www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

STATISTICHE DI SINTESI

Luca La Rocca
luca.larocca@unimore.it

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Una statistica di sintesi è una funzione dei dati che ne riassume un aspetto saliente.

Consideriamo, per esempio, i primi 12 paesi per valore dell'export emiliano-romagnolo (Regione Emilia Romagna, 2006):

Paese	Export (10^6 euro)	Paese	Export (10^6 euro)
Germania	4390	Fed. Russa	1021
Francia	4322	Belgio	948
Stati Uniti	4066	Paesi Bassi	934
Spagna	2561	Austria	840
Regno Unito	2396	Grecia	807
Svizzera	1068	Giappone	710

L'export totale (per i dodici paesi in questione) è pari a

$$\begin{aligned} 4390 + 4322 + 4066 + 2561 + 2396 + 1068 \quad &+ \\ 1021 + 948 + 934 + 840 + 807 + 710 \quad &= \quad 24063 \end{aligned}$$

milioni di euro ed è un esempio di statistica di sintesi.

Anche l'export massimo e l'export minimo (per i dodici paesi in questione) sono statistiche di sintesi e valgono rispettivamente 4390 e 710 milioni di euro.

Si noti come la parola "statistica" denoti sia la disciplina nel suo complesso (nel qual caso si scrive talvolta con l'iniziale maiuscola) sia una qualsiasi funzione dei dati.

L'export verso il Nord America non è una statistica (per i dati in questione) perché dipende dall'export verso il Canada (che non fa parte dei dati).

Se la popolazione di interesse è formata da tutti i paesi verso i quali l'Emilia Romagna esporta (popolazione della quale i 12 paesi considerati costituiscono un campione) l'export verso il Nord America è un parametro, vale a dire un (utile) riassunto dell'intera popolazione.

Sulla base dei dati disponibili, si può inferire che l'export verso il Nord America è almeno pari a 4066 milioni di euro (export verso gli Stati Uniti) e che l'export verso il Canada è inferiore a 710 milioni di euro, concludendo che l'export verso il Nord America potrebbe essere circa pari a 4400 euro.

Un indice di posizione è un valore che rappresenta, in un qualche senso, l'insieme dei valori assunti da un carattere in un collettivo.

Un indice di variabilità è un valore che misura la tendenza di un carattere ad assumere modalità diverse su unità diverse di un collettivo.

Se il collettivo in questione è la popolazione di interesse, avremo un parametro, se il collettivo è un campione, avremo una statistica. . .

. . . nel seguito, collocandoci nell'ambito della statistica descrittiva, supporremo che la popolazione coincida col campione e quindi ci disinteresseremo di questa distinzione, concentrandoci sui divesi modi in cui si possono riassumere i dati.

Introduzione

Indici di posizione

- Media aritmetica
- Media geometrica
- Media secondo Chisini
- Media troncata
- Mediana e altri percentili
- Moda

Indici di variabilità

A cosa può servire un indice di posizione?

Per esempio a confrontare due gruppi di unità statistiche e stabilire se un carattere si manifesti “tendenzialmente allo stesso livello” nei due gruppi: la variabilità del carattere, specie nel caso di elevate numerosità campionarie, rende il confronto tutt’altro che immediato (si pensi al caso in cui si vogliano confrontare gli scontrini emessi in due giorni diversi da un grande centro commerciale).

La seguente tabella riporta i valori di un indice di attività economica femminile (occupazione femminile come percentuale dell’occupazione maschile) per alcuni paesi dell’Europa Occidentale e Orientale, nel 1994. . .

Eastern Europe		Western Europe	
Country	Activity	Country	Activity
Bulgaria	88	Austria	60
Czech Republic	84	Belgium	47
Hungary	70	Denmark	77
Poland	77	France	64
Romania	77	Ireland	41
Slovakia	81	Italy	44
		Netherlands	42
		Norway	68
		Portugal	51
		Spain	31
		Sweden	77
		Switzerland	60
		United Kingdom	60
	477		722

Human Development Report 1995, United Nations Development Programme (Agresti & Finlay, 1997, Example 3.4). Dati non disponibili per la Germania.

. . . sembra evidente che “tendenzialmente” le donne siano più attive, da un punto di vita economico, nell’Europa Orientale (in questo senso i totali riportati in calce alla tabella sono fuorvianti) ma non è altrettanto chiaro l’ammontare di questa differenza.

Un altro esempio (Borra & Di Ciaccio, 2008, Esempio 3.2.1) può essere il confronto tra i tempi di percorrenza con due diversi mezzi di trasporto su uno stesso tragitto, avendo preso nota del tempo impiegato in un campione di giorni (es. 12 giorni metro, 12 giorni auto).

Oppure potremmo dovere decidere quanta merce (pane, vestiti, . . .) tenere in negozio sulla base della domanda registrata nei giorni passati: questo ci costringe a riassumere diverse osservazioni mediante un singolo valore.

La media aritmetica di un carattere quantitativo X è definita a partire dalla sua distribuzione unitaria $x_1, x_2, \ldots, x_{n-1}, x_n$ come

$$m(X) = \bar{x} = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

dove n è la numerosità campionaria e $x_1, x_2, \ldots, x_{n-1}, x_n$ sono modalità non necessariamente tutte distinte (né ordinate).

Nell'esempio sull'export emiliano-romagnolo si ha $n = 12$ e $x_1 = 4390$, $x_2 = 4322, \ldots, x_{n-1} = x_{11} = 807, x_n = x_{12} = 710$, di modo che

$$\bar{x} = \frac{\text{export totale}}{n} = \frac{24063}{12} = 2005.25 \quad \text{milioni di euro.}$$

Nell'esempio sull'attività economica femminile in Europa, denotando con X l'occupazione nell'Europa Orientale e con Y l'occupazione nell'Europa Occidentale, troviamo

$$\bar{x} = \frac{477}{6} = 79.5$$
$$\bar{y} = \frac{722}{13} \simeq 55.5$$

a conferma della prima impressione di una maggiore attività economica femminile in Europa Orientale (ora quantificata in 24 punti percentuali dell'occupazione maschile); per completezza vale la pena osservare che, sempre nel 1994, negli Stati Uniti e in Canada l'indice di attività economica femminile valeva rispettivamente 65 e 63 punti percentuali.

Nell'esempio sul confronto dei tempi di percorrenza con diversi mezzi di trasporto su uno stesso tragitto (Borra & Di Ciaccio, 2008, Esempio 3.2.1) si trova che mediamente (nel senso della media aritmetica) ci vogliono 4 minuti in meno con la metro.

In questo caso, ai fini pratici, occorre stabilire se la differenza osservata possa estendersi o meno ai giorni a venire (la differenza osservata potrebbe essere stata frutto del caso): si tratta di un tipico problema inferenziale che trova una soluzione nell'ambito della verifica di ipotesi (o della stima per intervalli).

Quando si calcolano le medie (aritmetiche) di un carattere in due o più gruppi si parla di medie di gruppo o medie condizionate (al fatto di appartenere a ciascun gruppo).

Se vogliamo calcolare la media aritmetica nell'unione di due gruppi, possiamo farlo a partire dalle medie (e dalle numerosità) di gruppo:

$$m_{globale} = \frac{n_a \times a + n_b \times b}{n_a + n_b}$$

dove a è la media aritmetica nel primo gruppo, avente numerosità n_a, e b è la media aritmetica nel secondo gruppo, avente numerosità n_b; nell'esempio sull'attività economica femminile in Europa si trova

$$m_{globale} = \frac{6 \times 79.5 + 13 \times 55.5}{6 + 13} \simeq 63.1$$

e infatti $(477 + 722)/19 \simeq 63.1$ (sempre alla prima cifra decimale).

Analogamente possiamo calcolare la media aritmetica nell'unione di tre o più gruppi:

$$m_{globale} = \frac{n_1 \bar{x}_1 + \cdots + n_k \bar{x}_k}{n_1 + \cdots + n_k}$$

dove k è il numero di gruppi, mentre n_i e $\bar{x}_i$ sono rispettivamente la numerosità campionaria e la media aritmetica dell'i-esimo gruppo, $i = 1, \ldots, k$.

Interessa in particolare il caso in cui i gruppi provengano da una suddivisione in classi, ovvero per calcolare la media aritmetica a partire da una distribuzione di frequenza...

Age	Freq.
$0 \vdash 30$	4
$30 \vdash 60$	2
$60 \vdash \infty$	3
Total	9

In questo caso le medie di gruppo non sono disponibili, ma possono essere approssimate dai valori centrali delle classi (questo corrisponde a supporre che le unità statistiche siano distribuite uniformemente all'interno di ogni classe); a tal fine sostituiamo ∞ con 100 anni (un limite ragionevole, anche se non assoluto, per l'età di una persona)...

i	Age class	n_i	$\bar{x}_i$	$n_i \times \bar{x}_i$				
1	$0 \vdash 30$	4	15	60				
2	$30 \vdash 60$	2	45	90				
3	$60 \vdash 100$	3	80	240				
	Total	9		390	/	9	=	43.3

In questo modo si ottiene un'approssimazione della media aritmetica di Age: si tratta di una buona approssimazione, nel caso specifico, in quanto il valore esatto è 43.7 anni, come si può verificare partendo direttamente dalla distribuzione unitaria riportata da Everitt (2005, p. 2) ed eliminando il dato mancante.

Vediamo un altro esempio:

Numero di Figli	%
Nessuno	27.60
Uno	16.80
Due	24.70
Tre	14.20
Quattro	8.40
Cinque	3.60
Sei	1.60
Sette	1.50
Otto o Più	1.10
Non Risponde	0.50
Totale	100.00

Fonte: General Social Survey 1991 (Bohrnstedt & Knoke, 1998).

Questa volta preliminarmente occorre:

- eliminare i non rispondenti;
- assegnare un valore convenzionale alla modalità Otto o Più (le altre modalità individuano univocamente la propria "media di classe").

Poiché otto figli sono già molti (per una famglia statunitense) conviene prendere 8 come valore convenzionale, accettando di ottenere una sottostima della media aritmetica; in assenza di ulteriori informazioni, l'uso di un valore diverso da 8 (maggiore di 8) produrrebbe un errore non necessariamente più piccolo e comunque di segno incognito (mentre così almeno sappiamo che si tratta di una sottostima).

i	Numero di Figli	$\bar{x}_i$	p_i^*	$p_i^* \times \bar{x}_i$				
1	Nessuno	0	27.60	0.00				
2	Uno	1	16.80	16.80				
3	Due	2	24.70	49.40				
4	Tre	3	14.20	42.60				
5	Quattro	4	8.40	33.60				
6	Cinque	5	3.60	18.00				
7	Sei	6	1.60	9.60				
8	Sette	7	1.50	10.50				
9	Otto o più	8	1.10	8.88				
	Totale		99.50	189.30	/	99.50	=	1.9030

Si noti come il contributo della modalità Otto o Più alla media aritmetica sia residuale e di conseguenza la scelta del valore convenzionale per tale modalità non sia critica.

In generale non va sempre bene come negli esempi visti, tuttavia è il meglio che si possa fare in assenza della distribuzione unitaria.

Si noti anche che nel caso in cui la distribuzione di frequenza sia data in termini di frequenze relative (che sommino a uno) la divisione finale per la numerosità campionaria è superflua.

Abbiamo visto che il calcolo della media aritmetica a partire da una distribuzione di frequenza può vedersi come un caso particolare della combinazione di più medie di gruppo in una singola media globale.

A sua volta la combinazione di più medie di gruppo in una singola media globale può vedersi come un caso particolare di media aritmetica pesata (ottenuta prendendo come pesi le numerosità di gruppo).

Più in generale una media pesata può essere utile quando i valori osservati non abbiano tutti la stessa importanza (nell'esempio sull'attività economica femminile in Europa si potrebbe volere pesare i valori dell'indice con le corrispondenti occupazioni maschili).

