

Capitolo 1. Gli scopi. Campione e popolazione. Tipi di variabili. Tabelle e istogrammi di frequenza. Distribuzioni. Skewness. Kurtosis. Modalità. Media e mediana. Devianza. Varianza. Deviazione standard. Distribuzione standardizzata. Imprecisione e vizio. Campionamento randomizzato e sistematico.

Scopi principali dei metodi statistici sono:

- rappresentare sinteticamente i dati
- valutare variazioni
- verificare ipotesi
- fare previsioni
- analizzare le relazioni tra le variabili
- analizzare le relazioni tra i soggetti
- discriminare gruppi, individuare tipologie, ecc.

Va pur detto che la statistica NON serve quando i fenomeni sono netti e la loro valutazione è univoca e non ammette dubbi. Non serve statistica per verificare se il agosto sia più caldo di gennaio o se il cianuro sia più tossico dell'ossigeno. Invece, in tutti quei casi in cui i fenomeni possono essere oggetto di discussione e di valutazioni diverse la statistica è senz'altro uno strumento utile di conoscenza.

Campione e popolazione

Il concetto di statistica è strettamente legato a quello di campione. Il campione è formato da un gruppo limitato di individui scelto a rappresentare l'intera popolazione. Dal campione è possibile ottenere informazioni sulla popolazione. La popolazione è spesso costituita da un numero infinito di individui, oppure anche da un numero fisicamente finito, ma praticamente irraggiungibile (es., tutte le foglie di un bosco, tutte le cellule di un organismo, ecc.) per cui, anche volendo, è quasi sempre impossibile disporre dei dati di tutti gli elementi della popolazione. Si parla di campioni osservazionali quando ci si limita ad osservare i fenomeni naturali, senza intervenire su di essi (es., il peso di ratti allevati in condizioni standard). Si parla invece di campioni sperimentali quando si interviene o si tenta di intervenire producendo o modificando i fenomeni artificialmente (es., il peso di ratti allevati con particolari diete o trattati con farmaci). E' comunque essenziale che la popolazione sia univocamente definita. Se ad es., si parla di animali di laboratorio è bene specificare - oltre alla specie - la varietà, il range di età e di peso medio, le condizioni di allevamento, ecc. Notare che la definizione di popolazione statistica non corrisponde al concetto di popolazione biologica. In statistica si considerano popolazioni di diametri, di valori di pressione, di pesi, ecc. Ovviamente le popolazioni statistiche non si riproducono.

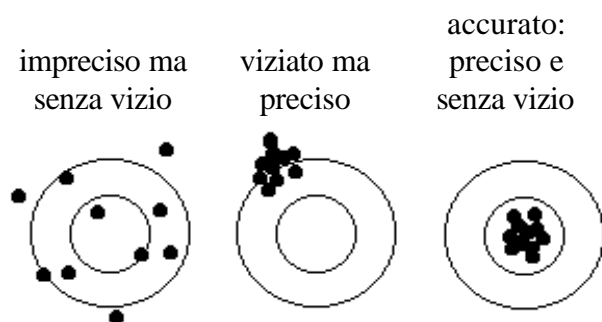
A seconda dei problemi, i metodi statistici possono prendere contemporaneamente in considerazione uno, due o più campioni.

Fonti di variabilità e campionamento. Imprecisione e vizio.

Come variabilità bisogna intendere l'effetto di due principali fattori: (a) la variabilità naturale, intrinseca dei fenomeni, e (b) la variabilità dovuta all'**imprecisione**. L'imprecisione determina un aumento della variabilità, ma di per sé non altera la media. L'imprecisione può essere valutata (ad es. ripetendo la

pesata di uno stesso oggetto) e in parte compensata da un maggior numero di misurazioni (ad es. facendo la media di più pesate). Un certo grado di imprecisione è comunque sempre presente; l'importante è che essa sia così piccola da essere trascurabile in rapporto al fenomeno da valutare.

Un altro tipo di errore, ma che non produce variabilità, è il **vizio**. Questo tipo di errore, molto più subdolo dell'imprecisione, consiste in un'alterazione costante dei dati (ad es. una bilancia non tarata). Il vizio, in assenza di campioni standard di riferimento (ad es. un oggetto di un certo peso, una soluzione di un certo pH, ecc.), è impossibile da riconoscere e quindi da evitare, ed ha conseguenze ben più negative della semplice imprecisione. Il vizio è ciò che determina sottostime o sovrastime. Inoltre il vizio altera i confronti con i dati ottenuti in altri laboratori e fa apparire differenze dove invece non c'è niente di diverso.



L'approssimazione numerica è un buon esempio di controllo di precisione e di assenza di vizio. Infatti, se da un lato le cifre meno significative rappresentano fluttuazioni non rilevanti (potenzialmente frutto di imprecisione) d'altro lato il troncamento di queste cifre porterebbe ad una diminuzione costante del valore (vizio di sottostima). Si interviene allora sul valore troncato incrementandone o riducendone l'ultima cifra a seconda del valore della parte troncata. Invece, un tipico esempio di vizio dovuto a troncamento del dato è quello dell'età. L'età infatti è aggiornata solo allo scadere del compleanno. Pertanto risulta che l'età è mediamente sottostimata di $\frac{1}{2}$ anno. Ma questa è una convenzione seguita da tutti e quindi non crea problemi.

