CENNI DI STATISTICA

MICHELE ANDREOLI

Le dimostrazioni semplificatte che seguono solo solo ad integrazione delle lezioni.[1]

1. Cos'è la statistica

La statistica si occupa dei metodi scientifici per raccogliere ed analizzare i dati (*statistica descrittiva*), ma anche per trarre conclusioni e prendere decisioni sulla base di tale analisi (*statistica induttiva o inferenziale*).

2. Variabili statistiche

2.1. **Definizione.** Una variabile è detta continua se può assumere tutti i valori reali compresi in un certo intervallo definito; è detta discreta se assume solo valori di tipo intero.

Per variabile casuale (o stocastica) si intende (detto in parole povere) una variabile collegata ad eventi di tipo probabilistico. Seguendo l'approccio più moderno (insiemistico), per variabile casuale si intende "*una funzione che ha come dominio lo spazio degli eventi S e come immagine una sottoinsieme dell'intervallo [0,1]*".

Esempio 1 Dette x ed y le due faccie che escono nel lancio di un dado, x ed y sono variabili casuali discrete, come anche x+y o xy ....

Esempio 2 Le coordinate (x,y) di inpatto di un proiettile su un bersaglio, sono variabili stocastiche continue.

2.2. **Indici riassuntivi.** Nella maggior parte dei casi pratici, la conoscenza completa del comportamento statistico di una variabile x non è disponibile. Allora si ricorre alla conoscenza di qualche parametro di tipo riassuntivo. I due parametri riassuntivi più usati sono (vedi pagine precedenti): la *media* e lo scarto *quadratico medio (*o la *varianza*, che è lo stesso).

$$\mu = \langle x \rangle \qquad \sigma^2 = \left\langle (x-\mu)^2 \right\rangle$$

[1]NOTA INFORMATIVA: questo documento non è stato redatto con Microsoft Word, ma col sistema di stampa denominato TEX (pronunicia: tek), lo stesso usato per la produzione di documentazione scientifica ad alta qualità tipografica. La maggior parte delle pubblicazioni scientifiche al mondo, compresi molti libri di matematica che comprerete all'Università, è realizzata con il TEX. TEX, e i suoi derivati quali LATEX (che sto ora usando), a differenza di MS Word, è un sofware libero e gratuito, creato da Leslie Davenport, e liberamente distribuito su Internet.

Lo scopo di questi parametri è quello di dare un'idea del *valore atteso* di x e di come i valori ottenibili *si disperdano* intorno a questo valore medio.

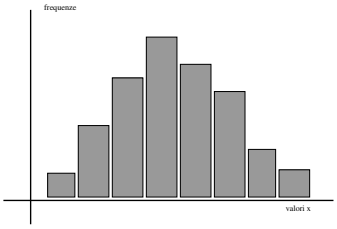
2.3. **Variabili standardizzate.** Una variabile stocastica si dice standard se la sua media è zero e lo scarto quadratico medio è 1.

Data una x non standard, se ne puo' sempre ricavare una standard z con la formula:[2] $z = \frac{x-\mu}{\sigma}$. Ma, in sostanza, cosa significa che z=-1.5? Ricavando la x che le corrisponde, trovo che $x = \mu - 1.5 \cdot \sigma$ e che quindi abbiamo ottenuto un valore inferiore alla media di "uno scarto e mezzo".

3. Distribuzione o densità di probabilità

3.1. **Istogrammi.** Sia x una variabile stocastica. Se, idealmente, potessimo misurare il suo valore un gran numero di volte, saremmo in grado di tracciare l'istogramma delle sue frequenze di distribuzione.

In sostanza, saremmo in grado di effettuare un grafico dove sull'asse delle Y riporteremmo le percentuali delle volte che il valore di x è caduto in un certi intervalli dell'asse X: quante volte tra [0,1], quante volte tra [1,2], etc.



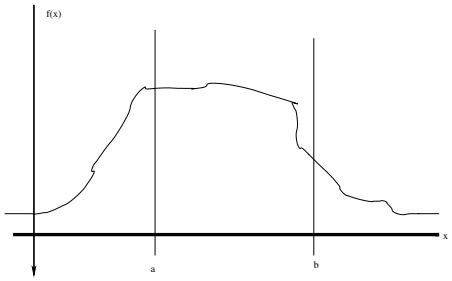
La probabilità che x cada nell'intervallo [a,b] viene indicata con il simbolo P(a<x<b). Affermare quindi che P(2<x<5)=0.20 significa che, se effettuassimo un alto numero di prove, il 20% devi valori di x si collocherebbe nell'intervallo [0,5].