In ambito universitario è una media aritmetica pesata il voto di laurea (al netto di eventuali bonus e dei punti attribuiti per la tesi): i voti degli esami sostenuti sono pesati con i corrispondenti crediti formativi.

I dati seguenti riguardano uno studente (di fantasia) in procinto di conseguire la Laurea Specialistica in Progettazione e Gestione della Comunicazione d'Impresa (proseguimento della Laurea Triennale in Comunicazione e Marketing, anche se naturalmente quest'ultima non era l'unica laurea triennale dalla quale vi si potesse accedere).

Primo Anno (Piano di Studio Standard)			
Esame	Crediti	Voto	C × V
Comunicazione finanziaria	8	30	240
Comunicazione pubblicitaria	8	22	176
Diritto della comunicazione…			
…e dell'informazione	8	23	184
Fondamenti di finanza	4	28	112
Fondamenti di strategia	4	22	88
Pianificazione e controllo	8	21	168
Risorse umane e marketing interno	8	23	184
Statistica per l'analisi di mercato	8	19	152
Strategie e politiche d'impresa	4	25	100
Totale	60		1404

Secondo Anno (Piano di Studio Standard)			
Esame	Crediti	Voto	C × V
Comunicazione e società	4	25	100
Costruzione d'immagine	8	23	184
Lingua inglese II	4	28	112
Organizzazione per progetti	4	21	84
Pubbliche relazioni	8	27	216
Strategia e gestione delle... ...relazioni di canale	4	24	96
Strategia e gestione delle... ...relazioni tra imprese	4	24	96
Totale	36		888

Pertanto il voto di partenza per l'esame di laurea sarà pari a

$$m_{30} = \frac{1404 + 888}{60 + 36} = \frac{2292}{96} = 23.875$$

trentesimi, ovvero pari a

$$m_{110} = 110 \times \frac{23.875}{30} = 87.542$$

centodecimi, da arrotondare a 88 / 110.

Se può interessare, i voti usati in questo esempio sono stati scelti "a caso" tra 18 e 30 trentesimi e danno luogo a una media aritmetica non pesata pari a 24.063 trentesimi (88.229 centodecimi).

In che senso la media aritmetica rappresenta l'insieme dei valori assunti da un carattere in un collettivo?

Preliminarmente allo studio di altri indici di posizione, vale la pena soffermarsi sulle seguenti proprietà della media aritmetica:

- la media aritmetica è il baricentro dei dati;
- la media aritmetica minimizza lo scarto quadratico medio;
- la media aritmetica conserva il totale.

Procediamo con ordine...

Si dicono scarti delle osservazioni $x_1, \ldots, x_n$ dal valore a le differenze

$$x_1 - a \quad \ldots \quad x_n - a$$

e in particolare ($a = \bar{x}$) interessano gli scarti dalla media aritmetica

$$x_1 - \bar{x} \quad \ldots \quad x_n - \bar{x}$$

perché la media aritmetica è il (solo) valore che annulla la somma degli scarti:

$$(x_1 - \bar{x}) + \cdots + (x_n - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

Female Economic Activity in Eastern Europe ($\bar{x} = 79.5$)

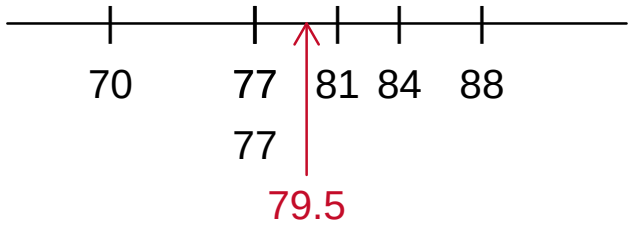
i	Country	x_i	$x_i - \bar{x}$
1	Bulgaria	88	$+ 8.5$
2	Czech Republic	84	$+ 4.5$
3	Hungary	70	$- 9.5$
4	Poland	77	$- 2.5$
4	Romania	77	$- 2.5$
6	Slovakia	81	$+ 1.5$
	Total	477	0.0

La media aritmetica, in quanto valore che annulla la somma degli scarti, è il baricentro dei dati:

> *se rappresentiamo concretamente la retta reale con un'asta sottile e applichiamo dei pesi di massa unitaria in corrispondenza dei punti di ascissa $x_1, \ldots, x_n$, la media aritmetica $\bar{x}$ è l'ascissa del punto dove appoggiare l'asta affinché sia in equilibrio.*

Il grafico seguente illustra questa proprietà della media aritmetica con riferimento ai dati della tabella precedente.

Baricentro



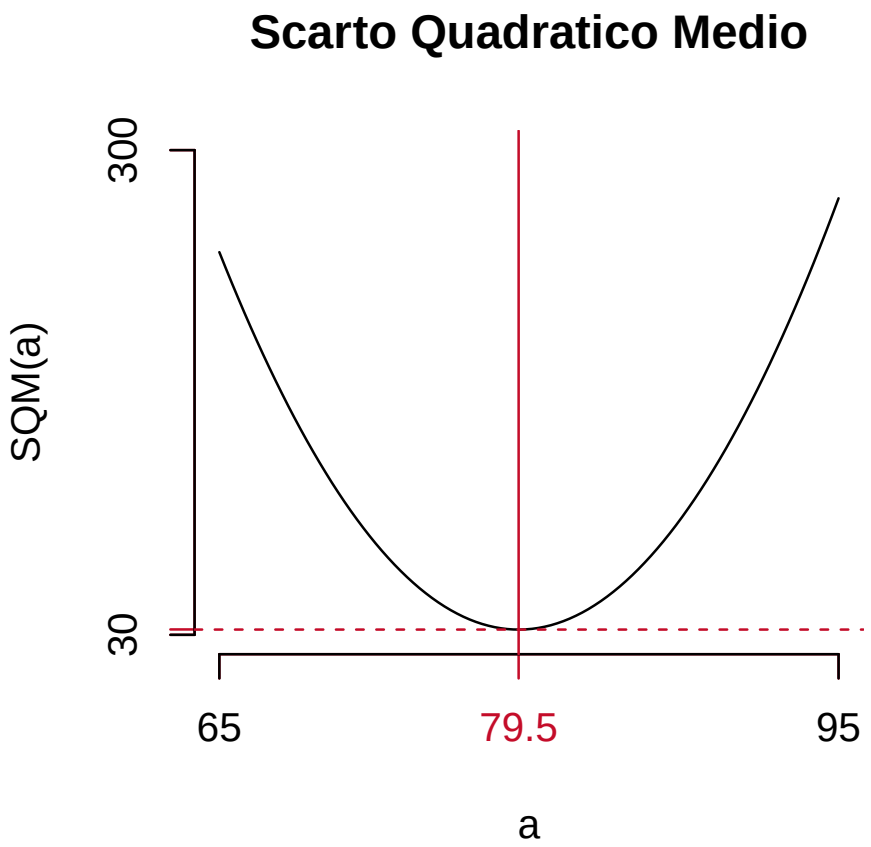
Si dicono scarti quadratici delle osservazioni $x_1, \ldots, x_n$ dal valore a le quantità

$$(x_1 - a)^2 \quad \ldots \quad (x_n - a)^2.$$

L'errore che si commette rappresentando $x_1, \ldots, x_n$ con a può misurarsi mediante lo scarto quadratico medio:

$$SQM(a) = \frac{(x_1 - a)^2 + \cdots + (x_n - a)^2}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2.$$

La media aritmetica è quel(l'unico) valore $a = \bar{x}$ che rende minimo lo scarto quadratico medio e dunque è la migliore rappresentazione possibile dei dati secondo questo criterio (funzione di perdita).


Scarto Quadratico Medio
300
30
SQM(a)
65
79.5
95
a

La media aritmetica può anche definirsi come quel(l'unico) valore $\bar{x}$ che, sostituito alle osservazioni $x_1, \ldots, x_n$, ne conserva il totale:

$$\underbrace{\bar{x} + \cdots + \bar{x}}_{n \text{ volte}} = n\bar{x} = x_1 + \cdots + x_n.$$

Nell'esempio dell'export emiliano-romagnolo è il valore dell'export per paese che si otterrebbe se l'export fosse ripartito uniformemente.

Nell'esempio del numero di figli per famiglia statunitense sono i figli che ogni famiglia manterrebbe se il mantenimento fosse su base collettiva.

Abbiamo visto che la media aritmetica conserva il totale, ma non è sempre il totale che interessa conservare...

...consideriamo per esempio la dinamica di un capitale

$$C_i = C_{i-1} \times (1 + r_i)$$

i cui interessi maturino ai tempi t_i ($i = 1, \ldots n$) dove:

- r_i è il tasso di interesse nel periodo tra t_{i-1} e t_i;
- C_0 è il capitale iniziale (al tempo t_0).

Il capitale finale, al tempo t_n, varrà

$$C_n = C_0 \times (1 + r_1) \times (1 + r_2) \times \cdots \times (1 + r_{n-1}) \times (1 + r_n).$$

Il seguente esempio (di fantasia, ma vedi anche Borra & Di Ciaccio, 2008, Esempio 3.3.1) illustra la situazione:

i	t_i (anno)	r_i	$1 + r_i$	C_i (euro a fine anno)
0	2006	–	–	10000.00
1	2007	4.2%	1.042	10420.00
2	2008	5.3%	1.053	10972.26
3	2009	5.1%	1.051	11531.85
4	2010	7.8%	1.078	12431.33
			1.243	

Il tasso di interesse complessivo (quadriennale) è pari al 24.3% e si calcola mediante un prodotto (non mediante una somma).

Quindi, se si vuole che il tasso di interesse medio $r_\star$, sostituito ai tassi di interesse $r_1, \ldots, r_n$, ne conservi gli effetti, si deve definirlo in modo che

$$\underbrace{(1 + r_\star) \times \cdots \times (1 + r_\star)}_{n \text{ volte}} = (1 + r_\star)^n = (1 + r_1) \times \cdots \times (1 + r_n)$$

ovvero $1 + r_\star$ deve essere la media geometrica di $1 + r_1, \ldots, 1 + r_n$:

$$1 + r_\star = \sqrt[n]{(1 + r_1) \times \cdots \times (1 + r_n)}.$$

Nell'esempio si trova $r_\star = 5.59\%$, laddove la media aritmetica di $r_1, \ldots, r_4$ vale 5.60%...

. . . una differenza piccola, senza dubbio, ma sistematica e crescente con il tempo (l'ultima riga riporta il tasso di interesse complessivo):

	Media Geometrica		Media Aritmetica	
Anno	Tasso	Capitale	Tasso	Capitale
2006	–	10000.00	–	10000.00
2007	5.59%	10559.00	5.60%	10560.00
2008	5.59%	11149.25	5.60%	11151.36
2009	5.59%	11772.49	5.60%	11775.84
2010	5.59%	12430.57	5.60%	12435.28
	24.31%		24.35%	

In pratica per calcolare la media geometrica conviene "passare ai logaritmi", in modo da ricondursi al calcolo di una media aritmetica:

$$\log(1 + r_\star) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + r_i) = \overline{\log(1 + r)}$$

dove $log(x)$ è definito da $10^{\log(x)} = x$ (lavorando in base dieci); es. $\log(1) = 0, \log(10) = 1, \log(100) = 2, \log(200) \simeq 2.303$.

Si troverà allora $1 + r_\star = 10^{\overline{\log(1+r)}} \ldots$

i	$1 + r_i$	$\log(1 + r_i)$
1	1.042	0.01786772
2	1.053	0.02160272
3	1.051	0.02242837
4	1.078	0.03261876
		0.09451757

$$\overline{\log(1+r)} = \frac{0.09451757}{4} = 0.02362939$$

$$1 + r_\star = 10^{0.02362939} = 1.055916$$

. . . e quindi $r_\star = 5.59\%$ (come già visto).

Nell'esempio è stato possibile calcolare la media geometrica sia dalla definizione che passando ai logaritmi perché

1. la numerosità campionaria era esigua ($n = 4$) e
2. i valori da mediare erano prossimi all'unità,

mentre in generale il passaggio ai logaritimi è praticamente inevitabile.