Imprecisione e vizio non risiedono solo nella fase di misurazione o di rappresentazione numerica dei dati. Il loro pericolo è presente soprattutto nella fase del campionamento. Se ad es. occorre formare un gruppo di soggetti da sottoporre a trattamento ed un altro gruppo da utilizzare come controllo, è sempre bene estrarre a sorte quali soggetti andranno a formare l'uno o l'altro gruppo. E' classico l'errore di chi, avendo a che fare con animali da laboratorio, destina al trattamento i primi 10 animali che riesce ad afferrare e lascia gli altri 10 come controllo. E' assai probabile la condizione fisica dei 10 animali che si lasciano afferrare per primi sia differente da quella degli altri 10 che si sottraggono alla cattura. Oppure il caso dei pazienti dell'illustre luminare con onorari milionari. Possiamo supporre che tali pazienti abbiano condizioni di vita, alimentazione, ecc. del tutto speciali. Pertanto il risultato dell'esperimento potrà essere deviato o influenzato da fattori estranei selezionati durante il campionamento scorretto. La **randomizzazione** (o **campionamento casuale**) è in genere in buon criterio di

campionamento. Ma migliori sono quei metodi attuano un campionamento **sistematico** o **stratificato**, che consiste nella scelta equilibrata dei soggetti alla luce di oggettive influenze ambientali. Ad es., dovendo estrarre un campione di abitanti di una città, è senz'altro meglio procedere in modo sistematico scegliendo individui di diversa età, sesso, professione, ecc. anziché procedere in modo del tutto casuale.

Le variabili

Le variabili quantificano l'espressione dei fenomeni osservati. Esistono diversi tipi di variabili, in quanto caratteri di natura diversa esigono diversi criteri di valutazione (es., qualitativi o quantitativi). Capire i tipi di variabili è essenziale per la scelta dei metodi statistici da applicare.

I tipi principali di variabili sono:

a) Il **semplice conteggio**. Ciò che interessa è il solo numero degli individui o degli eventi, ovviamente riferiti ad un certo ambito di spazio o di tempo (es., numero di batteri in un certo volume di liquido, numero di cellule per campo microscopico, numero di impulsi elettrici prodotti in un certo intervallo di tempo, numero di abitanti per Km²). La numerosità riferita all'unità di spazio è detta spesso densità; la numerosità riferita all'unità di tempo è detta spesso frequenza. I valori sono ovviamente discreti, cioè rappresentati da numeri interi.

b) Il **conteggio di oggetti**, questa volta **distinti per tipo**, carattere, categoria, qualità, modalità, attributo, ecc. Ciò che interessa non è più il semplice numero globale degli individui, ma la loro distinzione o ripartizione in diversi tipi. Esiste anche questa volta un riferimento ad un certo ambito spaziale o temporale, ma giusto per specificare le circostanze o la natura del campione. L'interesse principale sta ora nella ripartizione del carattere tra gli individui (es., quante cellule normali e non normali in una sezione istologica, quanti eritrociti immaturi (reticolociti) e quanti maturi in uno striscio di sangue, quanti individui di gruppo A, B, AB e 0 in una certa popolazione, quanti figli maschi e quante femmine in una certa famiglia, ecc.). Tali conteggi possono essere espressi come proporzioni o rapporti percentuali (es., il 15% delle cellule mostrano basofilia, il 5% dei globuli bianchi è rappresentato da monociti, ecc.) che comunque devono essere sempre riferiti al numero totale (es., sono state osservate in tutto 1427 cellule). Anche in questo caso i valori sono discreti.

Attenzione: le modalità' delle variabili qualitative devono sempre essere mutualmente esclusive (senza sovrapposizioni o ambiguità) ed esaustive (comprendere tutte le possibili tipologie) del campione. Il caso più semplice di modalità mutualmente esclusive ed esaustive è quello della presenza/assenza di un determinato carattere (es., malattia presente/malattia assente), oppure la presenza di un determinato carattere contro tutti gli altri (es., colore rosso/qualsiasi altro colore).

c) Il **conteggio di oggetti distinti per tipo** come sopra, ma **ordinabili** su una certa scala o grado di apprezzamento. L'esempio classico è quello dei giudizi o

punteggi (es., merito insufficiente, sufficiente, buono, ottimo; malattia lieve, moderata, grave; segnale di intensità debole, normale, forte, ecc.). I giudizi o punteggi nominali possono e sono spesso sostituiti da valori all'interno di scale convenzionali (es., scala dei voti scolastici, scala della forza del vento, ecc.). Le variabili nominali ordinabili possono essere in tal modo valutate quantitativamente con valori discreti, convenzionali, rappresentati in genere da piccoli interi (es., scarso: 0, mediocre: 1, buono: 2). I valori sono ancora discreti.

- d) Le **misure** strumentali di oggetti. Riguarda grandezze fisiche (es. peso, altezza, durata, massa, distanza, velocità, forza, lunghezza, superficie, ecc.) valutabili con un grado di precisione teoricamente infinito, anche se praticamente limitato dal livello di accuratezza dello strumento di misura. I valori sono quindi rappresentati da numeri su scala continua. Per praticità si usa spesso approssimare i dati alle prime 3 o 4 cifre significative, capaci di suddividere il range dei valori in mille o diecimila intervalli circa.

Esiste la possibilità di trasformare queste variabili continue nel tipo precedente di variabili qualitative discontinue ordinabili sostituendo al valore dei dati il **rango**, cioè la posizione occupata nella disposizione in ordine crescente.

Es.:

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

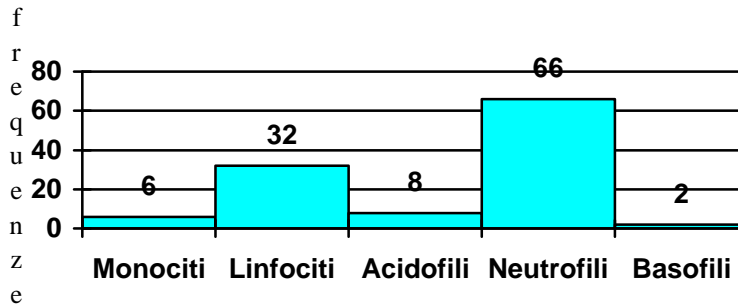
La trasformazione in rango comporta una perdita di informazione (aspetto negativo) ma elimina anche eventuali condizionamenti della distribuzione che potrebbero invalidare diverse analisi statistiche basate sul fatto o sull'ipotesi che la distribuzione dei dati sia normale (aspetto positivo). L'argomento sarà ripreso quando si parlerà dei cosiddetti metodi non-parametrici. Come semplice anticipazione, i metodi non-parametrici, generalmente applicati ai ranghi, sono svincolati dal tipo di distribuzione. A seconda dei problemi, i metodi statistici prendono in considerazione 1, 2 o più variabili (metodi mono- bi- e multi-variati).