3.2. **Densità di probabilità.** In generale, il valore di P viene rappresentato nella forma di integrale definito, in questo modo:

$$P(a < x < b) = \int_a^b f(x)\,dx$$

La funzione f(x) è detta "*funzione di distribuzione della variabile continua* x" ed è tale che $f(x)\,dx$ rappresenti la probabilità che $x \in [x, x + dx]$. Il caso discreto è analogo e l'integrale viene rimpiazzato da una sommatoria.

[2]Come si vede, z è un numero puro, cioè adimensionale



Quindi, l'area sotto la curva f(x) tra [a,b] fornisce la probabilità che x sia in questo intervallo. Se ne deduce che l'area di tutta la curva deve fare 1 (curva normalizzata): $\int_{-\infty}^{+\infty} f(x)\,dx = 1$.

3.3. **Distribuzione cumulativa.** più frequentemente, invece che f(x) si dà una funzione ausiliaria $\Phi(z)$ (detta *distribuzione cumulativa* o *funzione di ripartizione*), cosi' definita:

$$\Phi(z) = \int_0^z f(x)\,dx \quad e\,quindi \quad P(a < x < b) = \Phi(b) - \Phi(a)$$

Come si vede, $\Phi(z)$ non è altro che una delle infinite *primitive integrali di f(x);* in particolare, è l'unica primitiva che si annulla per z=0. Essa rappresenta l'area compresa nell'intervallo [0,z] ed è sufficiente a calcolare qualsiasi altra area. Per alcune funzioni importanti (come la gaussiana) viene fornita tabulata in fondo ai libri. In particolare, per la gaussiana e per tutte le distribuzioni simmetriche, $\Phi(z)$ è una funzione pari: $\Phi(-z) = \Phi(z)$.

3.4. **Uso delle tavole standard.** è in questo caso che si scopre l'utilità delle variabili standardizzate: le tavole che danno l'area sotto la curva f(x) (e cioè la distribuzione cumulativa $\Phi(z)$) suppongono che la variabile sia standard.

Cosi', se [x1,x2] sono i limiti di variazione di x, ed [z1,z2] i corrispondenti per z, calcolati con questa formula, invece che calcolare $P(x_1 < x < x_2)$, si calcolerebbe $P(z_1 < z < z_2)$: in effetti, l'area è la stessa. In pratica, il procedimento è il seguente: conoscendo x1 ed x2, si calcolano z1 ed z2 con la formula appena data, dopodicchè si vanno a vedere le tavole che danno l'area tra due punti standard.

La più importante delle funzioni cumulative è quella relativa alla distribuzione di Gauss, detta *distribuzione normale.*

4. L'approssimazione gaussiana

4.1. **La gaussiana come metodo di calcolo.** Uno degli utilizzi più comuni della gaussiana (una distribuzione continua) è quello di approssimare la binomiale (una

distribuzione discreta), e la ragione è questa: la binomiale contiene molti fattoriali e potenze e puo' dare problemi di calcolo se il numero di prove n è molto elevato.[3]

Ma se il numero di prove n è molto elevato, l'istogramma relativo alla binomiale è ben modellato dalla curva a campana di Gauss (ci sono molti più punti). Uno svantaggio si rivela in realtà un vantaggio, perchè è proprio la condizione necessaria per poter applicare Gauss (spesso $n > 30$ va già molto bene).

La binomiale pero' contiene i parametri (p, n) ; occorrono delle formule per ricavare i corrispondenti $(\mu\, \sigma)$, da usare con Gauss. Le formule, com'è noto, sono queste:[4]

$$\mu = n{\cdot}\, p \cdot \sigma^2 = n \cdot p \cdot (1-p)$$

4.2. **Esempio 1: valutare una singola probabilità bernouilliana.**

Problema 1. *Una moneta viene lanciata $n = 50$ volte; la probabilità a priori per le teste è $p = \frac{1}{2}$ (cioè: la moneta è non truccata). Qual'è la probabilità di avere esattamente 30 teste P(30)?*

Dal punto di vista della gaussiana, *esattamente 30 teste* non ha molto senso, perchè avremmo a che fare con un'area nulla. Si cercherà quindi di trovare $P(29.5 < x < 30.5)$, usando quindi un fascia larga=1 intorno al valore 30.

