

a') estremi $z_1 = 0$ e $z_2 = 1.25$ [Pr(0<z<1.25)];

b) estremi $z_1 = -0.68$ e $z_2 = 0$ [curva simmetrica, colonna con l'8, 0.2518];

c) estremi $z_1 = -0.46$ e $z_2 = 2.21$ [0.1772+0.4864 = 0.6636];

d) estremi $z_1 = 0.81$ e $z_2 = 2$ [0.48-0.2910 = 0.189];

e) a sinistra di $z = -0.6$ [area a sinistra di $z = 0$ (0.5)-area tra 0 e -0.6 = 0.5-0.2258 = 0.2742];

f) a destra di $z = -1.28$ [0.3997+0.5=0,8997];

g) a destra di $z = 2.05$ ed a sinistra di $z = -1.44$ [area totale - quella in mezzo].

h) a destra di $z = 2$ ed a sinistra di $z = 3$.

Ex9

Determinare il valore di z quando (disegnare la gaussiana):

a) l'area compresa tra 0 e z vale 0.3770 [z = +/-1.16].

b) l'area a sinistra di z vale 0.8621 [A=0.8621-0.5=0.3621, z=1.09].

Ex10

La lunghezza media delle foglie di una pianta vale 151 mm e lo scarto quadratico medio vale 15 mm. Assumendo che le lunghezze (x) siano distribuite normalmente, trovare quante foglie hanno una lunghezza:

a) compresa tra 120 e 155mm [l'area tra $z1$ e $z2$ vale circa 0.6];

$$Pr\{120<x<155\} = \int_{120}^{155} \varphi(x)\, dx = \int_{120}^{155} \left[\frac{1}{15\sqrt{2\pi}} e^{-\left(\frac{(x-151)^2}{2 \cdot 15^2} \right)} \right] dx$$

$z_1 = (120-151)/15 = -2.10$

$z_2 = (155-151)/15 = 0.30$

$$Pr\{120<x<155\} = \int_{-2.10}^{0.30} \psi(z)\, dz = \int_{-2.10}^{0.30} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz =$$

$$=0.1179+0.4821=0.6$$

b) inferiore a 151mm [(metà) Pr{x<151}=0.5];

c) maggiore di 166mm [z = (166-151)/15=15/15=1].

Ex11

Il peso medio degli animali di un allevamento vale 60 kg, con deviazione standard pari a 10 kg. Ipotizzando che la popolazione di pesi sia distribuita normalmente calcolare la probabilità che un animale scelto a caso pesi **a**) più/meno di 50 kg; **b**) più/meno di 60 kg; **c**) più/meno di 70 kg; **d**) tra 50 e 70 kg.

[Pr{x>50} ...]

Ex12

Il diametro medio di una partita di frutti è 50 mm, mentre lo scarto quadratico medio vale σ=6 mm. I limiti di tolleranza massimi ammessi dagli standard di qualità per tale prodotto vanno da 40 mm a 55 mm. Determinare la quota di prodotti difettosi assumendo che tali diametri siano distribuiti normalmente.

[$z1$=(40-50)/6=-1.66 e $z2$=(55-50)/6=0.83].

Ex13

Ad una distilleria arriva una partita di frutta destinata alla produzione di alcool. Si stabilisce che viene pagata solo la quota di frutta di grado zuccherino superiore a 10°Brix.

Viene così prelevato un campione rappresentativo il quale fornisce come valore medio 8°Brix con deviazione standard pari a 2°Brix.

Quale sarà la riduzione di prezzo?

Pr{x>10}==> Pr{z>(10-8)/2}= Pr{z>1}=0.5-0.3413=0.16, ovvero la riduzione sarà pari a circa 16%];

Ex14

Una variabile aleatoria continua x assume valori compresi nell'intervallo 0÷4. La funzione di densità di probabilità è $p(x)=0.5-a\cdot x$. Calcolare il valore di Pr{1<x<2}.

[Si calcola prima *a* imponendo che l'area valga 1. a=3/8, Pr=5/16].

Ex15

La distanza media di un elettrone dal nucleo di un orbitale *S* è 50Å, l'elettrone si trova a distanze maggiori di 10 Å con probabilità del 30%. Se la funzione d'onda ψ fosse normale, quanto varrebbe la probabilità di trovare l'elettrone a distanze dal nucleo superiori a 20 Å ?

Ex16

Una variabile aleatoria continua x assume valori compresi nell'intervallo 0÷4. La funzione di densità di probabilità è $p(x)=0.25-3a\cdot x$. Calcolare il valore di Pr{1<x<2}; Pr{1<x<3}; Pr{0<x<4};.

[Si calcola prima *a* imponendo che l'area valga 1. a=0, Pr=1/4; 2/4; 4/4].

Ex17

La densità di probabilità di una variabile aleatoria, definita nell'intervallo [1, 2], è proporzionale ad 1/x. Determinare il valore medio e la deviazione standard della distribuzione.

[La funzione di distribuzione è nota a meno di una costante moltiplicativa da ricavarsi dalla condizione di normalizzazione]

Ex18

Calcolare la deviazione standard di una popolazione di numeri reali distribuiti in modo perfettamente casuale tra due estremi.

08. INTERVALLO DI CONFIDENZA

Oltre a determinare quale livello di probabilità esiste che la variabile aleatoria x di distribuzione normale definita dai dati μ_x σ_x, assuma valori compresi nell'intervallo [x_1, x_2], è possibile anche affrontare un problema inverso, ovvero si può fissare un certo livello di probabilità e determinare gli estremi x_1 ed x_2 dell'intervallo, simmetrico rispetto al valore medio, entro il quale cade il valore assunto dalla variabile aleatoria x, col dato livello di probabilità.

Se affermiamo per esempio che, col 95% di probabilità (pag.22):

$$-1.96 \le z \le +1.96$$

allora possiamo immediatamente scrivere che:

$$-1.96 \le \frac{x - \mu_x}{\sigma_x} \le +1.96$$

e dunque risulta:

$$\mu_x - 1.96\cdot\sigma_x \le x \le \mu_x + 1.96\cdot\sigma_x$$

oppure, con una scrittura un po' diversa:

$$x = \mu_x \pm 1.96\cdot\sigma_x$$

E' cioè trovato l'intervallo, simmetrico rispetto a μ_x, entro il quale ricadono i valori della variabile x con il 95% di probabilità.

Per un qualsiasi altro livello di probabilità p, determinato il corrispondente valore $z(p)$ dalla tabellina della gaussiana, risulta:

$$x = \mu_x \pm z(p)\cdot\sigma_x$$

ovvero

$$x_1 = \mu_x - z\cdot\sigma_x \quad ed \quad x_2 = \mu_x + z\cdot\sigma_x$$

L'intervallo $[x_1, x_2]$ così determinato si chiama *intervallo di confidenza* o *intervallo fiduciale*, poiché confidiamo, al livello di probabilità p prescelto, che un valore scelto a caso dalla popolazione X ricada in tale intervallo.

Def. Il prodotto $1 \cdot \sigma_x$ si chiama deviazione probabile o deviazione standard della variabile aleatoria x. Per tale motivo è frequente trovare in letteratura che ci si riferisce allo scarto quadratico medio come alla deviazione standard.

09. ESERCIZI SUGLI INTERVALLI DI CONFIDENZA

Ex19

Calcolare l'intervallo di confidenza al 95% di una variabile aleatoria continua x avente distribuzione normale con valore medio 5 e deviazione standard 1.2.

Si tratta di trovare i valori x_1 ed x_2, simmetrici rispetto al valor medio tali che:

$$0.95 = \int_{x_1}^{x_2} \varphi(x)\,dx = \int_{-z}^{+z} \psi(z)\,dz \rightarrow x = \mu_x \pm 1.96 \cdot \sigma_x$$

i valori x_1 ed x_2 si ricavano semplicemente da:

$$x_{1,2} = \mu_x \pm z(p) \cdot \sigma_x = 5 \pm 1.96 \cdot 1.2 = 5 \pm 2.352$$

Ex20

Il peso medio di una varietà di frutti vale 50g, con una deviazione standard pari a 10g. Nell'ipotesi che la variabile aleatoria *peso di un frutto* x, sia distribuita normalmente, determinarne gli intervalli fiduciali al 95% ed al 99%. Sulla base dei coefficienti riportati p.e. in tabella a pag.22, risulta:

$$0.95 = \int_{x_1}^{x_2} \varphi(x)\,dx = \int_{-z}^{+z} \psi(z)\,dz \rightarrow x = 50g \pm 1.96 \cdot 10g$$
$$\rightarrow x_1 = 30.4g;\ x_2 = 69.6g$$

$$0.99 = \int_{x_1}^{x_2} \varphi(x)\,dx = \int_{-z}^{+z} \psi(z)\,dz \rightarrow x = 50g \pm 2.58 \cdot 10g$$
$$\rightarrow x_1 = 24.2g;\ x_2 = 75.8g$$

10. TEORIA ELEMENTARE DEI CAMPIONI

Il problema del campionamento

Le popolazioni oggetto di uno studio statistico (persone, cellule, batteri, prodotti industriali/agricoli) sono spesso troppo ampie perché si possa compiere sulla loro totalità il rilevamento delle grandezze che ci interessano (p.e. il valore medio o la distribuzione di frequenza): siamo costretti a ricorrere all'analisi di una porzione limitata della popolazione che viene detta *campione*.

Evidentemente le grandezze rilevate da un campione sono in generale diverse da quelle relative a tutta la popolazione.

Nasce dunque il problema di come scegliere il campione affinché rispecchi il più possibile le caratteristiche della popolazione dalla quale è estratto, e di valutare l'errore commesso.

Campioni casuali e numeri casuali

Per l'ottimizzazione della qualità del campione, i suoi elementi devono essere scelti dalla popolazione in modo *casuale*. A volte, per campioni piccoli (orientativamente meno di 10 elementi), in favore di una maggiore dispersione del campione nella popolazione è ammessa qualche procedura di campionamento sistematico (p.e. serbatoi per liquidi, come latte o vino; materiali in cumulo, come frutta e verdura; stratificazione e cenno al principio delle aree/baricentri/volumi; disposizioni ad *X, W* o secondo griglie ortogonali).

Data una popolazione, estrarre a caso alcuni suoi elementi significa sceglierli in modo che tutti abbiano l'identica probabilità di essere estratti.

Per l'estrazione degli elementi del campione si associano gli elementi della popolazione a numeri, e poi si scelgono i numeri secondo una serie casuale.

Problema della generazione delle serie di numeri casuali (tabelle, metodi fisici, elettronici, matematici, funzioni di generazione di numeri random e campionamento di Excel).

Esempi di cattivo campionamento (animali catturati in un recinto, produzione di latte della mattinata, prodotti venduti nel fine settimana).

Campionamento con e senza ripetizione e distribuzione della media campionaria: teorema del limite centrale

Se estraiamo più di un campione da una popolazione non infinita, già il secondo campione estratto trova una popolazione modificata rispetto a quella originaria in quanto privata degli elementi del primo campione. Diversa è la situazione se l'estrazione dei campioni successivi avviene dopo che ciascun campione estratto è stato reimmesso nella popolazione. In tal caso ogni estrazione trova la popolazione non modificata. *Estrazione con/senza reimmissione.*

Se dalla definizione di popolazione infinita discende la proprietà che essa non varia mentre estraggo i campioni, allora posso dire che nel campionamento con reimmissione la popolazione si comporta come infinita.

P.e. un prelievo di vino da un tino è un campionamento senza reimmissione, mentre una misura di lunghezza può essere considerata come un campione della infinita serie di determinazioni di lunghezza che è possibile condurre sul medesimo oggetto.

Anche se un campione è stato estratto in modo casuale dalla popolazione l'analisi statistica non fornisce evidentemente gli stessi valori ottenibili dall'analisi esaustiva della intera popolazione. Se estraggo da una stessa popolazione due campioni di numerosità *Nc*, questi forniranno infatti valori diversi degli indici statistici.

Tuttavia il Teorema del limite centrale (Lindeberg-Lévy 1922) afferma che:

La somma di Nc variabili aleatorie estratte da una popolazione con distribuzione qualsiasi, è anch'essa una variabile aleatoria, ma tendente ad assumere una distribuzione normale al crescere di Nc.

Dunque se si prelevano da una popolazione di *Np* individui dei campioni formati ciascuno da *Nc* elementi, e si calcola il valor medio di ciascuno di tali campioni, allora avrò una nuova popolazione del tipo $\bar{x}_1,\bar{x}_2,\bar{x}_3,\ldots$ Immaginando di proseguire all'infinito si costituisce una nuova popolazione caratterizzata da una propria distribuzione di frequenza detta *distribuzione delle medie campionarie*.

Essendo la media aritmetica di ciascun campione una somma di variabili aleatorie (la divisione per la costante *Nc* non modifica la forma della curva di distribuzione) allora, per il teorema del valore centrale, le medie campionarie tendono a distribuirsi normalmente attorno alla loro media che è quella vera della popolazione. Tale tendenza è tanto più accentuata quanto più grande è *Nc*, e quanto più la distribuzione della popolazione d'origine si avvicina a quella normale.

In altre parole il teorema del limite centrale può essere riformulato come:

Qualunque sia la distribuzione di probabilità di una popolazione d'origine, la distribuzione di probabilità di una popolazione di medie, ottenute da campioni di Nc elementi, può essere assimilata ad una distribuzione normale, purché Nc sia abbastanza grande.

O in maniera ancor più sintetica:

le popolazioni di medie campionarie sono asintoticamente normali

L'importanza del teorema risiede nel fatto che permette di ricorrere alla legge di probabilità normale, per le statistiche campionarie, anche quando la popolazione d'origine ha distribuzione di probabilità non normale (o addirittura sconosciuta).

Nella pratica possono essere ritenuti trascurabili gli errori commessi ritenendo la popolazione campionaria come normale se risulta che *Nc* è dell'ordine di grandezza di almeno qualche decina (orientativamente superiore a 30). In particolare tale soglia si abbassa se la popolazione di partenza è tendenzialmente normale.

Ex 21

Empio sul Teorema del limite centrale del menu DEMO del Programma Winks: *campionamento di una distribuzione uniforme.*

http://www.intuitor.com/statistics/CLAppClasses/CentLimApplet.htm

http://www.chem.uoa.gr/applets/appletcentrallimit/appl_centrallimit2.html

http://www.cs.uic.edu/~wilkinson/Applets/clt.html

http://onlinestatbook.com/stat_sim/index.html

I parametri della distribuzione delle medie campionarie

Il numero di campioni differenti (campionamento senza reimmissione) formati da *Nc* elementi che posso estrarre dalla popolazione di *Np* individui corrisponde al numero di combinazioni di *Np* elementi presi a gruppi di *Nc*.

Se considero tutte le combinazioni diverse, allora Per il teorema del limite centrale la distribuzione delle medie campionarie tende ad essere normale, e si dimostra inoltre che la media teorica della popolazione delle medie campionarie $\mu_{\bar{x}}$ tende a coincidere con quella della popolazione d'origine μ_x:

$$\lim_{Nc\to\infty} \mu_{\bar{x}} = \mu_x$$

Si dimostra anche che la distribuzione delle medie campionarie calcolate su tutti i campioni di numerosità *Nc* che è possibile formare con gli elementi di una popolazione di numerosità *Np*, ammette una dispersione attorno al valor medio data da:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{Nc}} \sqrt{\frac{Np - Nc}{Np - 1}}$$

Nel caso di popolazione infinita o, equivalentemente, se il campionamento ammettesse le ripetizioni (reimmissione degli elementi) la varianza campionaria diviene:

$$\sigma_{\bar{x}} = \lim_{Np\to\infty}\left[\frac{\sigma}{\sqrt{Nc}}\sqrt{\frac{Np-Nc}{Np-1}}\right] = \lim_{Np\to\infty}\left[\frac{\sigma}{\sqrt{Nc}}\sqrt{\frac{Np/Np-Nc/Np}{Np/Np-1/Np}}\right] = \frac{\sigma}{\sqrt{Nc}}$$

$$\lim_{Nc\to\infty} \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{Nc}}$$

La quantità $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{Nc}}$ è detta *errore standard* della media (secondo la notazione anglosassone SE: *standard error*).

Si può notare che:

1) $\sigma_{\bar{x}} < \sigma$, ovvero la deviazione standard della distribuzione delle medie campionarie è inferiore a quella della popolazione d'origine;

2) tanto maggiore è la variabilità della popolazione e tanto maggiore la variabilità delle medie campionarie;

3) la variabilità delle medie campionarie è tanto più piccola quanto più grande è il campione, ovvero all'aumentare della dimensione del campione aumenta la probabilità che la media del campione sia vicina a quella della popolazione (il campione diviene evidentemente più rappresentativo). Nel caso di misure ripetute riesco ad aumentare la precisione.

Distribuzione delle differenze campionarie

Date due popolazioni infinite X_A ed X_B caratterizzate da distribuzioni sconosciute. Si immagini di estrarre da queste due campioni, rispettivamente composti da Nc_A ed Nc_B elementi. Tali campioni avranno valori medi $\overline{x}_A$, e $\overline{x}_B$ e deviazioni standard $s_{\bar{x}A}$ e $s_{\bar{x}B}$.

E' naturalmente possibile calcolare la differenza tra tali valori medi ed in particolare possiamo immaginare di estrarre casualmente molte altre coppie di campioni dalle popolazioni A e B (al limite infinite: $\overline{x}_{A1}$, $\overline{x}_{A2}$, $\overline{x}_{A3}$... e $\overline{x}_{B1}$, $\overline{x}_{B2}$, $\overline{x}_{B3}$...), e di calcolare la relativa differenza tra le medie campionarie, ottenendo dunque una popolazione derivata di ampiezza infinita:

$\overline{x}_{A1} - \overline{x}_{B1}$, $\overline{x}_{A2} - \overline{x}_{B2}$, $\overline{x}_{A3} - \overline{x}_{B3}$, ..., $\overline{x}_{Ai} - \overline{x}_{Bi}$, ...

Per il teorema del limite centrale sappiamo che al crescere di Nc_A ed Nc_B la popolazione delle differenze campionarie, che possono appunto essere considerate come somme algebriche, tende ad essere distribuita normalmente con un valore medio ed una deviazione standard che è possibile esprimere in funzione dei corrispondenti parametri delle popolazioni di provenienza come:

$$\mu_{\bar{x}A-\bar{x}B} = \mu_{\bar{x}A} - \mu_{\bar{x}B} = \mu_A - \mu_B$$

$$\sigma_{\bar{x}A-\bar{x}B} = \sqrt{\sigma^2_{\bar{x}A} + \sigma^2_{\bar{x}B}} = \sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}}$$

La quantità $\sigma_{\bar{x}A-\bar{x}B}$ è detta *errore standard* per le differenze campionarie.

Risultati analoghi, anche se formalmente un poco più complessi, possono essere ottenuti nel caso di campionamento su popolazioni finite, utilizzando le relazioni viste al paragrafo precedente.

11. ESERCIZI SULLA TEORIA ELEMENTARE DEI CAMPIONI

Ex22

Trovare l'errore standard della media di un campione di 16 osservazioni la cui deviazione standard è risultata essere s=40.

Ex23

Ex24

Ex25

La produzione di un'azienda è costituita da una popolazione di polli da carne, caratterizzata da un peso medio di 1800 g ed una deviazione standard di 650 g.

Calcolare la probabilità che un lotto di 50 animali abbia un peso totale ***a)*** compreso tra 90 e 100 kg; ***b)*** superiore a 100 kg; ***c)*** inferiore a 90 kg.

(Notare che il peso medio del lotto vale evidentemente 1800g*50=90kg)

a) La popolazione di pesi (x) ha valor medio μ_x=1800 g, deviazione standard σ_x =650 g e distribuzione sconosciuta.

Invece la popolazione delle medie campionarie ($\bar{x}$) ha media $\mu_{\bar{x}} = \mu_x = 1800\ g$, deviazione standard $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{Nc}} = \frac{650}{\sqrt{50}} \cong 91.9\ g$ e distribuzione approssimativamente normale (poiché Nc=50>30).

Il problema viene affrontato cercando il seguente valore di probabilità:

$$Pr\{90'000\ g < 50 \cdot \bar{x} < 100'000\ g\},\ ovvero$$

$$Pr\{\frac{90'000\ g}{50} < \bar{x} < \frac{100'000\ g}{50}\} = Pr\{1800\ g < \bar{x} < 2000\ g\} =$$

$$= \int_{1800}^{2000} \varphi(\bar{x})\, d\bar{x} = \int_{z_1}^{z_2} \psi(z)\, dz$$

dove z_1 e z_2 rappresentano rispettivamente i valori 1800 e 2000 in unità standard:

$$z_1 = \frac{x_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \cong \frac{1800-1800}{91.9} = 0 \qquad z_2 = \frac{x_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \cong \frac{2000-1800}{91.9} \cong 2.17$$

Utilizzando le tabelle della curva normale standard si ottiene che l'area compresa tra z_1 = 0 e z_2 = 2.17 vale 0.485≈48% che è la probabilità richiesta.

b) la probabilità richiesta vale:

$$Pr\{100'000\ g < 50 \cdot \bar{x}\},\ ovvero$$

$$Pr\{\frac{100'000\ g}{50} < \bar{x}\} = Pr\{2000\ g < \bar{x}\} =$$

$$= \int_{2000}^{\infty} \varphi(\bar{x})\, d\bar{x} = \int_{z_1}^{z_2} \psi(z)\, dz$$

dove z_1 e z_2 rappresentano rispettivamente i valori 2000 ed ∞ in unità standard:

$$z_1 = \frac{x_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \cong \frac{2000-1800}{91.9} \cong 2.17 \qquad z_2 = \frac{x_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \cong \frac{\infty-1800}{91.9} \to \infty$$

dunque la probabilità cercata vale l'area a destra di z_1 = 2.17, ovvero p=0.015 = 1.5%

c) la probabilità richiesta vale:

$$Pr\{50\cdot\bar{x} < 90'000\ g\},\ ovvero$$

$$Pr\{\bar{x} < \frac{90'000\ g}{50}\} = Pr\{\bar{x} < 1800\ g\} =$$

$$= \int_{-\infty}^{1800} \varphi(\bar{x})\, d\bar{x} = \int_{z_1}^{z_2} \psi(z)\, dz$$

dove z_1 e z_2 valgono rispettivamente:

$$z_1 = \frac{x_1-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{-\infty-1800}{91.9} \to -\infty \qquad z_2 = \frac{x_2-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{1800-1800}{91.9} = 0$$

l'area a sinistra di $z_2 = 0$ vale 0.5, ovvero p=50%.

Ex26

Calcolare la probabilità che un cestino con 40 fragole pesi più di ***a)*** 400 g; ***b)*** 500 g; ***c)*** 600; ***d)*** 700g; sapendo che le fragole provengono da una popolazione, per la quale la variabile peso ha distribuzione sconosciuta, valore medio μ_x=15 g e deviazione standard σ=10 g.

a) la probabilità richiesta vale:

$$Pr\{400\ g < 40\cdot\bar{x}\}\ldots$$

Ex27

Una macchina destinata al riempimento automatico dovrebbe versare 240.0 cm^3 di birra in ogni bottiglia con uno scarto tipico di 15 cm^3. Il programma di manutenzione stabilisce che la macchina debba essere regolata quando la media campionaria di 30 bottiglie scelte a caso è inferiore a 235 o superiore a 245 cm^3. Qual è la probabilità di ottenere una media campionaria compresa entro tali limiti?

a) la probabilità richiesta vale:

$$Pr\{30\cdot 235\ cm^3 < 30\cdot\bar{x} < 30\cdot 245\ cm^3\}\ldots$$

L'esercizio si presta all'interpretazione delle *carte di controllo*, impiegate sia per i controlli di qualità all'accettazione dei semilavorati sia alle merci in uscita: non si può pretendere che tutti i lotti siano identici, ma quale ampiezza di variazione possiamo accettare? Con i metodi visti si calcola l'intervallo fiduciale, p.e. al 99%. Se un lotto in ingresso non rientra in questo intervallo allora viene rifiutato, se è in uscita allora è probabile che l'impianto necessiti di un intervento di manutenzione. Criteri analoghi sono utilizzati nella contrattualistica, nella redazione di norme tecniche e procedure relative alle politiche di controllo della qualità.

Ex 28

Estraiamo due campioni da una stessa popolazione e misuriamone i valori medi. A quale valore tende la differenza tra i due valori medi al crescere dell'ampiezza dei due campioni?

Ex29

Le mele di una certa azienda pesano in media 50 g, con una deviazione standard di 20 g. Calcolare la probabilità che due lotti contenenti 1000 frutti ciascuno differiscano in peso per più di 500 g.

Indichiamo con $\Delta\bar{x} = \bar{x}_A - \bar{x}_B$ il valore medio della differenza tra i pesi dei frutti dei due lotti. La probabilità da calcolare vale:

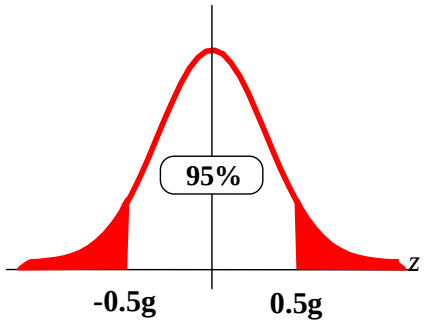
$$p = Pr\{1000\ |\Delta\bar{x}| > 500\ g\} = Pr\{|\Delta\bar{x}| > 0.5\ g\}$$

In altre parole una differenza globale di 500 g tra i due lotti equivale ad una differenza media di 0.5 g tra i valori medi campionari dei frutti.

Impostiamo il calcolo della probabilità, ricordando che, per il teorema del valore centrale, la variabile *differenza delle medie campionarie* tende ad essere distribuita normalmente, e che l'errore di approssimazione diviene trascurabile per campioni di elevata numerosità (in questo caso 1000>30).

Esplicitando il valore assoluto:

$$p = Pr\{\left(\Delta\bar{x}\cdot 1000 > 500\ g\right) \cup \left(\Delta\bar{x}\cdot 1000 < -500\ g\right)\}$$



$$p = \int_{-\infty}^{-0.5} \varphi(\Delta\bar{x})\, d\Delta\bar{x} + \int_{0.5}^{\infty} \varphi(\Delta\bar{x})\, d\Delta\bar{x} = 2\int_{z_1}^{z_2} \psi(z)\, dz$$

dove z_1 e z_2 valgono rispettivamente:

$$z_1 = \frac{0.5 - \mu_{\Delta\bar{x}}}{\sigma_{\Delta\bar{x}}} \qquad z_2 \to \infty$$

Dalla teoria sappiamo che il valor medio delle differenze campionarie vale:

$$\mu_{\Delta\bar{x}} = \mu_{\bar{x}A-\bar{x}B} = \mu_{\bar{x}A} - \mu_{\bar{x}B} = \mu_A - \mu_B$$

e poiché i campioni A e B provengono dalla medesima popolazione risulterà:

$$\mu_{\Delta\bar{x}} = \mu_A - \mu_A = 50\ g - 50\ g = 0\ g$$

e, per lo stesso motivo, la deviazione standard delle differenze campionarie, vale:

$$\sigma_{\Delta\bar{x}} = \sqrt{\sigma^2_{\bar{x}A} + \sigma^2_{\bar{x}B}} = \sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}} = \sqrt{\frac{20^2}{1000} + \frac{20^2}{1000}} \cong 0.894$$

e dunque, sostituendo nell'espressione di z_1:

$$z_1 = \frac{0.5 - \mu_{\Delta\bar{x}}}{\sigma_{\Delta\bar{x}}} = \frac{0.5-0}{0.894} \cong 0.56$$

L'area della curva normale standardizzata compresa tra i punti di ascissa 0.56 e $+\infty$, vale circa 0.288, e dunque la probabilità richiesta vale $p \cong 2*0.288 \cong 60\%$.

Calcolare quale dovrebbe essere la differenza minima per portare tale livello di probabilità sotto il 5%.

Ex30

Una popolazione di frutti ha un peso medio pari a 60 g, e deviazione standard di 18 g (con distribuzione sconosciuta). Calcolare la probabilità che due contenitori alveolati da 20 bacche differiscano in peso per più di 250 g. Calcolare poi la probabilità che due campioni casuali di 20 bacche differiscano in peso per più del 20%.

La probabilità da calcolare vale:

$$p = Pr\{20\ |\Delta\bar{x}| > 250\ g\}\ \ldots$$

e $p = Pr\{20\ |\Delta\bar{x}| > 0.20 \cdot 60 \cdot 20\ g\}\ldots$

Ex31

Una popolazione di bovini è caratterizzata da un peso medio pari a 400 kg, ed una deviazione standard di 80 kg. Determinare qual è la probabilità che due gruppi, composti da Nc animali ***a)*** differiscano in peso per più di 100 kg o ***b)*** differiscano in peso per più del 10%. Considerare Nc=1, 2, 5, 10, 20, 50.

La probabilità da calcolare vale:

$$p = Pr\{Nc\ |\Delta\bar{x}| > 100\ kg\}\ldots$$

$$p = Pr\{Nc\ |\Delta\bar{x}| > 0.1*400*Nc\}\ldots$$

Ripetere poi l'esercizio ***a*** supponendo invece che i due campioni di bovini provengano da due popolazioni differenti (A e B) per le quali risulti μ_A=350 kg, μ_B=450 kg, σ_A=σ_B=80 kg.

12. TEORIA STATISTICA DELLA STIMA

Stima dei parametri della popolazione da quelli campionari

La ricerca delle relazioni tra i parametri statistici della popolazione (valori *veri*, ma incogniti, come indici di posizione e dispersione) e quelli del campione (valori osservati) è oggetto della *teoria statistica della stima.*

A partire dai dati particolari misurati (osservazioni o statistiche campionarie) possiamo cercare di valutare il complesso di tutti i dati misurabili (popolazione). Non sarà generalmente possibile calcolare i valori veri, ma solo valori approssimati, compresi in un certo intervallo, con una determinata probabilità. In questo caso si dice che abbiamo *stimato* i parametri della popolazione dalla quale provengono i campioni.