Dall'istogramma di frequenze alla distribuzione di probabilità

Una utile forma di rappresentazione di gruppi di dati consiste nell'istogramma delle frequenze ripartite per categoria nel caso di variabili nominali, o per classi di intervalli nel caso di variabili quantitative. Ad es. consideriamo la tabella di frequenze di diversi tipi di leucociti trovati in uno striscio di sangue (variabile nominale non ordinabile):

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

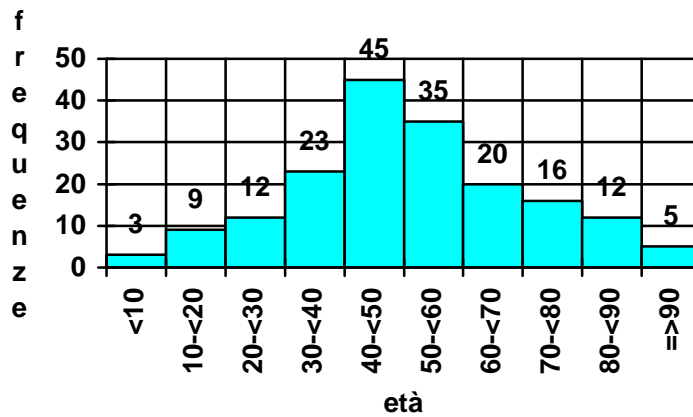
Tutta l'informazione della tabella può essere convertita in forma grafica in un istogramma di frequenze:



Attenzione che le categorie o classi nominali siano mutualmente esclusive (non ambigue) ed esaustive (comprendano tutti i tipi) !

Consideriamo ora la tabella di frequenze di classi di età di un campione di 180 soggetti:

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



In questo caso ogni classe rappresenta un certo intervallo di valori di ampiezza costante (tranne le classi in testa e coda). E' comunque molto importante che le classi non siano nè troppe (si otterrebbe un istogramma 'sdentato' irregolare) né troppo poche (si otterrebbe un istogramma concentrato senza dinamica). C'è una formula che suggerisce quante classi rappresentare:

$$N^{\circ} \text{ di classi} = 1 + 3.3 \log_{10}(n)$$

ove n indica il numero totale di dati da rappresentare in grafico.

Es.:

se $n = 100$, $N^{\circ} \text{ di classi} = 1 + 3.3 \log_{10}(100) \approx 8$

se $n = 500$, $N^{\circ} \text{ di classi} = 1 + 3.3 \log_{10}(500) \approx 10$

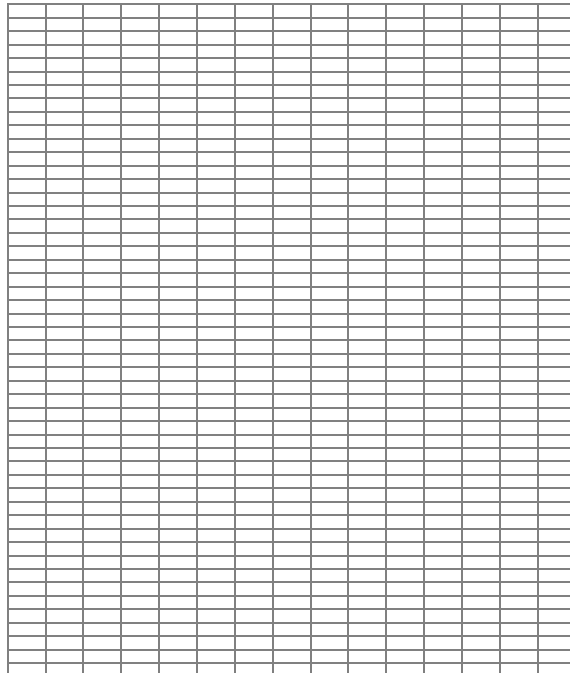
Esistono tuttavia situazioni pratiche che sconsigliano di seguire alla lettera tale norma. Ad es., se si vogliono ripartire dei soggetti per classi di età, dovremo fare in modo che gli intervalli delle classi corrispondano a valori interi corrispondenti agli anni di età. Sceglieremo quindi di volta in volta il compromesso che meglio si adatta alla situazione.

Ipotizzando di avere infiniti dati a disposizione, la formula espressa indicherebbe un numero di infinito di classi. Per cui il profilo dell'istogramma diventerebbe simile ad una curva continua. La curva della distribuzione di una popolazione infinita può pertanto considerarsi un istogramma suddiviso in infinite classi di intervallo infinitesimo.

Esercizio

Questi sono i valori di grigio di 196 pixels di una immagine digitale (di 14×14 pixels). Costruite sotto l'istogramma di frequenze di questi dati utilizzando un intervallo di classe di 5.