Una volta trovati i parametri corrispondenti alla gaussiana (che sono $\mu = 25$ e $\sigma = 3.54$) dobbiamo standardizzare i valori estremi, riducendoci al calcolo di:

$$P(1.27 < z < 1.56) = \Phi(1.56) - \Phi(1.27) = 0.0426 = 4.3\%$$

Quale sarebbe la risposta "esatta", cioè quella che fornirebbe la binomiale?

$$P(30) = \frac{50!}{30! \cdot 20!}(\frac{1}{2})^{30}(\frac{1}{2})^{20} == 0.04186$$

che è abbastanza vicino al valore trovato con l'approssimazione.

4.3. **Esempio 2: valutare un intero intervallo bernouilliano.**

Problema 2. *Nelle condizioni del problema precedente, qual'è la probabilità $P(10 < x < 30)$ che il numero di teste sia tra 10 e 30?*

Qui l'utilità dell'approssimazione gaussiana, come metodo di calcolo, è molto più evidente. Secondo la binomiale, infatti, ora dovremmo calcolare la quantità:

$$P(10 < x < 30) = \sum_{x=10}^{30} P(x) = P(10) + P(11) + ... + P(30)$$

[3]Lo scopo della distribuzione di Gauss non è solo quello di approssimare la binomiale, ma ha significati ben più profondi.

[4]Quella per μ è evidente: se faccio 80 prove e la probabilità di successo è $1/4$, avro' in media un numero di successi pari a $80 \cdot 1/4 = 20$.

un compito piuttosto laborioso!

Usando la gaussiana, al contrario, possiamo procedere esattamente come nell'altro caso, standardizzando gli estremi 9.5 e 30.5 e calcolando l'area con le tavole.

5. La popolazione e il Campione

5.1. **Concetto di campione.** Tutto quanto abbiamo visto relativo all'applicazione della distribuzione normale su basava sulla conoscenza dei valori $(\mu\,\sigma)$. C'è pero' un piccolo problema: in generale questi due valori $(\mu\,\sigma)$ (*parametri della popolazione*, o parametri "teorici") NON sono conosciuti a priori, ma siamo costretti a stimarli sulla base di un campione di N misure. Chiameremo (m,s) i parametri propri del campione (*parametri campionari*). L'assunsione fondamentale che facciamo (conseguenza del *teorema limite centrale*) è che, facendo tendere la numerosità del campione all'infinito, i parametri "campionari" tendano ai parametri "teorici".

è quello che fanno ogni giorno agenzie demoscopiche quali ABAQUS. Come fa un'agenzia demoscopica a stimare l'altezza media della popolazione italiana? Selezionerebbe un campione di N individui, li misurebbe e ne calcolerebbe la media aritmentica m , dando per scontato che il valore m approssimi tanto meglio il valore vero, quanto più è grande il campione ($N \to \infty$) .

5.2. **Stimatori corretti.** Le formule usate per la stima (detti *stimatori corretti*), si basano sulla media aritmetica e sono le seguenti:

$$m = \frac{1}{N}\sum_i x_i \cdot s^2 = \frac{1}{N-1}\sum_i (x_i - m)^2$$

Perchè compare N-1 invece che N nella formula per s? Perchè i valori su cui, in questa formula, si somma non sono più indipendenti, ma sono diventati N-1 (sono stati utilizzati per calcolare m stesso).

5.3. **Gli stimatori fluttuano!** C'è un'importante differenza tra un parametro campionario e un parametro di popolazione, ed occorre tenerla ben presente se si vuole comprendere il concetto di stimatore. La differenza è questa: mentre il parametro della popolazione è una quantità ben fissata, anche se ignota, il parametro della popolazione è una quantità soggetta a fluttuazione statistica: dipende infatti dal campione scelto. Ne consegue che (m, s^2) sono variabili stocastiche a tutti gli effetti, con una loro media ed una loro varianza (come tutte le variabili stocastiche).[5]

[5]Anche se, ammettiamolo, il fatto che una *media* abbia ... una *valore medio*, puo' confondere un po'.