Il punto 2. spiega anche perché la differenza tra $r_\star$ e $\bar{r}$ fosse piccola:

$$\log(1 + r) \simeq r,$$

se r è molto più piccolo di 1, di modo che $\log(1 + r_\star) = \overline{\log(1 + r)}$ diventa $r_\star \simeq \bar{r}$.

Media aritmetica o media geometrica?

È prassi comune usare di default la media aritmetica per riassumere un carattere quantitativo e infatti con il termine “media” (se non qualificato) si intende proprio la media aritmetica.

Tuttavia la media geometrica si è dimostrata più adatta nello studio della dinamica di un capitale. . .

. . . e d’altra parte il tasso di interesse medio che “conserva gli effetti” di un certo numero di investimenti contemporanei è una media aritmetica (pesata con i capitali).

Il punto è che in generale la media è una nozione che dipende dal contesto. . .

Una definizione generale di media è stata data dal matematico italiano Oscar Chisini (Bergamo 1889, Milano 1967)

http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Chisini.html

secondo il quale la media delle osservazioni $x_1, \ldots, x_n$ è definita, rispetto a una loro funzione f, come quel valore $\bar{x}$ che la conserva:

$$f(\bar{x}, \ldots, \bar{x}) = f(x_1, \ldots, x_n)$$

In particolare:

- se f è la somma, si ottiene la media aritmetica;
- se f è il prodotto, si ottiene la media geometrica;
- ...

Se percorrete

- prima 30 Km a 60 Km/h
- poi altri 30 Km a 120 Km/h

quale sarà la vostra velocità media?

Volete sostituire $v_1 = 60$ Km/h e $v_2 = 120$ Km/h con un unico valore $\tilde{v}$ che ne conservi gli effetti, ovvero che conservi il tempo impiegato per percorrere i 60 Km, sapendo che la relazione fra spazio, velocità e tempo è

$$s = v \times t.$$

Dunque volete conservare la funzione

$$f(v_1, v_2) = \frac{s_1}{v_1} + \frac{s_2}{v_2} = \frac{s_1}{\tilde{v}} + \frac{s_2}{\tilde{v}} = f(\tilde{v}, \tilde{v})$$

e troverete come media la media armonica

$$\tilde{v} = \frac{s_1 + s_2}{\frac{s_1}{v_1} + \frac{s_2}{v_2}} = \frac{60}{\frac{1}{2} + \frac{1}{4}} = 80 \text{ Km/h}$$

laddove la media aritmetica vale 90 Km/h.

Quest'ultima è invece la media di interesse (perché conserva lo spazio) se viaggiate per metà del tempo a 60 Km/h e per l'altra metà del tempo a 120 Km/h.

Un problema con la media aritmetica è la sua sensibilità ai valori anomali: un singolo valore molto grande (piccolo) può determinare, al limite quasi da solo, il valore della media aritmetica.

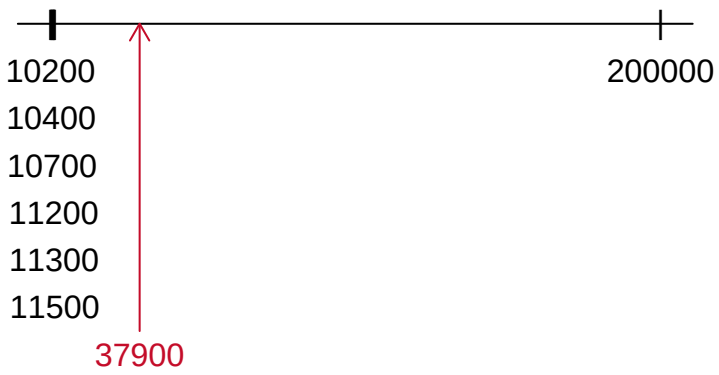
Se per esempio (Agresti & Finlay, 1997, Example 3.5) il proprietario di un negozio afferma che lo stipendio medio dei suoi dipendenti è pari a 37900 dollari, gli stipendi potrebbero essere

10200 10400 10700 11200 11300 11500 e 200000

dollari, dove l'ultimo stipendio è lo stipendio del figlio...

...in questo caso la media aritmetica non è una buona sintesi del collettivo.

Media determinata da un valore anomalo



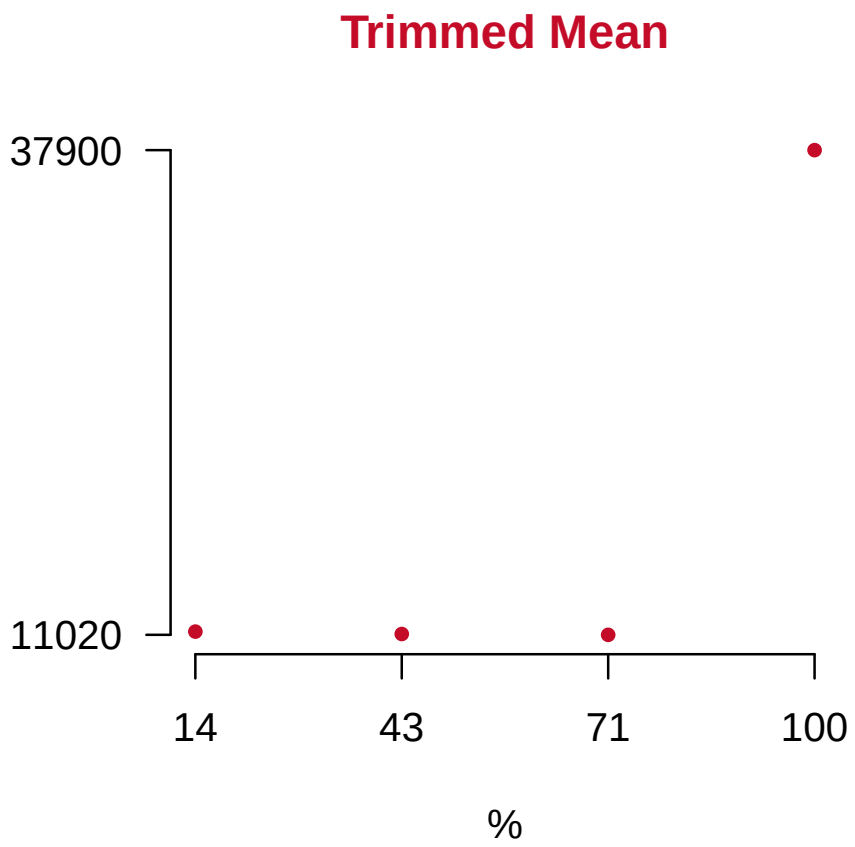
Una soluzione al problema dei valori anomali è la media troncata (trimmed mean): si scartano le osservazioni più grandi e più piccole, in una percentuale prefissata, per poi calcolare la media sulle osservazioni rimanenti.

Nell'esempio sugli stipendi la media troncata al 50% (calcolata sul 50%, circa, delle osservazioni centrali) vale

$$\frac{10700 + 11200 + 11300}{3} = \frac{33200}{3} = 11066.67$$

dollari ed è una buona sintesi degli stipendi pagati.

In alternativa, per non ridurre la numerosità campionaria più del necessario, si possono eliminare i valori anomali (una volta individuati).


Trimmed Mean
37900
11020
14
43
71
100
%

La mediana di un carattere quantitativo può vedersi come caso limite della media troncata, quando si scartino tutte le osservazioni tranne quella centrale (se le osservazioni sono in numero dispari) o tranne le due centrali (se le osservazioni sono in numero pari).

Nell'esempio sugli stipendi le osservazioni sono in numero dispari e la mediana è l'osservazione centrale:

10200 10400 10700 11200 11300 11500 200000.

Si tratta dell'unico valore che lascia alla sua sinistra e alla sua destra (incluso il valore stesso) almeno il 50% delle osservazioni: 4 su 7.

Consideriamo ora (Borra & Di Ciaccio, 2008, Esempio 3.5.2) il numero di capi venduti in un giorno da 6 negozi di abbigliamento:

15 20 11 18 27 6.

In questo caso, per calcolare la mediana, dobbiamo innanzi tutto ordinare le modalità osservate dalla più piccola alla più grande (nell'esempio precedente non era stato necessario compiere questa operazione perché le modalità erano già in ordine):

6 11 15 18 20 27.

Dopodiché, essendo le osservazioni in numero pari, calcoleremo la mediana come media aritmetica delle due osservazioni centrali...

$$6 \quad 11 \quad 15 \quad 18 \quad 20 \quad 27$$

...ottenendo il valore $(15 + 18)/2 = 16.5$.

In questo caso qualsiasi valore tra 15 e 18 lascia alla sua sinistra e alla sua destra (incluso il valore stesso) almeno il 50% delle osservazioni: 3 su 6 (4 su 6 a destra di 15 e a sinistra di 18); l'uso della media delle due osservazioni centrali (qui 15 e 18) è un'utile convenzione per individuare univocamente la mediana.

Ricapitolando, il calcolo della mediana consiste di due passi:

1. ordinare le modalità osservate;
2. individuare una modalità centrale.

Se la numerosità campionaria è dispari, la scelta al passo 2. è obbligata; se invece la numerosità campionaria è pari, si ricorre a una convenzione (media aritmetica delle due modalità centrali).

L'unico momento in cui si opera algebricamente sui dati è nella fase convenzionale (quando se ne presenta la necessità); pertanto, se si adotta una convenzione diversa, si può calcolare la mediana di un carattere qualitativo, purché ordinato...

. . . la mediana per un carattere qualitativo misurato su scala ordinale è definita come

> *la più piccola (grande) modalità che lascia sia alla sua sinistra che alla sua destra, incluso la modalità stessa, almeno il 50% delle osservazioni.*

Per esempio, le modalità osservate del carattere Health di Everitt (2005, p. 2) sono:

> *Very Good, Very Good, Average, Very Poor, Good, Good, Very Good, Average, Average, Good.*

Elencando le modalità osservate in ordine crescente,

Very Poor, Average, Average, Average, Good,
Good, Good, Very Good, Very Good, Very Good,

troviamo che (qualunque sia la convenzione adottata) la mediana è

Good.

In pratica, con molte più osservazioni che modalità distinte, raramente la convenzione (cui pure senz'altro ci atterremo) giocherà un ruolo.

La mediana si può anche calcolare a partire dalla distribuzione di frequenza cumulata relativa o percentuale: sarà la più piccola modalità con frequenza cumulata almeno pari al 50%...

Highest degree completed for a sample of Americans (Agresti & Finlay, 1997, p. 49): finding the median response.

Degree	Freq.	Rel.	Cum.
Not a high school graduate	38012	0.2140	0.2140
High school only	65291	0.3676	0.5816
Some college, no degree	33191	0.1869	0.7685
Associate's degree	7570	0.0426	0.8111
Bachelor's degree	22845	0.1286	0.9397
Master's degree	7599	0.0428	0.9825
Doctorate or professional	3110	0.0175	1.0000
Total	177618	1.0000	

Così come è robusta rispetto alla presenza di valori anomali, in quanto dipende solo da una o due modalità centrali, la mediana è insensibile alla presenza di valori rari: nell'esempio seguente (GSS 1991: Agresti & Finlay, 1997, p. 52) la media $(8.8 + 3.4 + 2.1 + 0.4)/100.1 = 0.147$ è più informativa della mediana 0 (es. confronto con un altro paese).