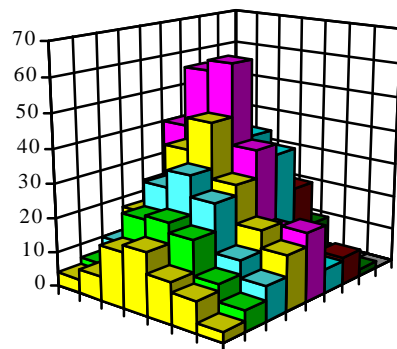
52 50 52 51 49 51 49 48 49 64 42 64 60 79
 53 52 53 49 48 49 51 49 51 65 65 65 73 72
 46 45 46 45 44 45 47 46 47 64 63 64 61 59
 61 59 61 58 57 58 52 51 52 57 56 57 73 72
 60 59 60 52 50 52 46 45 46 62 60 62 69 68
 46 45 46 44 43 44 46 45 46 62 61 62 61 59
 33 31 33 48 47 48 49 48 49 53 52 53 61 59
 37 35 37 53 53 53 45 44 45 49 48 49 59 58
 31 29 31 48 46 48 51 49 51 50 49 50 54 53
 22 21 22 35 34 35 53 52 53 50 49 50 48 46
 17 16 17 21 20 21 42 41 42 46 45 46 39 39
 16 15 16 12 11 12 27 26 27 39 38 39 35 35
 15 14 15 11 10 11 17 16 17 31 31 31 39 37
 15 14 15 17 16 17 12 12 12 19 18 19 34 33



Esistono anche **tabelle di frequenze a due entrate** come la seguente:

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

che possono ancora essere rappresentate graficamente (i cosiddetti grafici 3D). Il risultato è di un certo effetto anche se parte dell'informazione viene persa in quanto alcuni oggetti in secondo piano risultano nascosti:










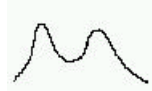
Sia la tabella a due entrate che l'istogramma 3D rappresentano una distribuzione **bivariata**, definita cioè da 2 variabili. Non è possibile invece rappresentare (decentemente) tabelle né istogrammi di frequenze con più di due variabili. E' tuttavia possibile estrarre importanti informazioni attraverso metodi cosiddetti di **statistica multivariata**. Ci occuperemo brevemente di questi metodi verso la fine del corso.

Riprendiamo comunque il discorso dell'istogramma di frequenze più semplice: dalla sua osservazione è possibile valutare alcune proprietà della **forma della distribuzione**. I parametri statistici che definiscono obiettivamente tali caratteristiche sono:

- Skewness**. Indica il grado di simmetria o di regolarità orizzontale della curva. La curva della distribuzione cosiddetta normale è anche simmetrica. Ma la simmetria non è condizione sufficiente per affermare che una distribuzione sia normale. Il valore di skewness standardizzata è zero se la distribuzione è simmetrica, negativo se vi è una cosiddetta **coda** a sinistra (presenza di classi con valori assai bassi) e positivo se vi è una coda a destra (presenza di classi con valori assai alti).

- b) **Kurtosis.** Indica il grado di appiattimento o di regolarità verticale della curva. La curva di distribuzione normale ha in ogni punto una caratteristica pendenza, in rapporto alla media ed alla deviazione standard (se ne parlerà più avanti, ma sin d'ora diciamo che media e deviazione standard definiscono esaurientemente una distribuzione normale). Per la distribuzione normale la kurtosis standardizzata è pari a zero. E' ovvio che dilatando o contraendo la scala dell'ordinata - come giocando con l'elastico - la curva potrà apparire ora sottile e slanciata ora larga e piatta, ma in questo caso i rapporti di pendenza resteranno invariati. Se invece alteriamo la curva avremo kurtosis negative se aumenta la concavità e kurtosis positive se aumenta la convessità. Ad esempio, se rubiamo alla base ed aggiungiamo alla sommità la kurtosis aumenta, mentre diminuisce nel caso contrario. Il caso più estremo di kurtosis positiva è quello di una distribuzione uniforme visualizzabile come un istogramma piatto.
- c) **Modalità.** La distribuzione normale (come del resto la stragrande maggioranza delle distribuzioni statistiche) è monomodale, cioè presenta un singolo picco di maggior frequenza (o probabilità) che degrada monotonamente a destra e sinistra. Se l'istogramma prodotto da un numero consistente di osservazioni dimostra la presenza di due o più picchi ben separati occorre verificare l'omogeneità del campione, se cioè esso non comprenda mescolanze di individui provenienti da più popolazioni. Questo è, in realtà, uno dei temi più interessanti della statistica, e numerose applicazioni che vedremo in seguito avranno appunto a che fare con la verifica di ipotesi di omogeneità o eterogeneità dei campioni.

Non è necessario soffermarsi ulteriormente sul significato di skewness e kurtosis né affrontarne il calcolo. E' tuttavia utile ricordare due fatti (1) i valori di skewness e kurtosis devono essere standardizzati, altrimenti non sono confrontabili, e pertanto risultano di scarsa utilità e (2) anche nei campioni tratti da distribuzioni normali molto raramente avremo skewness e kurtosis perfette, uguali a zero. Il loro valore quindi varierà con una certa libertà che dipenderà dalla grandezza del campione. Nei manuali di statistica più completi esistono tabelle che riportano la distribuzione dei valori di skewness e kurtosis per campioni tratti da distribuzioni normali. Il confronto tra i valori di skewness e kurtosis calcolati con quelli tabulati costituisce un semplice e rapido test per verificare l'ipotesi che i valori della popolazione da cui proviene il campione siano distribuiti normalmente (test di normalità).

	distribuzioni non normali	distribuzione normale	distribuzioni non normali
skewness	 distribuzione asimmetrica coda a destra skewness positiva es. peso di neonati alla nascita	 distribuzione simmetrica skewness nulla	 distribuzione asimmetrica coda a sinistra skewness negativa es. valori del visus corretto
kurtosis	 distribuzione convessa o uniforme kurtosis positiva es. numeri della tombola	 distribuzione normalmente flessa kurtosis nulla	 distribuzione concava kurtosis negativa
modalità		 distribuzione monomodale	 distribuzione bimodale es. statura di Pigmei e Vatussi

Media e mediana

Un altro utile elemento per giudicare se i nostri dati sono compatibili con la distribuzione normale è il confronto tra media e mediana. La media è data dalla somma dei dati diviso il numero: $m = (\sum x)/n$. La mediana invece riporta il valore del dato centrale, cioè del dato con rango uguale a $n/2+0.5$ se il numero dei dati è dispari, oppure $n/2$ e $n/2+1$ se il numero dei dati è pari (il rango di un dato è la sua posizione in classifica, ponendo i dati in ordine crescente). Quando la distribuzione è simmetrica, allora media e mediana coincidono. Se invece la distribuzione è vistosamente asimmetrica, è bene prendere come riferimento centrale la mediana al posto della media. La media in tal caso è fortemente influenzata dai valori estremi (piccolissimi o grandissimi) della coda. La mediana non è invece affatto influenzata dai valori degli estremi. Tale discussione tra media e mediana, con tutte le sue implicazioni, anticipa il confronto tra statistica parametrica e statistica non-parametrica che faremo al termine del corso. La mediana è un dato tipicamente non-parametrico.