5.4. **La media campionaria come variabile statistica.** Concentriamoci ora solo sul parametro m. Se io prendo N=100 italiani e calcolo la loro altezza media m e se, fatto cio', io scelgo altri 100 italiani e ne rifaccio la media, trovero' valori diversi, anche se di poco. La fluttuazione sarà tanto più piccola, quanto più grande è il campione N. Considerata come variabile stocastica, qual'è la media di m e qual'è la sua varianza? A questa domanda si puo' dare una risposta basandosi sulle proprietà generali dell'operatore di media.

Per le dimostrazioni, vedi Appendice

Il valore medio (teorico) della variabile m è proprio μ : benchè m fluttui a seconda del campione di italiani che ho scelto, siamo ragionevolmente sicuri che, se io mediassi questi valori, mi avvicinerei alla media teorica μ .

$$\langle m \rangle = \mu$$

Cosa si puo' invece dire della varianza s_m^2 di m , considerata come variabile statistica? Com'è intuitivo, m di per se varia meno di come varierebbe una singola misura. Si puo' dimostrare che:

$$s_m^2 = \langle \Delta m^2 \rangle = \langle (m-\mu)^2 \rangle = \frac{\sigma^2}{N}$$

formula che conferma il fatto che, se il campione N è molto grande, il valore medio di m deve tendere al valore "vero", al valore "teorico", la media vera della popolazione μ , e quindi lo scarto di m (come variabile statistica) deve tendere a zero.

5.5. **Discussione.** Supponiamo che misurando N=100 italiani si trovi un'altezza media $m = 170$ con uno scarto $s = 10$. Cosa significa $s = 10$? Significa che il grosso dei 100 italiani aveva pesi nel range 170 ± 10 , cioè tra 180 e 160 cm (questo verrà precisato più avanti).

I valori forniti sono TOTALMENTE inventati!

Cosa possiamo invece dire della "vera" altezza media degli italiani μ? Saremmo tentati di dire che $\mu = 170 \pm 10$, con cio' commettendo un grosso errore, perchè s=10 è lo scarto per il singolo peso x , non lo scarto per il peso medio m !

Se il singolo peso varia mediamente di 10 cm, il peso medio varierà molto di meno: l'operazione di media riduce la variabilità.

Quello che ci serve non è quindi s , ma s_m , lo scarto della media calcolabile con la formula

$$s_m = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{100}} = 1$$

La media teorica degli italiani è quindi $\mu = 170 \pm 1$. Insomma, più è grande il campione, più si riduce lo scarto, e quindi l'errore con cui valutiamo la media teorica.

Perchè ho usato il valore 10 per σ nella formula precedente? In realtà, σ non è noto. Ho dovuto usare uno stimatore per esso ed ho usato lo stimatore corretto dedotto dal campione.

6. Teoria statistica della stima

6.1. **Obbiettivo.** Valutare *l'affidabilità* con cui, a partire dai parametri del campione (m, s) , stimiamo i parametri della popolazione stessa $(\mu\, \sigma)$.

6.2. **Intervalli di confidenza.** Per concretezza, riprendiamo l'esempio dell'altezza media degli italiani. Abbiamo, in quel caso fornito il range $\mu = 170 \pm 1$ come il più probabile valore "vero" della media degli italiani, ben consci che vi sono italiani più alti di 171 cm e più bassi di 169.

Fortunatamente si puo' fare di meglio: si puo' introdurre il concetto di *intervallo di confidenza* per le nostre stime di μ. Ci si puo' chiedere, ad esempio, con quale probabilità il valore di μ cada negli intervalli $m \pm s_m$, oppure $m \pm 2s_m$ o anche $m \pm 3s_m$, etc.

Per quanto visto prima, queste probabilità possono essere valutate come l'area sotto la gaussiana, considerando la variabile x con media m e scarto s_m . Volendo ad esempio calcolare $P(m - s_m < x < m + s_m)$, dovremmo introdurre la variabile standard $z = \frac{x-m}{s_m}$ per riscrivere l'integrale nella forma $P(-1 < z < 1)$. Consultanto la tabella si trova:

$$P(-1 < z < 1) = 68.27\%$$

Questa è dunque l'affidabilità per un singolo scarto, in aumentare o diminuire. Analogamente si puo' fare per 2 scarti, 3 scarti, o anche per frazioni decimali di scarto.