Number of people you know who have committed suicide	%
0	88.8
1	8.8
2	1.7
3	0.7
4	0.1
Total	100.1

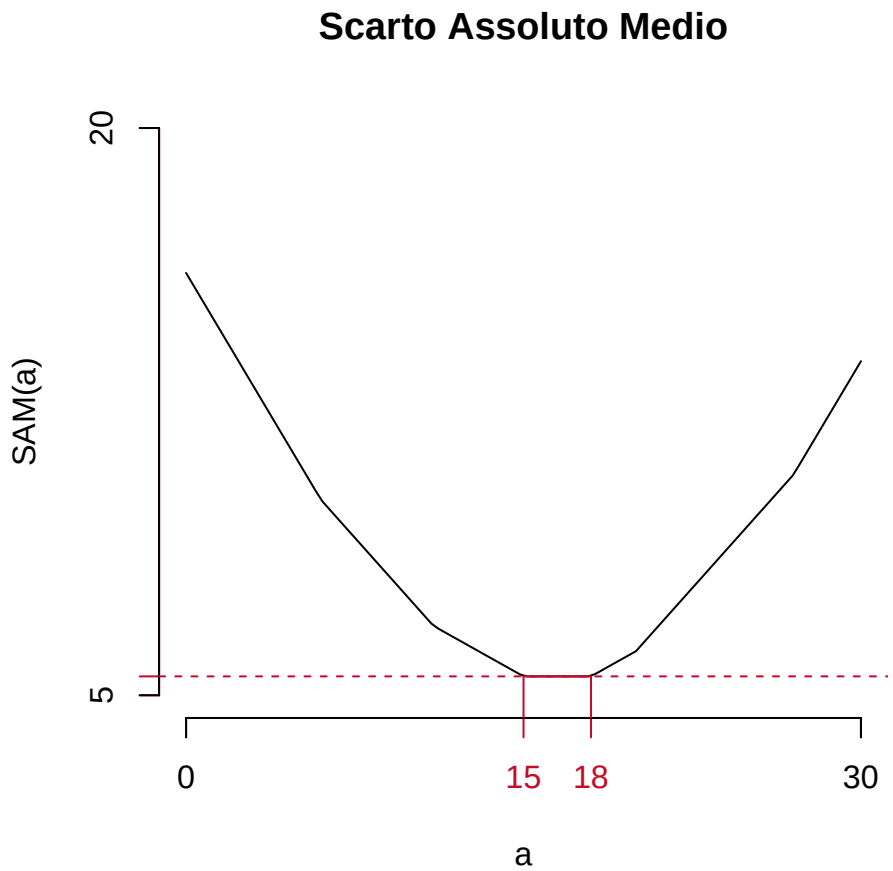
La mediana di un carattere quantitativo minimizza lo scarto assoluto medio (e dunque ne è la migliore rappresentazione possibile secondo questo criterio): gli scarti assoluti delle osservazioni $x_1, \ldots, x_n$ dal valore a sono le quantità

$$|x_1 - a| \quad \ldots \quad |x_n - a|$$

e lo scarto assoluto medio

$$SAM(a) = \frac{|x_1 - a| + \cdots + |x_n - a|}{n} = \frac{1}{n} \sum_{i=1}^{n} |x_i - a|$$

è una misura alternativa (allo scarto quadratico medio) dell'errore che si commette rappresentando $x_1, \ldots, x_n$ con a.



(dati sui capi di abbigliamento venduti)

Una generalizzazione della mediana è il k-esimo percentile:

> *un valore $x_{k\%}$ che lascia alla sua sinistra almeno il $k\%$ e alla sua destra almeno il $(100 - k)\%$ delle osservazioni, dove sinistra e destra includono il valore stesso.*

Evidentemente la mediana è il 50-esimo percentile: $x_{50\%}$ ($k = 50$).

In che senso, per esempio, il 90-esimo percentile rappresenta la distribuzione di un carattere? Non è un valore tipico (come la media e la mediana) ma è un valore elevato:

> *un'azienda potrebbe interessarsi agli studenti il cui voto di laurea è maggiore del 90-esimo percentile (dell'università dove si sono laureati).*

Nella descrizione di una distribuzione sono tipicamente di interesse:

- il 25-esimo percentile, $x_{25\%}$, detto anche primo quartile;
- il 75-esimo percentile, $x_{75\%}$, detto anche terzo quartile.

Va da sé che il secondo quartile è la mediana $x_{50\%}$.

Anche il minimo e il massimo delle osservazioni possono vedersi come percentili: $x_{0\%}$ e $x_{100\%}$. L'insieme dei cinque valori

$$x_{0\%} \quad x_{25\%} \quad x_{50\%} \quad x_{75\%} \quad x_{100\%}$$

costituisce il cosiddetto riassunto dei cinque numeri (five number summary) del carattere X.

Consideriamo, per esempio, la variabile Weight di Everitt (2005, p. 2) e ordiniamone le modalità:

Weight (lb)	100	105	*110*	110	120	120	135	*140*	150	160
Cum. Freq.	10%	20%	*30%*	40%	50%	60%	70%	*80%*	90%	100%

Troviamo:

- $x_{25\%} = 110$ lb e $x_{75\%} = 140$ lb;
- $x_{0\%} = 100$ lb e $x_{100\%} = 160$ lb;
- $x_{50\%} = (120 + 120)/2 = 120$ lb.

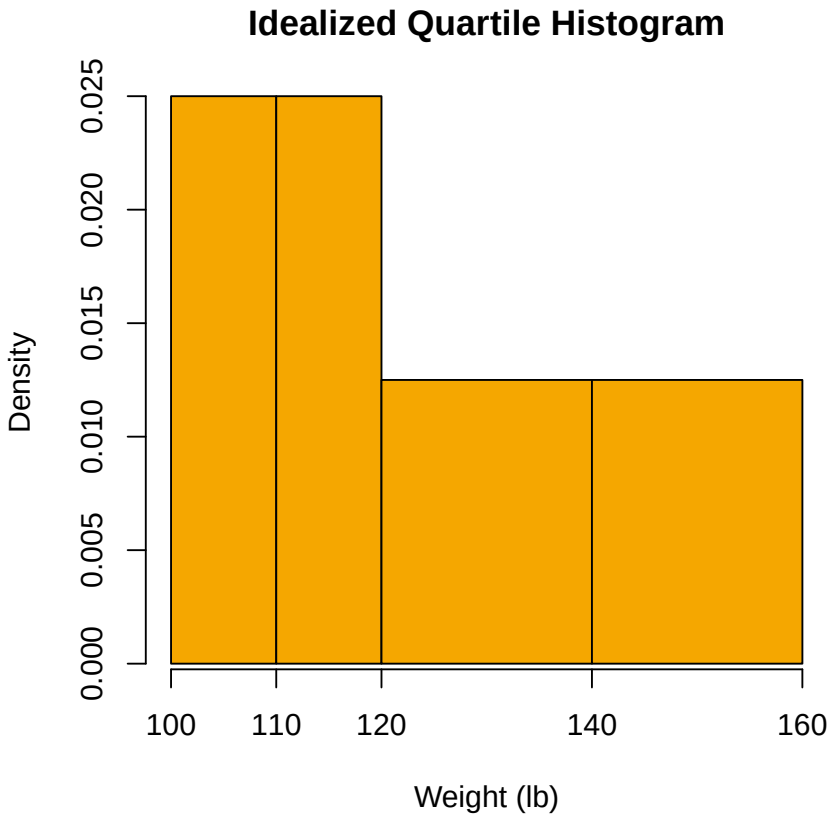
Il riassunto dei cinque numeri di Weight è pertanto

100 110 120 140 160

ed evidenzia una certa asimmetria a destra: la coda destra della distribuzione è più lunga di quella sinistra.

All'asimmetria a destra corrisponde un valore della media maggiore di quello della mediana: $m(\mathit{Weight}) = 125$ (la differenza tra media e mediana è una possibile misura dell'asimmetria. . . in alternativa al terzo momento centrato suggerito da Borra & Di Ciaccio, 2008, p. 97).

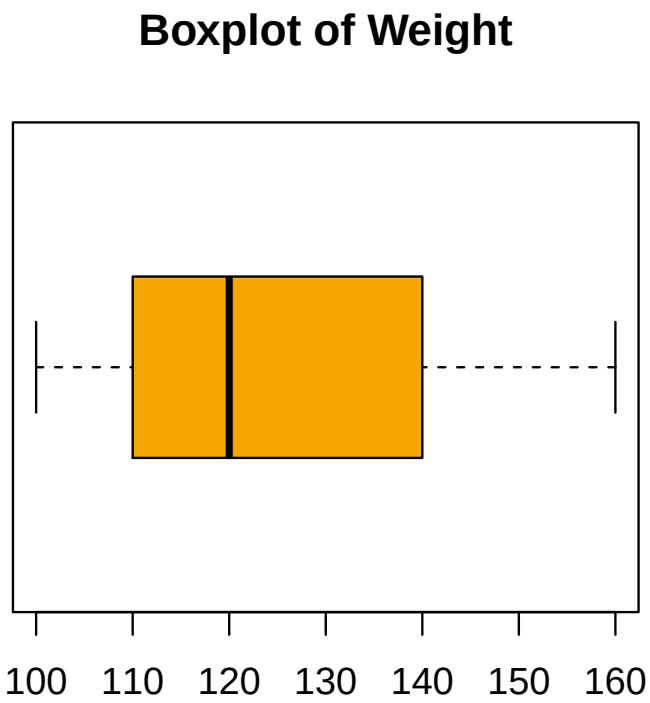
L'asimmetria di una distribuzione può essere evidenziata per via grafica con un "istogramma dei quartili idealizzato" o con un boxplot. . .



In questo istogramma

- la suddivisione è individuata dal riassunto dei cinque numeri e
- le barre hanno tutte area 25%

in modo da neutralizzare l'effetto delle osservazioni che assumono un valore di soglia (110, 120 o 140).




```
colE <- rgb(0.9609375,0.6562500,0.0000000)
boxplot(X$Weight,
        horizontal = TRUE,
        main = "Boxplot of Weight",
        col = colE)
```

Il boxplot permette di

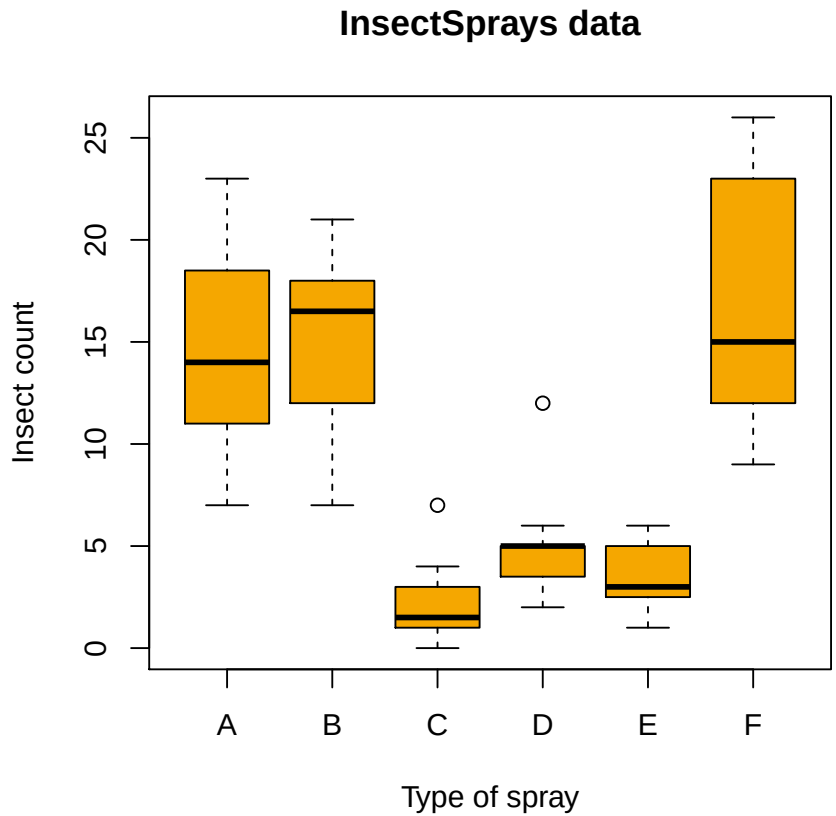
- evidenziare valori anomali (Borra & Di Ciaccio, 2008, p. 87);
- confrontare (la distribuzione di) due o più gruppi di osservazioni.