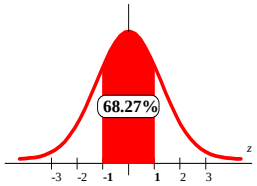
Intervallo di confidenza per la stima del valor medio

Ricordando che le medie campionarie tendono a distribuirsi normalmente all'aumentare dell'ampiezza del campione, possiamo considerarne la forma standardizzata:

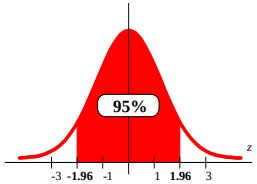
$$z = \frac{\overline{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{N_c}}$$

la variabile aleatoria continua z tende anch'essa a distribuirsi normalmente, e come sappiamo, risulta:

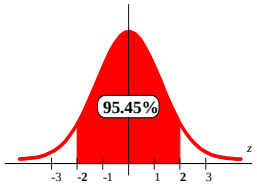
$$Pr\{-1<z<+1\} = \int_{-1}^{+1} \psi(z)\, dz \cong 68.27\%$$



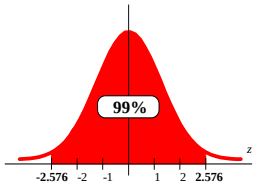
$$Pr\{-1.96<z<+1.96\} = \int_{-1.96}^{+1.96} \psi(z) dz \cong 95\%$$



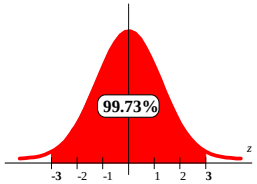
$$Pr\{-2<z<+2\} = \int_{-2}^{+2} \psi(z)\, dz \cong 95.45\%$$



$$Pr\{-2.576<z<2.576\} = \int_{-2.576}^{+2.576} \psi(z) dz \cong 99\%$$



$$Pr\{-3<z<+3\} = \int_{-3}^{+3} \psi(z)\, dz \cong 99.73\%$$



dunque, per le caratteristiche della curva normale standardizzata, possiamo affermare, ad esempio col 95% di probabilità che:

$$-1.96 \le z \le +1.96$$

allora possiamo immediatamente scrivere che:

$$-1.96 \le \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \le +1.96$$

e dunque, esplicitando il valore medio vero:

$$\bar{x} - 1.96\sigma_{\bar{x}} \leq \mu_{\bar{x}} \leq \bar{x} + 1.96\sigma_{\bar{x}}$$

tale disequazione viene poi generalmente scritta in modo un po' diverso:

$$\mu_{\bar{x}} = \bar{x} \pm 1.96\sigma_{\bar{x}}$$

E' dunque trovato, in funzione del dato campionario $\bar{x}$ (valori osservati), l'intervallo entro il quale ricade il valore vero $\mu_{\bar{x}}$ (sconosciuto) della popolazione, con il 95% di probabilità, ovvero l'**intervallo di confidenza** al 95% per la stima del valore medio vero della popolazione.

Il valore 1.96 è relativo ad un livello di probabilità del 95%, ma evidentemente i limiti fiduciali si possono calcolare in corrispondenza di un qualsiasi livello di probabilità, per esempio al 99%, impiegando il coefficiente 2.576.

I valori [-1,+1], [-1.96,+1.96], [-2.58,+2.58] si chiamano valori critici di z rispettivamente al 68.27%, al 95%, al 99%; genericamente li si indica con il simbolo z_c, o con $z_c(p)$, cioè come valori di z funzioni di un determinato livello di probabilità p. L'intervallo di confidenza può dunque essere sinteticamente indicato come:

$$\mu_x = \bar{x} \pm z_c(p)\sigma_{\bar{x}} = \bar{x} \pm z_c(p)\frac{\sigma_x}{\sqrt{Nc}}$$

livello di confidenza (p%)	99.73	**99**	98	95.45	**95**	90	80	68.27	50
$z_c(p$%)	**3.00**	2.58	2.33	**2.00**	1.96	1.64	1.28	**1.00**	0.67

E dunque, secondo tale tabellina, risulta:

- il valor medio vero della popolazione (μ_x), è compreso tra $\bar{x}$-$\sigma_{\bar{x}}$ e $\bar{x}$+$\sigma_{\bar{x}}$ con una probabilità pari al 68.27%;
- il valor medio vero della popolazione (μ_x), è compreso tra $\bar{x}$-*1.96*$\sigma_{\bar{x}}$ e $\bar{x}$+*1.96*$\sigma_{\bar{x}}$ con una probabilità pari al 95%;
- il valor medio vero della popolazione (μ_x), è compreso tra $\bar{x}$-*2.58*$\sigma_{\bar{x}}$ e $\bar{x}$+*2.58*$\sigma_{\bar{x}}$ con una probabilità pari al 99%.

Stima della varianza campionaria

Occorre far notare come per poter stimare il valore medio vero della popolazione μ_x, sulla base di quello campionario $\bar{x}$, occorre conoscere il valore della deviazione standard della popolazione σ_x.

σ_x non è sempre nota, è tuttavia possibile stimare anche tale valore sulla base del dato campionario s_x, purché il numero Nc di elementi del campione sia abbastanza grande (approssimativamente superiore a 30):

$$\sigma_x \approx s_x \cdot \sqrt{\frac{Nc}{Nc\text{-}1}}$$

Allora la formula per la stima del valore medio μ_x della popolazione, sulla base delle sole osservazioni campionarie risulta:

$$\mu_x = \bar{x} \pm z_c(p)\cdot\frac{\sigma_x}{\sqrt{Nc}} \approx \bar{x} \pm z_c(p)\cdot\frac{s_x}{\sqrt{Nc}}\cdot\sqrt{\frac{Nc}{Nc\text{-}1}} = \bar{x} \pm z_c(p)\cdot\frac{s_x}{\sqrt{Nc\text{-}1}}$$

Riassumendo: per prevedere con una determinata probabilità i limiti entro i quali cade la media di una popolazione della quale ignoriamo sia la media che la varianza, occorre:

- prelevare dalla popolazione di cui si ignora la distribuzione di probabilità, con la tecnica del campionamento casuale, un campione X di dimensione opportuna ($Nc > 30$);
- fissare un livello di confidenza (p.e. 95% a cui corrisponde $z_c \approx 1.96$, oppure 99% a cui corrisponde $z_c \approx 2.58$);
- calcolare media $\bar{x}$ e deviazione standard s_x del campione;
- calcolare i limiti di confidenza (o limiti fiduciali) della media, corrispondenti al livello di probabilità preassegnato: $\mu_x = \mu_{\bar{x}} = \bar{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc\text{-}1}}$

Intervalli di confidenza per la stima delle differenze delle medie

Poichè anche le differenze tra le medie campionarie $S=\Delta\overline{x}=\overline{x}_A-\overline{x}_B$ tendono a distribuirsi normalmente, possono anch'esse essere standardizzate e trattate come una variabile gaussiana. In particolare i limiti di confidenza per la stima della differenza tra le medie delle popolazioni A e B $\mu_A-\mu_B$, nota quella campionaria, sono dati da:

$$\mu_A-\mu_B = (\overline{x}_A - \overline{x}_B) \pm z_t(p)\cdot\sigma_{\overline{x}A-\overline{x}B}$$

essendo, come è già stato visto,

$$\sigma_{\overline{x}A-\overline{x}B} = \sqrt{\sigma^2_{\overline{x}A} + \sigma^2_{\overline{x}B}} = \sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}}$$

Anche in questo caso, per poter stimare il valore medio vero relativo alla popolazione, sulla base delle sole osservazioni campionarie, occorre conoscere i valori della deviazione standard delle popolazioni σ_{xA} e σ_{xB}, che non sono sempre noti a priori: purché la numerosità dei campioni sia sufficiente (approssimativamente superiore a 30), è possibile stimare le deviazioni standard della popolazione semplicemente *correggendo* quelle calcolate sui campioni:

$$\sigma_{xA} \approx s_{xA}\cdot\sqrt{\frac{N_{cA}}{N_{cA}-1}} \qquad \sigma_{xB} \approx s_{xB}\cdot\sqrt{\frac{N_{cB}}{N_{cB}-1}}$$

e dunque risulta:

$$\sigma_{\overline{x}A-\overline{x}B} = \sqrt{\sigma^2_{\overline{x}A} + \sigma^2_{\overline{x}B}} = \sqrt{\frac{\sigma_{xA}{}^2}{Nc_A} + \frac{\sigma_{xB}{}^2}{Nc_B}} \approx \sqrt{\frac{s_{xA}{}^2}{Nc_A-1} + \frac{s_{xB}{}^2}{Nc_B-1}}$$

Considerazioni aggiuntive

Ora siamo in grado di dimostrare la relazione che lega la distribuzione delle differenze campionarie alla distribuzione delle popolazioni di partenza: consideriamo la stima del valor medio vero di due popolazioni x_1 ed x_2:

$$\mu_1 = \overline{x}_1 \pm E_1 \qquad \mu_2 = \overline{x}_2 \pm E_2$$

con

$$E_1 = z_c\frac{\sigma_1}{\sqrt{N_{c1}}} \qquad E_2 = z_c\frac{\sigma_2}{\sqrt{N_{c2}}}$$

con z_t corrispondente ad un livello qualsiasi di probabilità.

Considerando tutti gli infiniti valori di z_c, risulta:

$$\Delta_1 = (\overline{x}_1 + E_1) - (\overline{x}_2 + E_2)$$
$$\Delta_2 = (\overline{x}_1 + E_1) - (\overline{x}_2 - E_2)$$
$$\Delta_3 = (\overline{x}_1 - E_1) - (\overline{x}_2 + E_2)$$
$$\Delta_4 = (\overline{x}_1 - E_1) - (\overline{x}_2 - E_2)$$

il valor medio delle differenze risulta:

$$\mu_\Delta = \frac{4\overline{x}_1 - 4\overline{x}_2}{4} = \overline{x}_1 - \overline{x}_2$$

e per quanto riguarda la deviazione standard si ottiene:

$$\sigma_\Delta = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(\Delta_i - \mu_\Delta)} =$$

$$\sqrt{\frac{1}{4}\left[(E_1 - E_2)^2 + (E_1 + E_2)^2 + (-E_1 - E_2)^2 + (-E_1 + E_2)^2\right]} =$$

$$\sqrt{\frac{1}{4}\left[2\cdot(E_1 - E_2)^2 + 2\cdot(E_1 + E_2)^2\right]} =$$

$$\sqrt{\frac{1}{4}\left[E_1^2 + 2E_1E_2 + E_2^2 + E_1^2 - 2E_1E_2 + E_2^2\right]} = \sqrt{\left[E_1^2 + E_2^2\right]}$$

13. ESERCIZI SULLA TEORIA STATISTICA DELLA STIMA

Ex32

Trovare l'intervallo di confidenza al 95% della media della popolazione dalla quale è stato prelevato il seguente campione di dati: 52, 48, 46, 41, 40, 37, 37, 32, 26, 24. (Nota: data la piccola ampiezza campionaria, ipotizziamo in prima approssimazione, che la popolazione di origine abbia distribuzione approssimativamente normale).

$$\overline{x} \approx 38.44 \quad Sx \approx 9.12 \quad Nc=10$$

$$\mu_x = \overline{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} \approx 38.44 \pm 1.96 \cdot \frac{9.12}{\sqrt{10-1}} \approx 38.44 \pm 6.96$$

Ex33

Le misure dei diametri medi di un campione casuale di 40 frutti, estratto da un determinato lotto di produzione, hanno fornito una media di 82 mm ed uno scarto quadratico medio di 12 mm. Determinare i limiti di confidenza al 95% ed al 99% per il diametro medio di tutti i frutti del lotto:

per i limiti al 95% risulta:

$$\overline{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} = 82 \pm 1.96 \cdot \frac{12}{\sqrt{40-1}} = 82 \pm 3.8 \ mm$$

e per i limiti al 99% risulta:

$$\overline{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} = 82 \pm 2.58 \cdot \frac{12}{\sqrt{40-1}} = 82 \pm 5.0 \ mm$$

Ex34

Stimare il peso medio dei polli di un allevamento che conta N_p=100'000 esemplari.

Piuttosto che effettuare centomila operazioni di pesatura si sceglie un campione casuale di N_c=50 animali e se ne determina il valor medio che risulta $\overline{x}$=2 kg con scarto quadratico medio Sx =0.5 kg. Vediamo allora quattro differenti modi, egualmente corretti, utilizzabili per esprimere sinteticamente il risultato della misura:

con il 99.73% di probabilità il valore cercato del peso medio vero sarà compreso nell'intervallo:

$$\overline{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} = 2 \pm 3 \cdot \frac{0.5}{\sqrt{50-1}} = \mathbf{2 \pm 0.214\ kg};$$

oppure:

$$\overline{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} = 2 \pm 3 \cdot \frac{0.5}{\sqrt{50-1}} = \mathbf{2\ kg \pm 214\ g};$$

oppure:

$$\overline{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} = 2 \pm 3 \cdot \frac{0.5}{\sqrt{50-1}} = \mathbf{2\ kg \pm 10.7\%};$$

oppure:

$$\mu_x \in [\mathbf{1.8 \div 2.2\ kg}].$$

Occorre notare come possa essere sufficiente semplicemente fornire il valore medio campionario con la deviazione standard. Tuttavia in tale forma, pur contenendo tutta l'informazione necessaria, risulta evidentemente più difficile da interpretare ed in definitiva meno espressivo, soprattutto in ambito industriale.

Ex35

Empio *95% confidence limit* dal menu DEMO del programma *Winks*.

http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/

Ex36

Le pere di una partita di 10'000 esemplari hanno un peso distribuito quasi normalmente, con valore medio pari a 120 g e deviazione standard di 30 g. Stimare (attraverso la determinazione di un intervallo fiduciale) il peso medio delle confezioni contenenti ciascuna 4 pere, con probabilità del 95%.

La popolazione di pesi (x) ha valor medio μ_x=120 g, deviazione standard σ_x =30 g e distribuzione approssimativamente normale.

Invece la popolazione delle medie misurate su campioni di numerosità 4 ($\bar{x}$) ha media $\mu_{\bar{x}} = \mu_x = 120\,g$, deviazione standard $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{Nc}} = \frac{30}{\sqrt{4}} = 15\,g$ e distribuzione approssimativamente normale (Nc<<30, ma la popolazione di partenza è tendenzialmente normale); allora i limiti al 95% del peso delle confezioni risultano:

$$\mu_{\bar{x}} = 4\left(\mu_x \pm z_c \cdot \sigma_{\bar{x}}\right) = 4\left(\mu_x \pm z_c \cdot \frac{\sigma_x}{\sqrt{Nc}}\right) = 4\left(120 \pm 1.96 \cdot \frac{30}{\sqrt{4}}\right) = 480 \pm 117.6\ kg$$

Ex37

Ad una cantina sociale arriva una grande quantità di uve differenti. Stabilire una procedura scientificamente corretta per la stima del grado zuccherino medio di tutta la quantità d'uva pervenuta.

Ex38

Nel corso dell'anno 1999, in alcuni giorni scelti a caso, la produzione giornaliera di detersivo in polvere è stata la seguente (in tonnellate): 74.4, 85.2, 88.2, 86.1, 82.6, 90.1, 93.9, 75.3, 79.8, 73.2, 77.9, 89.8. Stabilire, con una probabilità del 95%, l'intervallo di confidenza per stimare la produzione totale annua.

Ex39

Per un campione di 50 animali la dose efficace di un farmaco è stata trovata variabile da individuo a individuo, con distribuzione approssimativamente normale caratterizzata da un valore medio di 100 mg ed una deviazione standard s=20 mg. Si determini con un'affidabilità del 99%: ***a)*** la quantità di farmaco necessaria per fronteggiare un traffico annuo previsto di 1000 animali; ***b)*** la dose massima prevedibile.

Si confronti infine la quantità calcolata al punto ***b*** moltiplicata per 1000 con la quantità calcolata al punto ***a***.

a) Per stimare la dimensione della scorta occorre stimare l'intervallo fiduciale al 99% della dose media di anestetico:

$$\bar{x} \pm z_c \cdot \frac{s_x}{\sqrt{Nc-1}} = 100 \pm 2.58 \cdot \frac{20}{\sqrt{50-1}} = 100 \pm 7.37 mg$$

ovvero risulta che il valore medio della dose di anestetico per paziente dovrebbe risultare compreso nell'intervallo 92.6÷107.4 mg con il 99% di affidabilità. Dunque per 1000 animali ordineremo una quantità di principio certamente non inferiore a 1000·92.6 ≅ 92.6 kg, ma volendo essere maggiormente cautelativi ordineremo una quantità pari a 1000·107.4 ≅ 107.4 kg.

b) Dal campione di 50 individui è possibile stimare i parametri della popolazione, successivamente si tratta la variabile dose efficace di farmaco (x) come normale:

$$\mu_x \cong \bar{x} = 100\ mg \quad \text{e}$$

$$\sigma_x \cong \frac{s_x}{\sqrt{Nc-1}}\sqrt{Nc} = s_x\sqrt{\frac{Nc}{Nc-1}} = 20\sqrt{\frac{50}{50-1}} \cong 20.2$$

e dunque la dose massima prevedibile, con il 99% di probabilità, risulta:

$$x_{max} = \mu_x + z_{c(99\%)} \cdot \sigma_x = 100\ mg + 2.58 \cdot 20.2\ mg \cong 152\ mg$$

Volendo essere più cautelativi si può scegliere come valore stimato della media della popolazione il valore superiore dell'intervallo di confidenza per la stima del valor medio, ottenendo così un valore un poco più alto della dose massima prevedibile:

$$x_{max} = 107.4\ mg + 2.58 \cdot 20.2\ mg \cong 160\ mg$$

Ex40

Il tempo d'attesa medio alle casse di un supermercato, misurato su di un campione di 30 persone, è stato pari a 350 s, con una deviazione standard di 150 s. Stimare il tempo d'attesa medio, e quello massimo con probabilità del 95%, e del 99%, ipotizzando che la variabile aleatoria tempo d'attesa sia distribuita normalmente. Calcolare inoltre la probabilità di rimanere accodati per un tempo superiore a 5 minuti. Ripetere il calcolo supponendo di avere effettuato le misure su un campione di 90 persone.

Ex41

Ad un'azienda che lavora frutta biologica arriva una quantità di 1000 kg di prodotto. Un campione casuale di 10 kg viene prelevato e se ne determina il residuo di fitofarmaco: il valore medio risulta pari a 20 μg/kg, e la deviazione standard pari a 5 μg/kg. Determinare la quantità totale di fitofarmaco del lotto in arrivo, con un'affidabilità del 95%. Determinare inoltre la probabilità che su di un campione casuale di 1 kg si trovi un residuo di fitofarmaco superiore a 35 μg, nell'ipotesi che la variabile aleatoria residuo di fitofarmaco sia distribuita normalmente.

Ex42

Stimare la quantità di rete necessaria ad una macchina per impacchettare 10'000 confezioni di limoni, posto che su 20 confezioni è stato misurato un nastro di lunghezza media pari a 0.5 m/confezione con s=0.15 m.

Ex43

Una popolazione di bovini è caratterizzata da un peso medio pari a 400 kg, ed una deviazione standard di 80 kg. Determinare l'intervallo fiduciale per la differenza di peso tra due campioni di animali, estratti dalla medesima popolazione, di numerosità compresa tra 1 e 1000, con probabilità del 95%.

In questo caso sono noti i valori della popolazione, e si vogliono invece stimare quelli campionari. Allora dalla formula per il calcolo dell'intervallo di confidenza relativo alla differenza tra le medie campionarie ricaviamo:

$$(\bar{x}_A - \bar{x}_B) = \mu_A - \mu_B \pm z_t(p) \cdot \sigma_{\bar{x}A - \bar{x}B}$$

con $\sigma_{\bar{x}A-\bar{x}B} = \sqrt{\sigma^2_{\bar{x}A} + \sigma^2_{\bar{x}B}} = \sqrt{\frac{\sigma_{xA}^{\ 2}}{Nc_A} + \frac{\sigma_{xB}^{\ 2}}{Nc_B}}$

allora, sostituendo i numeri ai simboli, risulta:

$$\sigma_{\bar{x}A-\bar{x}B} = \sqrt{\frac{80^2}{Nc} + \frac{80^2}{Nc}} \quad \text{e} \quad (\bar{x}_A - \bar{x}_B) = (400 - 400) \pm 1.96 \cdot \sigma_{\bar{x}A - \bar{x}B}$$

Nc	$\sigma_{\bar{x}A-\bar{x}B}$	differenza tra i pesi medi di due campioni al 95% $\lvert\bar{x}_A - \bar{x}_B\rvert$	Valore medio della differenza tra i pesi totali di due campioni $Nc\lvert\bar{x}_A - \bar{x}_B\rvert$	Valore medio della differenza relativa tra i pesi di due campioni $\frac{Nc\lvert\bar{x}_A - \bar{x}_B\rvert}{Nc \cdot 400}$
1	113,14	221,7	221,7	0,554 (55%)
2	80,00	156,8	313,6	0,392 (39%)
5	50,60	99,2	495,8	0,248 (25%)
10	35,78	70,1	701,2	0,175 (17%)
20	25,30	49,6	991,7	0,124 (12%)
30	20,66	40,5	1214,6	0,101 (10%)
60	14,61	28,6	1717,7	0,072 (7%)
100	11,31	22,2	2217,5	0,055 (5%)
200	8,00	15,7	3136,0	0,039 (4%)
400	5,66	11,1	4435,0	0,028 (3%)
1000	3,58	7,0	7012,3	0,018 (2%)

- Osservazioni sulla *minima differenza significativa*: se due campioni di 30 animali differiscono in peso per meno del 10%, allora posso concludere con il 95% di probabilità che i due campioni sono prelevati dalla medesima popolazione. Viceversa, se la differenza di peso risultasse superiore, allora ciò potrebbe essere dovuto al caso solo per un 5% di probabilità. Ovvero, con probabilità del 95%, i due campioni rappresentano popolazioni di animali differenti (p.e. alimentati in maniera differente).

E' da notare dunque che in un esperimento con differenti razioni alimentari non basta osservare il fatto che un gruppo di animali abbia un peso medio superiore all'altro per concludere che esiste un effetto significativo indotto della diversità nel regime alimentare. Esistono infatti differenze anche tra campioni selezionati all'interno di una medesima popolazione. Ciò è evidentemente il frutto della normale diversità tra individui.

Ex44

Le pere di un ampio lotto frutti hanno un peso distribuito quasi normalmente, con valore medio pari a 120 g e deviazione standard di 30 g. Stimare la massima differenza di peso tra due confezioni contenenti ciascuna 20 pere, con probabilità del 95%. [Occorre determinare i parametri della popolazione di differenze campionarie, relativa a campioni di ampiezza 20]

Ex45

Stimare la differenza in peso tra due cestini, contenenti ciascuno 40 fragole, provenienti dalla medesima popolazione con livelli di probabilità del 10%, 50%, 90%, 95%, 99%. Le fragole provengano da un vasto lotto, nel quale la variabile peso ha distribuzione sconosciuta, valore medio μ_x=15 g e deviazione standard σ=10 g.

14. DETERMINAZIONE DELL'AMPIEZZA CAMPIONARIA

Ex46 Roletto 5.11

Si vuole stimare il livello medio dell'inquinamento da metalli pesanti dei reflui di un impianto zootecnico.

Viene ammesso un errore di stima massimo di ±5 μg/m^3 al livello del 95%. Quale deve essere la dimensione (minima) *Nc* del campione che soddisfa queste condizioni? Si tenga conto del fatto che la dispersione delle misure, predeterminata su un piccolo campione, si quantifica in uno scarto quadratico medio stimato σ=16μg/m^3?

I limiti di confidenza della stima, per il valore medio di inquinamento, sono dati in questo caso da

$$\bar{x} - z_\epsilon \frac{\sigma}{\sqrt{Nc}} \text{ e da } \bar{x} + z_\epsilon \frac{\sigma}{\sqrt{Nc}}$$

L'ampiezza di tale fascia d'incertezza vale dunque

$$\Delta = LimSup - LimInf = \left(\bar{x} + z_\epsilon \frac{\sigma}{\sqrt{Nc}}\right) - \left(\bar{x} - z_\epsilon \frac{\sigma}{\sqrt{Nc}}\right) = 2 \cdot z_\epsilon \frac{\sigma}{\sqrt{Nc}}$$

e deve essere contenuta nel margine di errore assegnato di ±5μg/m^3=10μg/m^3.

Allora assegnato il livello di confidenza del 95%, z_ϵ = 1.96 per cui:

$$\Delta = 10 \rightarrow 2 \cdot z_\epsilon \frac{\sigma}{\sqrt{Nc}} = 10 \rightarrow 2 \cdot z_\epsilon \frac{s}{\sqrt{Nc-1}} \cong 10 \rightarrow 2 \cdot 1.96 \frac{16}{\sqrt{Nc-1}} \cong 10$$

Da questa relazione si ricava dunque il valore cercato di *Nc*

$$Nc \cong \left(\frac{2 \cdot z_\epsilon \cdot s}{\Delta}\right)^2 + 1 = \left[2 \cdot z_\epsilon \cdot \left(\frac{s}{\Delta}\right)\right]^2 + 1 = \left(\frac{2 \cdot 1.96 \cdot 16}{10}\right)^2 + 1 \cong 40$$

La relazione ha validità generale, e può essere anche espressa per mezzo del rapporto r=Δ/s, (concettualmente simile al coefficiente di variazione) definito come rapporto tra l'incertezza ammessa nella stima del valor medio della popolazione, e l'errore tipico su ciascun campione di rilevazioni:

$$Nc \cong \left(\frac{2 \cdot z_\varepsilon \cdot s}{\Delta}\right)^2 + 1 = \left(\frac{2 \cdot z_\varepsilon}{r}\right)^2 + 1$$

per alcuni valori tipici, con riferimento ai livelli di confidenza del 95% (z_ε=1.96) e del 99% (z_ε=2.58) otteniamo:

r	***Nc 95%***	***Nc 99%***
0.2	385	667
0.5	62	107
1	16	28
1.5	8	13
2	5	8
2.5	3	5

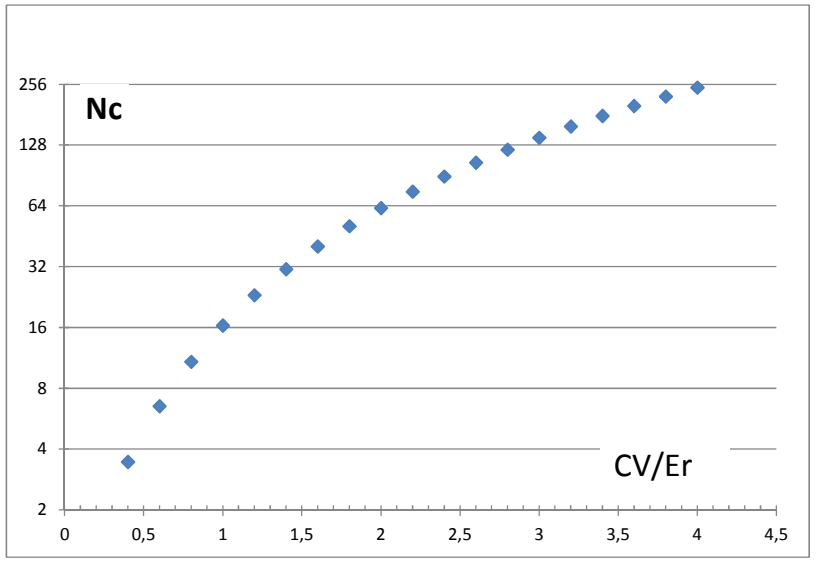
In alternativa è possibile utilizzare una formulazione modificata, derivata supponendo di dividere numeratore e denominatore per il valore medio della misura *m*:

$$Nc \cong \left(\frac{2 \cdot z_\varepsilon \cdot s/m}{\Delta/m}\right)^2 + 1 = \left(2 \cdot z_\varepsilon \cdot \frac{CV}{Er}\right)^2 + 1$$

Essendo *CV* il coefficiente di variazione che esprime la dispersione del dato ed *Er* l'errore ammissibile di stima, inteso come frazione della media.

Segue una tabellina di valori esemplificativi:

95%	***Er***		
CV	**0,05**	**0,1**	**0,15**
0,005	1	1	1
0,01	2	1	1
0,05	16	5	3
0,1	62	16	8
0,15	139	36	16
0,2	247	62	28
0,3	554	139	62
0,4	984	247	110
0,5	1538	385	172



La formulazione descritta, proprio a causa delle incertezze nella stima della varianza della popolazione, viene comunque più spesso utilizzata in fase di verifica dell'ampiezza campionaria, piuttosto che per la fase vera e propria di progetto di un esperimento. E' in ogni caso possibile verificare nuovamente l'adeguatezza di *Nc*, dopo avere osservato il campione.

Ex47

La Direzione Commerciale di una cooperativa di distribuzione che riunisce *Np*=2765 punti vendita piccoli e medi desidera conoscere a quanto ammontino le ordinazioni annuali di un prodotto di largo consumo. A questo scopo organizza una pre-indagine in 50 punti, a seguito della quale si osserva un valore medio, espresso in migliaia di euro, di 320 ed una deviazione standard di 90.

Sulla base di questa pre-indagine, quale dovrebbe essere la dimensione di una rete di campionamento fissa per avere una precisione di ±10k€ ad un livello fiduciale del 95%?

Possiamo scrivere:

$$10 \cdot 2 = 1.96 \cdot 2 \frac{\sigma}{\sqrt{Nc}}$$

da cui, stimando σ per mezzo del valore di s=90 ottenuto con il pre-sondaggio, otteniamo Nc=311.