Se lo facessimo, troveremmo la seguente tabella, con i valori più tipicamente usati:

Livello di confidenza	68.27 %	95%	95.45%	99 %	99.73%
scarto assoluto	1	1.96	2	2.58	3

Il più usato di questi corrisponde al 95% e quindi 1.96 scarti. In sostanza, si ha:

$$P(m - 1.96 \cdot s_m < x < m + 1.96 \cdot s_m) = 95\%$$

6.3. **Commenti.**

- Sappiamo che il valore medio m calcolato su un campio di 100 italiani non è la vera altezza media degli italiani, che resta per sempre sconosciuta.
- Sappiamo che questo valore medio m è in realtà una variabile aleatoria, con un suo proprio campo di variabilità (scarto s_m =1), per cui se prendiamo un campione di 100 italiani da 1000 città diverse siamo certi di trovare valori medi m sempre diversi.
- Sappiamo che (*siamo confidenti* che) il valore "vero" dell'altezza media degli italiani deve essere nell'intervallo $[m-1.96s_m, m+1.96s_m] = [168.04, 171.96]$, con una probabilità del 95% per cento.

Che significa *confidenti al 95%*? Significa che dei 1000 gruppi di italiani presi dalle 1000 città, solo 50 gruppi avranno altezza media fuori da questo intervallo, mentre per 950 gruppi la regola deve valere.

6.4. **Decisioni statistiche.** E se invece i gruppi "anomali" fossero 60, invece che 50? Che dovremmo dire dei 10 gruppi anomali in più?

Saremmo costretti a prendere una decisione e ad affermare cose del tipo:

- con probabilità al 95%, quei 10 gruppi in più non sono stati presi da città italiane!
- forse il campionamento non è stato corretto, ben distribuito sulla popolazione
- forse il campionamento era *realmente* pilotato ...
- forse ....

7. Esempio completo

7.1. **I dati del problema.** In una certa Università si analizza il peso di N=100 studenti maschi e si costruisce la seguente tabella:

Classe di peso (Kg)	Valore centrale x_i	Numero n_i
60-62	61	5
63-65	64	18
66-68	67	42
69-71	70	27
72-74	73	8

I valori hanno questo significato: 5 studenti hanno peso compreso tra 60 e 62; 18 hanno peso compreso tra 63 e 65, etc.

La colonna indicata con "valore centrale" fornisce invece il valore centrale nella classe dei pesi. In sostanza, possiamo dire che è come se 5 studenti avessero peso 61 Kg, 18 studenti avessero peso 64 Kg, etc.

7.1.1. *Stima della Media μ.* Vogliamo stimare la media μ della popolazione; useremo quindi il valore medio m ottenibile con questi particolari N=100 studenti.

Come calcolare la media m dei pesi? Chiaramente sarà:

$$m = \frac{5 \cdot 61 + 18 \cdot 64 + ...}{100} = \frac{\sum_{i=1}^{100} n_i \cdot x_i}{\sum_{i=1}^{100} n_i} = 67.45\, K$$

Se osservate la seconda forma, dovreste scoprire la media pesata dei valori centrali, pesati col numero di studenti che hanno quel peso.

7.1.2. *Stima dello Scarto σ.* Per lo scarto varrà una formula analoga alla media, tenendo pero' conto che nella formula degli stimatori corretti occorre usare N-1, e non N, al denominatore:

$$s^2 = \frac{\sum_{i=1}^{100} n_i \cdot (x_i - m)^2}{\sum_{i=1}^{100} n_i - 1} = 8.61$$

e quindi $s = 2.93$.

7.2. **Stima dello scarto della media.** Sarà, per quanto visto, $s_m = \frac{\sigma}{\sqrt{N}} = \frac{2.94}{10} = 0.29$

7.3. **Intervallo di confidenza al 95%.** Dobbiamo usare il fattore 1.96, per cui (con affidabilità del 95%) ci aspettiamo che la "vera" media per tutti gli studenti dell'università sia nell'intervallo

$$67.45 \pm 1.96 \cdot 0.29 = [66.88, 68.01]$$

7.4. **Esempio di decisione.** Mai scritto :-(