La matrice di dati `InsectSprays` è un oggetto built-in di R che contiene

> *The counts of insects in agricultural experimental units treated with different insecticides.*

come si può verificare con il comando `help(InsectSprays)`...

```
> head(InsectSprays)
  count spray
1    10     A
2     7     A
3    20     A
4    14     A
5    14     A
6    12     A
> tail(InsectSprays)
   count spray
67    13     F
68    10     F
69    26     F
70    26     F
71    24     F
72    13     F
```



. . . gli insetticidi di tipo C, D ed E appaiono funzionare meglio degli altri; in particolare l'insetticida di tipo C sembra il più efficace.

Si noti che:

- le osservazioni che distano dalla scatola più di una volta e mezza la lunghezza della scatola sono individuate come anomale e indicate con un pallino;
- i "baffi" si estendono dalla scatola sino al minimo e al massimo delle osservazioni calcolati escludendo le osservazioni anomale.

Si noti inoltre che la scatola relativa all'insetticida di tipo D è degenere, perché il terzo quartile coincide con la mediana.

La distribuzione di un carattere qualitativo misurato su scala nominale può essere riassunta dalla sua moda: la modalità che si presenta con maggiore frequenza.

Per esempio si può trovare che in Spagna la “religione modale” è quella cattolica:

Religion	%
Catholic	76.0
Irreligion	20.0
Islam	2.3
Others	1.8
Total	100.1

Source: Wikipedia (Religion in Spain) 23 Jan 2010

Se vi sono due o più modalità con la stessa frequenza, nel qual caso di parla di bimodalità o multimodalità, la moda è una sintesi poco utile.

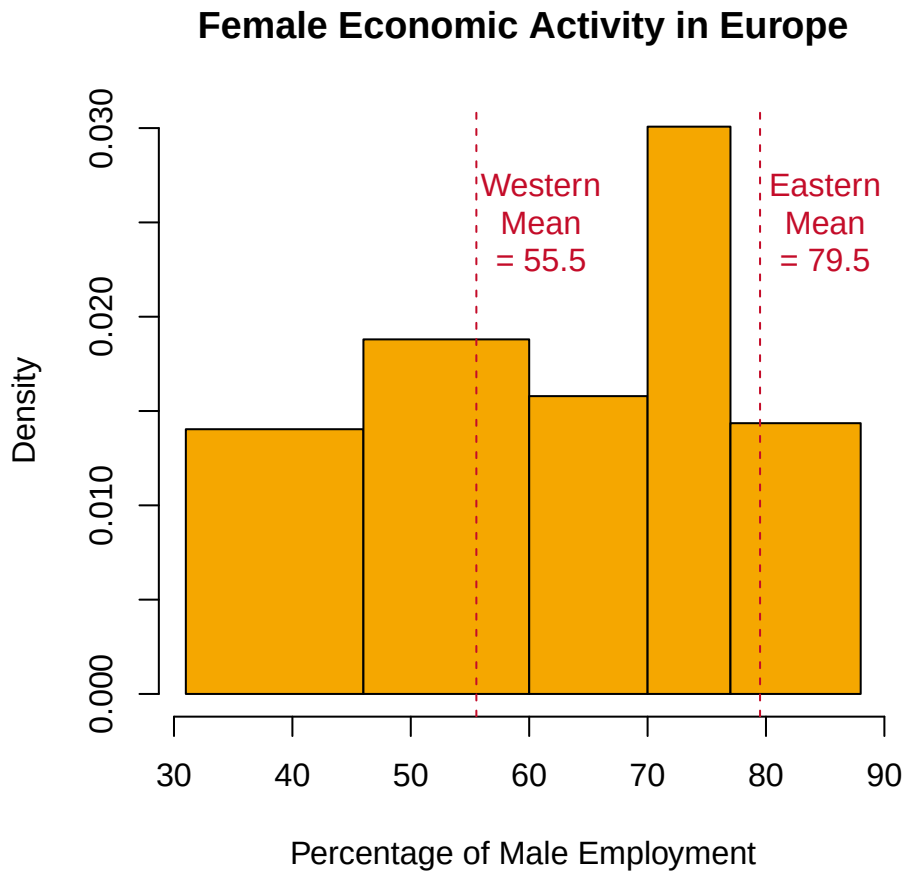
Per esempio il carattere Health di Everitt (2005, p. 2) è trimodale:

Health	Freq.
Very Poor	1
Poor	0
Average	3
Good	3
Very Good	3
Total	10

Per un carattere quantitativo continuo, la rilevazione della stessa modalità su due diverse unità statistiche è una mera coincidenza (tanto più probabile quanto più grossolana è la misura).

Converra indicare come classe modale, con riferimento a una data suddisione in classi, la classe che presenta la massima densità, ovvero la classe cui corrisponde la barra più alta dell'istogramma.

Si parlerà di moda locale quando una barra dell'istogramma è più alta delle barre adiacenti e si dirà multimodale un istogramma con più di una moda locale; la bimodalità dell'istogramma seguente segnala che il campione rappresentato è eterogeneo (Eastern vs Western Europe).


Female Economic Activity in Europe
Western
Mean
= 55.5
Eastern
Mean
= 79.5
Density
0.000
0.010
0.020
0.030
30
40
50
60
70
80
90
Percentage of Male Employment

Sai ched'è la statistica? È 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre uguale
puro co' la persona bisognosa.

Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:

e, se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due.

Come evidenziato dal precedente sonetto (Carlo Alberto Salustri, in arte Trilussa, 1871–1950) due distribuzioni con la stessa media possono essere piuttosto diverse.

Per fare un esempio meno estremo, si pensi (per fissare le idee) al consumo annuo di pasta (Kg pro capite) in due diversi gruppi di paesi e si confrontino le (ipotetiche) osservazioni

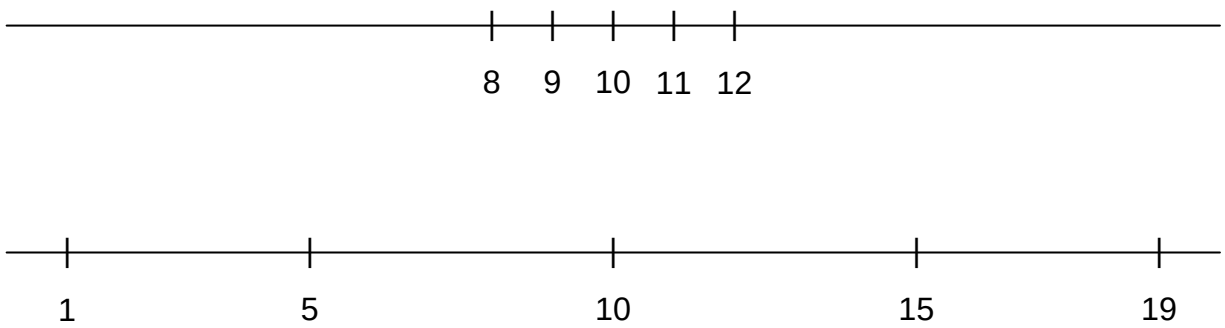
$$8 \quad 9 \quad 10 \quad 11 \quad 12$$

con le (altrettanto ipotetiche) osservazioni

$$1 \quad 5 \quad 10 \quad 15 \quad 19$$

notando che in entrambi i casi la media e la mediana valgono 10...

Distribution Variability



Un indice di variabilità è una statistica (funzione dei dati) che

- assume il suo valore minimo quando le modalità osservate sono tutte uguali;
- aumenta all'aumentare della "diversità" tra le modalità osservate.

In questo modo viene misurata la tendenza di un carattere ad assumere modalità diverse (in pratica l'uso di un particolare indice di variabilità precisa la nozione di variabilità).

In genere il valore minimo è zero, mentre non c'è un valore massimo (quindi non ha senso parlare di massima variabilità); fanno eccezione gli indici di concentrazione ed eterogeneità (che variano tra zero e uno).

Si possono costruire indici di variabilità in almeno due modi:

- confrontando due valori caratteristici della distribuzione (es. campo di variazione, differenza interquartile);
- confrontando le modalità osservate con un loro valore medio (es. deviazione standard, scostamento semplice dalla mediana).

Gli indici di concentrazione/eterogeneità sono invece ottenuti confrontando la distribuzione di quantità/frequenza osservata con una distribuzione di riferimento (distribuzione uniforme).

Il campo di variazione di un carattere quantitativo X è la differenza tra il massimo e il minimo delle modalità osservate:

$$range(X) = x_{100\%} - x_{0\%}.$$

Per esempio si trova

$$\begin{aligned} range(10, 10, 10, 10, 10) &= 10 - 10 = 0 \\ range(8, 9, 10, 11, 12) &= 12 - 8 = 4 \\ range(1, 5, 10, 15, 19) &= 19 - 1 = 18 \end{aligned}$$

per tre distribuzioni con (stessa media e) variabilità crescente.

Il campo di variazione

- è semplice da calcolare (il grosso del lavoro è ordinare le osservazioni) e di fatto si ottiene "gratis" perché non si può pensare di studiare la distribuzione di X senza calcolare $x_{0\%}$ e $x_{100\%}$;
- si basa su due sole osservazioni (la più piccola e la più grande) ignorando le rimanenti modalità osservate;
- è molto sensibile ai valori anomali (se sono presenti dei valori anomali, almeno uno di essi è determinante ai fini del calcolo).

Se il carattere è suddiviso in classi, la differenza fra gli estremi della suddivisione è una sovrastima del campo di variazione.

La differenza interquartile di un carattere quantitativo X è la differenza tra il terzo e il primo quartile del carattere stesso:

$$iqr(X) = x_{75\%} - x_{25\%}.$$

Per esempio si trova

$$\begin{aligned} iqr(10, 10, 10, 10, 10) &= 10 - 10 = 0 \\ iqr(8, 9, 10, 11, 12) &= 11 - 9 = 2 \\ iqr(1, 5, 10, 15, 19) &= 15 - 5 = 10 \end{aligned}$$

per le tre distribuzioni di prima.

La differenza interquartile

- è il campo di variazione del 50% delle osservazioni centrali (e come tale è sempre più piccola del campo di variazione di tutte le osservazioni);
- si basa su due sole osservazioni (il primo e il terzo quartile) ignorando le rimanenti modalità osservate;
- è robusta rispetto alla presenza di valori anomali (per definizione minoritari ed estremi).

Se il carattere è suddiviso in classi, occorre rappresentare le classi quartili con opportuni valori.

La varianza di un carattere quantitativo X è lo scarto quadratico medio delle sue modalità osservate $x_1, \ldots, x_n$ dalla loro media $\bar{x}$:

$$\mathcal{V}ar(X) = SQM(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2;$$

si tratta quindi del minimo scarto quadratico medio per $x_1, \ldots, x_n$.

Si può dimostrare che vale la formula

$$\mathcal{V}ar(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

e questa può essere sfruttata per il calcolo.

X = Female Economic Activity in Eastern Europe					
i		x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
1	Bulgaria	88	8.5	72.25	7744
2	Czech Republic	84	4.5	20.25	7056
3	Hungary	70	−9.5	90.25	4900
4	Poland	77	−2.5	6.25	5929
5	Romania	77	−2.5	6.25	5929
6	Slovakia	81	1.5	2.25	6561
	Total	477	0.0	197.50	38119

Troviamo $\bar{x} = 477/6 = 79.5$, quindi $\mathcal{V}ar(X) = 197.5/6 = 32.92$ dalla definizione, oppure mediante la formula alternativa ritroviamo $\mathcal{V}ar(X) = (38119/6) - (79.5)^2 = 6353.17 - 6320.25 = 32.92$.

Si può calcolare la varianza a partire da una distribuzione di frequenza, prendendo la media pesata degli scarti quadratici delle diverse modalità dalla media aritmetica (ogni modalità pesata con la propria frequenza).