Un altro modo di indicare la mediana è quello di definirla come **50° percentile**. Tale definizione, anche mai usata nel caso specifico della media, è utile poiché è generalizzabile. Ad esempio, il 5° percentile è quel valore-soglia che separa il 5% dei dati più piccoli dal restante 95%. Allo stesso modo, il valore-soglia che separa il 95% dei dati più bassi dal restante 5% è detto 95° percentile.

Es., i voti degli esami universitari sono chiaramente distribuiti in modo non normale (provate voi a fare un istogramma). In tal caso è bene riferire i risultati di una sessione non calcolando la media dei voti ma dicendo quanti studenti in percentuale, hanno superato l'esame. Il voto di 18/30 in tal caso è la soglia di scala a cui corrisponde quella percentuale, o meglio, quel percentile.

La ricerca di un parametro che descriva la dispersione

Nella distribuzione normale la media è un importante parametro in quanto rappresenta il riferimento centrale della distribuzione, ma non è sufficiente a rappresentare un'altra importante caratteristica della distribuzione: la dispersione.

Consideriamo ad esempio i due campioni:

campione A: 10, 50, 90 media = $(10+50+90)/3 = 50$

campione B: 45, 50, 55 media = $(45+50+55)/3 = 50$

Sia A che B hanno la stessa media. Ma è evidente che la distribuzione di A è assai più dispersa di quella di B. In altri termini, in A i dati sono in genere più distanti dalla media di quanto non lo siano in B. Cominciamo quindi a considerare le differenze tra i diversi dati e la media.

Es.:

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

Queste differenze tra i dati e la media sono dette in gergo **scarti**.

La somma degli scarti, positivi e negativi, è per definizione nulla, in quanto la media è il valore centrale della distribuzione (se non è nulla vuol dire che abbiamo fatto male i calcoli). Pertanto la semplice somma degli scarti non serve a niente. Le cose cambiano se eleviamo al quadrato gli scarti. In tal modo tutti i valori acquistano segno positivo.

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

In gergo, la somma dei quadrati degli scarti, cioè delle differenze tra i vari dati e la loro media, è detta semplicemente **somma dei quadrati** o **devianza**. Il simbolo più frequente è una **S** maiuscola, ma talora si trova anche il simbolo italiano **SQ** (Somma dei Quadrati) o l'inglese **SS** (Sum of Squares). Il termine abbreviato di somma dei quadrati si usa al posto della definizione completa, un po' più ingombrante: somma dei quadrati degli scarti tra ogni singolo dato x e la media m. In simboli algebrici:

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

La devianza è così un primo, grezzo indice di dispersione. Ma ha un grosso difetto: aumenta con l'aumentare del numero dei dati (n). Se ad es., raddoppiassimo il campione di sopra, duplicando i dati, otterremo la stessa media, ma devianza doppia:

dato x	differenza x-media	differenza (x-media) ²
3	3-5= -2	(3-5) ² = 4
6	6-5= +1	(6-5) ² = 1
4	4-5= -1	(4-5) ² = 1
7	7-5= +2	(7-5) ² = 4
3	3-5= -2	(3-5) ² = 4
6	6-5= +1	(6-5) ² = 1
4	4-5= -1	(4-5) ² = 1
7	7-5= +2	(7-5) ² = 4
somma = 40 media = 5	somma = 0	somma = 20 devianza

E' estremamente scomodo il fatto che, pur mantenendo la stessa media e dispersione dei dati, la devianza aumenti in funzione di n. Per ottenere pertanto un parametro stabile, confrontabile, dobbiamo quindi dividere la devianza per il numero di dati. Il risultato di tale operazione è detto **scarto quadratico medio** o **devianza media** o, meglio, **varianza** indicata col simbolo s^2 :

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

Nei campioni poco numerosi la varianza così stimata è inferiore alla vera varianza della popolazione (si parla di sottostima). Per correggere tale vizio è sufficiente dividere la devianza per n-1 anziché per n. Tale correzione è necessaria ed efficace quando applicata ai piccoli campioni, mentre è irrilevante, nel senso che non ha effetto né è necessaria, quando applicata a grandi campioni con $n > 100$.

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

Nel nostro primo caso con $n = 4$, $s^2 = 10/(4-1) = 3.333$.

n-1 si definiscono **gradi di libertà** della varianza, nel senso che i dati che concorrono a determinare la varianza sono n-1, in quanto un dato è già vincolato dalla media (la varianza l'abbiamo calcolata dagli scarti dalla media). Il concetto di gradi di libertà non è sempre chiaro dal primo istante, per cui è utile rinforzarlo ora affermando che è possibile modificare a proprio piacimento qualsiasi media aggiungendo un solo dato scelto da noi (provare per credere). Questo dato che alla fine determina la media è quello sottratto ad n nei gradi di libertà della varianza. D'ora in poi capiterà spesso di considerare i gradi di libertà di diversi

parametri statistici: si tratterà semplicemente di assegnare al parametro statistico il corretto numero di dati che concorrono *liberamente* a definirlo.