Il valore di Nc non sembra trascurabile rispetto alla numerosità della popolazione (Np=2765) e difatti utilizzando la formulazione meno approssimata, relativa ad una popolazione finita, otteniamo:

$$2 \cdot 10 = 2 \cdot 1.96 \cdot \frac{\sigma}{\sqrt{Nc}} \sqrt{\frac{Nc - Np}{Nc - 1}} \quad \textit{da cui } \mathrm{Nc}=279$$

Ex48

Un agronomo desidera stabilire, con una precisione di ±0.5 mg e con un livello di fiducia del 99%, il contenuto medio in ferro (mg/kg di suolo secco) di un terreno agricolo che deve essere destinato alla coltura di spinaci. Da esperienze precedenti sa che la deviazione standard di questa variabile vale circa 1.3 mg. Da quante osservazioni deve essere costituito il campione per ottenere la precisione desiderata? Quale dovrebbe essere la dimensione del campione per un livello fiduciale del 95%?

$$Nc = \left(\frac{2 z_c s}{\Delta}\right)^2 + 1 = \left(\frac{2 \cdot 1.96 \cdot 1.3}{1}\right)^2 + 1 = 27$$

Ex49

Il titolare di un'impresa agricola, specializzata nella produzione di patate, pensa di migliorare la qualità del prodotto. A tal fine vuole stimare con un livello fiduciale del 99% il tempo medio che intercorre tra la semina ed il raccolto di una nuova varietà. Una piccola campagna di precampionamento su pochi esemplari ha fornito un valore di deviazione standard pari a 5 giorni. Quale dovrebbe essere l'ampiezza campionaria minima per eseguire la stima con una precisione di ±2 giorni ?

$$Nc = \left(\frac{2 z_c s}{\Delta}\right)^2 + 1 = \left(\frac{2 \cdot 2.58 \cdot 5}{4}\right)^2 + 1 = 43$$

Ex50

Per valutare la concentrazione del piombo nel sangue degli allievi di una scuola, si sono scelti con una tecnica campionaria adeguata 50 allievi. I risultati sono i seguenti: media 10.1 ng/cm^3; deviazione standard 0.6 ng/cm^3. Calcolare i limiti fiduciali al 95% per il contenuto medio di piombo nel sangue di tutti gli allievi della scuola. Indicare inoltre quale avrebbe dovuto essere la dimensione del campione per ridurre l'ampiezza dei limiti fiduciali a ±0.1 ng/cm^3.

$$Nc = \left(\frac{2 z_c s}{\Delta}\right)^2 + 1 = \left(\frac{2 \cdot 1.96 \cdot 0.6}{0.2}\right)^2 + 1 = 139$$

15. IL TRATTAMENTO STATISTICO DELLE MISURE

Regole di scrittura delle quantità numeriche (a.a.)

Una chiacchierata introduttiva ed un esempio

Esempio: ha senso chiedere ad un fornitore una macchina in grado di tagliare un nastro lungo un metro? Si voglia p.e. misurare la lunghezza di un certo oggetto. La definizione operativa di misura implica il confronto tra l'oggetto da misurare ed un campione del metro. Da tale confronto potrà risultare che la lunghezza *l* dell'oggetto è inferiore al metro, vale a dire:

$0 < l < 1$ m

Supponiamo di ottenere dal confronto dell'oggetto con i vari sottomultipli del metro la seguente successione di risultati:

0.2m < *l* < 0.3m, se eseguiamo la misura con la precisione del decimetro;

0.24m < *l* < 0.25m, se eseguiamo la misura con la precisione del centimetro;

0.247m < *l* < 0.248m, se eseguiamo la misura con la precisione del millimetro;

0.2473m < *l* < 0.2474m, avendo assunto di poter risolvere il decimo di millimetro.

Siamo interessati a capire fino a quale punto possiamo spingere questo procedimento: è facile convincersi che non potremo giungere a determinare la lunghezza che ci interessa come un determinato numero reale. A mano a mano che proviamo a determinare meglio la nostra misura incontriamo nuove difficoltà: inizialmente sarà la rugosità delle superfici, poi la dilatazione termica, la lunghezza d'onda finita della luce con la quale si illumina l'oggetto, fino ad arrivare, assumendo di poter utilizzare un microscopio immaginario per il confronto, a problemi legati alla natura non continua della materia.

Ma prima ancora di arrivare a questi limiti concettuali sorge il dubbio se veramente il nostro campione è lungo 1 metro, ovvero si dovrà affrontare il problema della riproducibilità e costanza del campione di misura. In definitiva non è difficile convincersi che il meglio che si potrà fare sarà dire che la lunghezza di interesse è compresa fra due valori:

$l_{min} < l < l_{max}$ (oppure $l = l_{medio} \pm \Delta l$)

A questo punto sorgono immediate delle domande:

Quale significato dobbiamo attribuire all'espressione *essere compresa*? É sempre vera ?

Se si effettua un secondo esperimento e si trova che i due intervalli l_{min}-l_{max} differiscono cosa succede? A quale dei due credere ?

E' possibile trattare le misure fisiche con metodi statistici, cioè parlare della *probabilità* che il valore vero della quantità fisica misurata sia compreso in un certo intervallo.

Quando si dice che una molecola di azoto è formata da 2 atomi di azoto, che un libro ha 100 pagine o che 1 metro corrisponde a 1000 mm, i numeri 2, 100, 1 e 1000 sono esatti.

Quando invece i numeri sono il risultato di una misura fisica (p.e. volume, massa, temperatura) essi non sono esatti, perciò devono essere riportati con un numero di cifre né maggiore né minore di quello necessario per esprimere l'accuratezza della misura fisica. Questo numero di cifre è detto numero di cifre significative.

Cifre significative

Si ammetta di pesare un corpo con una bilancia che possiede la sensibilità di un milligrammo. Il numero di cifre con le quali si deve riportare il risultato della pesata deve mostrare che la misura è stata fatta con l'approssimazione di 1 mg. Sia ad esempio il risultato 1.245g: scritto con 4 cifre indica che le prime tre sono esatte, mentre la quarta (il 5) è stata ottenuta per approssimazione al mg più vicino, e dunque il valore della misura sarà dunque compreso tra 1.244g e 1.246g.

Spesso si scrive 1.245±0.001*g*, ed un modo alternativo di scrivere può essere 1245±1 mg.

Se i numeri provengono da misure, dunque, sono espressione della precisione propria della metodologia e della strumentazione di misura impiegata. Tale precisione è espressa attraverso il *numero di cifre significative*, inteso come il numero minore di cifre necessarie per esprimere una quantità con la precisione voluta.

Regola le cifre necessarie, esclusi gli zeri necessari per localizzare la posizione del punto decimale, sono dette *cifre significative.*

$389.5 = 38.95 \cdot 10^{1} = 3.895 \cdot 10^{2} = 3895 \cdot 10^{-1}$ ha quattro cifre significative.

389.5 m è uguale a 0.3895 km, ha sempre 4 cifre significative.

La precisione di una misura è evidentemente indipendente dall'unità di misura scelta per esprimerla.

$3.8950\ 10^{-3}$ ha cinque cifre significative.

175.4 ha 4 cifre significative.

175.400 ha 6 cifre significative.

0.29 (= 2.9e-1) ha 2 cifre significative.

0.029 (= 2.9e-2) ha 2 cifre significative.

0.0029 (= 2.9e-3) ha 2 cifre significative.

0.002900 (= 2.900e-3) ha 4 cifre significative.

Calcoli

Nelle operazioni l'operando meno preciso limita superiormente la precisione del risultato.

Il risultato di una operazione di somma o sottrazione non può avere più cifre significative, dopo la virgola, di quante ne abbia l'operando con il minor numero di cifre significative dopo la virgola.

2.432 + 3.421 = 5.853

2.43 + 3.421 = 5.85(1)

Il risultato di una operazione di moltiplicazione, divisione o estrazione di radice non può avere più cifre significative di quante ne abbia l'operando con il minor numero di cifre significative.

$82.43 \cdot 3.42 = 281.9106 \rightarrow 282$

$2.43872 / 0.042 = 58.0647619 \rightarrow 58$

$48.6^{1/2} \approx 6.971370023 \rightarrow 6.97$

200 mm / 300 pagine = 0.6666… =$6.67\ 10^{-1}$ mm/pagina

Regola di arrotondamento

Arrotondamento all'intero più vicino: 72.8 diventa 73;

Arrotondamento ai primi due decimali:

72.8146 diventa 72.81;

72.460 => 72.46 arrotondamento per difetto;

72.461 => 72.46 arrotondamento per difetto;

72.462 => 72.46 arrotondamento per difetto;

72.463 => 72.46 arrotondamento per difetto;

72.464 => 72.46 arrotondamento per difetto;

72.466 => 72.47 arrotondamento per ecceso;

72.467 => 72.47 arrotondamento per ecceso;

72.468 => 72.47 arrotondamento per ecceso;

72.469 => 72.47 arrotondamento per ecceso.

72.465, togliendo o aggiungendo la quantità 0.5, potrebbe diventare rispettivamente 72.46 o 72.47. La normativa stabilisce di arrotondare alla cifra pari più vicina, e quindi diventa 72.46.

72.465 => 72.46 arrotondamento è per difetto, ma se avessi il numero 72.455 allora, decidendo di togliere o aggiungere la quantità 0.5, arrotondato alla seconda cifra decimale, potrebbe diventare 72.45 o 72.46; secondo l'arrotondamento alla cifra pari più vicina diventa 72.46, e dunque in questo caso l'arrotondamento è per eccesso. In questo modo si tende a minimizzare l'accumulo degli errori di arrotondamento.

72.450 => 72.45 arrotondamento per difetto;

72.451 => 72.45 arrotondamento per difetto;

72.452 => 72.45 arrotondamento per difetto;

72.453 => 72.45 arrotondamento per difetto;

72.454 => 72.45 arrotondamento per difetto;

72.455 => 72.46 arrotondamento per eccesso;

72.456 => 72.46 arrotondamento per eccesso;

72.457 => 72.46 arrotondamento per eccesso;

72.458 => 72.46 arrotondamento per eccesso;

72.459 => 72.46 arrotondamento per eccesso.

Notazione scientifica

Non è certamente chiaro scrivere un numero come 0.001245 kg: volendo esprimere tale misura in kg occorre usare la notazione scientifica: $1.245 \cdot 10^{-3}$ kg.

L'uso delle potenze intere di 10 chiarifica l'indicazione di quantità molto "grandi" o molto "piccole".

864000000 diviene 8.64 10^{+8} oppure 8.64e+8 oppure 8.64E+8.

Come esponenti della base 10 occorre preferire i multipli interi di 3 (0, ±3, ±6, ±9, ±12).

Nella normale pratica di laboratorio, ***1-*** si esprimono le quantità numeriche impiegando il Sistema Internazionale di unità di misura; ***2-*** si adotta la notazione scientifica, con esponenti che siano preferibilmente multipli interi di tre; ***3-*** si indicano come cifre significative tutte le cifre certe più la prima incerta arrotondata.

Gli errori di misura

Una **misura diretta** consiste nel confronto diretto della grandezza in esame con la sua unità di misura, come ad esempio la misura delle dimensioni di un frutto eseguita con un calibro.

La maggior parte delle misure eseguite nell'industria si avvale di **strumenti tarati** nei quali, su un quadrante, appare il risultato della misura, senza che si richiedano operazioni manuali di confronto (p.e. igrometro, termometro).

Gli errori di misura

Nell'effettuazione delle misure si possono commettere errori di diverse specie.

Errori sistematici: dovuti ad una o più cause che agiscono sempre con una determinata legge. Si tratta di errori mediamente costanti per un certo strumento (come in un orologio preciso che sia stato anticipato di 10 minuti).

L'errore sistematico, solitamente dovuto ad imperfetta calibrazione dell'apparato di misura, è molto temibile perché non si hanno mezzi per accorgersene se non confrontando lo strumento con un altro corretto.

Errori accidentali: ripetendo la misurazione di una grandezza diverse volte, nelle "stesse" condizioni, si ottengono valori in generale fra loro diversi. Le differenze fra detti valori individuano la presenza di errori imprevedibili, detti accidentali.

Ogni errore accidentale è causato dalla concomitante azione di molte cause diverse (p.e. modificazioni delle condizioni ambientali, vibrazione degli strumenti, cambiamento nello sperimentatore, distrazioni, fluttuazioni nell'alimentazione elettrica, disturbi elettromagnetici, ...) tra loro non interagenti, influenti secondo cause sconosciute, tali che il loro effetto sia mediamente nullo (altrimenti si avrebbe un errore sistematico) generalmente tali da indurre errori per eccesso e per difetto con identica probabilità, e che comunque siano più probabili effetti prossimi a quello medio.

Va osservato come all'errore di misura venga spesso a sovrapporsi anche l'eventuale naturale diversità tra gli individui di una popolazione (naturale fluttuazione statistica, p.e. campioni di latte prelevati da uno stesso serbatoio): nella pratica si tende a confondere tali due sorgenti di variazione.

Le prove ripetute

Si ipotizzi di avere ripetuto più volte la determinazione di una stessa grandezza x, nelle stesse condizioni, ottenendo i valori $x_1, x_2, ..., x_{Nc}$: quale dobbiamo ritenere sia sinteticamente il risultato di queste determinazioni?

La popolazione di misure sarà caratterizzata da un qualche tipo di distribuzione attorno al proprio valor medio.

Quando le cause di tali scostamenti sono molte, piccole, sconosciute (ovvero casuali) e indipendenti, allora si verifica che la distribuzione dei dati si avvicina a quella di Gauss, e lo studio dei risultati di una serie di misure si riconduce a quello di una variabile aleatoria con distribuzione normale.

Siano $\bar{x}$ e s_x, rispettivamente il valore medio e la deviazione standard delle Nc misurazioni; tali valori costituiscono una stima dei corrispondenti valori veri μ_x e σ_x, allora in base a quanto visto sulla teoria elementare del campionamento, possiamo scrivere:

$$\bar{x} - z_c(p) \cdot \frac{s_x}{\sqrt{Nc-1}} \le \mu_x \le \bar{x} + z_c(p) \cdot \frac{s_x}{\sqrt{Nc-1}}$$

$$\mu_x = \bar{x} \pm z_c(p) \cdot \frac{s_x}{\sqrt{Nc-1}}$$

E' ovvio che qualora fosse già noto il valore vero della varianza della x, al valore stimato s_x si sostituirebbe il valore vero σ_x.

Va notato come al crescere di Nc, si restringe l'intervallo in cui si localizza μ_x. Facendo infinite misure ($Nc \to \infty$) determiniamo $\mu_x = \bar{x}$. Ciò non significherebbe che le misurazioni avrebbero portato ad un risultato esatto, ma solo che sarebbe stato eliminato completamente l'errore accidentale, il risultato della misura può cioè essere comunque lontano dal vero valore della grandezza in esame per la presenza di errori sistematici non eliminabili con la ripetizione delle prove.

Occorre sottolineare il fatto che lo scopo delle prove ripetute può essere doppio: 1) prove ripetute sullo stesso campione, al fine di diminuire l'influenza dell'errore aleatorio; 2) prove ripetute su campioni differenti, per stimare il valor medio della popolazione d'origine.

La propagazione degli errori

Molte grandezze fisiche non possono di solito essere misurate in una singola misura diretta, ma vengono invece determinate in due passi distinti. In primo luogo, occorre misurare una o più grandezze *x, y,..*.che possono essere misurate direttamente e dalle quali la grandezza che ci interessa può essere calcolata.

P.e. per trovare l'area di un rettangolo occorre misurarne le lunghezze dei lati e poi moltiplicarle tra loro. Altri esempi riguardano misure di velocità, energia, portata, massa volumica, pressione,...

In questi casi anche la stima degli errori viene fatta in due passi. Qui si descrive come stimare il modo nel quale le incertezze sulle singole misure influiscono sul risultato finale.

Supponiamo di avere misurato due grandezze x, y con gli errori Δx e Δy. La grandezza che interessa sia $q=x+y$ (oppure $q=x-y$). Le quantità Δx e Δy possono essere gli scarti tipici associati ad una certa probabilità (p.e. le deviazioni standard).

I valori più alti e più bassi di x ed y sono evidentemente $x \pm \Delta x$, ed $y \pm \Delta y$. La media tra il valore più alto e quello più basso fornisce $q=x+y$, e l'errore, ovvero la differenza tra il valore più alto e quello più basso, risulta $\Delta x + \Delta y$. Ovvero nel calcolo di somme e differenze gli errori assoluti si sommano.

Analogamente si dimostra che nel caso di prodotti/quozienti si sommano gli errori relativi:

$\Delta q/q = \Delta x/x + \Delta y/y$

Più in generale supponiamo di avere misurato una grandezza $x=x_0$ nella forma standard $x_0 \pm \Delta x$ e di voler calcolare una qualche funzione nota $q(x)$. Poiché in generale Δx sarà piccolo rispetto ad x, allora i valori $q(x_0 \pm \Delta x)$ saranno vicini, possiamo così commettere un piccolo errore sostituendo alla curva $q(x)$ la retta tangente in x_0. L'equazione di tale retta è:

$$q(x) \approx q(x_0) + \left.\frac{dq(x)}{dx}\right|_{x0} \cdot (x - x_0)$$

il coefficiente angolare K di tale retta vale dunque la derivata della $q(x)$ calcolata nel punto x_0. Allora una variazione (ovvero un'incertezza) Δx della variabile indipendente si riflette amplificata del termine K sulla variazione di q.

$$\Delta q(x) = \left|\frac{dq(x)}{dx}\right|_{x0} \cdot \Delta x$$

☞Se è stato misurato un angolo come 20±3 gradi, determinare la incertezza nella stima del coseno.

☞Se si commette un errore del 5% nella misura del diametro medio di un frutto, quale errore si commette nella misura del suo volume ?

☞Valutare l'incertezza nella determinazione di densità di un frutto.

Si dimostra come in generale essendo $x,y,...,z$ variabili indipendenti misurate con incertezze Δx, Δy,..., Δz, utilizzate per calcolare la funzione $q(x,y,...,z)$. Se le incertezze sono pure tra loro indipendenti e casuali, allora l'incertezza in q vale circa la somma vettoriale degli scarti:

$$\Delta q = \sqrt{\left(\frac{dq}{dx}\Delta x\right)^2 + \ldots + \left(\frac{dq}{dz}\Delta z\right)^2}$$

ed in ogni caso non è mai superiore alla somma:

$$\Delta q \le \left|\frac{dq}{dx}\right|\Delta x + \ldots + \left|\frac{dq}{dz}\right|\Delta z$$

☞Studiare il caso della misura di una pressione ($p=F/F$).

Definizioni caratterizzanti le metodologie di misura

Accuratezza: è un indice della rispondenza del valor medio di una serie di misure ripetute con il valore vero (*errore sistematico*);

Precisione: è un indice della dispersione tra una serie di misure ripetute (*errore accidentale*);

Ripetibilità: è un indice della precisione ottenuta in misure eseguite più volte nello stesso laboratorio;

Riproducibilità: è un indice della precisione ottenuta in misure eseguite in laboratori diversi.

A parità di *range*, il costo di uno strumento può essere più che decuplo rispetto ad un altro a causa dei minori errori di misura ottenibili. Nella valutazione degli strumenti a scopo commerciale non si distingue fra errore di accuratezza e di precisione, perché si suppone che uno strumento in vendita sia tarato, ed essi vengono valutati globalmente come errore complessivo di misura.

Per qualificare uno strumento se ne dà l'errore complessivo percentuale. Tale errore è generalmente proporzionale all'indicazione: si divide il valore assoluto della massima deviazione standard per il fondoscala (ottenendo p.e. 0.02) e si dirà che l'errore è del ±2% sul fondo scala.

La **classe** di uno strumento è data dal suo errore percentuale: uno strumento di classe 2 è uno strumento che presenta una incertezza di ±2% (senza altre indicazione si intenderà sul fondo scala).

Strumenti di classe 5 sono da considerarsi commerciali. Un buono strumento avrà classe 1, sarà ottimo se avrà classe 0.5.

Riferimenti normativi:

ISO 10012-1:1992, Requisiti di assicurazione della qualità relativi agli apparecchi per le misurazioni - Sistema di conferma metrologica di apparecchi per misurazioni.

La verifica degli strumenti

Per eliminare gli errori sistematici (accuratezza) di uno strumento (operazione da eseguire periodicamente, perché nel tempo le risposte variano) si può, o sottoporlo alla misura di una grandezza che sia nota (grandezza campione), o confrontarne la risposta con quella ottenuta con uno strumento di classe superiore.

Una verifica completa comprende:

verifica dello zero;

verifica di scala o di range (dopo avere corretto lo zero, si fa una misura verso il fondoscala);

verifica di linearità ed isteresi.

16. ESERCIZI SUL TRATTAMENTO STATISTICO DELLE MISURE

Ex51

Da un grosso serbatoio vengono prelevati 5 campioni di una determinata bevanda e ne viene determinato il valore di pH che risulta: 5, 5.2, 5.4, 4.8, 5.1. Esprimere sinteticamente il risultato dell'analisi.

Ex52

Da un deposito di granaglie vengono prelevati 5 campioni e ne viene determinato il valore di concentrazione di piombo. Esprimere correttamente il risultato dell'analisi.

Ex53

Una stessa determinazione di grado zuccherino viene eseguita da 5 diversi operatori, ciascuno dei quali ripete due volte la misura. Esprimere sinteticamente il risultato dell'analisi.

Ex54

Da un camion sono state prelevate Nc bottiglie di latte (Nc>20) le quali hanno fornito i seguenti valori di massa volumica: $x_1, x_2, \ldots, x_{Nc}$.

Determinare la media e la deviazione standard campionari;

stimare la deviazione standard della popolazione;

calcolare l'intervallo di confidenza al 95% per la stima del valore medio della massa volumica di tutto il latte trasportato;

calcolare la probabilità che il latte contenuto in una bottiglia fornisca un valore di massa volumica superiore ad x^*;

calcolare la quantità di bottiglie per le quali la massa volumica assume un valore compreso tra $x1$ ed $x2$;

calcolare i limiti di massa volumica entro i quali è compreso il 90% delle bottiglie;

calcolare la probabilità che la massa volumica media valutata su di un campione di Nc bottiglie sia superiore ad x^*;(noti i parametri della distribuzione delle medie campionarie, valutate su campioni di ampizza Nc, si procede normalmente)

calcolare i limiti di massa volumica entro i quali è compreso il 90% dei campioni di mpiezza Nc;(limiti=media+/-z(90%)*Sigma della ditribuzione delle medie campionarie con Nc)

calcolare la probabilità che la differenza tra le masse volumiche medie calcolate su due campioni di ampiezza Nc sia superiore a Dx^*(occorre calcolare i parametri della distribuzione delle differenze e poi si procede normalmente);

calcolare la probabilità che la differenza tra le masse volumiche medie calcolate su due campioni di ampiezza Nc sia superiore al d%.(idem)

17. TEORIA DELLE DECISIONI STATISTICHE. TEST DI SIGNIFICATIVITÀ

Ipotesi statistica e livello di significatività

Nel controllare una produzione industriale spesso occorre capire se i cambiamenti (ovvero le differenze) che si misurano sui prodotti sono dovute a fattori aleatori, ovvero alla variabilità propria del prodotto (p.e. tutti i prodotti vegetali o animali sono sensibilmente diversi tra loro), oppure ad effettivi cambiamenti nel processo (cambiamento di fattori climatici, deterioramento delle macchine, invecchiamento dei componenti, diversità negli operatori, variazioni di processo).

Es.1 per confrontare l'efficacia di due tipi di mangime (*A* e *B*) viene somministrato il prodotto *A* ad una metà dei bovini di un grande allevamento, ed all'altra il prodotto *B*. Gli esemplari sono differenti per età, stato di salute, razza, ecc. Dopo un tempo adeguato si misurano le rese in latte degli animali registrando per un campione del gruppo *B* un volume medio di latte superiore del 15% a quello di un campione del gruppo *A*.

A causa della naturale diversità tra gli individui (e quindi dei campioni), anche nel caso che i due gruppi *A* e *B* venissero trattati esattamente allo stesso modo, difficilmente si otterrebbe uno stesso valore del volume medio. Allora nasce l'esigenza di valutare se la differenza tra le due medie è legata solamente a fenomeni aleatori (fluttuazione statistica) oppure traduce una reale diversità nelle popolazioni *A* e *B* dalle quali sono stati prelevati i due campioni (ovvero la differenza è statisticamente significativa)?

Es.2 si sperimenta un nuovo processo di pastorizzazione del latte. Una quantità A viene trattata con metodo tradizionale ed una B con quello nuovo. Alla fine si trovano cariche microbiche residue inferiori. Concludiamo immediatamente che la nuova tecnologia è efficace?

Se invece la stessa quantità di latte viene divisa in due gruppi, A e B trattati allo stesso modo, e poi si misura la carica batterica. Si trova che è identica? Certamente no. Concludiamo allora che A e B sono differenti, pur sapendo che i campioni A e B provengono dalla medesima popolazione?

Es.3 – Abbiamo ricevuto 2 campioni A e B, simili ma caratterizzati da medie e deviazioni differenti: provengono dalla stessa popolazione?

Quando si deve decidere se e in quale misura i cambiamenti di un processo hanno modificato un prodotto, si calcola la probabilità che le differenze misurate siano dovute solo alla normale *fluttuazione statistica*, cioè alla normale diversità tra i campioni estratti dalla stessa popolazione.

Tale calcolo viene generalmente presentato in forma di ipotesi da verificare. L'ipotesi secondo la quale le differenze rilevate tra due campioni sono dovute alla sola fluttuazione statistica, e non ad una diversità nel trattamento dei due campioni, si dice in genere *ipotesi nulla* (si indica con H_0) e la procedura di confronto si dice *test di significatività*.

L'ipotesi simmetrica (ovvero quella secondo la quale esiste invece una reale differenza) si chiama ipotesi alternativa (H_1).

La probabilità che la differenza tra due campioni sia dovuta solo al caso, ovvero la probabilità con la quale l'ipotesi nulla è verificata, è detta *livello di significatività* del test.

Nella pratica si usano frequentemente livelli di significatività dello 0.05 (5%) per misure ordinarie di valenza industriale e dello 0.01 (1%) per misure di laboratorio.

Test basati sulla distribuzione normale. Medie e differenza di medie

Nell'esempio del mangime per mucche è impossibile stabilire di primo acchito se i due regimi di alimentazione hanno efficacia differente, in situazioni di questo genere si formula l'ipotesi che l'efficacia dei due mangimi non presenti alcuna differenza significativa (ipotesi nulla H_0).

La grandezza che è stata scelta per esprimere l'efficacia del mangime è il peso degli animali. Se gli effetti dei due mangimi A e B non differiscono in modo significativo, le due popolazioni di pesi devono avere media uguale. Quindi l'ipotesi nulla sarà per il caso dei mangimi H_0: $\mu_A = \mu_B$. Ciò ha il significato di affermare che i due campioni di mucche sono estratti da una stessa popolazione.

Di conseguenza l'ipotesi alternativa diviene H_1: $\mu_A \neq \mu_B$.

Ricordando le espressioni viste a proposito della distribuzione delle differenze tra le medie campionarie, risulta:

$$\mu_{\bar{x}A-\bar{x}B} = \mu_{\bar{x}A} - \mu_{\bar{x}B} = \mu_A - \mu_B = 0$$

$$\sigma_{\bar{x}A-\bar{x}B} = \sqrt{\sigma^2_{\bar{x}A} + \sigma^2_{\bar{x}B}} = \sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}} \approx \sqrt{\frac{s^2_A}{Nc_A - 1} + \frac{s^2_B}{Nc_B - 1}}$$

Se i campioni sono grandi tale distribuzione tende ad essere normale, ed allora considerandone la forma standardizzata:

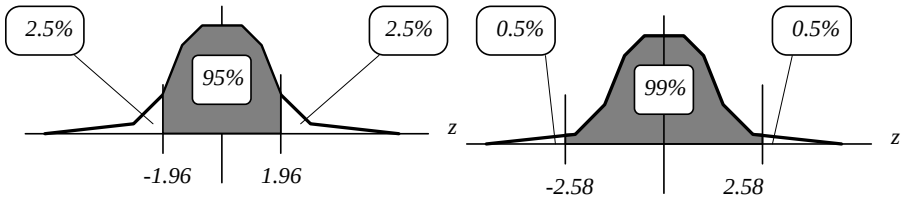
$$z = (S - \mu_S)/\sigma_S$$

si ottiene:

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sigma_{\bar{x}A-\bar{x}B}} = \frac{(\bar{x}_A - \bar{x}_B) - 0}{\sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}}} \approx \frac{(\bar{x}_A - \bar{x}_B) - 0}{\sqrt{\frac{s^2_A}{Nc_A - 1} + \frac{s^2_B}{Nc_B - 1}}}$$

Se il valore z così calcolato cade nell'intervallo $[-z_\alpha, +z_\alpha]$ allora il valore calcolato della differenza è un valore molto comune, ed effettivamente possiamo concludere che non c'è differenza significativa tra le medie, ovvero i due campioni di mucche sono considerabili come campioni provenienti da una medesima popolazione.

Se invece il valore calcolato di z cade al di fuori dei limiti di confidenza preassegnati allora potremmo ancora trovarci di fronte ad una normale fluttuazione aleatoria, però estremamente rara (tipicamente ci si riferisce ai livelli del 5% per misure di campo o industriali; dell' 1% per misure di laboratorio in condizioni ben controllate).



Area di accettazione dell'ipotesi nulla (la differenza non è significativa);

Area di rifiuto dell'ipotesi nulla (la differenza è significativa).

- Rifiutiamo l'ipotesi al livello di significatività dello 0.05 se il valore di z cade al di fuori degli estremi -1.96 +1.96 (cioè z<-1.96 oppure z>+1.96, ovvero l'area delle code è inferiore al 5%).
- Rifiutiamo l'ipotesi al livello di significatività dello 0.01 se il valore di z cade al di fuori degli estremi -2.58 +2.58 (cioè z<-2.58 oppure z>+2.58, ovvero l'area delle code è inferiore all'1%).

Oltre che sulle differenze tra medie si possono condurre test di significatività su statistiche campionarie riguardanti uno qualsiasi degli indici di posizione o di dispersione già visti (mediana, quantili, C.V., ecc.).

Nel caso delle medie risulta evidentemente:

$$z = \frac{S - \mu_S}{\sigma_S} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{x - \mu_x}{\sigma_x/\sqrt{Nc}}$$

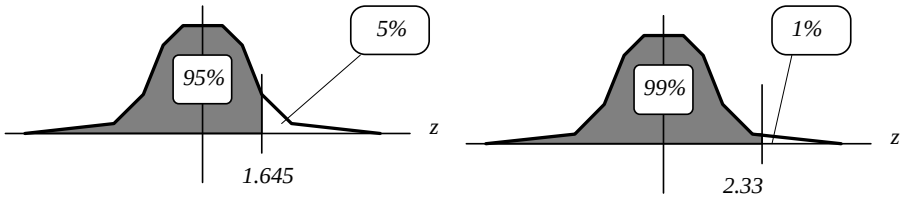
dove come al solito, se non è nota la varianza della popolazione, questa può essere sostituita con quella campionaria purché il campione sia grande.