Si trova

$$\begin{aligned} \mathcal{V}ar(X) &= \frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2 \quad = \quad \overline{x^2} - \bar{x}^2 \end{aligned}$$

dove n_i è la frequenza della modalità x_i, per $i = 1, \ldots, k$,
e $n = n_1 + \cdots + n_k$.

$X =$ Numero di Figli					
x_i	f_i	$f_i x_i$	x_i^2	$f_i x_i^2$	
0	0.2774	0.0000	0	0.0000	
1	0.1688	0.1688	1	0.1688	$(1.903)^2 = 3.6214$
2	0.2482	0.4964	4	0.9928	
3	0.1427	0.4281	9	1.2843	$6.7312 -$
4	0.0844	0.3376	16	1.3504	$3.6214 =$
5	0.0362	0.1810	25	0.9050	3.1098
6	0.0161	0.0966	36	0.5796	
7	0.0151	0.1057	49	0.7399	$\mathcal{V}ar(X) = 3.1098$
8	0.0111	0.0888	64	0.7104	
Totale	1.0000	1.9030		6.7312	

La varianza

- è nulla se (e solo se) tutte le osservazioni sono uguali;
- si basa su tutte le modalità osservate (nessuna esclusa);
- è piuttosto sensibile ai valori anomali (gli scarti sono elevati al quadrato e dunque amplificati);
- è misurata sulla scala del quadrato del carattere (es. $\mathcal{K}g^2$, se il carattere è misurato in $\mathcal{K}g$).

Se il carattere è suddiviso in classi, rappresentando le classi con le medie di classe si approssima la varianza tendenzialmente per difetto.

Cosa vuol dire che la varianza dell'indice di attività economica femminile in Europa Orientale è pari a 33 punti percentuali quadri?

Poiché la varianza manca di interpretazione pratica, si introduce la deviazione standard, definita come radice quadrata della varianza:

$$sd(X) = \sqrt{\mathcal{V}ar(X)};$$

in questo modo si ottiene un indice che è espresso nelle stesse unità di misura del carattere X (di cui ha le stesse dimensioni fisiche).

Per esempio la deviazione standard dell'indice di attività economica femminile in Europa Orientale è pari a $\sqrt{32.92} = 5.74$ punti percentuali.

Cosa vuol dire che la deviazione standard dell'indice di attività economica femminile in Europa Orientale è pari a 5.7 punti?

Vuol dire che scarti dell'ordine di 5.7 punti dal valore medio di 79.5 punti sono tipici per i paesi dell'Europa Orientale.

Apprendendo che $m(X) = 79.5$ impariamo che le modalità osservate sono "prossime" a 79.5, mentre apprendedo che $sd(X) = 5.7$ impariamo che cosa significhi "prossime"; per esempio

- ci aspettiamo di avere tra le modalità osservate un valore come $79.5 + 5.7 \simeq 85$ (infatti l'indice vale 84 per la Repubblica Ceca)
- ma non ci aspettiamo di avere tra le modalità osservate un valore come $79.5 - 57.4 \simeq 22$ (infatti il minimo valore osservato è 70).

Quante deviazioni standard può distare un'osservazione dalla media?

Qualunque sia la distribuzione di X:

- al massimo il 25% delle osservazioni dista due o più deviazioni standard dalla media;
- al massimo il 12% delle osservazioni dista tre o più deviazioni standard dalla media;
- al massimo il 7% delle osservazioni dista quattro o più deviazioni standard dalla media;
- ...

Theorem (Chebyshev)

Se k è un numero positivo e $\widetilde{P}_k$ è la frequenza percentuale delle osservazioni x di X che soddisfano la disuguaglianza

$$|x - m(X)| \geq k \times sd(X)$$

allora, qualunque sia la distribuzione di X, si ha

$$\widetilde{P}_k \leq \frac{100}{k^2}.$$

Es. $\widetilde{P}_2 \leq 100/4 = 25$, $\widetilde{P}_3 \leq 100/9 \lesssim 12$, $\widetilde{P}_4 \leq 100/16 \lesssim 7, \ldots$

Il teorema di Chebyshev sovrastima la percentuale di osservazioni che distano almeno k deviazioni standard dalla media; nell'esempio sui figli

X = Numero di Figli										
x_i	0	1	2	3	4	5	6	7	8	Totale
p_i	27.74	16.88	24.82	14.27	8.44	3.62	1.61	1.51	1.11	100.00

si ha $m(X) = 1.903$ e $sd(X) = \sqrt{3.1098} = 1.7635$ cosicché

$$\begin{aligned} \widetilde{P}_2 &= 1.61 + 1.51 + 1.11 = 4.23 < 25 \\ \widetilde{P}_3 &= 1.11 = 1.11 < 12 \\ \widetilde{P}_4 &= = 0.00 < 7 \end{aligned}$$

mentre per $k = 1$ il teorema di Chebyshev afferma un'ovvietà: $\widetilde{P}_1 = p_0 + p_4 + p_5 + p_6 + p_7 + p_8 < 100/1 = 100$.

Lo scostamento semplice dalla mediana di un carattere quantitativo X è lo scarto assoluto medio delle sue modalità osservate $x_1, \ldots, x_n$ dalla loro mediana $x_{50\%}$:

$$ss_{50\%}(X) = SAM(x_{50\%}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - x_{50\%}|;$$

si tratta quindi del minimo scarto assoluto medio per $x_1, \ldots, x_n$.

Lo scostamento semplice dalla mediana è più robusto della deviazione standard (ma meno della differenza interquartile) rispetto alla presenza di valori anomali, perché gli scarti non sono amplificati dall'elevamento al quadrato...

Stipendio dei dipendenti di un negozio (migliaia di dollari)					
i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$x_i - x_{50\%}$	$\lvert x_i - x_{50\%} \rvert$
1	10.2	−27.7	767.29	−1.0	1.0
2	10.4	−27.5	756.25	−0.8	0.8
3	10.7	−27.2	739.84	−0.5	0.5
4	11.2	−26.7	712.89	0.0	0.0
5	11.3	−26.6	707.56	0.1	0.1
6	11.5	−26.4	696.96	0.3	0.3
7	200.0	162.1	26276.41	188.8	188.8
Totale	265.3	0.0	30657.20	186.9	191.5

Si trova $\bar{x} = 265.3/7 = 37.9$ e $x_{50\%} = 11.2$, quindi si ha $sd(X) = \sqrt{30657.20/7} = 66.2$ e $ss_{50\%}(X) = 191.5/7 = 27.4$.

Sia X la quantità di cenere inquinante (g/min) emessa da un campione di fabbriche dove è installato un certo tipo di filtro e Y la corrispondente quantità in un campione di fabbriche dove è installato un altro tipo di filtro (Borra & Di Ciaccio, 2008, Esempio 4.3.3); se abbiamo osservato

$$\begin{aligned} m(X) &= 64.67 \\ sd(X) &= 13.65 \\ m(Y) &= 34.22 \\ sd(Y) &= 12.02 \end{aligned}$$

sembra evidente che il secondo tipo di filtro è più efficiente, ma davvero è anche più "regolare" (meno variabile)?

Poiché $m(X) \gg m(Y)$ (diversi livelli di emissione) conviene adoperare il coefficiente di variazione (rapporto tra deviazione standard e media):

$$\begin{aligned} cv(X) &= \frac{sd(X)}{m(X)} = \frac{13.65}{64.67} = 21\% \\ cv(Y) &= \frac{sd(Y)}{m(Y)} = \frac{12.01}{34.22} = 35\% \end{aligned}$$

si verifica così che la variabilità di X è in termini relativi minore della variabilità di Y (potrebbe volere dire che il secondo tipo di filtro è maggiormente sensibile alle condizioni operative); se invece di media e deviazione standard avessimo il riassunto dei cinque numeri, potremmo misurare la variabilità relativa rapportando $iqr(X)$ a $x_{50\%}$.

In generale, ovviamente, il confronto della variabilità di due caratteri fornisce un risultato che dipende dall'indice di variabilità adottato.

Quando ha senso calcolare il coefficiente di variazione?

Quando i caratteri sono misurati su scala di rapporti:

- in tal caso sia la media che la deviazione standard sono definiti a meno di uno stesso fattore costante, di modo che il coefficiente di variazione non dipende dall'unità di misura adottata;
- in caso contrario (scala a intervalli) la media è definita a meno di un termine costante e il coefficiente di variazione perde senso...

...es. $m(C) = 20$, $sd(C) = 3$ e $m(F) = 68$, $sd(F) = 5.4$, come visto, possono rappresentare la stessa distribuzione di temperatura (in gradi Celsius e Fahrenheit) ma producono diversi coefficienti di variazione:

$$cv(C) = 15\% \neq 7.9\% = cv(F).$$

Diremo che un carattere quantitativo Z è misurato in unità standard (u.s.) se Z ha media nulla e deviazione standard unitaria:

$$\begin{aligned} m(Z) &= 0, \\ sd(Z) &= 1. \end{aligned}$$

Per un carattere misurato in unità standard il teorema di Chebyshev afferma che la percentuale di osservazioni in valore assoluto maggiori o uguali a k vale al massimo $100/k^2$ (es. al massimo il 25% delle osservazioni non si colloca tra -2 e $+2$ unità standard).

È sempre possibile standardizzare un carattere quantitativo:

1. si sottrae la media dalle osservazioni, ottenendone gli scarti;
2. si dividono gli scarti così ottenuti per la deviazione standard.

Prima vediamo un esempio, approfondendo l'utilità delle unità standard, poi giustifichiamo la procedura di standardizzazione. . .

. . . per esempio possiamo standardizzare l'indice di attività economica femminile in Europa Orientale (denotiamolo con X) trovando:

i		x_i	$x_i - m(X)$	$\frac{x_i - m(X)}{sd(X)}$	$\left[\frac{x_i - m(X)}{sd(X)}\right]^2$
1	Bulgaria	88	8.5	1.48	2.1904
2	Czech Republic	84	4.5	0.78	0.6084
3	Hungary	70	−9.5	−1.66	2.7556
4	Poland	77	−2.5	−0.44	0.1936
5	Romania	77	−2.5	−0.44	0.1936
6	Slovakia	81	1.5	−0.26	0.0676
	Total	477	0.0	−0.02	6.0092

$$m(X) = 79.5, \quad sd(X) = 5.74$$

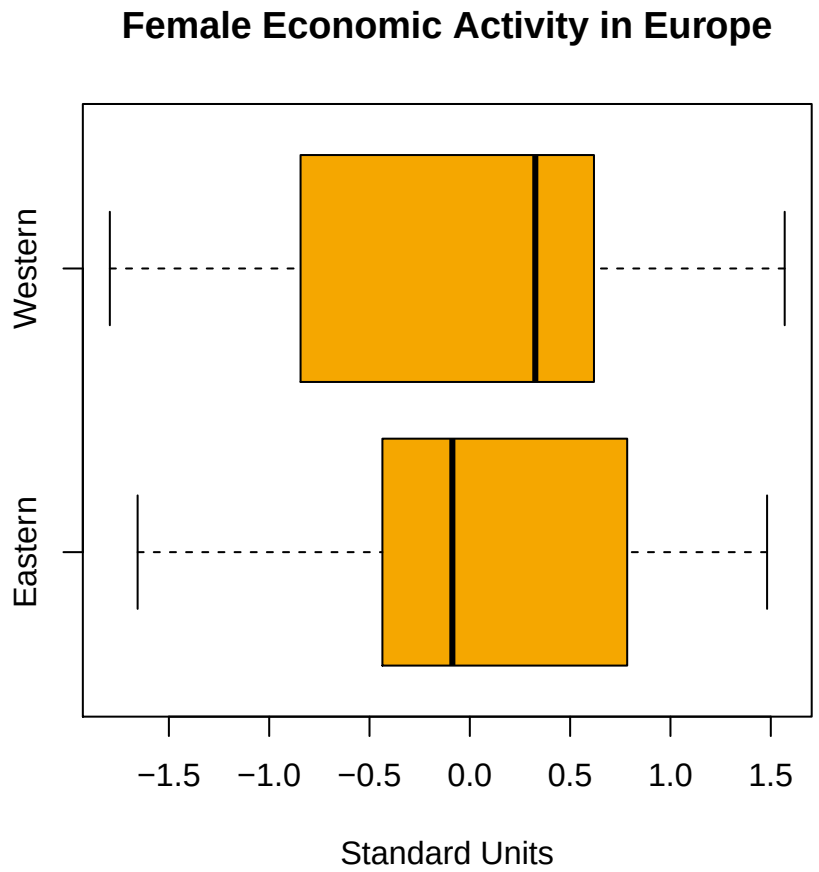
Analogamente possiamo standardizzare l'indice di attività economica femminile in Europa Occidentale (denotiamolo con Y):