Anche la varianza ha però un problema: quello di mantenere il carattere quadratico della scala (essendo ottenuta dai quadrati degli scarti). Ad es., la varianza della statura è espressa in m², la varianza dell'età in anni². Per riferire il parametro di dispersione alla scala di misura dei dati dobbiamo ora eliminare tale fattore quadratico, estraendo la radice quadrata della varianza. Il risultato di tale operazione è detto **deviazione standard**, indicata col simbolo s :

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

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

In conclusione, con la deviazione standard abbiamo trovato un parametro che

- rappresenta la dispersione dei dati attorno alla media
- tiene conto di tutti i dati del campione (a differenza del range minimo - massimo)
- non è influenzato dalla numerosità del campione (a differenza della devianza)
- è valutato con la stessa scala dei valori originari (a differenza della varianza).

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

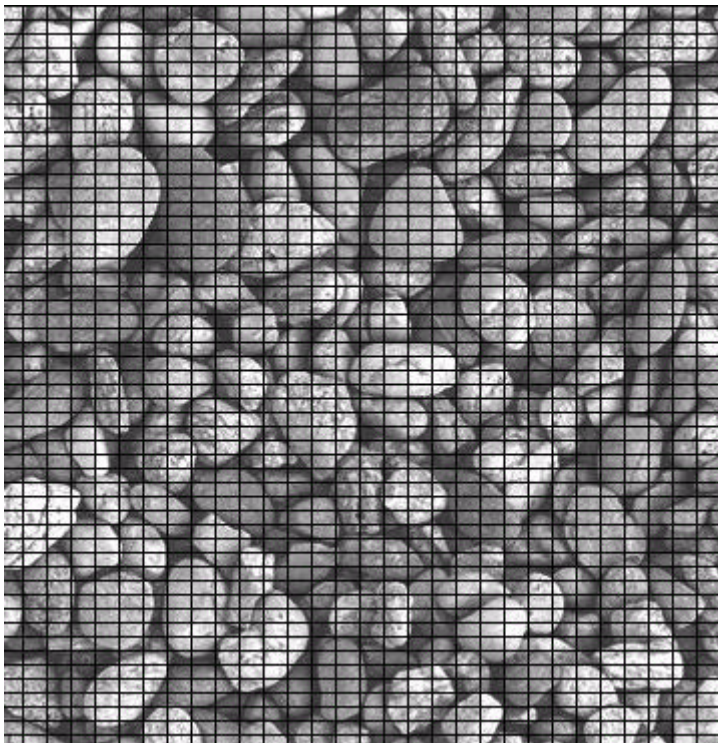
Esercizio

L'area di una forma 2D vista sotto un reticolo regolare è proporzionale al numero di incroci del reticolo che cadono sulla forma stessa. Supponendo che l'immagine abbia complessivamente una area di 10 cm^2 , calcolate l'area di 10 forme.

1	2	3	4	5	6	7	8	9	10


Infine, calcolate l'area media, la deviazione standard ed i limiti fiduciali della media


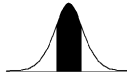
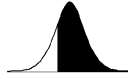
m	s	LF



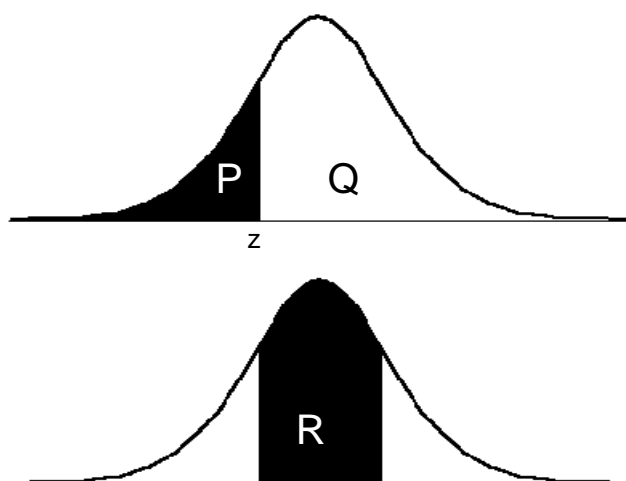
La distribuzione normale standardizzata

I valori di probabilità della distribuzione normale corrispondono a segmenti di area o integrali della curva gaussiana. Il loro calcolo, riferito di volta in volta ad una certa media m e deviazione standard s , per quanto non difficile, è piuttosto laborioso. Per questo è preferibile riferirsi sempre ad una distribuzione standard, cioè con media $= 0$ e deviazione standard $= 1$, i cui valori di probabilità, riferiti a diversi intervalli di scala, sono riportati in tutti i manuali.

CURVA DI DISTRIBUZIONE NORMALE			
media = 0			
deviazione standard = 1			
area totale sotto la curva = 1			
Ascissa valori negativi	Area della curva a sinistra dell'ascissa 	Area della curva compresa tra \pm ascissa 	Area della curva a destra dell'ascissa 
z	P(z)	R(z)	Q(z)
-4.000	0.0000	0.9999	1.0000
-3.500	0.0002	0.9995	0.9998
-3.250	0.0006	0.9988	0.9994
-3.000	0.0013	0.9973	0.9987
-2.750	0.0030	0.9940	0.9970
-2.576	0.0050	0.9900	0.9950
-2.500	0.0062	0.9876	0.9938
-2.250	0.0122	0.9756	0.9878
-2.000	0.0227	0.9545	0.9772
-1.960	0.0250	0.9500	0.9750
-1.950	0.0256	0.9488	0.9744
-1.900	0.0287	0.9426	0.9713
-1.850	0.0322	0.9357	0.9678
-1.800	0.0359	0.9281	0.9641
-1.750	0.0401	0.9199	0.9599
-1.700	0.0446	0.9109	0.9554
-1.650	0.0495	0.9011	0.9505
-1.645	0.0500	0.9000	0.9500
-1.600	0.0548	0.8904	0.9452
-1.550	0.0606	0.8789	0.9394
-1.500	0.0668	0.8664	0.9332
-1.450	0.0735	0.8529	0.9265
-1.400	0.0808	0.8385	0.9192
-1.350	0.0885	0.8230	0.9115
-1.300	0.0968	0.8064	0.9032
-1.250	0.1056	0.7887	0.8944
-1.200	0.1151	0.7699	0.8849
-1.150	0.1251	0.7499	0.8749
-1.100	0.1357	0.7287	0.8643
-1.050	0.1469	0.7063	0.8531
-1.000	0.1587	0.6827	0.8413
-0.950	0.1711	0.6579	0.8289
-0.900	0.1841	0.6319	0.8159
-0.850	0.1977	0.6047	0.8023
-0.800	0.2119	0.5763	0.7881
-0.750	0.2266	0.5467	0.7734
-0.700	0.2420	0.5161	0.7580
-0.650	0.2578	0.4843	0.7422
-0.600	0.2743	0.4515	0.7257
-0.550	0.2912	0.4177	0.7088
-0.500	0.3085	0.3829	0.6915
-0.450	0.3264	0.3473	0.6736
-0.400	0.3446	0.3108	0.6554
-0.350	0.3632	0.2737	0.6368
-0.300	0.3821	0.2358	0.6179
-0.250	0.4013	0.1974	0.5987
-0.200	0.4207	0.1585	0.5793
-0.150	0.4404	0.1192	0.5596
-0.100	0.4602	0.0797	0.5398
-0.050	0.4801	0.0399	0.5199
0.000	0.5000	0.0000	0.5000