Test ad una e due code

Se interessa solo il campo positivo o solo quello negativo dei valori di z allora il test è detto *ad una coda*, in quanto l'intervallo di confidenza non è più simmetrico, ma coinvolge parzialmente un ramo della gaussiana.

Ciò può avvenire se oltre alle relazioni uguale/diverso si utilizzano anche maggiore/minore. P.e. nel caso dell'alimentazione delle mucche l'ipotesi alternativa potrebbe anche essere espressa in una forma del tipo del tipo: "il mangime A è migliore di quello B: H_1: $\mu_A > \mu_B$".

In tale caso cambiano i valori dei limiti di confidenza, che si possono comunque facilmente determinare a partire dalla tavola delle aree della curva normale:



Area di accettazione dell'ipotesi nulla (la differenza non è significativa);

Area di rifiuto dell'ipotesi nulla (la differenza è significativa).

livello di confidenza	0.9	**0.95**	**0.99**	0.995	0.998
livello di significatività	0.1	**0.05**	**0.01**	0.005	0.002
z_c per test a due code	±1.645	**±1.96**	**±2.58**	±2.81	±3.08
(il simbolo ± è da intendersi come + **e** -)					
z_c per test ad una coda	±1.28	**±1.645**	**±2.33**	±2.58	±2.88
(il simbolo ± è da intendersi come + **o** -)					

18. ESERCIZI SULLA TEORIA DELLE DECISIONI STATISTICHE - IL TEST Z

Ex55

Le mucche di un allevamento producono mediamente 20 litri di latte al giorno, con uno scarto quadratico medio di 6 litri. Trovare quale tra i valori di produzione di latte, ottenuti da differenti bovine, sono irregolari: 4, 6, 8, 10, 16, 20, 26, 28 ?

Come regoletta sulla base della quale decidere la "regolarità" dei valori misurati stabiliamo che un animale è in buona salute se la quantità di latte che produce rientra nell'intervallo di confidenza al 90% per la stima del valore medio della quantità di latte prodotta dagli animali di tutto l'allevamento.

Svolgimento 1)

Media=20, deviazione standard=6;

$z_c(90\%)$ per un test a due code vale 1.645, dunque l'intervallo di confidenza al 90% è dato da:

$$\mu_x \pm z_c \cdot \sigma_x = 20\ l \pm 1.645 \cdot 6\ l$$

ovvero

$$x_{min} = 10.13\ kg\ ,\ x_{max} = 29.87\ kg$$

Risultano dunque fuori intervallo per difetto i valori 4, 6, 8 e 10 *l*.

Svolgimento 2) Si controlla che il valore di z risulti interno all'intervallo $[-z_c, +z_c]$, essendo $z_c(90\%)=1.645$:

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{x - 20}{6}$$

x	*Z*	*test*
4	$z = \frac{4-20}{6} \cong -2{,}67$	esterno ($x << -1.645$)
6	$z = \frac{6-20}{6} \cong -2.33$	esterno ($x < -1.645$)
8	$z = \frac{8-20}{6} \cong -2.00$	esterno ($x \leq -1.645$)
10	$z = \frac{10-20}{6} \cong -1.67$	esterno ($x < -1.645$)
16	$z = \frac{16-20}{6} \cong -0.67$	valore nella norma
20	$z = \frac{20-20}{6} \cong 0$	valore nella norma
26	$z = \frac{26-20}{6} \cong 1.00$	valore nella norma

Ex56

Problemi di controllo continuo della produzione industriale (*carte di controllo*).

Si considera una linea di lavorazione destinata a selezionare ed impacchettare frutti in cestini. Sulla macchina in condizioni di perfetto funzionamento è stato rilevato un peso medio dei cestini pari a 574 g con una deviazione standard di 80g. Per controllare lo stato della macchina, ogni giorno viene prelevato un campione di 6 cestini, e ne viene calcolato il peso medio. Si determini una metodologia per stabilire la necessità di una revisione.

Con una confidenza del 99.73% si può dire che la media campionaria $\bar{x}$ deve essere compresa tra gli estremi $\mu_x - 3 \cdot \sigma_{\bar{x}}$ e $\mu_x + 3 \cdot \sigma_{\bar{x}}$. La deviazione standard della media campionaria vale $\sigma_{\bar{x}} \cong \sigma_x / Nc = 80/\sqrt{6}\ g$, dunque ad un livello di confidenza del 99.73% la media campionaria dovrebbe essere compresa nell'intervallo $574\ g \pm 240/\sqrt{6} \cong [476\ g \div 672\ g]$.

Così potremmo stabilire come regola che se più dell 1% delle medie campionarie cade all'esterno di tale intervallo dobbiamo sottoporre la macchina ad una revisione.

Media Campionaria	**Lun.**	**Mart.**	**Merc.**	**Giov.**	**Ven.**
0.672	*				
0.574		*		*	
0.476			*		*

Nella realtà aziendale i test di questo genere sono generalmente più articolati, in modo da evidenziare con maggiore efficacia il fatto di trovarsi di fronte ad un'anomalia sistematica e non ad una (ancorchè rara) fluttuazione statistica: per esempio il capitolato di manutenzione (o il software di gestione di una linea di lavorazione) potrebbe prevedere come condizione di fuori linea il ripetersi per k volte del superamento dei limiti su un campione di n osservazioni.

Ex57

Un gruppo di pomodori da mensa è stato raccolto con una macchina raccoglitrice: di 32 tra questi pomodori è stata misurata la resistenza allo schiacciamento, ottenendo un valore medio di 20 N, con una deviazione standard di 12 N. Dalla medesima coltivazione e nello stesso periodo sono stati raccolti altri pomodori con procedura manuale: per un gruppo casuale di 32 tra questi la resistenza meccanica risultava caratterizzata da un valore medio di 26 N, con una deviazione standard di 15 N.

Si chiede di verificare se la raccolta meccanica ha indebolito la struttura delle bacche. Ripetere il calcolo considerando però campioni di ampiezza pari a 60.

Secondo l'ipotesi nulla non c'è differenza significativa, ovvero le due classi sono campioni tratti da una stessa popolazione di studenti. In tale ipotesi, essendo i campioni grandi, possiamo stimare gli scarti della popolazione con quelli campionari, allora la variabile standardizzata, relativa alla distribuzione della popolazione delle differenze, assume il valore:

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sigma_{\bar{x}A-\bar{x}B}} \approx \frac{(20 - 26) - 0}{\sqrt{\frac{12^2}{32-1} + \frac{15^2}{32-1}}} \approx -1.81$$

Tale valore di z corrisponde ad una probabilità di poco superiore al 7%, ovvero possiamo dire che esiste una probabilità pari a circa il 7% che la differenza osservata tra i campioni A e B sia dovuta al caso, o in altre parole possiamo affermare che la raccolta meccanica induce un indebolimento delle bacche, ammettendo però una probabilità di sbagliare pari a circa il 7%. Generalmente tale livello di probabilità è ritenuto non sufficiente per sostenere un'ipotesi, dunque nel caso attuale concludiamo che i due metodi di raccolta non inducono differenze significative.

Se invece le differenze fossero state osservate su campioni più ampi (N=80), allora avremmo ottenuto:

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sigma_{\bar{x}A-\bar{x}B}} \approx \frac{(20 - 26) - 0}{\sqrt{\frac{12^2}{80-1} + \frac{15^2}{80-1}}} \approx -2.8$$

che corrisponde ad un livello di probabilità superiore all'1%, dunque in tal caso potremmo dire che il procedimento di raccolta influenza le caratteristiche dei frutti (almeno al livello dell'1%).

Ex58

Due differenti trattamenti di decontaminazione sono stati applicati a due campioni di uova (A e B) composti rispettivamente da 40 e 50 elementi. La carica batterica residua sul gruppo A, rispetto a quella di partenza, espressa in punti percentuali è risultata 7.4% con e una deviazione standard 0.8%. Nella seconda classe, la carica residua è stata misurata pari al 7.8% di quella originale con una deviazione standard di 0.7%. C'è una differenza significativa tra i due trattamenti ai livelli dello 0.05 e dello 0.01?

La variabile standardizzata, relativa alla distribuzione della popolazione delle differenze, assume il valore:

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sigma_{\bar{x}A-\bar{x}B}} \approx \frac{(7.4 - 7.8) - 0}{\sqrt{\frac{0.8^2}{40-1} + \frac{0.7^2}{50-1}}} \approx -2.49$$

Tale valore cade a sinistra di z = -1.96, è cioè lontano dalla media, dai valori più probabili, e dunque è significativo al livello dello 0.05.

Però per un livello di confidenza del 99%, z_c = -2.58, che è un poco più piccolo di -2.49, dunque tale differenza è significativa, come visto, al livello del 5%, ma non al livello dell'1%.

In questi casi si conclude che la differenza tra i due trattamenti è *probabilmente* significativa.

E' comune indicare i risultati di osservazioni statistiche come segue:

- al livello dello 0.01 come *molto significativi (**)*,
- quelli con livello di significatività compreso tra 0.01 e 0.05 *probabilmente significativi (*),*
- ed i restanti *non significativi*.

Molti codici di calcolo fanno corrispondere a tali intervalli 1, 2 o 3 asterischi.

Ex59

Il peso medio di 50 studenti che hanno partecipato ai corsi di atletica è di 68.2kg, con una deviazione standard di 2.5kg, mentre il peso medio di 50 studenti che non si sono interessati al corso di atletica è di 67.5kg, con una deviazione standard di 2.8kg. Determinare il livello di significatività della differenza.

Calcolati al solito i parametri della popolazione di differenze, la variabile standardizzata, relativa alla distribuzione della popolazione delle differenze, assume il valore:

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sigma_{\bar{x}A-\bar{x}B}} \approx \frac{(68.2 - 67.5) - 0}{\sqrt{\frac{2.5^2}{50-1} + \frac{2.8^2}{50-1}}} \approx 1.32$$

Tale valore corrisponde ad un livello di significatività di circa 0.18. Ovvero possiamo dire che ci sono differenze tra i gruppi con una probabilità di sbagliare di circa il 20%, che risulta troppo alta per poter confutare l'ipotesi nulla; ovvero riteniamo che l'aumento osservato di circa un kg sia dovuto alla naturale diversità tra i campioni.

Ex60

Un panel test, composto da cinque assaggiatori, volto alla valutazione di due succhi di frutta di diversa composizione ha fornito i seguenti punteggi:

	media	*deviazione standard*
composizione A	5.4	3
composizione B	6.8	3.6

esprimersi sulla significatività della differenza trovata. Provare a ripetere le medesime valutazioni nel caso di un gruppo di assaggiatori composto da 15 membri.

Ex61

A seguito di una estesa campagna di misura, il residuo medio μ di fitofarmaci rilevato sulla superficie di alcuni ortaggi è risultato essere pari a 1800 ppm, con una deviazione standard di 100 ppm.

Si sperimenta una modifica nel processo di lavaggio dei prodotti, al fine di economizzare il trattamento abbassando le temperature e le portate d'acqua; si teme tuttavia che in tal modo la qualità dei frutti possa soffrirne. Per provarlo si preleva un campione di 50 ortaggi e si trova che il residuo medio $\bar{x}$ di fitofarmaco è salito a 1850 ppm, con la medesima deviazione standard. Possiamo quindi pensare che tale piccola variazione sia indice di una normale fluttuazione statistica o un vero peggioramento, al livello di significatività dello 0.01?

Si deve decidere tra le due ipotesi:

H_0: *1850* ≈ 1800ppm, non c'è in realtà nessun peggioramento nella qualità dei prodotti.

H_1: *1850* > 1800ppm, c'è effettivamente un peggioramento nella qualità dei prodotti.

Con riferimento alla teoria sulla distribuzione delle medie campionarie, si procede calcolando la probabilità che la media di un campione, estratto da una popolazione di media nota, possa differire da questa di una quantità assegnata:

$$z = \frac{S - \mu_S}{\sigma_S} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_x/\sqrt{Nc}} = \frac{1850\text{-}1800}{100/\sqrt{50}} \cong 3.55$$

Dobbiamo fare uso del test ad una coda: al livello di significatività dello 0.01 la regola di decisione è: se il valore z osservato è maggiore di $z_{99\%}$=2.33, il risultato è significativo al livello dello 0.01, e l'ipotesi nulla viene rifiutata. Dunque concludiamo che il risultato è altamente significativo: c'è dunque un reale scadimento nella qualità degli ortaggi.

Tale osservazione corrisponde ad un livello di significatività superiore a 0.01. Ovvero ammettiamo che ci sia uno scadimento della qualità degli ortaggi con una probabilità di sbagliare inferiore all'1%.

Se la dispersione nelle misure del residuo di fitofarmaco fosse stata più elevata, p.e. con deviazione standard pari a 250ppm, allora cosa si sarebbe concluso ?

Questo caso si presta inoltre ad una osservazione interessante: il valore di 1800 ppm e la sua deviazione standard di 100 ppm, sono stati misurati durante il normale funzionamento dell'impianto per un tempo molto lungo (se comparato con i tempi della sperimentazione) dunque su un campione molto ampio. Allora possiamo ritenere che i valori 1800 ppm e 100 ppm, siano stati ricavati da un campione di ampiezza tendente ad infinito. Ricavando in tale ipotesi una forma semplificata dell'espressione standardizzata della differenza tra le medie di due campioni otteniamo un risultato già noto per altra via:

poiché i due campioni A e B sono estratti, per ipotesi, da una medesima popolazione, risulta:

$$\mu_{\bar{x}A\text{-}\bar{x}B} = \mu_{\bar{x}A} - \mu_{\bar{x}B} = \mu_A\text{-}\mu_B = 0$$

inoltre, poichè:

$$\lim_{N_{cA}\to\infty} \bar{x}_A = \mu_A$$

e

$$\lim_{N_{cA}\to\infty} \sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}} = \sqrt{\frac{\sigma^2_B}{Nc_B}} = \frac{\sigma_B}{\sqrt{Nc_B}}$$

risulta, allora come già è stato visto:

$$\lim_{N_{cA}\to\infty} z = \lim_{N_{cA}\to\infty} \frac{(\bar{x}_A\text{-}\bar{x}_B) - (\mu_A\text{-}\mu_B)}{\sqrt{\frac{\sigma^2_A}{Nc_A} + \frac{\sigma^2_B}{Nc_B}}} = \frac{\mu_A\text{-}\bar{x}_B}{\frac{\sigma_B}{\sqrt{Nc_B}}}$$

Ex62

Una popolazione di bovini è caratterizzata da un peso medio pari a 400 kg, ed una deviazione standard di 80 kg. Determinare ***a)*** qual è la probabilità che un animale differisca in peso per più (meno) di 100 kg dal valore medio, e ***b)*** la probabilità che il peso medio degli animali di un campione di 10 differisca in peso per più di 100 kg dal valore medio della popolazione.

a) $$z = \frac{S - \mu_S}{\sigma_S} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{x - \mu_x}{\sigma_x} = \frac{100}{80} \cong 1.25$$

che corrisponde ad una probabilità del 10.6%.

b) $$z = \frac{S - \mu_S}{\sigma_S} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{x - \mu_x}{\sigma_x/\sqrt{Nc}} = \frac{100}{80/\sqrt{10}} \cong 3.95$$

che corrisponde ad una probabilità dello 0.004%.

Ex63

Un fabbricante dichiara di produrre una soluzione acida con concentrazione dell'8%. Un campione di 30 confezioni viene esaminato prima dell'acquisto, determinando però un valore medio di concentrazione $\bar{x}$=7.5% con uno scarto quadratico medio s_x=0.76%. Determinare la veridicità delle dichiarazioni del commerciante.

L'ipotesi nulla è che la media del campione sia statisticamente uguale alla media della popolazione. La variabile standardizzata z risulta:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{Nc-1}}} = \frac{7.5 - 8.0}{\frac{0.76}{\sqrt{30-1}}} \approx -3.54$$

Dalla tabellina osserviamo come il valore di z critico per il test a due code al livello del 99.8%, vale 3.08. Dunque il prodotto verrà rifiutato ammettendo una la probabilità di sbagliare inferiore allo 0.2%, o in altre parole, la probabilità che per effetto della sola fluttuazione statistica la soluzione di tale campione abbia un valore medio di concentrazione inferiore al valore 8, è inferiore allo 0.2%.

Rifiutiamo in sostanza l'ipotesi nulla al livello p=0.002=2‰.

Ex64

In relazione all'acquisto di reagenti per il laboratorio, stabilire una regola per l'accettazione o il rifiuto, basata sulle misure ottenute da un campione delle merci in ingresso.

Ex65 66

Un'azienda costruttrice dichiara che il consumo di carburante di un nuovo modello di automobile, misurato secondo le norme in vigore, è di 11.5 *l*//100 km. Una prova condotta con la collaborazione di 40 acquirenti ha permesso di rilevare un consumo medio di 13.1 *l*//100 km con deviazione standard di 4.4 *l*//100 km. Qual è la probabilità di ottenere un risultato uguale o superiore a questo ? Si può ritenere che il consumo medio valutato dagli acquirenti sia realmente diverso da quello indicato dal costruttore ?

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{Nc-1}}} = \frac{13.1 - 11.5}{\frac{4.4}{\sqrt{40-1}}} \approx 2.27$$

La probabilità che la discordanza sia dovuta ad una fluttuazione statistica è di poco superiore al valore 1.16%, dunque qualche indagine ulteriore sui risultati ottenuti potrebbe essere opportuno.

Ex67

Dati i seguenti problemi formulare l'ipotesi nulla, precisare se si tratta di un test ad una o due code, discutere come organizzare la raccolta e l'analisi dei dati al fine di prendere una decisione.

- I tecnici di una unità sanitaria locale sono convinti che in un lago, destinato alla balneazione, il numero di coliformi presenti in 100 cm^2 di acqua sia superiore a 2400 unità. Se questo sospetto fosse fondato, sarebbe necessario prendere misure drastiche per eliminare le fonti di inquinamento.
- Il centro di ricerche agronomiche *Peperone* ha messo a punto una nuova varietà di fragole che sembrano presentare una resa sensibilmente superiore a quella delle varietà precedenti. Dato che la nuova varietà deve essere messa in vendita ad un prezzo abbastanza elevato, il centro ricerche vuole verificare che la resa della nuova varietà sia effettivamente più alta delle precedenti.
- Un orticoltore è convinto che utilizzando un certo fertilizzante organico ottiene melanzane di peso medio superiore ad 1kg.

Ex68

19. L'ANALISI DEI DATI CON *MICROSOFT EXCEL*

Analisi statistica di misure ripetute

E' stato misurato il residuo di fitofarmaco sulla superficie di un campione di 20 frutti, determinando i valori della seguente tabella *(valori in μg/kg frutto)*:

Frutto	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
Residuo	18	16	12	14	17	18	18	13	14	16	16	13	15	14	14	15	13	17	16	12

- Esprimere sinteticamente il risultato dell'analisi.

Traccia di soluzione:

a) Calcolare l'intervallo di confidenza al 99% per la stima del valore medio. Occorre dunque calcolare il valore medio e lo scarto quadratico medio campionario con le opportune funzioni di *Excel* e calcolare gli estremi superiore ed inferiore dell'intervallo fiduciale come $\mu_x = \overline{x} \pm \Delta$ *essendo* $\Delta = z_t(0.99) \cdot s_x/\sqrt{Nc-1}$. (per i calcoli si possono usare le funzioni diretta ed inversa sulla distribuzione normale: *p=2*(distrib.norm.st(z)-distrib.norm.st(0))* oppure *z=inv.norm.st(p)* per coda sinistra, oppure *z=inv.norm.st[(p+1)/2]* per intervalli di confidenza; con $p \in]0,1[$) [$\mu_{min99\%}$ = 13.9; $\mu_{min99\%}$ = 16.2].

b) Per il calcolo della fascia d'incertezza Δ è possibile utilizzare la funzione statistica di Excel *confidenza(1-p, Sigma, Nc)*. Per esempi e suggerimenti d'uso si può utilizzare il menù di *help* (*?*, digitare *confidenza*, pulsante *cerca*).

c) Esprimere infine l'ampiezza della fascia d'incertezza Δ come valore percentuale. [$15.1 \pm 2.58 \cdot 1.9/\sqrt{19} = 15.1 \pm 1.1 = 15.1 \pm 7.5\%$]

d) Ripetere i calcoli precedenti in relazione a differenti livelli di affidabilità (p.e. 90%, 95%).

e) Valutare l'adeguatezza del campione utilizzato se si vuole contenere l'errore di stima entro i 2μg/kg, con affidabilità del 99%. [$Nc=1+(2z \cdot Sx/2)^2$]

f) utilizzare i valori della distribuzione *t-Student*, più adatta per i piccoli campioni, al posto dei valori critici di *z*: *t=inv.t(1-p; Nc-1) ==> es. t=inv.t(1-0.95; 20-1)*

Test di significatività

Nella tabella seguente sono riportati i valori di tensione di rottura a flessione misurati su due gruppi di 20 biscotti. Si tratta di due campioni scelti a caso tra tutti gli elementi prodotti da due differenti linee di produzione: una tradizionale ed una leggermente modificata nei parametri di tempo e temperatura di cottura, con l'obiettivo di conseguire una maggiore economia di processo.

C.1	21	27	21	28	30	18	24	28	20	20
	21	23	26	12	27	19	14	19	19	23

C.2	24	26	26	24	23	25	26	23	25	27
	23	24	26	28	25	26	23	25	26	25

Utilizzando le opportune funzionalità del programma *MS Excel*, si chiede di:

- valutare la significatività delle differenza tra i valori medi dei due gruppi, ovvero la probabilità che provengano dalla medesima popolazione.

Inoltre,

- selezionare un campione casuale di 5 elementi della stessa classe, e calcolarne la stima per lo scarto quadratico medio ed il valor medio della popolazione di origine. Confrontare i risultati stimati con quelli dell'intero gruppo;
- selezionare due campioni casuali di 5 elementi (dello stesso gruppo) e calcolare la significatività della differenza tra le due medie campionarie;
- selezionare due campioni casuali di 5 elementi (di gruppi diversi) e calcolare la significatività della differenza tra le medie campionarie;

Suggerimento: *è possibile utilizzare sia le funzioni standard di Excel che il modulo di analisi dei dati (Calcolare prima le varianze corrette e poi Dati→AdD→Test Z per medie).*

20. IL CONTROLLO STATISTICO DI PROCESSO

Il piano di campionamento

Le tecniche di randomizzazione dei campioni vengono applicate nell'industria alimentare, sia al controllo dei prodotti in uscita, sia ai controlli di accettazione delle merci in ingresso.

Un documento minimo che descriva il **piano di campionamento** deve specificare:

il lotto della merce in analisi;

la numerosità del campione;

una stima dell'affidabilità del test, in funzione della numerosità del campione;

le condizioni di accettazione e di rifiuto (eventualmente concordate tra le due parti al momento della stesura del capitolato d'acquisto);

modalità e condizioni delle misure.

Carte di controllo

Le **carte di controllo** costituiscono uno strumento analitico e grafico per stabilire il rispetto delle specifiche nel processo produttivo, e per controllarne (e valutarne criticamente) l'andamento nel tempo. Esistono carte di controllo per variabili e per attributi.

La carta di controllo fissa in forma grafica il valor medio atteso, della grandezza in osservazione, ed i limiti di variazione ammissibili.

	T_1	T_2	T_3	T_4	...
$\overline{X} + \Delta Xs$	*				
$\overline{X}$		*		*	
$\overline{X} - \Delta Xi$			*		*

Generalmente gli scostamenti superiore ed inferiore sono di modulo uguale. Se sono note a priori le caratteristiche del processo, allora se ne conosce la variabilità standard, e dunque in generale gli scostamenti vengono assunti pari a Z volte la deviazione standard.

A tal proposito, nell'ipotesi che la variabile osservata x, abbia distribuzione normale con media μ_x e deviazione standard σ_x, si riportano nella tabellina che segue, rispettivamente gli intervalli di variazione per un valore estratto a caso dalla popolazione e per la media di un campione di ampiezza Nc:

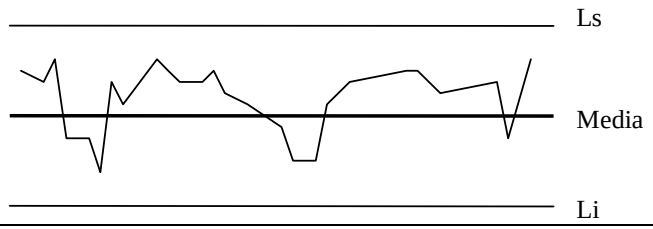
Probabilità	Singolo valore	Media campionaria
p	x	$\overline{x}$
$\approx 68.3\%$	$\mu_x \pm 1\ \sigma_x$	$\mu_x \pm 1 \frac{\sigma_x}{\sqrt{Nc}}$
$\approx 95.4\%$	$\mu_x \pm 2\ \sigma_x$	$\mu_x \pm 2 \frac{\sigma_x}{\sqrt{Nc}}$
$\approx 99.7\%$	$\mu_x \pm 3\ \sigma_x$	$\mu_x \pm 3 \frac{\sigma_x}{\sqrt{Nc}}$

Ad intervalli irregolari T_i, o comunque secondo un piano di campionamento, si preleva un campione di prodotto, se ne calcola il valor medio e lo si colloca sulla carta di controllo.

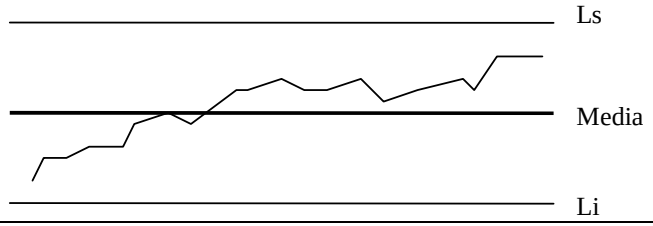
Se il campione è grande, ed è stato scelto $Z=3$, allora un punto ha una probabilità inferiore al 100-99.73≈0.3% di collocarsi all'esterno della carta di controllo. Per campioni piccoli, occorre rifarsi alle tabelle della variabile t di Student.

Si possono anche costruire carte di controllo riferite non al valor medio del campione ma alla sua variabilità.

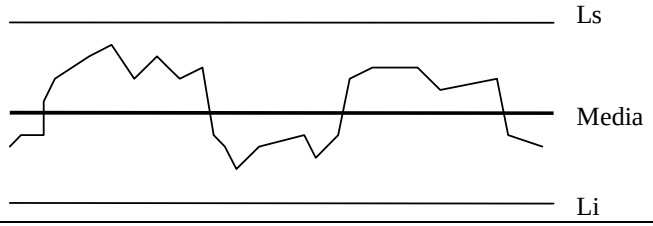
Esiste una morfologia tipica delle carte di controllo, in grado di denunciare malfunzionamenti nelle linee di lavorazione:



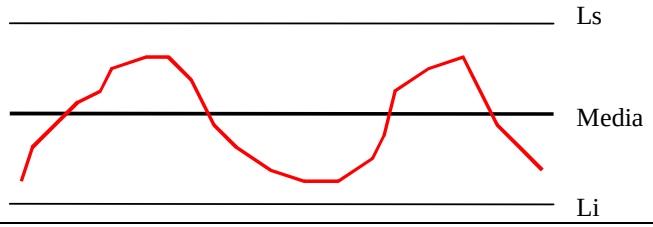
I punti al di sopra della linea centrale sono più numerosi di quelli al di sotto.



La successione di punti tende a crescere.



I punti tendono a formare successioni quasi costanti, alternativamente sopra e sotto la linea centrale: bistabilità.



I punti si susseguono in successioni alternativamente crescenti e decrescenti con periodo mediamente costante: variazioni cicliche nel processo.

Tutti questi modi anomali di funzionamento sono individuati da una regola del tipo: il processo è da verificare quando *n* punti consecutivi stanno dalla stessa parte rispetto alla linea media.

La redazione di carte di controllo per variabili è normata dalla tabella UNI 4728-66: *Metodi statistici per il controllo della qualità: Carte di controllo per variabili.*

	Diagramma delle medie		**Diagramma delle dispersioni**	
Numerosità dei campioni N	*linea centrale*	*limiti di controllo*	*linea centrale*	*limiti di controllo*
25<N	$\overline{X}$	$\overline{X} \pm 3\frac{\sigma}{\sqrt{N}}$	σ	$\sigma \cdot \left(1 \pm \frac{3}{\sqrt{2N}}\right)$
11>N>25	$\overline{X}$	$\overline{X} \pm A \cdot \sigma$	σ	$B_1 \cdot \sigma$
N<10	$\overline{X}$	$\overline{X} \pm A_2 \cdot \sigma$	σ	$D_1 \cdot \sigma$

N	A	A_2	D_1	B_1
2	2.121	1.880	0.000	0.000
3	1.732	1.023	0.000	0.000
4	1.500	0.729	0.000	0.000
5	1.342	0.577	0.000	0.000
6	1.225	0.483	0.000	0.085
7	1.134	0.419	0.205	0.158
8	1.061	0.373	0.387	0.215
9	1.000	0.337	0.546	0.262
10	0.949	0.308	0.687	0.302
12	0.866			0.365
14	0.802			0.414
16	0.750			0.454
18	0.707			0.468
20	0.671			0.513
22	0.640			0.536
24	0.612			0.556
>25	$3/\sqrt{N}$			$1 \pm 3/\sqrt{2N}$

21. TEORIA DEI PICCOLI CAMPIONI

Piccoli campioni e distribuzione t di Student

Per campioni di grande ampiezza (*Nc* indicativamente superiore a 30) detti *grandi campioni*, le distribuzioni campionarie della media, delle differenze e della deviazione standard sono approssimativamente normali, con approssimazione tanto migliore quanto più elevato risulta *Nc*.

Per campioni piccoli occorre invece utilizzare una teoria differente, detta teoria campionaria esatta, valida per campioni di qualsiasi numerosità.

Quando i campioni sono piccoli siamo innanzitutto lontani dalle ipotesi del teorema del valore medio, e la sostituzione di σ_x con s_x costituisce una imprecisione notevole: dunque i valori critici di z calcolati con riferimento alla curva normale diventano inaffidabili. William Gosset (*The application of the law of error to the work of the Brewery*, 1904) ha studiato il problema della distribuzione del valor medio e della varianza di piccoli campioni, determinando teoricamente la distribuzione della variabile standardizzata *t*:

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{Nc\text{-}1}}$$

tale variabile corrisponde alla z, ma approssimata sulla base di soli parametri campionari, risulta inoltre, come la variabile standardizzata z, adimensionale.