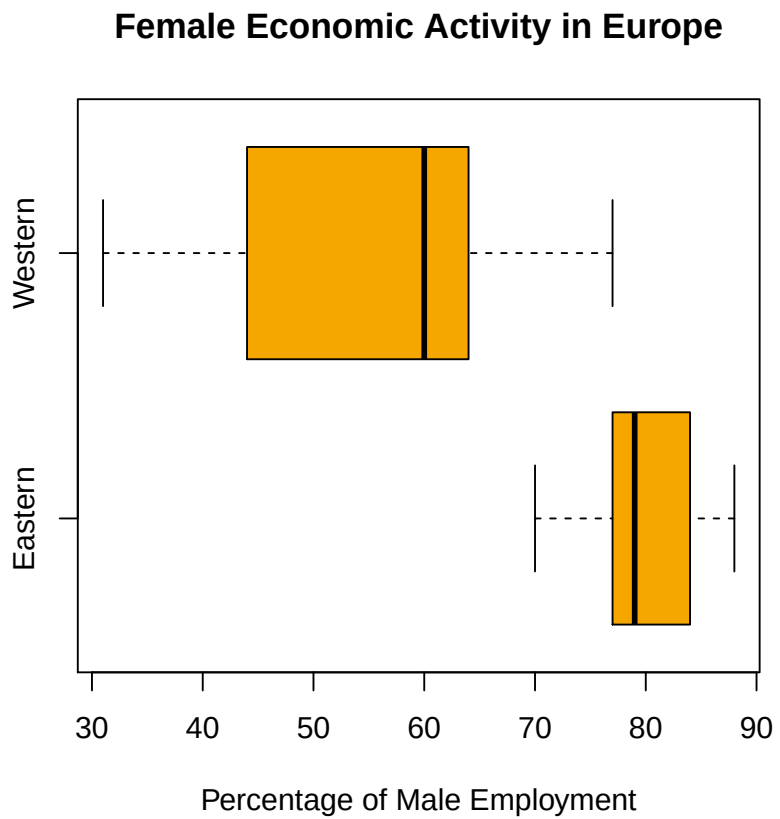
Country	y_i	$\frac{y_i - m(Y)}{sd(Y)}$	Country	y_i	$\frac{y_i - m(Y)}{sd(Y)}$
Austria	60	0.33	Norway	68	0.91
Belgium	47	−0.62	Portugal	51	−0.33
Denmark	77	1.57	Spain	31	−1.79
France	64	0.62	Sweden	77	1.57
Ireland	41	−1.06	Switzerland	60	0.33
Italy	44	−0.84	United Kingdom	60	0.33
Netherlands	42	−0.99	Total	722	0.03

$$m(X) = 55.5, \quad sd(X) = 13.7$$

Le unità standard consentono di prescindere da posizione e scala nel confrontare distribuzioni con medie e deviazioni standard diverse; in questo modo ci si può concentrare su

- pesantezza delle code (es. il minimo in Europa Occidentale, Spagna, -1.79 u.s., è più estremo del minimo in Europa Orientale, Ungheria, -1.66 u.s., ma in entrami i casi si hanno code leggere, visto che in valore assoluto non si arriva nemmeno a due unità standard, contro un possibile 25% indicato da Chebyshev);
- asimmetria (a destra in Europa Orientale, a sinistra e più marcata in Europa Occidentale)...


Female Economic Activity in Europe
Western
Eastern
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Standard Units




```
> load("dataFemActivity.rda")
> femact[12:15,]
                Activity  Europe
Switzerland           60 Western
United Kingdom        60 Western
Bulgaria              88 Eastern
Czech Republic        84 Eastern
> colE <- rgb(0.9609375,0.6562500,0.0000000)
> boxplot(Activity ~ Europe,
+         data = femact,
+         main = "Female Economic Activity
+                 in Europe",
+         horizontal = TRUE,
+         col = colE,
+         xlab = "Percentage of
+                 Male Employment")
```

La procedura di standardizzazione è giustificata dall'equivarianza di media e deviazione standard rispetto a trasformazioni affini positive:

se

$$X = aZ + b, \quad a > 0, \quad b \in \mathbb{R},$$

allora

$$\begin{aligned} m(X) &= a\, m(Z) + b, \\ sd(X) &= a\, sd(Z). \end{aligned}$$

Discutiamo prima un'applicazione e poi la standardizzazione...

...scegliendo $a = 9/5$ e $b = 32$ otteniamo la relazione tra gradi Celsius e gradi Fahrenheit:

$$F = \frac{9}{5}C + 32$$

- se la temperatura media in gradi Celsius vale $m(C) = 20$, la stessa in gradi Fahrenheit vale $m(F) = \frac{9}{5} \times 20 + 32 = 68$
- se la deviazione standard in gradi Celsius vale $sd(C) = 3$, la stessa in gradi Fahrenheit vale $sd(F) = \frac{9}{5} \times 3 = 5.4$

Per comprendere la standardizzazione riformuliamo l'equivarianza:

se

$$Z = \frac{X - b}{a}, \quad a > 0, \quad b \in \mathbb{R},$$

allora

$$m(Z) = \frac{m(X) - b}{a},$$
$$sd(Z) = \frac{sd(X)}{a}.$$

Prendendo $b = m(X)$ e $a = sd(X)$ si trova $m(Z) = 0$ e $sd(Z) = 1$.

Si parla di concentrazione quando la variabilità di un carattere trasferibile è intesa come non uniformità della sua distribuzione di quantità nel collettivo di interesse.

La distribuzione di minima concentrazione è dunque quella uniforme (tutte le unità statistiche posseggono la stessa quantità di carattere) e in questo caso il carattere si dice equidistribuito.

È anche definita una distribuzione di massima concentrazione (unica a meno dell'ordine di elencazione delle unità) nella quale tutto il carattere è posseduto da una sola unità statistica.

Per valutare il grado di concentrazione di un carattere trasferibile conviene ordinare le unità osservate dalla più "povera" alla più "ricca"

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

e considerarne la distribuzione unitaria di quantità cumulata relativa

$$Q_{(i)} = \frac{A_{(i)}}{A_{(n)}}$$

dove $A_{(i)} = x_{(1)} + x_{(2)} + \cdots + x_{(i)}$ è la quantità di carattere posseduta dalle i unità più "povere" e n è la numerosità campionaria.

Per esempio, nel caso delle esportazioni emiliano-romagnole (in milioni di euro) verso i primi dodici paesi per valore, troveremo...

i	Paese	$x_{(i)}$	$A_{(i)}$	$Q^{\%}_{(i)}$	$F^{\%}_{(i)}$	$F^{\%}_{(i)} - Q^{\%}_{(i)}$
1	Giappone	710	710	2.95	8.33	5.38
2	Grecia	807	1517	6.30	16.67	10.37
3	Austria	840	2357	9.80	25.00	15.20
4	Paesi Bassi	934	3291	13.68	33.33	19.65
5	Belgio	948	4239	17.62	41.67	24.05
6	Fed. Russa	1021	5260	21.86	50.00	28.14
7	Svizzera	1068	6328	26.30	58.33	32.03
8	Regno Unito	2396	8724	36.25	66.67	30.42
9	Spagna	2561	11285	46.90	75.00	28.10
10	Stati Uniti	4066	15351	63.80	83.33	19.53
11	Francia	4322	19673	81.76	91.67	9.91
12	Germania	4390	24063	100.00	100.00	0.00
	Totale	24063	102798		650.00	222.78

. . . e confronteremo (per ogni i da 1 a n) le quantità osservate $Q_{(i)}$ con le quantità di riferimento

$$F_{(i)} = \frac{i}{n}$$

relative al caso di equidistribuzione (distribuzione uniforme).

Avremo sempre (per ogni i da 1 a n)

$$Q_{(i)} \leq F_{(i)}$$

perché abbiamo ordinato le unità statistiche dalla più "povera" alla più "ricca" e in particolare (banalmente) avremo $Q_{(n)} = F_{(n)} = 1$; quindi. . .

...la somma delle differenze $F_{(i)} - Q_{(i)}$, al variare di i da 1 a $n-1$, sarà

1. sempre positiva (o nulla nel caso in cui le quantità osservate siano quelle dell'equidistribuzione);
2. al massimo pari a alla somma delle $F_{(i)}$, al variare di i da 1 a $n-1$, nel caso in cui si osservi $Q_{(1)} = Q_{(2)} = \cdots = Q_{(n-1)} = 0$ e $Q_{(n)} = 1$ (le quantità osservate siano quelle della massima concentrazione).

Possiamo pertanto definire l'indice di concentrazione relativa

$$G = \frac{\sum_{i=1}^{n-1} (F_{(i)} - Q_{(i)})}{\sum_{i=1}^{n-1} F_{(i)}}$$

a valori tra zero (equidistribuzione) e uno (massima concentrazione)...

. . . per esempio, nel caso delle esportazioni emiliano-romagnole verso i primi dodici paesi per valore, troviamo

$$G = \frac{222.78}{550} = 0.41$$

vale a dire una concentrazione pari al 41% del massimo possibile (un "moderato grado di concentrazione").

In pratica il grado di concentrazione sarà giudicato "elevato" o "ridotto" in relazione ad altre realtà (luoghi/tempi); si veda per esempio il grafico intitolato Gini Index - Income Disparity since World War II su Wikipedia (`http://en.wikipedia.org/wiki/Gini_coefficient`).

L'indice di concentrazione relativa può riscriversi come

$$G = 1 - \frac{2\sum_{i=1}^{n-1} A_{(i)}}{(n-1)A_{(n)}}$$

sfruttando l'identità $F_{(1)} + F_{(2)} + \cdots + F_{(n-1)} = (n-1)/2$;
nell'esempio delle esportazioni emiliano-romagnole verso i primi dodici paesi per valore si ritrova

$$G = 1 - \frac{2 \times (102798 - 24063)}{11 \times 24063} = 0.41$$

con meno calcoli rispetto all'uso diretto della definizione.