CURVA DI DISTRIBUZIONE NORMALE			
media = 0			
deviazione standard = 1			
area totale sotto la curva = 1			
Ascissa valori positivi	Area della curva a sinistra dell'ascissa 	Area della curva compresa tra \pm ascissa 	Area della curva a destra dell'ascissa 
z	P(z)	R(z)	Q(z)
0.000	0.5000	0.0000	0.5000
0.050	0.5199	0.0399	0.4801
0.100	0.5398	0.0797	0.4602
0.150	0.5596	0.1192	0.4404
0.200	0.5793	0.1585	0.4207
0.250	0.5987	0.1974	0.4013
0.300	0.6179	0.2358	0.3821
0.350	0.6368	0.2737	0.3632
0.400	0.6554	0.3108	0.3446
0.450	0.6736	0.3473	0.3264
0.500	0.6915	0.3829	0.3085
0.550	0.7088	0.4177	0.2912
0.600	0.7257	0.4515	0.2743
0.650	0.7422	0.4843	0.2578
0.700	0.7580	0.5161	0.2420
0.750	0.7734	0.5467	0.2266
0.800	0.7881	0.5763	0.2119
0.850	0.8023	0.6047	0.1977
0.900	0.8159	0.6319	0.1841
0.950	0.8289	0.6579	0.1711
1.000	0.8413	0.6827	0.1587
1.050	0.8531	0.7063	0.1469
1.100	0.8643	0.7287	0.1357
1.150	0.8749	0.7499	0.1251
1.200	0.8849	0.7699	0.1151
1.250	0.8944	0.7887	0.1056
1.300	0.9032	0.8064	0.0968
1.350	0.9115	0.8230	0.0885
1.400	0.9192	0.8385	0.0808
1.450	0.9265	0.8529	0.0735
1.500	0.9332	0.8664	0.0668
1.550	0.9394	0.8789	0.0606
1.600	0.9452	0.8904	0.0548
1.645	0.9500	0.9000	0.0500
1.650	0.9505	0.9011	0.0495
1.700	0.9554	0.9109	0.0446
1.750	0.9599	0.9199	0.0401
1.800	0.9641	0.9281	0.0359
1.850	0.9678	0.9357	0.0322
1.900	0.9713	0.9426	0.0287
1.950	0.9744	0.9488	0.0256
1.960	0.9750	0.9500	0.0250
2.000	0.9772	0.9545	0.0228
2.250	0.9878	0.9756	0.0122
2.500	0.9938	0.9876	0.0062
2.576	0.9950	0.9900	0.0050
2.750	0.9970	0.9940	0.0030
3.000	0.9987	0.9973	0.0013
3.250	0.9994	0.9988	0.0006
3.500	0.9998	0.9995	0.0002
4.000	1.0000	0.9999	0.0000

Il significato di $P(z)$, $R(z)$ e $Q(z)$ è chiarito nel grafico seguente:



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

Ovviamente,
 $P(z)$ e $Q(z)$ sono complementari ad 1.
 $P(z) = 1 - Q(z)$

$R(z)$ invece è l'area compresa tra $\pm z$.

Ovviamente,
 $R(z) = 1 - 2P(z)$ se $z \leq 0.5$
 $R(z) = 1 - 2Q(z)$ se $z \geq 0.5$

La relazione tra un certo valore di ascissa z e l'area di curva a sinistra di z o $P(z)$ è spesso espressa come **percentile**. Ad esempio, a $z = -1.750$ corrisponde un valore di $P(z) \approx 0.04$. Diremo allora che -1.750 rappresenta il 4° percentile della distribuzione (solo il 4% dei soggetti avranno valori uguali o inferiori a -1.750). A $z = 0.850$ corrisponde un valore di $P(z) \approx 0.8$. Diremo allora che 0.850 rappresenta l'80° percentile (l'80% dei soggetti avranno valori uguali o inferiori a 0.850).

Ma che cosa è z ? È ciò che ci evita di calcolare nuovi integrali di probabilità per ogni diversa media e deviazione standard. Si tratta quindi di un'unica tabella applicabile a qualsiasi campione. Ma per far ciò dobbiamo trasformare i nostri dati in modo tale che la loro distribuzione abbia media = 0 e deviazione standard = 1. Tale operazione è detta **standardizzazione** e consiste semplicemente nel sottrarre ad ogni dato la media e dividere poi per la deviazione standard:

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

z non è altro che lo scarto di un dato dalla media espresso in unità di deviazioni standard.