Sotto ipotesi largamente applicabili, si dimostra che la distribuzione di probabilità di *t* è definita da:

$$\varphi(t) = \frac{Y_0(Nc)}{\left(1 + \frac{t^2}{Nc\text{-}1}\right)^{Nc/2}}$$

dunque la *distribuzione t di Student* dipende da *Nc* (questo non avviene per la distribuzione di Gauss) e, per grandi valori di *Nc*, tende a quella normale standardizzata:

$$\lim_{Nc\to\infty} \varphi(t) = \psi(t) = \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}$$

Tale distribuzione risulta inoltre un poco più appiattita di quella normale, ovvero è caratterizzata da una maggiore dispersione.

Il questa trattazione semplificata si intenderà per *gradi di libertà* la quantità *Nc-1*.

Intervalli di confidenza

Come è stato descritto per la distribuzione normale, si possono definire intervalli di confidenza attraverso i coefficienti di confidenza $\pm t_c$, per esempio del 95% attraverso la:

$$0.95 = \int_{-tc}^{+tc} \varphi(t)\, dt$$

ovvero per tale livello di probabilità risulta:

$$-t_c \le t \le t_c$$

da cui, procedendo come già visto, esplicitando la definizione di t, risulta:

$$-t_c(95\%) \le \frac{\bar{x} - \mu_x}{s_x/\sqrt{Nc\text{-}1}} \le +t_c(95\%)$$

e dunque possiamo stimare il valore medio della popolazione attraverso i soli dati campionari, con probabilità assegnata:

$$\mu_x = \bar{x} \pm t_c(p)\cdot\frac{s_x}{\sqrt{Nc\text{-}1}}$$

Si nota dunque che quando ci si trova ad analizzare piccoli campioni, o anche solo se si vuole utilizzare comunque la teoria esatta, è sufficiente sostituire a z_c il valore t_c, corrispondente alla medesima probabilità.

Per quanto riguarda l'integrazione numerica della distribuzione *t di Student*, occorre notare che, a differenza della z di Gauss, dipende anche dalla numerosità del campione. Per evitare ogni difficoltà nell'uso pratico, dunque, i valori critici di *t* sono generalmente espressa in valori percentili. Per esempio si scrive $t_{0.975}$ per designare il 97.5-esimo percentile.

Con questa notazione t_{100} indica quel valore alla cui sinistra si trova il 100% dei valori di t (generalmente $+\infty$), ovvero corrisponde a tutta l'area sottesa dalla curva di distribuzione.

t_0 corrisponde invece al valore alla cui sinistra non cade nessun valore di t.

Si può dare un significato anche al segno del percentile intendendo p.e. con $-t_{20}$ il valore alla cui destra cade il 20% dei valori.

Allora $t_0=-t_{100}$, ovvero il valore alla cui sinistra si trova lo 0% dei valori di t, o il valore alla cui destra si trova il 100% dei valori di t.

Così $-t_{0.975}$ individua il valore alla cui destra cade il 97.5% dei valori, ovvero equivale a $t_{0.025}$. In questo modo a destra di $+t_{0.975}$ rimane confinato il 2.5% dei valori di t, così come nella coda di sinistra, a sinistra di $-t_{0.975}$ rimane confinato il 2.5% dei valori di t.

Allora tra $-t_{0.975}$ $+t_{0.975}$ rimane confinato il 100-2.5-2.5=95% dei valori di t. Utilizzare $t_c=t_{0.975}$ significa riferirsi, in un test a due code, ad una probabilità del 95%.

Funzioni di Excel:

- area compresa tra $-tc$ e tc: ***1-distrib.t(tc; Nc-1; 2)***
- intervallo fiduciale per t: ***inv.t(1-p; Nc-1)***

Es.

1 - DISTRIB.T(1; 10-1; 2) = 0,66

INV.T(1-0,95; 10-1) = 2,26

Area sottesa dalla curva di *Student* tra *-t* e *+t*

Nc-1	*valori di t*									
	0,7	***0,8***	***1,0***	***1,5***	***2,0***	***2,5***	***3,0***	***4,0***	***5,0***	***6,0***
2	0,444	0,492	0,577	0,728	0,816	0,870	0,905	0,943	0,962	0,973
3	0,466	0,518	0,609	0,769	0,861	0,912	0,942	0,972	0,985	0,991
4	0,477	0,531	0,626	0,792	0,884	0,933	0,960	0,984	0,993	0,996
5	0,485	0,540	0,637	0,806	0,898	0,946	0,970	0,990	0,996	0,998
6	0,490	0,546	0,644	0,816	0,908	0,953	0,976	0,993	0,998	0,999
7	0,493	0,550	0,649	0,823	0,914	0,959	0,980	0,995	0,998	0,999
8	0,496	0,553	0,653	0,828	0,919	0,963	0,983	0,996	0,999	1,000
9	0,498	0,556	0,657	0,832	0,923	0,966	0,985	0,997	0,999	1,000
10	0,500	0,558	0,659	0,835	0,927	0,969	0,987	0,997	0,999	1,000
15	0,505	0,564	0,667	0,846	0,936	0,975	0,991	0,999	1,000	1,000
20	0,508	0,567	0,671	0,851	0,941	0,979	0,993	0,999	1,000	1,000
25	0,510	0,569	0,673	0,854	0,944	0,981	0,994	1,000	1,000	1,000
30	0,511	0,570	0,675	0,856	0,945	0,982	0,995	1,000	1,000	1,000
40	0,512	0,572	0,677	0,859	0,948	0,983	0,995	1,000	1,000	1,000
50	0,513	0,573	0,678	0,860	0,949	0,984	0,996	1,000	1,000	1,000
60	0,513	0,573	0,679	0,861	0,950	0,985	0,996	1,000	1,000	1,000
80	0,514	0,574	0,680	0,862	0,951	0,986	0,996	1,000	1,000	1,000
100	0,514	0,574	0,680	0,863	0,952	0,986	0,997	1,000	1,000	1,000
150	0,515	0,575	0,681	0,864	0,953	0,987	0,997	1,000	1,000	1,000
200	0,515	0,575	0,681	0,865	0,953	0,987	0,997	1,000	1,000	1,000

Intervalli di confidenza della variabile *t* di *Student*

Nc-1	*area sottesa dalla curva di Student tra –t e +t*									
	99,5%	***99,0%***	***97,5%***	***95,0%***	***90,0%***	***80,0%***	***75,0%***	***70,0%***	***60,0%***	***55,0%***
2	14,089	9,925	6,205	4,303	2,920	1,886	1,604	1,386	1,061	0,931
3	7,453	5,841	4,177	3,182	2,353	1,638	1,423	1,250	0,978	0,866
4	5,598	4,604	3,495	2,776	2,132	1,533	1,344	1,190	0,941	0,836
5	4,773	4,032	3,163	2,571	2,015	1,476	1,301	1,156	0,920	0,819
6	4,317	3,707	2,969	2,447	1,943	1,440	1,273	1,134	0,906	0,808
7	4,029	3,499	2,841	2,365	1,895	1,415	1,254	1,119	0,896	0,800
8	3,833	3,355	2,752	2,306	1,860	1,397	1,240	1,108	0,889	0,794
9	3,690	3,250	2,685	2,262	1,833	1,383	1,230	1,100	0,883	0,790
10	3,581	3,169	2,634	2,228	1,812	1,372	1,221	1,093	0,879	0,786
15	3,286	2,947	2,490	2,131	1,753	1,341	1,197	1,074	0,866	0,776
20	3,153	2,845	2,423	2,086	1,725	1,325	1,185	1,064	0,860	0,771
25	3,078	2,787	2,385	2,060	1,708	1,316	1,178	1,058	0,856	0,767
30	3,030	2,750	2,360	2,042	1,697	1,310	1,173	1,055	0,854	0,765
40	2,971	2,704	2,329	2,021	1,684	1,303	1,167	1,050	0,851	0,763
50	2,937	2,678	2,311	2,009	1,676	1,299	1,164	1,047	0,849	0,761
60	2,915	2,660	2,299	2,000	1,671	1,296	1,162	1,045	0,848	0,760
80	2,887	2,639	2,284	1,990	1,664	1,292	1,159	1,043	0,846	0,759
100	2,871	2,626	2,276	1,984	1,660	1,290	1,157	1,042	0,845	0,758
150	2,849	2,609	2,264	1,976	1,655	1,287	1,155	1,040	0,844	0,757
200	2,838	2,601	2,258	1,972	1,653	1,286	1,154	1,039	0,843	0,757

Test di ipotesi e di significatività

I test di ipotesi e di significatività già visti per le statistiche campionarie, possono essere estesi a problemi implicanti piccoli campioni, semplicemente sostituendo al valore z un corrispondente valore ottenuto dalla distribuzione t.

Medie

Per provare l'ipotesi nulla secondo la quale una popolazione (tendenzialmente normale) ha media μ_x, disponendo di piccoli campioni, si utilizza il valore t:

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{Nc-1}}$$

Differenza di medie

Volendo verificare l'ipotesi nulla che due campioni casuali di ampiezza Nc_1 ed Nc_2, non presentino differenze significative, ovvero siano estratti dalla stessa popolazione, si utilizza un valore t dato dalla espressione vista a proposito della differenza di medie, dove si rapporta la differenza delle medie all'errore standard della popolazione delle differenze:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{Nc_1} + \frac{1}{Nc_2}}} \quad \text{con } s = \sqrt{\frac{Nc_1 \cdot s_1^2 + Nc_2 \cdot s_2^2}{Nc_1 + Nc_2 - 2}}$$

Con il valore di t ottenuto si calcola l'area sottesa dalla curva di distribuzione di probabilità fra gli estremi $-t$ e $+t$, per il caso Nc_1+Nc_2-2 gradi di libertà, che fornisce il livello di significatività della differenza fra i campioni.

22. CRITERI NON PARAMETRICI

I test di significatività che abbiamo visto finora non possono prescindere da alcune ipotesi sulla distribuzione della popolazione dalla quale vengono estratti i campioni in analisi.

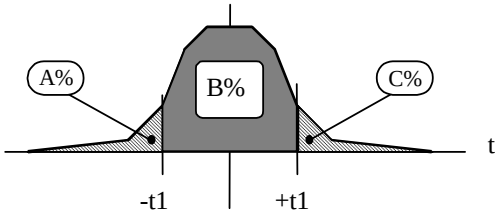
Possono verificarsi condizioni nelle quali tali ipotesi siano difficilmente formulabili, in tali casi possono applicarsi metodi di stima detti **non parametrici**, che sono indipendenti dai parametri che descrivono la distribuzione statistica della popolazione (asimmetria, media, deviazione standard), ed in tal caso si parla di statistica non parametrica.

23. ESERCIZI SULLA TEORIA DEI PICCOLI CAMPIONI

Ex69

Le tabelle dei valori numerici relativi alla distribuzione *t* di Student si riferiscono ad un intervallo finito dei valori campionati, e dunque non all'intervallo [+∞, -∞] come per la distribuzione normale (teoricamente relativa ad un numero infinito di campioni), ma all'intervallo di variazione misurato, espresso in percentili. Inoltre la tabella III, diversamente dalla II, fornisce non l'area sottesa dalla curva in funzione della variabile indipendente, ma l'inverso.

Determinare gli intervalli di confidenza relativi ad una distribuzione *t* di Student a 9 gradi di libertà in modo che:



- l'area della coda di destra sia 0.05.

In tal caso l'area complementare vale 1-0.05=0.95, ed il valore *+t1* rappresenta dunque il 95-esimo percentile. Dalla tabella III si ricava un valore di $t_{0.95}$=1.83;

- l'area totale di entrambe le code sia 0.05.

La curva di Student è simmetrica, dunque l'area di ciascuna coda vale 0.025. Allora l'area a sinistra di *+t1* vale (1-0.025)=0.975, e *+t1* rappresenta il 97.5esimo percentile. Il valore di *t1* corrispondente a 9 g.l. risulta 2.26; -t1 vale evidentemente -2.26;

- l'area tra *-t1* e *+t1* sia 0.99.

L'area delle code vale (1-0.99)=0.01, e l'area di ciascuna coda vale 0.01/2=0.005. Dunque *t1* è il 1-0.005=99.5-esimo percentile, e risulta $t_{0.995}$=>*t1*=3.25;

- l'area della coda di sinistra sia 0.01.

t = $-t_{0.99}$ = -2.82;

- l'area a sinistra di *t1* sia 0.9.

t1 è il 90-esimo percentile, e dunque *t1*=1.38.

è in ogni caso possibile riferirsi alle funzioni di MS-Excel:

distrib.t(t, Nc-1, 2) che fornisce l'area delle code, esterne all'intervallo [*-t; t*]

inv.t(p, Nc-1) che fornisce l'intervallo fiduciale per la variabile *t*, associato al livello di probabilità *p*.

Ex70

Calcolare i valori critici della variabile *t* per i quali l'area della coda di destra della distribuzione *t* vale 0.05 quando il numero di gradi di libertà vale rispettivamente 16, 27, 200.

Se l'area della coda di destra vale 0.05 allora l'area rimanente vale 1-0.05=0.95, e il valore cercato di *t* sarà il 95° percentile. Dalla tabella III risulta:

ν	$t_{0.95}$
16	1.75
27	1.70
200	1.645

Dalla tabella dei valori dell'area sottesa dalla curva normale si ricava che quando z=1.645 l'area della coda di destra vale 0.45, ovvero l'area totale compresa sotto alla porzione di curva z<1.645 vale 0.45+0.5=0-95.

Ex71

Confrontare i coefficienti di confidenza al 95% a due code relativi alla distribuzione normale ed alla distribuzione *t* di Student.

Per la distribuzione normale occorre cercare sulle tabelle il valore corrispondente ad un'area pari a 0.95/2=0.475 che vale 1.96, dunque i valori cercati sono z=±1.96.

Per la distribuzione t occorre cercare i valori che isolano due code ciascuna pari al 5%/2=2.5% dell'area totale. Allora i valori percentili saranno $t_{0.025}$ e $t_{(1-0.025)}=t_{0.975}$. Poiché la curva è simmetrica possiamo considerare solo quest'ultimo, e dalla tabella risulta:

ν	$t_{0.975}$	z
10	2.23	1.96
20	2.09	1.96
40	2.02	1.96
120	1.98	1.96
160	1.97	1.96

Ex72

Su di un campione di 10 misure del diametro di altrettanti ortaggi è stata determinata una media $\bar{x} = 4.38\ cm$ ed uno scarto quadratico medio s_x=0.86 cm. Determinare i limiti di confidenza al 95% ed al 99% per la misura del diametro.

Riferendosi in prima approssimazione al modello di distribuzione normale, relativo a grandi campioni, si determina:

$$\mu_x = \bar{x} \pm z_c(p)\cdot\frac{s_x}{\sqrt{Nc-1}}$$

con z_c(95%)=1.96 e z_c(99%)=2.58, si ottiene rispettivamente:

$$\mu_{x(95\%)} = 4.38 \pm 1.96\cdot\frac{0.86}{\sqrt{10-1}} \approx 4.38 \pm 0.56 \approx 4.38 \pm 12.8\%$$

$$\mu_{x(99\%)} = 4.38 \pm 2.58\cdot\frac{0.86}{\sqrt{10-1}} \approx 4.38 \pm 0.74 \approx 4.38 \pm 16.9\%$$

O in altre parole siamo confidenti al 95% ed al 99% che il valor medio del diametro della popolazione sarà compreso rispettivamente negli intervalli [3.82÷4.94] e [3.64÷5.12].

Riferendosi invece, come è più corretto fare, alla teoria dei piccoli campioni, per quando visto all'esercizio precedente, i limiti di confidenza risultano:

$$\mu_{x(95\%)} = \bar{x} \pm t_{0.975}\cdot\frac{s_x}{\sqrt{Nc-1}} \quad \text{e} \quad \mu_{x(99\%)} = \bar{x} \pm t_{0.995}\cdot\frac{s_x}{\sqrt{Nc-1}}$$

con N-1=10-1=9 risulta $t_{0.975}$=2.26 e $t_{0.995}$=3.25, dunque otteniamo:

$$\mu_{x(95\%)} = 4.38 \pm 2.26\cdot\frac{0.86}{\sqrt{10-1}} \approx 4.38 \pm 0.65 \approx 4.38 \pm 14.8\%$$

$$\mu_{x(99\%)} = 4.38 \pm 3.25\cdot\frac{0.86}{\sqrt{10-1}} \approx 4.38 \pm 0.93 \approx 4.38 \pm 21.3\%$$

O in altre parole siamo confidenti al 95% ed al 99% che il valor medio del diametro della popolazione sarà compreso rispettivamente negli intervalli [3.73÷5.03] e [3.45÷5.31].

Si nota che gli intervalli di confidenza calcolati in base al modello gaussiano sono più ristretti, ma ciò non significa affatto che la stima sia più precisa.

Ex73

Da una linea di produzione sono state prelevate Nc (Nc<20) bottiglie di latte le quali hanno fornito i seguenti valori di acidità: ……

Determinare la media e la deviazione standard campionari;

calcolare l'intervallo di confidenza al 95% per la stima del valore medio dell'acidità di tutto il latte trasportato;

stimare la deviazione standard della popolazione di bottiglie;

calcolare la probabilità che il latte contenuto in una bottiglia fornisca un valore di acidità superiore ad x^*;

calcolare la quantità di bottiglie per le quali l'acidità assume un valore compreso tra $x1$ ed $x2$;

calcolare i limiti di acidità entro i quali è compreso il 90% delle bottiglie;

calcolare la probabilità che l'acidità media valutata su di un campione di Nc bottiglie sia superiore ad x^*;

calcolare i limiti di acidità media entro i quali è compreso il 90% dei campioni di mpiezza Nc;

calcolare la probabilità che la differenza tra le acidità medie calcolate su due campioni di ampiezza Nc sia superiore a Dx^*;

Vengono prelevate altre Nc bottiglie di latte da un altro camion. Verificare che il latte proviene dalla stessa linea, ovvero valutare il livello di significatività della differenza tra le medie campionarie;

valutare la probabilità che due campioni casuali di ampiezza Nc, prelevati dai due camion differisca di una quantità superiore a Dx^*.

Ex74

Un fabbricante dichiara di produrre cavi con una resistenza media alla trazione pari a μ_x=8kN. Un campione di 6 cavi viene provato determinando una resistenza media alla trazione di $\bar{x}$=7750 N con uno scarto quadratico medio di s_x=145 N. Determinare la veridicità delle dichiarazioni del commerciante al livello di significatività dello 0.01.

L'ipotesi nulla è che la media del campione sia statisticamente uguale alla media della popolazione, ovvero:

$$8\ kN = \mu_x \approx \bar{x}$$

poiché il campione è piccolo ci riferiamo al modello di distribuzione t di Student. La variabile standardizzata t risulta:

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{Nc\text{-}1}} = \frac{7750\text{-}8000}{145}\sqrt{6\text{-}1} = -3.86$$

poiché ci interessano i valori superiori ad 8 kN si esegue il test sulla coda di destra della curva di distribuzione di t, allora il valore relativo ad un livello di significatività di 0.01 è $t_{0.99}$ che per ν=6-1 vale -3.36. Essendo -3.86 inferiore a -3.36 rifiutiamo l'ipotesi nulla.

Ex75

Relativamente alla determinazione della minima ampiezza campionaria, per la stima dell'intervallo fiduciale di un valore medio, supponendo che, il campione sia grande, che la popolazione sia infinita e distribuita quasi normalmente allora, come è già stato illustrato, stabilito un livello di probabilità, risulta:

$$\Delta\mu = 2z_c\frac{\sigma}{\sqrt{N_c}}$$

se N_c non è considerabile grande e lo scarto quadratico medio della popolazione non è noto, allora occorre introdurre qualche complicazione per riferirsi alla teoria di Student: con procedimento già illustrato, dalla $\mu = \bar{x} \pm t_c(N_c)\frac{s}{\sqrt{N_c - 1}}$ si ricava la $\Delta\mu = 2t_c(N_c)\frac{s}{\sqrt{N_c - 1}}$.

Il fatto che t_c sia funzione di N_c rende l'equazione non risolvibile in Nc con metodi analitici, però il valore di N_c può essere verificato con metodo numerico iterativo:

a) si parte ipotizzando un valore di tentativo di N_c e si raccoglie un campione causale di tale numerosità. Su tale campione di determina lo scarto quadratico medio s;

b) si calcola il valore di t_c corrispondente ad N_c;

c) si calcola un nuovo valore di N_c con la formula $N_c = 1 + \left(2t_c(N_c)\frac{s}{\Delta\mu}\right)^2$;

d) se il nuovo valore di N_c è molto diverso dal precedente si ritorna al punto **2**;

e) il valore così stabilizzato può essere utilizzato come confronto per verificare l'adeguatezza del campione selezionato.

Se inoltre non è possibile considerare la popolazione di provenienza dei campioni come infinita, occorre complicare ulteriormente il metodo:

anche l'equazione $\Delta\mu = 2t_c(N_c)\frac{s}{\sqrt{N_c - 1}}\sqrt{\frac{N_p - N_c}{N_p - 1}}$ non è risolvibile in N_c con metodi analitici, però tale valore può essere ricavato con metodo numerico iterativo:

[...]

- si calcola il valore di t_c corrispondente ad N_c;
- si calcola un nuovo valore di N_c con la formula $N_c = 1 + \left(2t_c(N_c) \frac{s}{\Delta\mu} \sqrt{\frac{N_p - N_c}{N_p - 1}} \right)^2$;
- se il nuovo valore di N_c è molto diverso dal precedente si ritorna al punto **2**.
- [...]

24. L'ANALISI DEI DATI CON *MICROSOFT EXCEL*

Test di significatività

Viene sperimentata una piccola modifica nel processo di produzione di biscotti, al fine di conseguire una maggiore efficienza energetica. Al fine di verificare che tale piccola modifica non abbia alterato significativamente la qualità del prodotto, vengono prelevati due campioni di ampiezza 10 (C1 e C2), rispettivamente dalla linea di produzione standard e da quella modificata Il comportamento meccanico dei biscotti viene caratterizzato con un test penetrometrico (T1) e con un test di flessione (T2):

T1	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
C1	21	23	27	12	27	19	14	19	19	23
C2	24	22	25	20	15	30	21	28	30	21

T2	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
C1	21	26	21	25	23	19	24	24	24	22
C2	25	27	26	28	30	25	26	28	27	28

utilizzando le opportune funzioni del programma *MS Excel*, si chiede di:

- valutare la significatività delle differenza tra C1 e C2. (Cosa possiamo dire circa la capacità dei test di differenziare le due classi?)
- selezionare due campioni casuali di 5 elementi (della stessa classe) e calcolare la significatività della differenza tra le due medie campionarie;
- selezionare due campioni casuali di 5 elementi (di classi diverse) e calcolare la significatività della differenza tra le medie campionarie;

Suggerimento: *è possibile utilizzare sia le funzioni standard di Excel che il modulo di analisi dei dati (Test t con varianze diverse per valutare la significatività della differenza tra i valori medi campionari, ovvero la probabilità che i due campioni provengano da una stessa popolazione). In particolare è possibile valutare la capacità discriminante dei due test (C1 - C2 sia con T1 che con T2). Si ottengono risultati identici con un test ANOVA.*

25. IL TEST χ^2

Frequenze osservate e teoriche

Nelle determinazioni di laboratorio esiste sempre uno scarto tra le misure effettuate e le previsioni statistiche. Supponiamo che in un certo campione si sia osservato un insieme di eventi $E_1, E_2,...,E_k$, rispettivamente con frequenza (cumulativa o no) $o_1, o_2,...,o_k$ (dette *frequenze osservate*), e che secondo la teoria tali frequenze siano invece $e_1, e_2,...,e_k$ (dette *frequenze teoriche o attese*). Occorre dunque stabilire se tali differenze sono significative, ed in particolare decidere se la nostra distribuzione osservata è significativamente vicina quella attesa.

Definizione di χ^2

Se è vera l'ipotesi che le nostre misure siano distribuite come prevede la teoria allora possiamo aspettarci che le deviazioni o_i-e_i siano piccole, al contrario la nostra ipotesi è sbagliata.

Per rendere più preciso il senso dei termini *piccolo* oppure *grande* si divide lo scarto per la radice di e_i.

Si dimostra infatti che se le o_k sono distribuite normalmente attorno al loro valore medio e_k, allora la loro deviazione standard risulta $\sqrt{e_k}$. Allora si considera il rapporto $\frac{o_k - e_k}{\sqrt{e_k}}$.

Per diminuire poi l'influenza degli scarti più piccoli (dovuti a piccoli errori sperimentali) ed esaltare quelli più significativi, il rapporto viene elevato al quadrato. Dunque un indicatore di discordanza tra distribuzioni osservate e distribuzioni teoriche è dato dalla statistica χ^2:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + ... + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j}$$

Tanto più χ^2 è superiore a zero, e tanto più le frequenze osservate differiscono da quelle teoriche.

Se k è almeno superiore a 5, allora la distribuzione campionaria di χ^2 si approssima bene alla distribuzione del χ^2 di Pearson con un numero di gradi di libertà dato da $\nu = k-1-m$, essendo m il numero di parametri della popolazione stimati attraverso le statistiche campionarie, determinati al fine di calcolare le frequenze teoriche:

$$Y = Y_0 \cdot \chi^{\nu-2} e^{-\chi^2/2}$$

Il test-χ^2 può essere usato per verificare l'adattamento di una distribuzione teorica (normale, di Student, di Fisher, ..) a quella ottenuta dai campioni. É anche possibile confrontare distribuzioni di frequenza discrete con distrib. continue, ricorrendo anche al concetto di frequenza cumulata.

In generale ci si aspetta che il valore dei singoli termini della somma sia circa 1, e poiché ci sono k termini, se $\chi^2 < k$, in prima approssimazione, la distribuzione osservata e quella attesa si accordano bene.

Un metodo migliore consiste nell'applicazione dell'usuale metodo di verifica delle ipotesi statistiche: si formula l'ipotesi $H0$ che tra la frequenza teorica calcolata e quella teorica non ci siano differenze al livello di significatività dello 0.05 o dello 0.01, poi si calcola il valore dell'indice χ^2 e lo si confronta con i valori critici $\chi^2_{0.95}$ e $\chi^2_{0.99}$.

Il test è ad una coda essendo i valori di χ^2 evidentemente solo positivi, dunque se il valore calcolato è maggiore del valore critico rifiutiamo l'ipotesi H_0, e possiamo affermare che c'è differenza al prescelto livello di significatività.

.

26. ESERCIZI SUL TEST $\chi 2$

Ex76

Viene lanciata una moneta per 200 volte, e si registrano 115 teste e 85 croci. Verificare che la moneta sia un generatore random ad un livello di significatività dello 0.05 e dello 0.01.

Le frequenze teoriche (attese=expected) sono rispettivamente e_1=100 ed e_2=100, mentre quelle osservate sono o_1=115 ed o_2=85.
Allora:

$$\chi^2 = \sum_{j=1}^{2} \frac{(o_j - e_j)^2}{e_j} = \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100} \approx 4.50$$

Il numero di gradi di libertà vale $k-1=2-1=1$, ed il valore critico $\chi^2_{0.95}$ vale 3.84 (tabella IV). Dunque poiché tale valore è inferiore a 4.50 se ne deduce che la moneta non è buona al livello di significatività dello 0.05.

27. ANALISI DELLA VARIANZA

La distribuzione F

Per analizzare le differenze campionarie occorre confrontare, oltre alle differenze tra le medie, anche le differenze tra le varianze. Per semplificare la matematica di tale sviluppo, tradizionalmente si considera invece il rapporto S_1^2/S_2^2 . Se tale rapporto è vicino ad 1, allora indica una piccola differenza tra i campioni, e v.v.
La distribuzione ottenuta da tutti i rapporti che è possibile ottenere da tutti i campioni delle due popolazioni è detta *distribuzione F di Fisher.*

Se consideriamo due campioni rispettivamente di dimensioni N1 ed N2, ottenuti da popolazioni approssimativamente normali, si definisce la statistica F:

$$F=\frac{\frac{\tilde{S}_1^2}{\sigma_1^2}}{\frac{\tilde{S}_2^2}{\sigma_2^2}}=\frac{S_1^2\cdot\frac{N_1}{N_1-1}\cdot\frac{1}{\sigma_1^2}}{S_2^2\cdot\frac{N_2}{N_2-1}\cdot\frac{1}{\sigma_2^2}}$$

tale statistica ammette una distribuzione data da:

$$Y=Y_0\frac{F^{\frac{\nu_1}{2}-1}}{\left(\nu_1\cdot F+\nu_2\right)^{\frac{\nu_1+\nu_2}{2}}}$$

essendo $\nu_i=N_i-1$, ed Y_0 una costante di normalizzazione. I valori calcolato dall'integrazione della Y, sono tabulati in funzione dei gradi di libertà e dei limiti di confidenza. Così si possono confrontare le varianze S_1^2 ed S_2^2 risultano o meno significativamente differenti.

Esperimenti ad un fattore

Si pone spesso il problema di valutare l'influenza di una o più variazioni combinate in un processo, sulle variazioni rilevate campionando i prodotti. I valori misurati su campioni provenienti da trattamenti diversi sono generalmente diversi. Ma anche i valori misurati su campioni provenienti da una stessa popolazione sono generalmente diversi. Occorre così essere in grado di valutare la significatività di tali differenze, ovvero di separare la componente di variazione dovuta ad una fluttuazione statistica da quella dovuta alle differenze di trattamento.