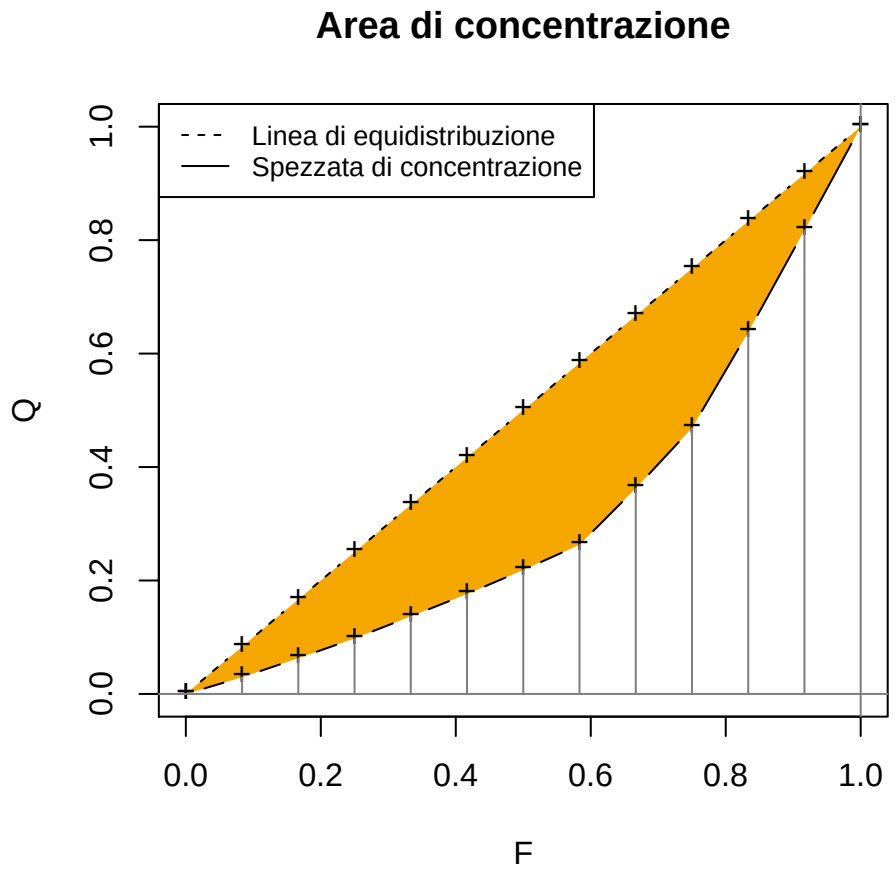
L'indice di concentrazione relativa può anche riscriversi come

$$G = \frac{2n}{n-1}\left\{\frac{1}{2} - \sum_{i=0}^{n-1}\frac{Q_{(i)} + Q_{(i+1)}}{2n}\right\}, \quad \text{con } Q_{(0)} = F_{(0)} = 0,$$

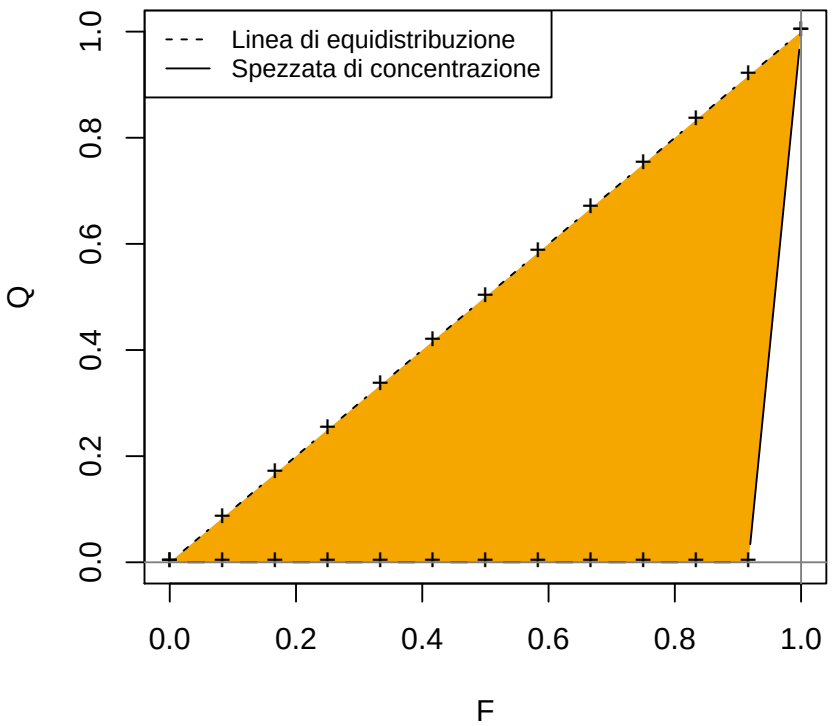
di modo che $\frac{(n-1)}{n}G$ è pari al doppio dell'area di concentrazione; di conseguenza, per n "grande", si ha

$$G \simeq 1 - \sum_{i=0}^{n-1}(Q_{(i+1)} + Q_{(i)}) \times (F_{(i+1)} - F_{(i)})$$

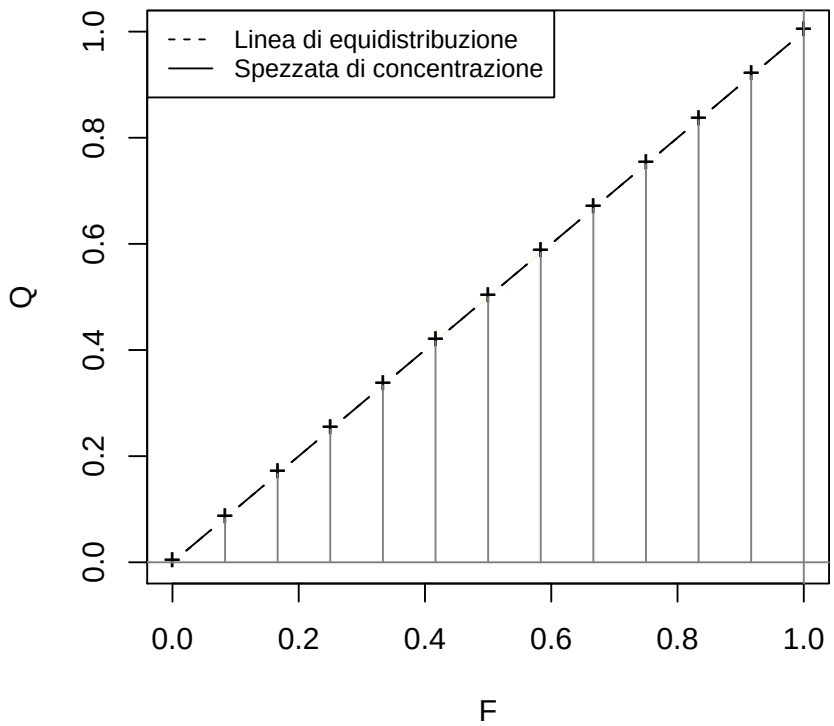
e questa approssimazione può essere utile quando si abbia una distribuzione di quantità rispetto a una suddivisione in classi...

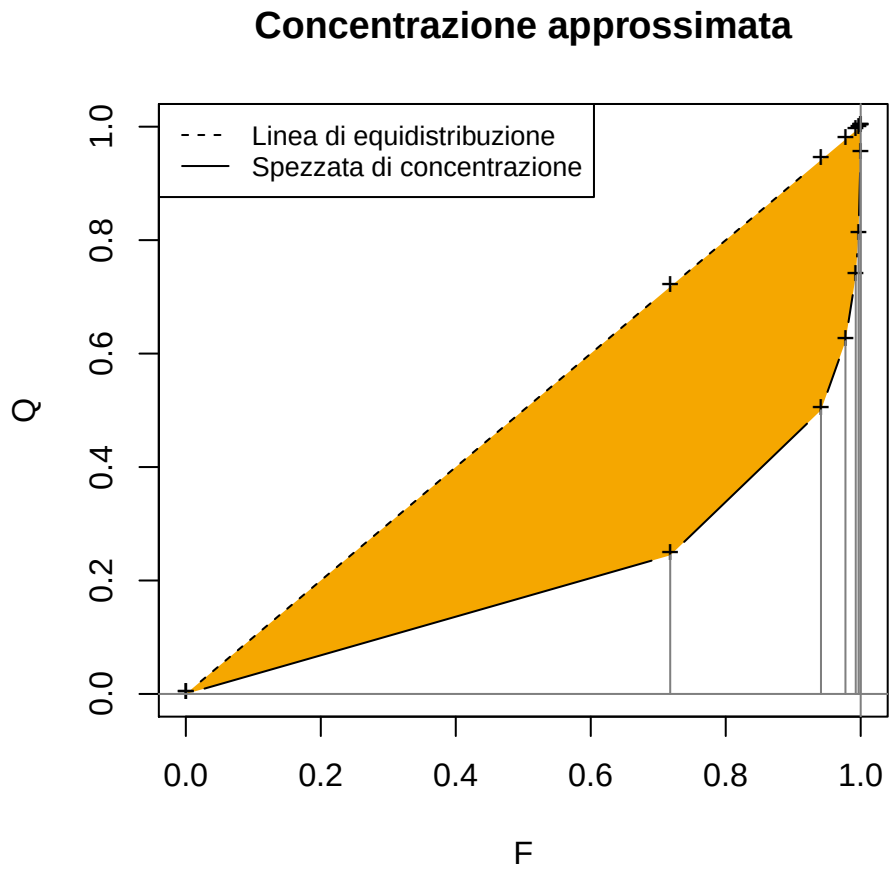

Area di concentrazione
Linea di equidistribuzione
Spezzata di concentrazione
Q
F
0.0
0.2
0.4
0.6
0.8
1.0

Massima concentrazione



Concentrazione nulla




Concentrazione approssimata
Linea di equidistribuzione
Spezzata di concentrazione
Q
F
0.0
0.2
0.4
0.6
0.8
1.0

. . . es. distribuzione degli addetti nelle imprese italiane nel 1988 (Borra & Di Ciaccio, 2008, Esempio 4.7.3):

Addetti (x)	Imprese (n)	Tot. Addetti $\approx \bar{x} \times n$		F	Q
0–2	2043.0		2718.3	0.7177	0.2444
3–9	636.0		2845.6	0.9412	0.5002
10–19	103.2		1352.0	0.9774	0.6217
20–49	43.4		1281.2	0.9927	0.7369
50–99	11.8		808.7	0.9968	0.8096
100–499	8.3		1588.3	0.9997	0.9524
500–999	0.8		529.4	1.0000	1.0000
Totale	2846.5	migliaia	11123.5	migliaia	

si trova $G = 58\%$ con la formula approssimata.

Si parla di eterogeneità quando la variabilità di un carattere qualitativo (misurato su scala nominale) è intesa come uniformità della sua distribuzione di frequenza nel collettivo di interesse.

Il grado di eterogeneità di un carattere con k modalità distinte si può, per esempio, valutare mediante l'indice di entropia

$$H = -\frac{1}{\log k} \sum_{i=1}^{k} f_i \log f_i$$

dove $f_1, f_2, \ldots, f_k$ sono le frequenze relative delle k modalità; si rinvia a Borra & Di Ciaccio (2008, p. 95) per un indice di eterogeneità alternativo.

L'entropia

- è funzione di tutte le modalità osservate (attraverso le loro frequenze relative);
- assume il suo valore minimo (zero) quando tutte le osservazioni sono uguali: es. $f_1 = 1$ e $f_2 = \cdots = f_k = 0$;
- assume il suo valore massimo (uno) quando la distribuzione del carattere è uniforme: $f_1 = \ldots f_k = 1/k$.

Dunque, in generale, l'entropia è un numero tra zero e uno:

> *per la variabilità misurata mediante entropia si può parlare di variabilità massima (distribuzione uniforme) oltre che di variabilità minima/nulla (distribuzioni degeneri).*

Religion	f_i	$-\log f_i$	$-f_i \log f_i$
Catholic	0.760	0.1192	0.0906
Irreligion	0.200	0.6990	0.1398
Islam	0.023	1.6383	0.0377
Others	0.018	1.7447	0.0314
Total	1.001		0.2995

Poiché $\log 4 = 0.6021$ si trova

$$H = \frac{0.2995}{0.6021} = 0.4974$$

circa pari al 50% del massimo possibile.

AGRESTI, A. & FINLAY, B. (1997).
Statistical Methods for the Social Sciences.
Prentice-Hall, Upple Saddle River.

BOHRNSTEDT, G. W. & KNOKE, D. (1998).
Statistica per le Scienze Sociali.
Il Mulino, Bologna.

BORRA, S. & DI CIACCIO, A. (2008).
Statistica: Metodologie per le Scienze Economiche e Sociali (Seconda Edizione).
McGraw-Hill, Milano.

EVERITT, B. (2005).
An R and S-PLUS® Companion to Multivariate Analysis.
Springer-Verlag, London.

REGIONE EMILIA ROMAGNA (2006).
I Numeri dell'Emilia Romagna.
CLEUB, Bologna.