Per fare un'applicazione pratica,

- prendiamo in esame il valore x di un soggetto di un campione che assumiamo provenga da una popolazione distribuita in modo normale
- in base alla media e deviazione standard del campione standardizziamo x e ricaviamo z
- $P(z)$ corrispondente a z ci dà la probabilità di trovare nel campione e nella popolazione soggetti con valori uguali o inferiori a x
- $Q(z)$ corrispondente a z ci dà la probabilità di trovare nel campione e nella popolazione soggetti con valori uguali o superiori a x
- $R(z)$ corrispondente a z ci dà la probabilità di trovare nel campione e nella popolazione soggetti con valori compresi tra x ed il suo valore speculare rispetto alla media [sarebbe $m + (m - x)$, cioè $2m - x$].

Se vogliamo valutare le probabilità relative ad un certo intervallo di scala compreso tra due valori qualsiasi x_1 e x_2 , dobbiamo prima standardizzare questi in z_1 e z_2 e poi trovare in tabella l'area inclusa nell'intervallo (con qualche semplice operazione). Se invece ci interessa trovare

l'intervallo di scala attuale in relazione ad un dato livello di probabilità definito da valori di z , trasformeremo al z in x mediante la relazione inversa:

$$x = (z \cdot s) + m$$

E' importante ricordare alcuni **valori critici** della distribuzione normale standardizzata:

- a) l'intervallo media ± 1 deviazione standard comprende il 68% dei dati, essendo:
media - 1 deviazione standard = 16° percentile, $16 = (100 - 68)/2$
media + 1 deviazione standard = 84° percentile, $84 = 100 - (100 - 68)/2$
- b) l'intervallo media ± 1.96 deviazioni standard comprende il 95% dei dati, essendo:
media - 1.96 deviazioni standard = 2.5° percentile
media + 1.96 deviazioni standard = 97.5° percentile
- c) per quanto la curva della distribuzione sia asintotica, nella pratica possiamo limitare il nostro interesse all'intervallo:
media ± 4 deviazioni standard comprendente il 99.99% dei dati, essendo.
media - 4 deviazioni standard < 0.1° percentile
media + 4 deviazioni standard > 99.9° percentile

Il valore critico 1.96, nei calcoli che non esigono troppa accuratezza, può essere approssimato a 2. Bisogna comunque tenere conto che 2 non significa semplicemente 'il doppio' ma proviene dal quel 1.96 della distribuzione normale.

Esercizi

Parametri di dispersione

		x		x-m		(x-m) ²
somma	Σx		$\Sigma(x-m) =$		devianza	$S = \Sigma(x-m)^2 =$
numerosità	n					
media	$\Sigma x/n = m =$					
gradi di libertà GDL = n - 1 =						

$$\text{varianza} = s^2 = \frac{S}{\text{GDL}} = \frac{\Sigma(x-m)^2}{n-1} =$$

$$\text{deviazione standard} = s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x-m)^2}{n-1}} =$$

$$\text{errore standard} = s_m = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{\Sigma(x-m)^2}{n-1}}}{\sqrt{n}} =$$

Standardizzazione

	x	$z = \frac{x-m}{s}$
m =		0
s =		1

Errore standard

<table><tr><td>4</td></tr><tr><td>5</td></tr><tr><td>6</td></tr><tr><td>4</td></tr><tr><td>6</td></tr><tr><td>7</td></tr><tr><td>5</td></tr><tr><td>3</td></tr></table>	4	5	6	4	6	7	5	3	sottocampione	media = 5.000									
4																			
5																			
6																			
4																			
6																			
7																			
5																			
3																			
<table><tr><td>5</td></tr><tr><td>5</td></tr><tr><td>6</td></tr><tr><td>5</td></tr><tr><td>3</td></tr><tr><td>7</td></tr><tr><td>5</td></tr><tr><td>7</td></tr></table>	5	5	6	5	3	7	5	7	sottocampione	media = 5.375									
5																			
5																			
6																			
5																			
3																			
7																			
5																			
7																			
<table><tr><td>4</td></tr><tr><td>4</td></tr><tr><td>6</td></tr><tr><td>5</td></tr><tr><td>4</td></tr><tr><td>7</td></tr><tr><td>4</td></tr><tr><td>6</td></tr></table>	4	4	6	5	4	7	4	6	sottocampione	media = 5.000									
4																			
4																			
6																			
5																			
4																			
7																			
4																			
6																			
<table><tr><td>6</td></tr><tr><td>5</td></tr><tr><td>4</td></tr><tr><td>5</td></tr><tr><td>6</td></tr><tr><td>4</td></tr><tr><td>5</td></tr><tr><td>6</td></tr></table>	6	5	4	5	6	4	5	6	sottocampione	media = 5.125									
6																			
5																			
4																			
5																			
6																			
4																			
5																			
6																			
<table><tr><td>5</td></tr><tr><td>4</td></tr><tr><td>6</td></tr><tr><td>5</td></tr><tr><td>6</td></tr><tr><td>5</td></tr><tr><td>5</td></tr><tr><td>6</td></tr></table>	5	4	6	5	6	5	5	6	sottocampione	media = 5.250									
5																			
4																			
6																			
5																			
6																			
5																			
5																			
6																			
<table><tr><td colspan="2">dai 40 dati del campione</td></tr><tr><td>n =</td><td>40</td></tr><tr><td>m =</td><td>5.15</td></tr><tr><td>s =</td><td>1.051</td></tr><tr><td>s_m =</td><td></td></tr></table>	dai 40 dati del campione		n =	40	m =	5.15	s =	1.051	s _m =		<table><tr><td colspan="2">dalle 5 medie dei sottocampioni</td></tr><tr><td>n_m =</td><td></td></tr><tr><td>m_m =</td><td></td></tr><tr><td>s_m =</td><td></td></tr></table>	dalle 5 medie dei sottocampioni		n _m =		m _m =		s _m =	
dai 40 dati del campione																			
n =	40																		
m =	5.15																		
s =	1.051																		
s _m =																			
dalle 5 medie dei sottocampioni																			
n _m =																			
m _m =																			
s _m =																			