Se l'esperimento viene condotto variando il valore di una sola grandezza allora viene detto ad un fattore. P.e. possiamo valutare le rese in frumento derivanti da 4 diversi tipi di lavorazione del terreno. La variabile che viene modificata è il solo "tipo di lavorazione del terreno".
Per ciascuno dei 4 tipi di lavorazione si raccolgono in generale più campioni, ovvero gli appezzamenti di terreno variamente lavorati saranno in generale più di 4. Ciò viene fatto al fine di aumentare l'affidabilità del campione, diminuendo l'influenza di altre variabili di disturbo, come per esempio la non uniformità del terreno.
Se per ciascun tipo di lavorazione si preparano *b* appezzamenti di terreno, possiamo raccogliere i dati delle rese per ettaro in una tabella del tipo:

	Campo 1	**Campo 2**	**Campo 3**
Lavorazione 1	*resa 11*	*resa 12*	*resa 13*
Lavorazione 2	*resa 21*	*resa 22*	*resa 23*
Lavorazione 3	*resa 31*	*resa 32*	*resa 33*
Lavorazione 4	*resa 41*	*resa 42*	*resa 43*

In generale da un esperimento ad un fattore le osservazioni si ricavano da *a* gruppi indipendenti di campioni, ciascuno ripetuto *b* volte (nel nostro esempio *a*=4 e *b*=3). I dati sono così organizzabili in una struttura del tipo:

Trattamento 1	*X11*	*X12*	...	*X1b*	$\overline{X}_1$
Trattamento 2	*X21*	*X22*	...	*X2b*	$\overline{X}_2$
...	...	...	...	...	...
Trattamento 3	*Xa1*	*Xa2*	...	*Xab*	$\overline{X}_a$

Con il simbolo $\overline{X}_j$ è stata indicata la media delle misure riportate sulla riga j-esima:

$$\overline{X}_j=\frac{1}{b}\sum_{k=1}^{b}X_{jk}\qquad j=1,2,\ldots,a$$

Tali valori si chiamano *medie di gruppo* o *medie di trattamento*. Si definisce poi una **media generale** o **complessiva** come:

$$M=\overline{X}=\frac{1}{ab}\sum_{j=1}^{a}\sum_{k=1}^{b}X_{jk}$$

Ex77

Da due popolazioni distribuite normalmente, aventi varianza rispettivamente pari a $\sigma_1^2=16$ e $\sigma_2^2=25$, vengono estratti due campioni di dimensioni N1=9 e N2=12. Se le varianze dei campioni sono $S_1^2=20$ e $S_2^2=8$, determinare se la differenza tra le varianze è casuale o significativa al livello dello 0.05.

La variabile standardizzata di Fisher risulta:

$$F=\frac{\frac{\tilde{S}_1^2}{\sigma_1^2}}{\frac{\tilde{S}_2^2}{\sigma_2^2}}=\frac{S_1^2\cdot\frac{N_1}{N_1-1}\cdot\frac{1}{\sigma_1^2}}{S_2^2\cdot\frac{N_2}{N_2-1}\cdot\frac{1}{\sigma_2^2}}=\frac{20\cdot\frac{9}{9-1}\cdot\frac{1}{16}}{8\cdot\frac{12}{12-1}\cdot\frac{1}{25}}\approx 4.03$$

Il numero di gradi di libertà risultano $\nu1$=N1-1=8 e $\nu2$=N2-1=11. Dalla tabella V si ricava che F0.95=2.95.

Poiché la F calcolata vale 4.03 ed è maggiore di 2.95, concludiamo che la varianza del primo campione è significativamente più grande di quella del secondo.

28. ORGANIZZAZIONE DEGLI ESPERIMENTI A PIÙ FATTORI

Il piano sperimentale con classificazione gerarchica

Si possono presentare situazioni in cui i fattori di variazione sono tra loro concatenati in modo più o meno evidente. In tali casi si ricorre ad un particolare schema sperimentale che prende il nome di schema con classificazione gerarchica (*nested classification*).

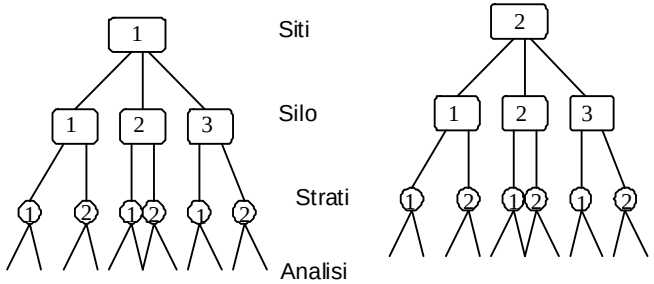
P.e. supponiamo di voler stabilire il grado di inquinamento da metalli pesanti in granaglie ottenute da coltivazioni ubicate in prossimità di industrie potenzialmente inquinanti (fonderie di piombo, lavorazioni su ceramica o vetro). Il confronto dei **siti** di coltivazione risulta essere il livello gerarchico principale: avremo un **testimone** (località lontana da ogni fonte di piombo) e p.e. 3 località nelle quali ci si aspetta una ricaduta di polveri pericolose.

Peraltro i grani raccolti nei diversi siti non sono conservati in un unico contenitore, ma in vari silos, dei quali non sappiamo nulla ne' sullo stato di conservazione ne' sulla possibilità di cedere metalli ai materiali in essi conservati. Se prelevassimo le granaglie da analizzare da un solo silo per ciascun sito, non potremmo evitare l'incertezza derivante dal fatto che il possibile inquinamento registrato derivi dal silo e non dalla località. L'unico modo per separare questi due effetti è di eseguire prelievi da due o più silos per ogni località. Se non dovessimo trovare nessuna differenza tra silos della stessa località, oltre a quella che potremmo attenderci sulla base della variabilità osservata entro lo stesso silos, allora potremmo ascrivere le differenze osservate ai diversi siti. Decidiamo allora di eseguire tre prelievi da tre diversi silos per ciascun sito.

Si presenta un altro problema: dato che i silos sono molto capienti, sono stati riempiti con partite giunte in tempi successivi e nulla ci garantisce che gli strati inferiori, quelli raccolti prima, abbiano il medesimo grado di inquinamento di quelli mediani o superiori. P.e. tra le diverse raccolte potrebbe esserci stata una pioggia. Decidiamo quindi di prelevare da ciascun silos due porzioni da esaminare: una nella metà inferiore ed una in quella superiore. Le eventuali differenze ci daranno una misura della variabilità all'interno di uno stesso silos.

Infine, sulle granaglie prelevate all'interno di un silo si esegue un'analisi in doppio, ottenendo due risultati per ogni porzione prelevata. Quest'ultima fase deve fornirci la varianza residua.

In totale i livelli gerarchici risultano tre: i siti; i silos; gli strati:



Si potrebbe pensare ad un quarto livello nel caso si impiegassero due diversi metodi di analisi (p.e. un incenerimento a secco ed un'ossidazione a umido). Si deve comunque tener presente che per lo stadio finale sono sempre necessari almeno due risultati per porzione analizzata, al fine, come già si è detto di valutare la varianza residua, quella imputabile ai fattori accidentali.

L'ANOVA a replicazione gerarchica è particolarmente adatta negli esperimenti preliminari, perché consente di chiarire a quale punto dell'esperimento si ha un ampliamento della varianza residua.

Nel caso delle granaglie risultano come si è detto tre livelli: a) le località; b) i silos; c) gli strati all'interno dei silos. Ogni porzione è stata analizzata in doppio: vi sono pertanto due replicazioni (n=2). Nulla vieta che le replicazioni siano più di due, in funzione del tipo di esperimento, della variabilità del carattere in esame, ecc.

Il contenuto in metalli pesanti delle granaglie (p.e. in microgrammi per 100 grammi di sostanza secca) può essere collezionato in una struttura del tipo:

Località	**Silos**	**Strato**	**Replicazioni**		**Totali**		
					Strato	Silos	Località
Testimone	1	L1					
		L2					
	2	L1					
		L2					
	3	L1					
		L2					
A	1	L1					
		L2					
	2	L1					
		L2					
	3	L1					
		L2					
B	1	L1					
		L2					
	2	L1					
		L2					
	3	L1					
		L2					
C	1	L1					
		L2					
	2	L1					
		L2					
	3	L1					
		L2					

29. L'ANALISI DELLA VARIANZA COMPORTANTE UN'INTERAZIONE TRA I FATTORI

Esperimenti fattoriali

Spesso si conducono esperimenti al fine di valutare l'effetto di due o più fattori di variazione, sia quando questi agiscono isolati che quando operano congiuntamente. Così p.e. se si studia l'effetto di concimazioni con azoto, fosforo e potassio su di una coltura, l'esperimento può consentire di accertare gli effetti imputabili a dosi diverse di azoto, di fosforo e di potassio impiegati isolatamente (**effetti principali**). Ma è evidente che in esperimenti di questo tipo ha interesse anche valutare l'effetto imputabile a tutte le combinazioni possibili delle tre concimazioni: avremo così **interazioni del primo ordine**, quando sono contemporaneamente presenti due dei fattori considerati e interazioni del secondo ordine, quando siano presenti tutti e tre i fattori principali.

I piani sperimentali destinati a studiare contemporaneamente l'effetto di due o più fattori vengono chiamati **fattoriali**.

Supponiamo che la concimazione azotata da sola provochi, su una certa specie di cereali, un aumento medio di produzione di granella pari a 0.5t/ha e che quella fosforica produca un incremento medio di 0.3t/ha. Se impiegando contemporaneamente i due fertilizzanti si avesse un aumento medio di 1.2t/ha rispetto al testimone non concimato, dovremmo dedurre che i due fattori si potenziano a vicenda (**interazione positiva**).

Con gli esperimenti fattoriali si studiano in generale due o più fattori principali, ognuno dei quali ad una diversa concentrazione. Per organizzare e condurre un esperimento in base ad un piano fattoriale, lo sperimentatore deve fissare un certo numero di **livelli** per ciascun fattore controllato e poi eseguire delle prove per tutte le combinazioni possibili. Se quindi vi sono l1 livelli per il primo fattore, l2 per il secondo, l3 per il terzo, occorrerà condurre l1xl2xl3 prove per ciascuna replicazione.

Il piano fattoriale 2x2

Si tratta del piano fattoriale più semplice, nel quale due fattori (A e B) vengono studiati a due livelli (A1, A2, B1, B2):

Prova	Fattore A	Fattore B	Risposta
1	A1	B1	r1
2	A2	B1	r2
3	A1	B2	r3
4	A2	B2	r4

La procedura d'analisi dei dati è simile a quella già vista: si calcolano la variazione totale, quella dovuta ai trattamenti e quella residua.

Si calcolano le devianze dovute ai due fattori isolati A e B e le si sottraggono a quella dovuta ai trattamenti. Quella rimasta è la variazione dovuta all'interazione dei due fattori:

Sorgente di variazione	**Devianza**	**G.L.**	**Varianza**	**F**
Fattore A				
Fattore B				
Interazione AxB				
Residuo				

Il valore di F calcolato, a confronto con quello tabulato fornisce il livello di significatività dell'interazione.

30. ESERCIZI SULL'ANALISI DELLA VARIANZA

Ex78

Nella tabella che segue sono riportate le rese (in tonnellate/ettaro) di un certo tipo di frumento trattato con i prodotti A, B, C.

31. L'ANALISI DEI DATI CON *MICROSOFT EXCEL*

Test di significatività

Delle uova di gallina sono state trattate con due procedimenti sperimentali (P1 e P2) per la decontaminazione del guscio. Successivamente è stata analizzata la carica microbica residua sul guscio dopo 1, 5, 10, 20 giorni di conservazione, su 5 repliche. Stabilire se i trattamenti sono stati efficaci.

I valori riportati sono misurati come log10 CFU e sono intesi come differenza tra un campione testimone e quello trattato. Evidentemente se il trattamento decontaminante fosse completamente inefficace, allora la popolazione microbica si comporterebbe come quella sul testimone e la tabellina riporterebbe sostanzialmente una serie di valori nulli.

t.conserv.- P1	1	2	3	4	5
1	1.0	1.1	0.9	1.2	1.0
5	1.2	1.2	1.0	1.3	1.1
10	1.4	1.5	1.2	1.3	1.5
20	1.4	1.4	1.1	1.5	1.6

t.conserv.-P2	1	2	3	4	5
1	0.9	1.0	0.9	1.0	1.0
5	1.3	1.3	1.2	1.4	1.2
10	1.5	1.6	1.4	1.5	1.6
20	1.7	1.6	1.6	1.7	1.8

Verificare la significatività dell'effetto del tempo e del trattamento.

Suggerimento: è possibile utilizzare sia le funzioni standard di Excel che il modulo di analisi dei dati (ANOVA ad un fattore), prima nel confronto 1-5-10-20, all'interno dello stesso trattamento, e poi nel confronto tra P1 e P2, in corrispondenza di uno specifico valore di tempo di conservazione.

32. ANALISI DELLE SERIE TEMPORALI

Introduzione

Se un sistema varia nel tempo, e ne viene prelevato un campione ad intervalli costanti, se ne ottiene una serie temporale (o serie storica).

P.e. una serie temporali sono costituite dall'andamento delle vendite di frutta durante l'anno, dalla temperatura media in una serra, dal numero di nuovi iscritti al corso di STA.

Matematicamente una serie temporale è costituita dunque da una serie discreta di osservazioni del tipo:

$y(t_1), y(t_2), y(t_3), ...$

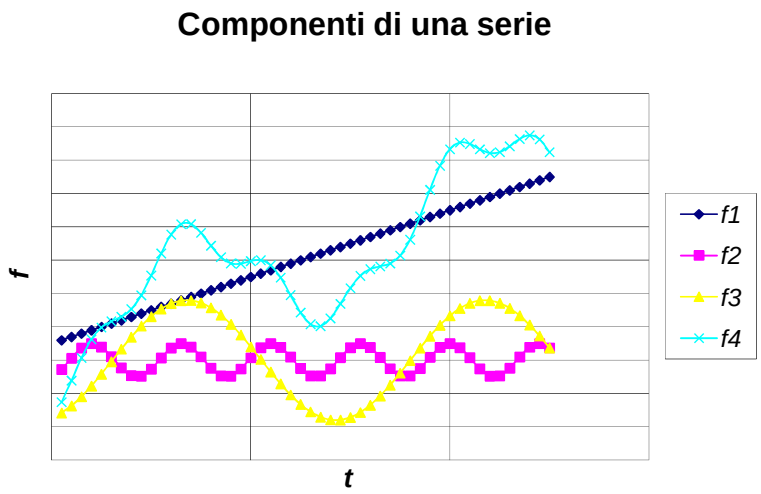
Spesso è possibile individuare delle periodicità all'interno delle serie temporali. Cioè avviene che valori quasi uguali, si ripresentino ad intervalli quasi costanti.

Componenti o movimenti di una serie temporale

Consideriamo le tre equazioni:

$f1(t)=a\cdot t+b, \quad f2(t)=A\cdot sin(\omega\cdot t+\phi), \quad f3(t)=B\cdot sin(\omega\cdot t\cdot K+\phi)$

i cui grafici risultano (qualitativamente):



la funzione *f4* è la risultante dalla somma delle *f1, f2* ed *f3*.

L'andamento di *f4* risulta abbastanza rappresentativo di una serie temporale realistica e ne contiene alcuni elementi caratteristici.

Movimenti secolari o a lungo termine: (componente *f1*), si riferiscono alla direzione generale che il grafico di una serie temporale sembra prendere in un lungo intervallo di tempo. Tale componente (o movimento) è detta variazione secolare o tendenza (*trend*).

Movimenti ciclici: (componente *f3*), si riferiscono alle oscillazioni semiperiodiche attorno ad un valore medio. Tale componente individua generalmente i movimenti ciclici di più lungo periodo come i cicli di recessione, depressione, ripresa e prosperità.

Movimenti stagionali: (componente *f2*), sono movimenti semiperiodici di frequenza più alta. Possono sovrapporsi molte di queste componenti con periodo annuale, stagionale, mensile, settimanale, giornaliero. Individuano fenomeni come i cicli vegetativi delle piante, la variazione della tipologia dei consumi alimentari al variare della stagione, ecc.

Movimenti casuali: alle componenti precedenti si sovrappone sempre una fluttuazione irregolare, con distribuzione generalmente quasi-normale. Se tale componente ha piccola intensità, corrisponde ad un disturbo, se invece l'influenza è notevole allora individua fenomeni eccezionali come inondazioni, scioperi, elezioni, epidemie, ecc. Generalmente gli effetti eccezionali di tali variazioni, anche se di ampiezza notevole, tendono ad attenuarsi dopo un periodo caratteristico.

Analisi delle serie temporali

Data una serie storica si pone dunque il problema di calcolarne le componenti (armoniche). Generalmente ci si riferisce ad un modello semplificato come quello già visto, di tipo additivo, o di tipo moltiplicativo:

$Y=T+C+S+I$, *oppure* $Y=T\cdot C\cdot S\cdot I$

essendo *T* la componente del trend, *C* la componente ciclica di fondo, *S* i movimenti stagionali ed *I* quelli irregolari.

Il problema ha un grande interesse applicativo sia per la possibilità di interpretazione di un fenomeno (dai dati si risale al meccanismo che li ha generati) che per quella di effettuare estrapolazioni (previsioni).

Campi di applicazione tipici sono il controllo delle scorte di magazzino, la pianificazione delle vendite, previsioni economiche, previsioni meteorologiche, dispositivi di controllo.

Medie mobili

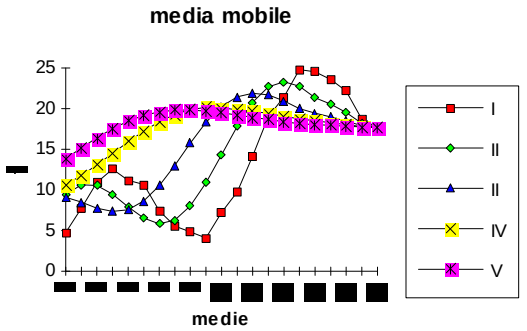
Data la serie temporale $Y_1, Y_2, Y_3,...$

si definisce come media mobile d'ordine *N* la sequenza di medie aritmetiche:

$$\frac{1}{N}\sum_{i=1}^{N} Y_i \quad \frac{1}{N}\sum_{i=2}^{N+1} Y_i \quad \frac{1}{N}\sum_{i=3}^{N+2} Y_i \quad ....$$

L'operatore media mobile funziona come un *filtro passa-basso*, ovvero ha la proprietà di ridurre l'intensità della variazione presente in un insieme di dati.

Con applicazioni ripetute di un operatore mobile di qualsiasi ordine, si riescono ad eliminare dapprima le fluttuazioni casuali, e poi si attenuano, tutte le componenti a frequenza più alta:



A volte si utilizzano medie mobili ponderate. P.e. un sistema automatico di acquisizione della temperatura in una cella frigorifera, per evitare brusche manovre del sistema di controllo dovrebbe poter eliminare le fluttuazioni casuali di alta frequenza (p.e. il passaggio di un operatore). Così un modo possibile di acquisire una nuova misura (T_n), mediandola con quelle già acquisite potrebbe essere del tipo:

$$T_n \leftarrow \frac{W_1\cdot T_n + W_2\cdot T_{n-1}}{W_1+W_2} \text{ oppure } T_n \leftarrow W\cdot T_n + (1-W)\cdot T_{n-1}$$

Occorre notare che già il valore T_{n-1} contiene l'informazione di tutte le misure pregresse T_{n-2}, T_{n-3},....., ma attenuate di un fattore W^n.

E' semplice rendersi conto del fatto che l'operatore media mobile tronca le serie all'inizio o alla fine:

Consumo di latte

Mese	Valori registrati	Media su 3 mesi	Media su 5 mesi
gen	200	.	.
feb	135	.	.
mar	195	176.7	.
apr	197	175.8	.
mag	310	234.2	207.5
giu	175	227.5	202.5
lug	155	213.3	206.5
ago	130	153.3	193.5
set	220	168.3	198.0
ott	277	209.2	191.4
nov	235	244.2	203.5
Media	*197.4286*	*197.3714*	*200.4143*
Dev. Std.	**57.78944**	**31.38597**	**6.271743**

Stima del trend

Una stima della componente *T* può essere ottenuta generalmente attraverso un procedimento di regressione lineare ai minimi quadrati applicato ai dati grezzi, oppure con applicazioni ripetute dell'operatore media mobile.

L'operazione aumenta di efficacia se sono note a priori le frequenze delle variazioni cicliche.

Stima delle variazioni stagionali e cicliche

Sviluppo in serie di Fourier: una qualsiasi funzione del tempo *f(t)* continua su [*-T/2, T/2*] può essere ricondotta ad una funzione del tipo:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[A_n \cdot \cos(n \cdot \omega_0 \cdot t - \Phi_n)\right]$$

con:

$$\omega_0 = \frac{2\pi}{T} \quad A_n = \sqrt{a_n^2 + b_n^2} \quad \Phi_n = \operatorname{atan}\left(\frac{b_n}{a_n}\right)$$

dove il valore dei diversi coefficienti può essere ricavato per integrazione:

$$a_n = \frac{2}{T} \int_{-T/2}^{+T/2} \left[f(t) \cdot \cos(n \cdot \omega_0 \cdot t)\right] dt \qquad b_n = \frac{2}{T} \int_{-T/2}^{+T/2} \left[f(t) \cdot \operatorname{sen}(n \cdot \omega_0 \cdot t)\right] dt$$

p.e. lo sviluppo in serie di Fourier di $f(x)=x^2$ diviene:

$$x^2 = \frac{\pi^2}{3} - 4\left(\cos(x) - \frac{\cos(2x)}{2^2} + \frac{\cos(3x)}{3^2} - \frac{\cos(4x)}{4^2} + \ldots.\right)$$

Rappresentando in un grafico le ampiezze A_n in funzione dell'ordine di armonica si ottiene lo *spettro di frequenza* della funzione *f(t)*.

Se in particolare la funzione *f(t)* deriva dalla somma di una serie di funzioni periodiche, allora è possibile estrarre tutte le componenti armoniche fino ad un ordine prefissato semplicemente calcolando per integrazione i coefficienti A_n.

Se in particolare la funzione *f(t)* è nota per punti, come nel caso delle serie storiche, allora occorrerà calcolare i coefficienti A_n con un metodo di integrazione numerica, spingendosi fino all'ordine di armonica desiderato.

É evidentemente possibile prima calcolare la retta di tendenza e sottrarla ai dati grezzi sui quali applicare l'analisi di Fourier, oppure sottrarre dai dati grezzi i risultati dell'analisi di Fourier, per ricavare un dato di tendenza generale.

Aliasing: poiché disponiamo solo di un segnale campionato nel tempo, e non di una funzione continua, perdiamo risoluzione nel campo delle frequenze più alte. Si dimostra che l'armonica di più alta frequenza che riusciamo ad individuare è quella il cui periodo vale la metà dell'intervallo del tempo di campionamento (teorema di Shannon).

33. L'APPROSSIMAZIONE E L'INTERPOLAZIONE AI MINIMI QUADRATI

Relazioni tra variabili

Finora abbiamo considerato il caso di analisi statistica di una sola variabile aleatoria, quando invece si raccolgono misure di tipo diverso può esistere qualche forma di legame, più o meno marcata, tra le variabili che si osservano.

Per esempio il colore di un frutto è legato verosimilmente al grado zuccherino della polpa o allo stato di maturazione; la durezza di un terreno è collegata alla sua massa volumica; la massa di una pianta alla la sua età; l'altezza dei padri a quella dei figli. Pur in assenza di precisi legami fisici, le due variabili possono in qualche modo ritenersi collegate.

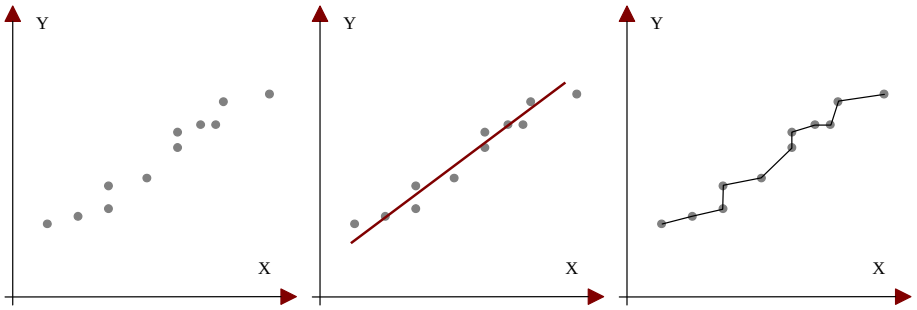
Altri esempi:

- l'altezza di caduta di un frutto e l'effetto del danneggiamento;
- la temperatura di una cella frigorifera e lo stato di conservazione dei prodotti;
- la temperatura di un processo industriale e la carica batterica residua sul prodotto;

Interpolazione ed approssimazione

Se riteniamo che esista un collegamento tra diverse serie di dati, ottenute campionando un certo fenomeno fisico, possiamo cercare di esprimere tale relazione in forma matematica, attraverso un'equazione.

Fissiamo ora l'attenzione su di un problema con due sole variabili, p.e. il volume (x) ed il peso (y) di una popolazione di N frutti. Riportando su di un piano x-y tutti gli N valori rilevati (x_i, y_i) si ottiene un insieme di punti detto **diagramma a dispersione** (*scattergram*). La nuvola di punti spesso si dispone in modo da rendere evidente un qualche andamento preferenziale.



Una curva che passi esattamente per ciascun punto (x_i,y_i), ovvero un'equazione $y(x)$, tale che:

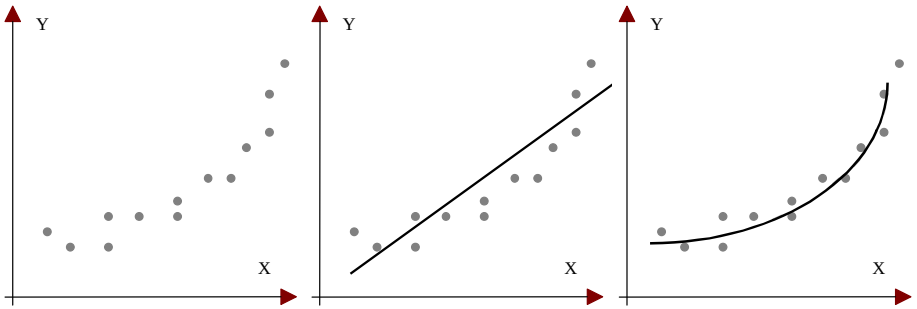
$$y(x_i)=y_i \quad \text{con} \quad i=[1 \div N]$$

viene detta curva **interpolante**.

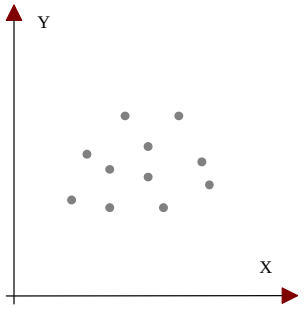
Una curva invece che non passa necessariamente per ciascun punto della nuvola, ma *abbastanza vicino*, ovvero:

$$y(x_i) \cong y_i \quad \text{con} \quad i=[1 \div N]$$

viene detta curva **approssimante**. Nella pratica, così come in molta manualistica, i due termini vengono comunemente confusi.



Se i dati misurati possono essere ben approssimati da una retta, si dice che tra le variabili X ed Y esiste una **relazione lineare**. Viceversa o esiste una relazione di tipo **non lineare**, o non esiste alcun tipo di relazione.



Se i punti sul diagramma a dispersione tendono a formare una nuvola con la stessa densità in ogni direzione, allora significa che i dati *y* tendono ad essere indipendente da *x*, si dice allora che i dati sono *incorrelati*.

Nel passaggio dal discreto al continuo otteniamo:

- un'espressione sintetica in grado di riassumere anche grandi insiemi di numeri;
- un'equazione che ci permette di fare osservazioni diverse come previsioni (estrapolazioni); valutazioni sul fenomeno in studio anche in corrispondenza di quei valori delle variabili che non sono stati rilevati (interpolazioni); o anche operazioni diverse come la ricerca di condizioni di massimo o di minimo;
- un modo per capire quali dei parametri che descrivono lo stato del sistema sono più importanti.

Esempi da tesi di laurea: grafici *Dino.ppt*

Varie curve interpolanti/approssimanti

Le relazioni analitiche tra i dati sperimentali possono essere espresse mediante una forma polinomiale:

$$y = \sum_i a_i \cdot x^i = a_0 \cdot x^0 + a_1 \cdot x^1 + a_2 \cdot x^2 + \ldots + a_n \cdot x^n$$

ovvero $$y = \sum_i a_i \cdot x^i = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_n \cdot x^n$$

che come caso particolare ammette il polinomio di grado zero:

$$y = a_0$$

e la retta:

$$y = a_0 + a_1 \cdot x$$

Oppure mediante altre espressioni:

Curva esponenziale (Popolazioni di microrganismi; decadimento radioattivo, accrescimento piante)	$y = a_0 + a_1 \cdot e^{a_2 \cdot x}$
razionale	$y = \dfrac{1}{a_0 + a_1 \cdot x^{a_2}}$
potenza	$y = a_0 + a_1 \cdot x^{a_2}$
logistica	$y = \dfrac{1}{a_0 + a_1 \cdot e^{a_2 \cdot x}}$
logaritmica	$y = a_0 + a_1 \cdot \log(a_2 \cdot x)$
serie di Fourier (serie storiche)	$y = a_0 + \sum_{i=1}^{N} [a_i \cdot \sin(\omega_i \cdot x + \varphi_i)]$

- Esempio con *MS-Excel* o *CurvExpert->BeanRoot.Dat* - *Tools*: Curve Finder

Generalmente i programmi di calcolo che si impiegano per determinare i coefficienti di tali forme riescono a trovare il *miglior* tipo di compromesso (*best fitting*), oppure come primo orientamento ci si può aiutare riportando i propri valori su un grafico *semi/bi*-logaritmico. Infatti, ricordando le proprietà della funzione logaritmo naturale:

$\log(x \cdot y) = \log(x) + \log(y)$

$\log(x / y) = \log(x) - \log(y)$

$\log(x^y) = y \cdot \log(x)$

(p.e. $\log\left(\frac{A^p B^q C^r}{D^s E^t}\right) = p \cdot \log A + q \cdot \log B + r \cdot \log C - s \cdot \log D - t \cdot \log E$)

ed applicando la funzione logaritmo ad entrambi i membri della equazione esponenziale si ottiene:

$\mathrm{Log}(y - a_0) = \mathrm{Log}(a_1) + x \cdot a_3 \cdot \mathrm{Log}(a_2) = A_1 + x \cdot A_2$

ovvero risulta che se si riportano sull'ascissa di un grafico i valori *x* e in ordinata i valori *Log(y-a0)* (*grafico semilogaritmico*) si ottiene una retta.

Dunque se da un diagramma a dispersione si evidenzia che la relazione tra *Log(Y)* ed *x* è lineare, allora l'equazione approssimante sarà del tipo esponenziale.

Similmente applicando la funzione logaritmo ad entrambi i membri della curva geometrica otteniamo:

$\mathrm{Log}(y - a_0) = \mathrm{Log}(a_1) + a_2 \cdot \mathrm{Log}(x) = A_1 + a_2 \cdot \mathrm{Log}(x)$

ovvero risulta che se si riportano sull'ascissa di un grafico i valori *Log(x)* e in ordinata i valori *Log(y-a0)* (*grafico bilogaritmico*) si ottiene una retta.

E parimenti applicando la funzione logaritmo ad entrambi i membri della espressione della curva iperbolica otteniamo:

$\mathrm{Log}(1/y - a_0) = A_1 + a_2 \cdot \mathrm{Log}(x)$

ovvero risulta che se si riportano sull'ascissa di un grafico i valori *Log(x)* e in ordinata i valori *Log(1/y-a₀)* (*grafico bilogaritmico*) si ottiene una retta.

Dunque se si evidenzia che la relazione tra *Log(Y)* e *Log(x)* è lineare, allora l'equazione approssimante sarà del tipo della progressione geometrica o iperbolica.

Simmetricamente si ottengono rette applicando la funzione esponenziale ad espressioni di tipo logaritmico.

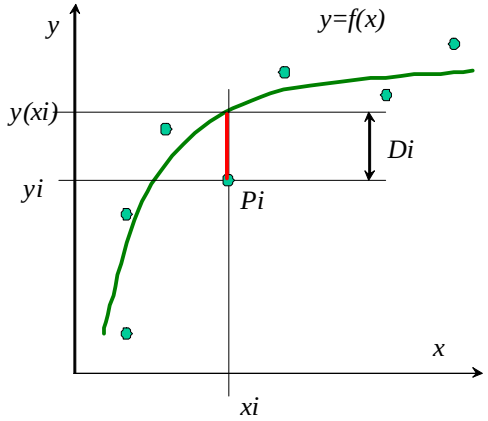
Il metodo dei minimi quadrati

Esistono diversi criteri, a volte arbitrari a volte sostenuti da un significato fisico, in base ai quali stabilire in che misura una curva si adatta ad una nuvola di punti meglio di un'altra. In generale sono criteri basati sulla minimizzazione del valore di un qualsiasi parametro di dispersione, ovvero un parametro che sintetizza in un unico valore una sorta di scostamento medio tra la curva approssimante e le misure sperimentali.

Data dunque la popolazione di N punti (x_i,y_i) ed una curva $y(x)$,

per ciascun punto si definisce lo *scarto* D_i come $D_i=y(x_i)-y_i$.

Se $D_i=0$ l'adattamento in $x=x_i$ è perfetto.

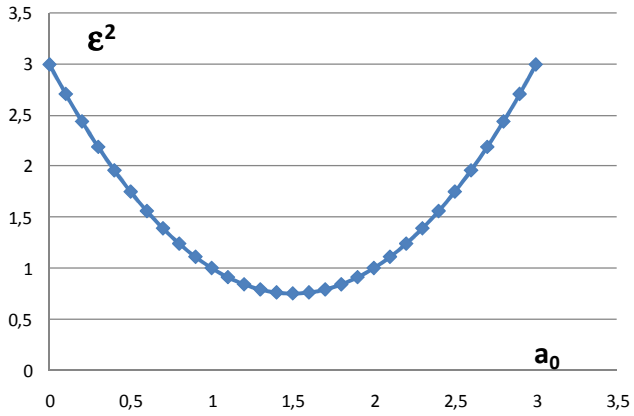


Dunque la bontà dell'adattamento della curva ai punti, potrebbe essere espressa per esempio come una media della popolazione degli scarti. Si nota che usando una media aritmetica, avendo sia scarti con segno positivo che scarti con segno negativo, avremmo piccoli valori anche per adattamenti non buoni. Possiamo riferirci ai valori assoluti, i quali però danno lo stesso peso sia alle piccole che alle grandi differenze (e si presta male ai trattamenti analitici) così a volte viene adottata una media pesata dove il peso dipende dalla stessa distanza tra punto e curva.

Un metodo assai impiegato è quello della somma dei quadrati, si utilizza cioè come indicatore della bontà dell'adattamento l'espressione:

$$\varepsilon^2 = \sum_{i=1}^{N} D_i^2 = \sum_{i=1}^{N} [y_i - y(x_i)]^2$$

Poiché tanto più piccolo è il valore di ε^2 e tanto migliore è l'adattamento del modello ai dati sperimentali, si definisce come modello ai minimi quadrati, quella espressione $y(x)$ che rende minimo il valore ε^2.

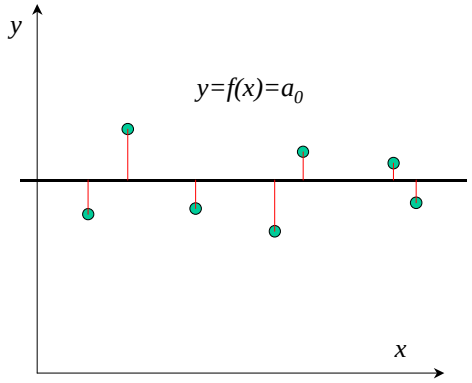


- Esempio con una retta: *WinStats.exe ->Demos->LastSquares*

La retta ai minimi quadrati

Dati gli N punti (x_i,y_i) determiniamo la retta orizzontale che meglio approssima tali punti, secondo il criterio dei minimi quadrati. (E' semplice mostrare intuitivamente che tale retta deve esistere e che è unica).

Il modello è dunque $y(x)=a_0$.



http://hadm.sph.sc.edu/courses/J716/demos/leastsquares/leastsquaresdemo.html

Lo scarto i-esimo, misurato in ordinata, è evidentemente:

$$D_i = y_i - y(x_i) = y_i - a_0 \qquad \text{e dunque} \qquad D_i^2 = y_i^2 - 2\,y_i a_0 + a_0^2$$

ed ε^2 risulta:

$$\varepsilon^2 = \sum_i D_i^2 = \sum_i (y_i^2 - 2y_i\,a_0 + a_0^2)$$

Per determinare il valore di a_0 che rende minimo il valore di ε^2 occorre cercare i valori che ne rendono nulla la derivata prima:

$$\frac{d(\varepsilon^2)}{da_0} = \frac{d}{da_0}\sum_i D_i^2 = 0 \rightarrow \sum_i \frac{d}{da_0}(y_i^2 - 2y_i\,a_0 + a_0^2) = 0 \rightarrow$$

$$\sum_i (-2y_i + 2a_0) = 0 \rightarrow \sum_i (-2y_i) + \sum_i (2a_0) = 0 \rightarrow$$

$$-2\sum_i (y_i) + 2\sum_i (a_0) = 0 \rightarrow 2\sum_i y_i = 2\sum_i a_0$$

e dunque ricavando il valore di a_0 risulta:

$$\sum_i y_i = N \cdot a_0 \rightarrow a_0 = \frac{1}{N}\sum_i y_i = \bar{y}$$

cioè risulta che la media aritmetica è la miglior stima secondo il criterio dei minimi quadrati.

Se invece si vuole determinare l'espressione di una retta comunque inclinata $y=a_0+a_1 \cdot x$ in grado di approssimare i dati sperimentali occorre determinare il valore dei coefficienti a_0 ed a_1 che rendono minima la somma degli scarti quadratici:

$$D_i^2 = (y_i - y(x_i))^2 = (y_i - (a_0 + a_1 \cdot x_i))^2$$

$$\varepsilon^2 = \sum_i D_i^2 = \sum_i \left[y_i^2 - 2y_i(a_0 + a_1 \cdot x_i) + a_0^2 + 2a_0 \cdot a_1 \cdot x_i + a_1^2 \cdot x_i^2\right]$$

e dunque risulta:

$$\begin{cases} \dfrac{d\varepsilon^2}{da_0} = 0 \rightarrow \sum_i \left[-2y_i + 2a_0 + 2a_1 \cdot x_i\right] = 0 \\ \dfrac{d\varepsilon^2}{da_1} = 0 \rightarrow \sum_i \left[-2y_i x_i + 2a_0 x_i + 2a_1 x_i^2\right] = 0 \end{cases}$$

il valore di a_0 ricavato dalla prima equazione è:

$$\sum_i 2a_0 = \sum_i 2y_i - \sum_i 2a_1 \cdot x_i \rightarrow a_0 = \frac{\sum_i 2y_i - \sum_i 2a_1 x_i}{2N} = \bar{y}_i - a_1\bar{x}_i = \bar{Y} - a_1 \cdot \bar{X}$$

ed il valore di a_1 contenuto nella seconda è:

$$\sum_i \left(-2y_i x_i + 2a_0 \cdot x_i + 2a_1 x_i^2\right) = 0 \rightarrow \sum XY = a_0 \sum X + a_1 \sum X^2 = 0$$

Dal sistema tra queste due equazioni (dette equazioni normali della retta dei minimi quadrati) si ricavano i valori a_0 ed a_1:

$$a_0 = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - \left(\sum X\right)^2} \qquad a_1 = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

Il punto di coordinate $\left(\overline{X}=\overline{x}_i=\frac{\sum_i x_i}{N} \quad \overline{Y}=\overline{y}_i=\frac{\sum_i y_i}{N}\right)$ è detto *centro* del sistema di punti (x_i,y_i). Si dimostra che tale punto appartiene alla retta ai minimi quadrati.

La parabola ai minimi quadrati

La parabola d'ordine 2

Dati gli N punti (x_i,y_i) determiniamo il ramo di parabola passante per l'origine che meglio approssima tali punti, secondo il criterio dei minimi quadrati:

il modello è dunque $y(x)=a_0 \cdot x^2$.

Lo scarto i-esimo, misurato in ordinata, è evidentemente:

$$D_i = y_i - y(x_i) = y_i - a_0 \cdot x_i^2 \text{ e dunque } D_i^2 = y_i^2 - 2\, y_i\, a_0 \cdot x_i^2 + a_0^2 \cdot x_i^4$$

ed ε^2 risulta:

$$\varepsilon^2 = \sum_i D_i^2$$

Per determinare il valore di a_0 che rende minimo il valore di ε^2 occorre cercare i valori che ne rendono nulla la derivata prima:

$$\frac{d(\varepsilon^2)}{da_0} = \frac{d}{da_0}\sum_i D_i^2 = 0 \rightarrow \sum_i (-2y_i x_i^2 + 2a_0 x_i^4) = 0 \rightarrow$$

$$\rightarrow 2\sum_i y_i x_i^2 = 2\sum_i a_0 x_i^4$$

e dunque ricavando il valore di a_0 risulta:

$$a_0 = \frac{\sum_i y_i x_i^2}{\sum_i x_i^4}$$

La parabola d'ordine qualsiasi

Dati gli N punti (x_i,y_i) determiniamo il ramo di parabola d'ordine n passante per l'origine che meglio approssima tali punti, secondo il criterio dei minimi quadrati:

il modello è dunque $y(x)=a_0 \cdot x^n$.

Lo scarto i-esimo, misurato in ordinata, è evidentemente:

$$D_i = y_i - y(x_i) = y_i - a_0 \cdot x_i^n \text{ e dunque } D_i^2 = y_i^2 - 2\, y_i\, a_0 \cdot x_i^n + a_0^2 \cdot x_i^{2n}$$

ed ε^2 risulta:

$$\varepsilon^2 = \sum_i D_i^2$$

Per determinare il valore di a_0 che rende minimo il valore di ε^2 occorre cercare i valori che ne rendono nulla la derivata prima:

$$\frac{d(\varepsilon^2)}{da_0} = \frac{d}{da_0}\sum_i D_i^2 = 0 \rightarrow \sum_i (-2y_i x_i^n + 2a_0 x_i^{2n}) = 0$$

e dunque ricavando il valore di a_0 risulta:

$$a_0 = \frac{\sum_i y_i x_i^n}{\sum_i x_i^{2n}}$$

E' semplice verificare che per n=2 si ritrovano i risultati del caso precedente.

Il quartetto di Anscombe

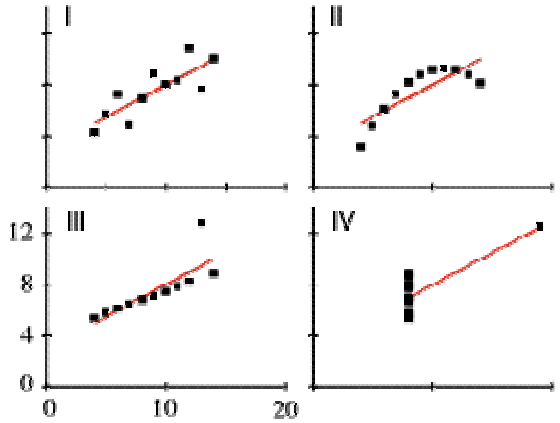
Consideriamo le 4 serie di dati sotto riportate, composte ciascuna da 11 punti (F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician,* 27 [February 1973], 17-21):

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

I quettro set di dati sono simili per molti aspetti statistici:

- media aritmetica dei valori $x = 9.0$
- media aritmetica dei valori $y = 7.5$
- equazione della retta ai minimi quadrati: $y = 3 + 0.5x$
- somma dei quadrati degli scarti = 110.0
- varianza dei valori $x = 27.5$
- coefficiente di correlazione = 0.82
- coefficiente di determinazione = 0.67

Tuttavia se visualizziamo il diagramma a dispersione si rivelano evidenti differenze tra le quattro serie di dati:



Conclusione: è necessario osservare i dati in analisi e capire il significato delle tecniche statistiche.

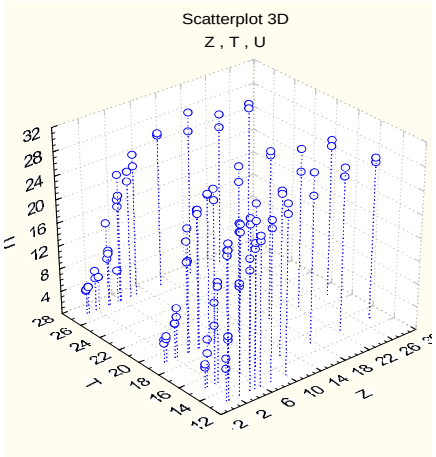
☞ Esempio con *OutlierEffect.zip*

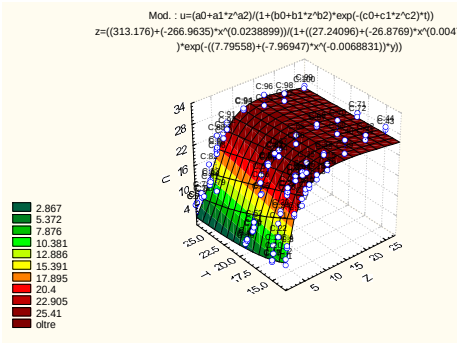
Problemi in più variabili

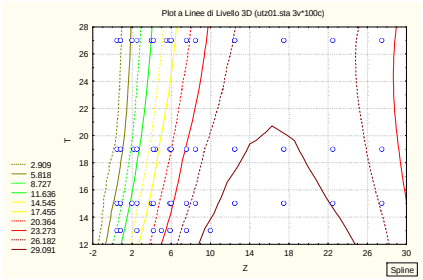
Il metodo della regressione ai minimi quadrati può applicarsi anche a problemi in più variabili (*regressione multipla*), ovvero dove si disponga dei dati sperimentali (x_i,y_i,z_i) si può cercare l'espressione di una superficie di regressione $z=f(x,y)$ tale da approssimare al meglio la nuvola di punti. Il procedimento è identico a quello visto nel caso dei problemi in sole due variabili:

- si scrive l'espressione dello scarto i-esimo in funzione di un numero k di parametri incogniti;
- si costruisce la sommatoria dei quadrati di tali scarti ε^2;
- si imposta un sistema costituito dalle k equazioni (generalmente non lineari) che esprimono l'annullarsi delle derivate prime di ε^2 calcolate rispetto a ciascuno dei k parametri incogniti.

```
t: Tempo (in giorni nel mese di luglio)
z: Profondità (in cm)
u : Umidità (in punti percentuali sul secco)
```







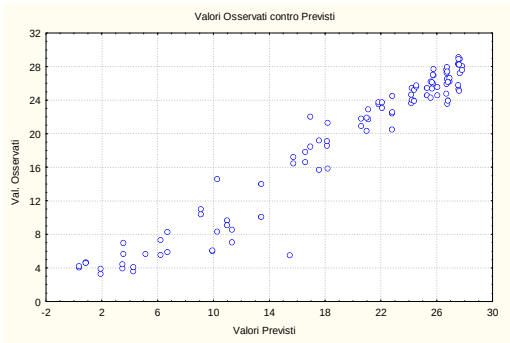
- Eq. logistica in *t*, con parametri dipendenti da *z* secondo una potenza:

$$u = \frac{A0 + A1z^{A2}}{1 + \left(B0 + B1z^{B2}\right)e^{-\left(C0 + C1z^{C2}\right)t}}$$

Mod. : u=(a0+a1*z^a2)/(1+(b0+b1*z^b2)*exp(-(c0+c1*z^c2)*t)...

Continua... Var. dip: U Perd.:(OBS-PRED)**2
Perd fin.: 432.60578525 R=.96987 Varianza spiegata: 94.065%

N=100	A0	A1	A2	B0	B1	B2	C0
Stima	313.1762	-266.964	.023890	27.24097	-26.8769	.004783	7.795580



La regressione lineare nel caso generale

Supponiamo che in corrispondenza di (m+1) valori x_i, anche non tutti distinti, siano rilevati i valori y_i. Supponiamo inoltre di avere scelto n+1 (<m+1) funzioni (polinomiali, esponenziali, trigonometriche, ecc.), dette basi, $\varphi_0(x)$, $\varphi_1(x)$,...,$\varphi_n(x)$, e di voler approssimare il fenomeno in esame rappresentato dai dati (x_i,y_i) con una combinazione lineare delle funzioni $\varphi_i(x)$. Ovvero il modello f(x) del tipo:

$$f(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + ... + c_n\varphi_n(x) = \sum_{i=0}^{n} c_i \cdot \varphi_i(x)$$

i coefficienti c_i si determinano imponendo che risulti minimo il residuo tra dati e modello, espresso come somma degli scarti quadratici:

$$\varepsilon^2 = \sum_{i=0}^{m}\left[y_i - f(x_i)\right]^2 = \sum_{i=0}^{m}\left[y_i - \sum_{k=0}^{n} c_k\varphi_k(x_i)\right]^2 = \sum_{i=0}^{m}\left[r_i\right]^2$$

Perché il metodo dia risultati utili è importante scegliere bene il modello, ovvero le funzioni $\varphi_i(x)$. Tale scelta è in genere guidata dalle possibili informazioni note sul comportamento del fenomeno in esame, oppure semplicemente dalla distribuzione dei dati stessi. Una delle scelte più frequenti è certamente $\varphi_k(x) = x^k$ ma evidentemente non è sempre la più adeguata.

Per ciascuno degli m+1 punto del piano può essere scritta un'equazione del tipo:

$$f(x_i) - y_i = r_i \rightarrow c_0\varphi_0(x) + c_1\varphi_1(x) + ... + c_n\varphi_n(x) - y_i = r_i$$

l'insieme di tali equazioni può essere organizzata in forma matriciale:

$$\begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ \dots & \dots & \dots & \dots \\ \varphi_0(x_m) & \varphi_1(x_m) & \dots & \varphi_n(x_m) \end{bmatrix} \begin{Bmatrix} c_0 \\ c_1 \\ \dots \\ c_n \end{Bmatrix} - \begin{Bmatrix} y_0 \\ y_1 \\ \dots \\ y_m \end{Bmatrix} = \begin{Bmatrix} r_0 \\ r_1 \\ \dots \\ r_m \end{Bmatrix}$$

o in forma sintetica:

$$[\varphi]\{c\} - \{y\} = \{r\}$$

L'aspetto essenziale della funzione f(x) è la linearità nei parametri incogniti c_i, per questo motivo il modello viene detto lineare ed è affrontabile con i metodi dell'algebra lineare:

Per esempio il modello $f(x) = c_1 + c_2 \cdot e^{c_3 x}$ è non lineare.

La soluzione del problema lineare dei minimi quadrati, ovvero la determinazione del valore dei parametri c_i, è ottenuta minimizzando la quantità ε^2 ovvero $\frac{\partial(\varepsilon^2)}{\partial c_i} = 0 \quad i = 0,1,...,n$.

Impostando tale condizione, e riordinando le equazioni si perviene alla forma:

$$\left[[\varphi]\cdot[\varphi]^t\right]\{c\} = \begin{Bmatrix} \sum_{i=0}^{m}\left(y_i \cdot \varphi_0(x_i)\right) \\ \dots \\ \sum_{i=0}^{m}\left(y_i \cdot \varphi_n(x_i)\right) \end{Bmatrix} = [\varphi]^t \cdot \{y\}$$

che è il sistema lineare (le cui equazioni sono dette *equazioni normali del problema*) la, le cui n+1 soluzioni rappresentano appunto i valori cercati dei parametri c_i.

Si può dimostrare che, se il determinante della matrice dei coefficienti non è nullo, il problema lineare dei minimi quadrati ammette sempre una soluzione e che questa è unica.

34. ESERCIZI SULLA REGRESSIONE SEMPLICE

Ex79

Dire se è possibile rappresentare su scala logaritmica una funzione continua che assume valori positivi e negativi. Motivare la risposta.

Ex80

Una retta di correlazione passa per i punti P(0,2) e Q(5,17). Trovare il valore di y in corrispondenza di x=3.

Ex81

Calcolare le rette di regressione dei seguenti tre punti (1,1), (2,4), (3,2); considerando sia la regressione di x su y che quella di y su x.

Ex82

Comprimendo una determinata massa di aria sono stati misurati i seguenti valori di pressione e volume:

V *[dm³]*	54.3	61.8	72.4	88.7	118.6	194.0
P *[atm]*	61.2	49.5	37.6	28.4	19.2	10.1

Stimare il valore della pressione per V=100 dm^3.

Se il gas è perfetto tali dati devono adattarsi perfettamente alla legge politropica:

$$P \cdot V^n = C$$

E' naturalmente possibile determinare direttamente il valore dei parametri C ed n utilizzando del software specifico, tuttavia in questo caso è possibile trasformare il problema in modo da poter applicare la soluzione vista a proposito della retta ai minimi quadrati. Infatti, applicando l'operatore logaritmo ad entrambi i membri della legge politropica si ottiene:

$$\log P + n \cdot \log V = \log C \quad \text{ovvero} \quad \log P = \log C - n \cdot \log V$$

ponendo $\log P = y$, $\log C = a_0$, $\log V = x$, ed $a_1 = -n$

si ottiene la forma lineare $y = a_0 + a_1 x$

Si riportano in una tabella i valori di x ed y:

x=logV	1.73	1.79	1.85	...
y=logP	1.78	1.69	1.57	...

e si calcolano i coefficienti a_0 ed a_1 della retta di regressione con le formule canoniche, ottenendo: $a_0 = 4.2$, $a_1 = -1.40$

Poiché $a_0 = \log C$, si ricava $C = 1.6e+4$; e poiché $a_1 = -n$ si ricava $n = 1.4$, dunque risulta:

$$P \cdot V^{1.4} = 16'000$$

Dunque per V=100 si ottiene: P≈25.36 atm.

Ex83

Sulla base del CMQ, ricavare il parametro a_0 per i semplici modelli seguenti:

$y = a_0^x$; $y = x^a$; $y = a_0 \log(x)$.

35. TEORIA DELLA CORRELAZIONE

Regressione e coefficiente di correlazione

Poiché spesso non è chiaro il grado di dipendenza tra le variabili x ed y si può ricorrere a diverse definizioni delle quantità D_i: per esempio la distanza tra punti e curva potrebbe essere misurata anche in direzione orizzontale piuttosto che lungo la verticale, oppure anche sulla direzione normale alla curva $y(x)$. Le curve interpolanti ottenute in questi modi risultano generalmente differenti.

Una equazione $y(x)$ calcolata con una qualsiasi procedura di approssimazione o di interpolazione viene detta *curva di regressione di y su x* se si valutano le distanze tra dati e modello in direzione parallela all'asse delle ordinate. Simmetricamente si può determinare una curva di regressione di x su y se si valutano le distanze tra dati e modello in direzione parallela all'asse delle ascisse.

L'operazione di regressione di una o più variabile su un'altra ha senso quando sia noto il rapporto di dipendenza (o di *correlazione*) tra le variabili (p.e. peso-volume di frutti). In assenza di leggi fisiche non si può avere la certezza di un tale legame anche tra variabili con andamenti apparentemente collegati. Per esprimere il grado di dipendenza tra variabili sarebbe preferibile un indice di correlazione, in grado di determinare la bontà dell'adattamento di un modello ai dati rilevati, svincolato dalla scelta del tipo di scarto (direzione x, y o normale alla curva).

Vediamo come si costruisce un tale, indice. Innanzitutto si definisce come ***varianza totale*** il termine seguente, che non dipende da $y(x)$ ed è in grado di esprimere la naturale dispersione dei dati osservati (varianza dei dati sperimentali y_i):

$$V_T = \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2$$

Si definisce poi come ***varianza residua*** il termine seguente, che corrisponde alla varianza degli scarti (corrisponde a ε^2/N). In particolare si nota che se $y(x)$ è un'interpolante la varianza residua è nulla:

$$V_R = \frac{1}{N}\sum_{i=1}^{N} (y_i - y(x_i))^2$$

Si definisce come ***coefficiente di determinazione*** (spesso designato semplicemente come *R-quadro*) la quantità $\boldsymbol{R^2}$ definita come:

$$R^2 = \frac{Vt - Vr}{Vt} = 1 - \frac{Vr}{Vt}$$

Si definisce inoltre come ***coefficiente di correlazione*** la quantità $\boldsymbol{R}$, definita evidentemente come:

$$R = \sqrt{R^2} = \sqrt{1 - \frac{Vr}{Vt}}$$

Si può dimostrare che le quantità R ed R^2 rimangono le stesse sia che come variabile indipendente si assuma x oppure y.

R o R^2 sono comunemente adottati come indicatori della bontà della stima dei dati osservati (x_i, y_i) con il modello $y(x)$, ovvero del fatto che esista una relazione tra x ed y.

Quanto più è piccola la dispersione degli scarti attorno alla linea di regressione (ovvero la varianza residua), tanto migliore è l'adattamento del modello. Al limite se x e y sono perfettamente correlate allora non esisterà varianza residua ed il rapporto fra le varianze sarà 0, e di conseguenza $R^2 \rightarrow 1$.

Si dimostra inoltre che se invece non vi è nessuna relazione tra le variabili allora il rapporto tra la varianza residua e la varianza totale originaria tende ad 1, e di conseguenza $R^2 \rightarrow 0$.

Se si ottiene un valore $R^2=0.8$ allora risulta $1-Vr/Vt=80\%$, da cui $Vr/Vt=1-80\%=20\%$ ovvero la variabilità dei valori $y(x)$ attorno alla linea di regressione vale il 20% della varianza originaria; in altre parole si è riusciti a *spiegare* con il modello $y(x)$ l'80% della variabilità originaria, mentre rimane un 20% di varianza residua.

Idealmente, si vorrebbe spiegare il più possibile della varianza originaria (se non tutta). Un valore di R^2 prossimo ad 1 indica che si riusciti a capire l'origine della dispersione dei dati sperimentali con le sole variabili specificate nel modello.

Significatività di R

Resta da stabilire quando R è *sufficientemente* vicino ad 1, così da stabilire che x ed y sono *probabilmente correlate*.

Se disponiamo di due sole osservazioni (x_1,y_1 ed x_2,y_2) otteniamo un valore R=1, anche se le variabili x ed y non sono affatto correlate. D'altronde, anche per variabili perfettamente incorrelate, all'aumentare di N, difficilmente si otterrà un valore di R esattamente uguale a zero.

- Show examples from *Eng/Winstats->Demos->Correlations* ($0<R<1$).

Per stabilire se un valore calcolato di R è significativamente vicino al valore 1, è stata calcolata la *probabilità che N misure di due variabili, perfettamente incorrelate, forniscano un valore di R superiore ad un valore determinato*. Lo sviluppo di tale calcolo è piuttosto complesso e se ne riporta qui solo qualche risultato:

N	***R_0***										
	0	***0.1***	***0.2***	***0.3***	***0.4***	***0.5***	***0.6***	***0.7***	***0.8***	***0.9***	***1***
3	*100*	*94*	*87*	*81*	*74*	*67*	*59*	*51*	*41*	*29*	*0*
6	*100*	*85*	*70*	*56*	*43*	*31*	*21*	*12*	*6*	*1*	*0*
10	*100*	*78*	*58*	*40*	*25*	*14*	*7*	*2*	*.5*	-	*0*
20	*100*	*67*	*40*	*20*	*8*	*2*	*.5*	*.1*	-	-	*0*
50	*100*	*49*	*16*	*3*	*.4*	-	-	-	-	-	*0*

La tabella riporta la probabilità che N misure di due variabili x ed y, perfettamente incorrelate producano un coefficiente di correlazione $R>R_0$. $\Pr\{R>R_0\}$ con N misure casuali tra insiemi incorrelati. In altre parole la tabella riporta la probabilità di sbagliare affermando che le variabili sono tra loro correlate.

P.e. un valore di $R=0.6$ risulta non significativo per un campione di numerosità $N=10$, mentre lo diviene per $N=20$. La probabilità di sbagliare dicendo che esiste correlazione vale rispettivamente 7% e 0.5%.

I valori dati sono probabilità percentuali, le caselle vuote indicano valori inferiori allo 0.5%.

In generale, per ragioni evidenti, la prima e l'ultima colonna (corrispondenti ad $R=0$ e ad $R=1$) sono omesse.

Una scelta piuttosto comune è quella di considerare una correlazione R_0 come *significativa* se la probabilità di ottenere un coefficiente R superiore ad R_0 da variabili incorrelate è minore del 5% (*). *Molto significativa* se la probabilità corrispondente è inferiore all'1% (**).

Dunque per evidenziare una correlazione tra due variabili occorre che R sia elevato (tipicamente superiore a 0.5) e che tale valore sia significativo

- Esempi da tesi di laurea: tabelle *Dino.ppt & GLBarbi.ppt*

	R										
N	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**0.95**	**0.99**
3	92.63%	85.13%	77.38%	69.20%	60.42%	50.77%	39.93%	27.46%	13.09%	5.58%	0.59%
5	86.86%	73.81%	60.94%	48.38%	36.32%	25.06%	15.03%	6.90%	1.59%	0.33%	0.01%
10	78.20%	57.65%	39.46%	24.53%	13.35%	5.99%	1.97%	0.37%	0.02%	0.00%	0.00%
15	72.21%	47.31%	27.46%	13.64%	5.49%	1.63%	0.30%	0.02%	0.00%	0.00%	0.00%
20	67.44%	39.67%	19.71%	7.89%	2.36%	0.47%	0.05%	0.00%	0.00%	0.00%	0.00%
30	59.88%	28.87%	10.65%	2.80%	0.47%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%
40	53.91%	21.56%	5.96%	1.04%	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
50	48.95%	16.35%	3.41%	0.39%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
60	44.70%	12.53%	1.98%	0.15%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
80	37.74%	7.52%	0.68%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
100	32.22%	4.60%	0.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
120	27.71%	2.85%	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
150	22.34%	1.41%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
200	15.89%	0.45%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
250	11.47%	0.15%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
300	8.38%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

I valori riportati in tabella possono essere approssimati dalla relazione semplificata che segue:

$$p \approx a + \frac{b}{N^c \cdot R^d}$$

con: $a= -0.141$; $b= 0.468$; $c= 0.347$; $d= 0.641$.

Coefficiente di correlazione lineare (Pearson)

La definizione di coefficiente di correlazione è evidentemente dipendente dalla forma della funzione approssimante $y(x)$, e dunque a questa si riferisce.

Nel caso particolare di regressione lineare (ovvero quando si intenda adattare ai dati osservati un modello del tipo $y(x)=a_0+a_1 \cdot x$) l'espressione di R si semplifica particolarmente, e viene indicata di solito con il simbolo r:

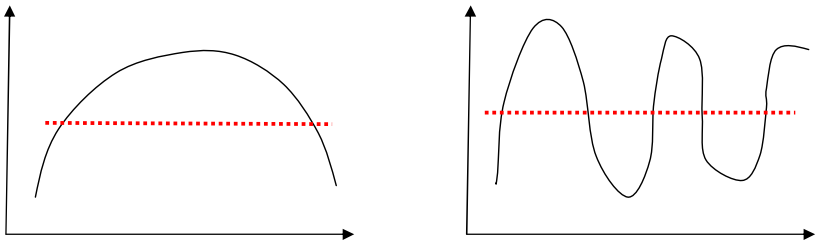
$$r = \frac{\sum_{i=1}^{N} [(x_i - \bar{x}) \cdot (y(x_i) - \bar{y})]}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{N} (y(x_i) - \bar{y})^2}}$$

Un basso coefficiente di correlazione lineare non esclude la possibilità che esista una relazione tra x ed y, ma solo che probabilmente non esiste una relazione di tipo lineare.

Questa formulazione è in grado di esprimere anche il segno della correlazione: un valore positivo di r individua il fatto che al crescere dei valori x crescono anche i valori di y, viceversa un valore negativo di r individua il fatto che al crescere dei valori x i valori di y calano.

- Es. Quanto valgono r, R ed R^2 nel caso di una retta orizzontale?

Un alto valore di r significa che la correlazione tra le due variabili è ben descritta da una relazione lineare; un basso valore invece significa che non c'è correlazione o che, semplicemente, la correlazione non è lineare. Come per esempio nel casi di forte non linearità in figura:



36. CORRELAZIONE MULTIPLA E PARZIALE

Quando le variabili da correlare sono due ($y=f(x)$) si parla di correlazione semplice, quando sono più di due di correlazione multipla. Consideriamo ora il caso in cui si debba verificare un collegamento fra tre variabili ($z=f(x,y)$), p.e. la carica microbica su di un alimento in funzione della temperatura e del tempo di trattamento.

I modelli di regressione più semplici applicabili al caso di 3 variabili sono, come nel caso di due variabili, quelli polinomiali di grado basso, p.e. un piano:

$$z = a_0 + a_1 \cdot x + a_2 \cdot y$$

Una forma polinomiale generalmente applicata al caso di regressione su due variabili è del tipo:

$$z = z_0 + \sum_{i=0}^{n} a_i \cdot x^i \cdot y^{n-i}$$

Analogamente a quanto già visto, con più variabili si definisce il coefficiente di correlazione, sulla base di Vt e Vr, essendo ora:

$$V_T = \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}_i)^2 \qquad V_R = \frac{1}{N} \sum_{i=1}^{N} (z_i - z(x_i, y_i))^2$$

- Esempi di tabelle di coefficienti di correlazione:

azioni (utz01.sta)

...a...	T	Z	U
T	1.000000	.107190	-.209117
Z	.107190	1.000000	.675753
U	-.209117	.675753	1.000000

Correlazioni (utz01.sta)

Continua... Correl. marcate significative al liv. p < .0
100 (Eliminaz. casewise dati mancanti)

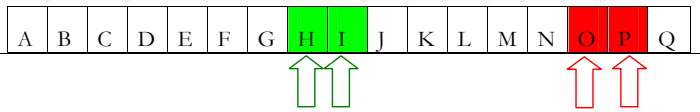
Variabile	T	Z
U	-.2091	.6758
	p=.037	p=.000

	N	***Dm***	***Dcv***
Coverage	$r = 0{,}499$	$r = 0{,}802$	$r = 0{,}121$
	$p<0{,}01$	$p<0{,}01$	$p=0{,}61$

- http://www.unibo.it/qualita

Coefficiente di correlazione per Agraria: corsi di laurea

Domande con risposta su scala 0-100 (3)	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A. Lezioni frequentate																	
B. Chiarezza scopi e programma	0,36																
C. Pertinenza per corso di laurea/diploma	0,29	0,36															
D. Chiarezza dei nessi fra parti	0,22	0,54	0,42														
E. Interesse degli argomenti trattati	0,21	0,35	0,56	0,51													
F. Utilità lezioni ai fini dell'apprendimento	0,21	0,37	0,44	0,46	0,58												
G. Facile reperibilità di materiali didattici e testi	0,07	0,28	0,12	0,35	0,16	0,22											
H. Stimolazione interesse e coinvolgimento	0,23	0,43	0,43	0,42	0,57	0,64	0,29										
I. Capacità di esposizione del docente	0,15	0,43	0,35	0,46	0,45	0,66	0,32	0,75									
J. Puntualità del docente a lezione	0,20	0,31	0,13	0,33	0,14	0,18	0,25	0,21	0,25								
K. Puntualità del docente a ricevimento	0,17	0,37	0,22	0,38	0,27	0,32	0,42	0,34	0,35	0,60							
L. Chiarezza modalità di accertamento e di esame	0,10	0,36	0,29	0,37	0,23	0,28	0,33	0,34	0,38	0,30	0,42						
M. Disponibilità del docente al dialogo	0,08	0,26	0,34	0,27	0,35	0,37	0,21	0,47	0,42	0,31	0,50	0,35					
N. Adeguatezza del materiale didattico usato in aula	0,17	0,34	0,29	0,43	0,32	0,44	0,46	0,52	0,53	0,33	0,52	0,46	0,40				
O. Adeguatezza delle aule di lezione	0,14	0,09	0,11	0,16	0,15	0,14	0,14	0,10	0,14	0,12	0,23	0,09	0,17	0,15			
P. Lezioni svolte dal titolare	0,10	0,02	0,12	0,00	0,04	0,14	0,00	0,12	0,11	0,14	0,20	0,02	0,22	0,09	0,04		
Q. Utilità delle eventuali attività didattiche integrative	0,13	0,20	0,34	0,26	0,33	0,26	0,19	0,26	0,23	0,30	0,36	0,29	0,45	0,26	0,09	0,17	
S. Grado complessivo di soddisfazione	0,17	0,39	0,43	0,47	0,63	0,64	0,32	0,69	0,71	0,26	0,37	0,39	0,51	0,52	0,15	0,11	0,36



Correlations Pearson index (cog.sta)
Marked correlations are significant at p < .05000

N=96 (Casewise deletion of missing data)

		H7	*H6*	*D3*	*H0*	*B6*	*D2*	*L2*	*B7*	*S*	*M*	*Ma*	*Mp*	*Ca*	*B0*	*D0*	*L6*	*Ms*	*Md*
h	r^2	0,85	0,73	0,71	0,64	0,60	0,56	0,54	0,47	0,43	0,42	0,41	0,41	0,40	0,22	0,09	0,03	0,02	0,01
	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.09	0.19	0.45

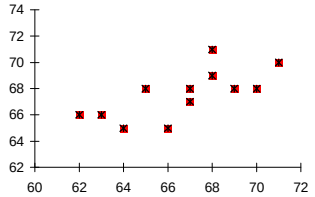
37. ESERCIZI SULLA CORRELAZIONE LINEARE

Ex84

La tabella riporta i pesi x ed y di un campione di 12 padri e dei loro figli:

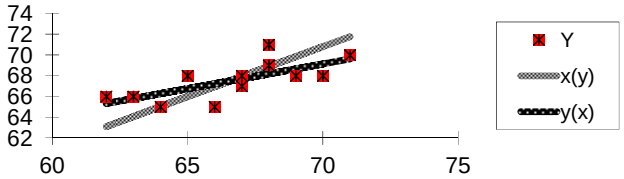
x	*65*	*63*	*67*	*64*	*68*	*62*	*70*	*66*	*68*	*67*	*69*	*71*
y	*68*	*66*	*68*	*65*	*69*	*66*	*68*	*65*	*71*	*67*	*68*	*70*

Costruito il diagramma a dispersione determinare le rette di regressione di x su y, e di y su x.



Si ricavano, con le formule già viste, i coefficienti delle due rette di regressione:

$$y=0.476x+35.82 \quad \text{e} \quad x=1.036y-3.38$$



La varianza totale vale Vt=3.24, mentre la varianza residua risulta Vr=1.64, di conseguenza il coefficiente di correlazione diviene R=0.702, che risulta il medesimo sia che come variabile indipendente si assuma x oppure y.

Poiché in questo caso il modello adottato è lineare, si può ottenere immediatamente il valore di R, dalla sua particolarizzazione al caso della regressione lineare. In più la relazione particolarizzata fornisce anche il segno di r, che in questo caso individua il fatto che al crescere di x cresce anche y.

Il valore 0.7, ottenuto da un campione di numerosità pari a 12, risulta significativo al livello ≈0.015 (1.5%).

Ex85

Un coefficiente di correlazione ricavato da un campione di 18 elementi vale 0.32. A quale livello di significatività possiamo ipotizzare l'esistenza della correlazione?

	r_0										
N	*0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*	*1*
10	*100*	*78*	*58*	*40*	*25*	*14*	*7*	*2*	*.5*	*-*	*0*
20	*100*	*67*	*40*	*20*	*8*	*2*	*.5*	*.1*	*-*	*-*	*0*

Nel nostro caso si vede che per un campione di 20 elementi, un valore di r pari a 0.3, individua l'esistenza di una correlazione con una probabilità dell'80%. Per i valori 18 e 0.32 tale valore sarà, come si vede dalla tabella, un poco più alto. In ogni caso la correlazione sarebbe ritenuta non significativa.

Oppure applicando la formuletta approssimata, otteniamo:

$$p \approx -0.141 + \frac{0.468}{18^{0.347} \cdot 0.32^{0.641}} \approx 0.22$$

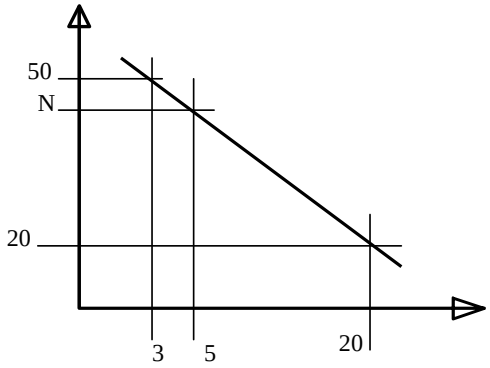
Ex86

Qual è la minima ampiezza campionaria richiesta affinché si possa concludere che un coefficiente di correlazione pari a 0.3 sia significativo al livello dello 0.05?

	r_0										
N	*0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*	*1*
3	*100*	*94*	*87*	*81*	*74*	*67*	*59*	*51*	*41*	*29*	*0*
6	*100*	*85*	*70*	*56*	*43*	*31*	*21*	*12*	*6*	*1*	*0*
10	*100*	*78*	*58*	*40*	*25*	*14*	*7*	*2*	*.5*	*-*	*0*
20	*100*	*67*	*40*	*20*	*8*	*2*	*.5*	*.1*	*-*	*-*	*0*
50	*100*	*49*	*16*	*3*	*.4*	*-*	*-*	*-*	*-*	*-*	*0*

si percorre la colonna 0.3, fino a superare il valore 5%, e si ricava un valore approssimativamente vicino a 50.

Interpolando linearmente i valori mancanti, con l'aiuto di un grafico *N-p%*, si ottiene:



impostando la similitudine tra i triangoli otteniamo:

$\frac{50 - 20}{20 - 3} = \frac{N - 20}{20 - 5}$ da cui ricavando *N*:

Oppure si può invertire la formula approssimata:

$$N \approx \left(\frac{b}{(p - a) \cdot r^d} \right)^{1/c}$$

38. L'ANALISI DEI DATI CON *MICROSOFT EXCEL*

Approssimazione di dati sperimentali e correlazione:

Nella tabella seguente sono riportati i volumi di vendita *V* di due determinati prodotti alimentari *P1* e *P2*, riferiti ad un decennio:

t	1	2	3	4	5	6	7	8	9	10
V_{P1}	10	8	18	29	36	59	96	115	158	179
V_{P2}	12	09	21	27	37	65	88	100	145	158

- Utilizzando la tecnica di regressione ai minimi quadrati, approssimare i dati relativi a ***P1*** con una equazione, e con questa stimare i dati di vendita per l'anno *11*. Verificare inoltre l'eventuale correlazione tra le due serie di dati, e tra queste e la variabile tempo.

Traccia di soluzione:

a) tracciare il grafico *t-V* utilizzando il tipo di grafico a dispersione (*scatter*);
b) inserire una linea di tendenza; [$R^2 \cong 0.93$, $V = 19.87 \cdot t - 38.55 \Rightarrow V(11) \cong 180$]
c) utilizzare l'equazione della linea di tendenza per stimare i dati di vendita per l'anno *11*;
d) sperimentare altre forme dell'equazione di regressione, fino a determinare quella che corrisponde al miglior valore di R^2. Confrontare tra loro le previsioni per l'anno *11* ottenute con il modello lineare e quello polinomiale d'ordine 2 [$V(11) \cong 228$];
e) verificare l'efficacia di un polinomio approssimante d'ordine 5, nello stimare i dati per i prossimi 1, 2, 3 anni;
f) utilizzare la funzione *correlazione* del modulo di *analisi dei dati* per la verifica del grado di correlazione tra le variabili *t*, *V1* e *V2* [si nota che il volume di vendita del prodotto *P1* può essere meglio previsto sulla base di *P2* piuttosto che della variabile tempo];
g) perché la matrice di correlazione ha tutti gli elementi della diagonale principale pari ad 1? E perché le caselle sopra alla diagonale sono vuote (o, in altre parole, perché la matrice è simmetrica)?
h) utilizzando il modulo *analisi dei dati => regressione*; valutare la significatività di R^2 e calcolare le rette di regressione *Vp1(Vp2)* e *Vp2(Vp1)* e confrontare il valore di R^2 con quelli calcolati al punto precedente;
i) tracciare un grafico *Vp1(Vp2)* e determinare la retta di regressione.

39. DIFETTOSITÀ CAMPIONARIA

Dimensionamento di un campione

1 - un fornitore afferma di poter vendere un lotto di 1000 pezzi, caratterizzato da una difettosità inferiore al 10%.

I difetti possono riguardare ad esempio le confezioni: etichetta storta, ammaccature, tappo avvitato male, trafilaggio;

2 - quale deve essere la dimensione minima di un campione adatto a sostenere una tale affermazione?

3 - dipende da *p*, ovvero dal livello di affidabilità che si vuole dare all'affermazione *1*, che potrebbe essere così modificata per un ipotetico contratto: non posso dire con certezza quanti pezzi difettosi conterrà il lotto (per saperlo dovrei controllare tutti gli elementi del lotto) però posso dire con probabilità *p*=95% che il lotto di 1000 pezzi contiene meno del 10% di pezzi difettosi se su un campione da *Nc* pezzi non sono stati osservati elementi difettosi.

Come determiniamo *Nc*?

Distribuzione ipergeometrica

Supponiamo di avere una popolazione di *N* elementi. Un certo numero *D*<*N* di essi sia difettoso. Si estragga dalla popolazione un campione casuale di *n* elementi, allora la probabilità *P(x)* che *x* elementi del campione siano difettosi vale:

$$P(x)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, \ldots, \min(n, D)$$

in cui $\binom{a}{b}$ è il numero di combinazioni di *a* elementi presi in numero *b* alla volta: $\binom{a}{b}=\frac{a!}{b!(a-b)!}$

Esempio: Si supponga che un lotto contenga 100 elementi, dei quali 5 non siano conformi alle specifiche tecniche. Se si forma un campione casuale di 10 elementi, senza sostituzione, allora la probabilità di trovarne nessuno o uno non conforme nel campione è rispettivamente:

$$P\{x=0\}=\frac{\binom{5}{0}\binom{95}{5}}{\binom{100}{10}}\cong 0.584 \qquad P\{x=1\}=\frac{\binom{5}{1}\binom{95}{9}}{\binom{100}{10}}\cong 0.339$$

$$P\{x\le 1\}=P\{x=0\}+P\{x=1\}=\frac{\binom{5}{0}\binom{95}{5}}{\binom{100}{10}}+\frac{\binom{5}{1}\binom{95}{9}}{\binom{100}{10}}\cong 0.584+0.339=0.923$$

Excel: *distrib.ipergeom(x; n; D; N)*

Consideriamo una popolazione costituita da un lotto di 1000 elementi con una difettosità del 10%. Valutiamo la probabilità che un campione di ampiezza *n*, estratto da tale popolazione sia privo di difetti:

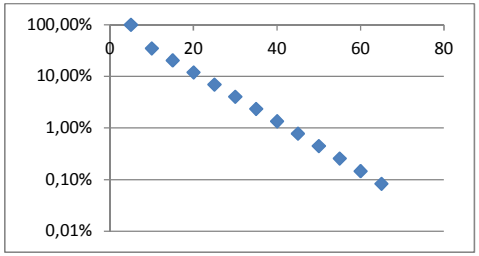
p=distrib.ipergeom(0; *n*; 100; 1000)

N	1000	1000	1000	1000	1000	1000	1000	1000	1000
D	100	100	100	100	100	100	100	100	100
n	5	10	15	20	25	30	35	40	45
x	0	0	0	0	0	0	0	0	0
p	58,98%	34,69%	20,35%	11,90%	6,94%	4,03%	2,34%	1,35%	0,78%

campioni piccoli, privi di difetti, possono essere molto comuni (e viceversa campioni grandi privi di difetti tendono ad essere rari).

Con un'ampiezza pari almeno a 30 esiste una probabilità inferiore al 5% di non trovare difetti. Ovvero se su un campione di 30 elementi non troviamo elementi difettosi, esiste una probabilità superiore al 95% che la difettosità del lotto sia inferiore al 10%.

Similmente osserviamo che se vogliamo restringere l'incertezza al livello dell'1%, dobbiamo considerare un'ampiezza campionaria pari ad almeno 45.



Distribuzione Binomiale

Se la quota di prodotti difettosi è molto diluita nella popolazione, allora quest'ultima può ritenersi infinita e la trattazione si semplifica un po', facendo riferimento alla distribuzione binomiale. In tale ipotesi la difettosità non è più espressa in riferimento ad una precisa dimensione di lotto, ma semplicemente come frazione p di elementi difettosi sul totale. La probabilità $P(x)$ di estrarre, con n estrazioni indipendenti, x pezzi difettosi vale:

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \qquad x = 1, 2, \ldots, n.$$

in cui $\binom{n}{x}$ sono i coefficienti binomiali, ovvero le combinazioni x a x di n elementi: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Excel: *distrib.binom(x; n; p; falso)*

Esempio: Si abbia un lotto molto grande, il quale contenga il 10% di pezzi difettosi. Allora p=0.10 e la probabilità di estrarre esattamente 2 pezzi difettosi su un campione di 10 pezzi estratti, cioè di estrarre 8 pezzi buoni su 10, è: $P(8) = \binom{10}{2} \cdot 0.10^2 \cdot 0.90^8 \cong 0.1937$

che si calcola con Excel come: *=Distrib.Binom(2; 10; 0,1; Falso)*

Con una difettosità del 3%, la probabilità di estrarre un campione di ampiezza n privo di difetti (x=0) vale, per n variabile tra 30 e 270:

d	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x	0	0	0	0	0	0	0	0	0
n	30	60	90	120	150	180	210	240	270
P	40,10%	16,08%	6,45%	2,59%	1,04%	0,42%	0,17%	0,07%	0,03%

Ex: Ripetere i calcoli in riferimento ad una difettosità dell'1% e dello 0.1%. Calcolare la probabilità che il campione di ampiezza n contenga 0, 1 o 2 pezzi difettosi. Qual'è la dimensione giusta del campione, ovvero l'ampiezza minima che permette di contenere l'errore sotto alla soglia del 5% con probabilità del 95%?

40. RIFERIMENTI NORMATIVI

ASTM D4131, (R 2005) Standard Practice for Sampling Fish with Rotenone

ASTM D4211, Classification for Fish Sampling

ASTM D4638, Standard Guide for Preparation of Biological Samples for Inorganic Chemical Analysis

ASTM D4687, (R 2006) Standard Guide for General Planning of Waste Sampling

ASTM D6063, Standard Guide for Sampling of Drums and Similar Containers by Field Personnel

ASTM D6299, Standard Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance

ASTM D6699, Standard Practice for Sampling Liquids Using Bailers

ASTM D75/D75M, Standard Practice for Sampling Aggregates

ASTM E122, (E 2011) Standard Practice for Calculating Sample Size to Estimate, With Specified Precision, the Average for a Characteristic of a Lot or Process

ASTM E2819, Standard Practice for Single- and Multi-Level Continuous Sampling of a Stream of Product by Attributes Indexed by AQL

ASTM E2870, Standard Test Method for Evaluating Relative Effectiveness of Antimicrobial Handwashing Formulations using the Palmar Surface and Mechanical Hand Sampling

ASTM-STP 15 D, Statistical interpretation of data — Part 4: Detection and treatment of outliers

ISO 11024, General Guidance on Chromatographic Profiles - Part 2: Utilization of Chromatographic Profiles of Samples of Essential Oils

ISO 11648-1, Statistical Aspects of Sampling from Bulk Materials - Part 1: General Principles

ISO 13307, Microbiology of food and animal feed - Primary production stage - Sampling techniques

ISO 14001, Environmental management systems — Requirements with guidance for use

ISO 16269, Statistical Interpretation of Data - Part 7: Median - Estimation and Confidence Intervals

ISO 17604, Microbiology of food and animal feeding stuffs — Carcass sampling for microbiological analysis AMENDMENT 1: Sampling of poultry carcasses

ISO 22000, Food safety management systems Requirements for any organization in the food chain

ISO 2602, Statistical Interpretation of Test Results - Estimation of the Mean - Confidence Interval

ISO 2854, Statistical Interpretation of Data - Techniques of Estimation and Tests Relating to Means and Variances

ISO 2859-1, Sampling procedures for inspection by attributes — Part 1: Sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection

ISO 2859-10, Sampling procedures for inspection by attributes Part 10: Introduction to the ISO 2859 series of standards for sampling for inspection by attributes; Supersedes ISO 2859-0:1995

ISO 3301, Statistical Interpretation of Data - Comparison of Two Means in the Case of Paired Observations First Edition

ISO 3494, Statistical Interpretation of Data - Power of Tests Relating to Means and Variances First Edition

ISO 3534-2, Statistics Vocabulary and symbols Part 2: Applied statistics

ISO 3863, Cylindrical Cork Stoppers - Dimensional Characteristics, Sampling, Packaging and Marking

ISO 3951, Sampling Procedures and Charts for Inspection by Variables for Percent Nonconforming

ISO 3951-1, Sampling procedures for inspection by variables - Part 1: Specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL

ISO 3951-4, Sampling procedures for inspection by variables — Part 4: Procedures for assessment of declared quality levels

ISO 3951-5, Sampling procedures for inspection by variables Part 5: Sequential sampling plans indexed by acceptance quality limit (AQL) for inspection by variables (known standard deviation)

ISO 4072, Green Coffee in Bags - Sampling

ISO 5479, Statistical Interpretation of Data - Tests for Departure from the Normal Distribution

ISO 5538, Milk and Milk Products - Sampling - Inspection by Attributes

ISO 5555, Details Draft 2013.04.18 Animal and vegetable fats and oils - Sampling

ISO 6497, Details Active 2002.11.15 Animal feeding stuffs Sampling

ISO 661, Animal and vegetable fats and oils Preparation of test sample

ISO 664, Oilseeds — Reduction of laboratory sample to test sample

ISO 6887, Microbiology of food and animal feeding stuffs preparation of test samples, initial suspension and decimal dilutions for microbiological examination Part 2: Specific rules for the preparation of meat and meat products

ISO 874, Fresh Fruits and Vegetables - Sampling

ISO 9001 CORR 1, Quality management systems — Requirements TECHNICAL CORRIGENDUM 1 - Fourth Edition

ISO 948, Spices and Condiments - Sampling

ISO 950, Superseded by: ISO 13690, Cereals - Sampling (as Grain)

ISO DIS 3534-4, Statistics - Vocabulary and symbols - Part 4: Survey sampling

ISO TR 10017, Guidance on statistical techniques for ISO 9001:2000

ISO TR 13519, Guidance on the development and use of ISO statistical publications supported by software

ISO TR 18532, Guidance on the application of statistical methods to quality and to industrial standardization

ISO TR 8550, Guidance on the selection and usage of acceptance sampling systems for inspection of discrete items in lots — Part 1: Acceptance sampling

ISO TS 19036 AMD 1, Microbiology of food and animal feeding stuffs — Guidelines for the estimation of measurement uncertainty for quantitative determinations AMENDMENT 1: Measurement uncertainty for low counts

ISO/TR 13425, Guida per la scelta di metodi statistici nella normazione e nelle specifiche.

Manuale ISO 3: 1989, Metodi statistici.

UNI 4724, Metodi statistici per il controllo della qualità. Rappresentazione tabellare, numerica e grafica di dati aventi carattere di variabile. Distribuzione di frequenza. Attendibilità dei dati. Calcolo delle stime dei parametri.

UNI 4726, Metodi statistici per il controllo della qualità. Grafico di probabilità normale.

UNI 4842, Allegato metodi statistici per il controllo della qualità. Procedimento di collaudo statistico per attributi. Istruzioni per l'impiego.

41. PROVE DI ACCERTAMENTO

Pentalogo dell'esaminando modello

1. l'esame di statistica ed informatica è unico e consta di una prova scritta;
2. il testo del compito riguarda sia esercizi numerici che aspetti teorici. In ogni caso per poter affrontare il compito è necessario avere studiato tutta la materia svolta durante il corso;
3. durante gli esami è necessario dotarsi di penna, calcolatrice portatile, tessera con numero di matricola. Non è invece ammessa la consultazione di libri e appunti;
4. farsi trovare già ben spaziati, evitando concentrazioni di natura strategica, aiuta ad impiegare razionalmente il tempo a disposizione;
5. l'iscrizione alle liste d'esame, la diffusione dei risultati delle prove e la registrazione degli esiti si effettuano attraverso il servizio *Almaesami*.

Prova parziale di accertamento di statistica

A.A.2012/13 CdL specialistica

Nome	*Cognome*	*#Matricola*	*e-mail*

Tema A - Il laboratorio di analisi della qualità di un'azienda riceve un lotto di 1000 frutti. Per verificare il rispetto degli accordi contrattuali, relativamente al grado di maturazione, vengono estratti con un procedimento casuale alcuni elementi dal lotto, dei quali si misura la resistenza, con un semplice penetrometro, ottenendo i seguenti valori (in N):

6 + n; 6 + n/2; 6 + n/1.2; 6 + (n+9)/3; 8 + n

A1 - stimare la deviazione standard del lotto;
A2 - stimare l'intervallo fiduciale al 99% per il valore medio di resistenza per l'intero lotto;
A3 - calcolare la probabilità che una campione di 5 individui abbia una resistenza media inferiore all'85% del valore medio campionario;
A4 - calcolare la probabilità che un elemento della popolazione abbia una resistenza inferiore all'85% del valore medio. Quale ipotesi supplementare è necessaria?
A5 – Si preleva un altro campione proveniente dallo stesso fornitore, ma da una partita successiva:

$$n + 1;\ n/1.2;\ n;\ (n+9)/2;\ n + 3$$

verificare la significatività della differenza tra i due campioni al livello del 95%.
A6 – Sulla base dei dati calcolati sul primo campione, determinare con un'affidabilità del 95%, la minima ampiezza campionaria necessaria per stimare il valore medio della popolazione con un errore del 10%.

Note: i campioni sono piccoli per questioni di praticità, applicare tuttavia ugualmente la teoria dei grandi campioni. **-** essendo ***a*** l'ultima cifra del proprio numero di matricola, e ***b*** il numero di lettere del proprio cognome, calcolare ***n*** come: $n = a + b$

Note sullo svolgimento degli esercizi:

- sui fogli utilizzati per lo svolgimento dei temi occorre indicare chiaramente il proprio nome e numero di matricola. Si raccomanda di scrivere in modo ordinato e sostanzialmente comprensibile.
- Nonostante non sia richiesta la consegna dell'eventuale brutta copia, non trascurare di indicare chiaramente tutti i passaggi algebrici e logici. In particolare, soprattutto nel caso esistano diversi modi possibili di procedere, motivare brevemente le proprie scelte.
- Per quanto riguarda i calcoli aritmetici, scrivere prima le espressioni in forma simbolica e poi sostituire i valori numerici ai simboli.
- La prova d'esame si conclude in due ore.
- Non è permessa la consultazione di libri ed appunti.
- L'esito della prova d'esame rimane valido per il corrente A.A. e verrà pubblicato entro una settimana su AlmaEsami.

■

42. SOMMARIO

STA/VE/TR Cesena - appunti di statistica - Rev. lun 25 nov 2013