Analisi dei Dati Sperimentali e Statistica

Premesse alla statistica

Luca Mari, versione 7.10.15

Contenuti

I principali concetti introdotti in questo capitolo

I principali simboli introdotti in questo capitolo

$\#A$: cardinalità dell'insieme A (pag. 2)

$\{a_1, a_2, \ldots a_n\} = \{a_i\}_{i=1,\ldots,n}$: insieme di n elementi a_i (pag. 2)

$\langle a_1, a_2, \ldots a_n \rangle = \langle a_i \rangle_{i=1,\ldots,n}$: successione di n elementi a_i (pag. 2)

$f : X \to Y$: funzione f dall'insieme dominio X all'insieme codominio Y (pag. 2)

$v_{rel\%}(x_i, x_j) = 100\dfrac{[x_j - x_i]}{x_i}$: tasso di variazione per la successione $\langle x_i \rangle$ (pag. 2)

$[x(t_i + \Delta t) - x(t_i)]/\Delta t$: rapporto incrementale per la successione temporale $\langle x_i \rangle = \langle x(t) \rangle$ (pag. 4)

$x'(t) = \lim_{\Delta t \to 0} \dfrac{x(t + \Delta t) - x(t)}{\Delta t}$: derivata della funzione $x(t)$ (pag. 4)

$x(t)=\int_{t_1}^{t} x'(\tau)d\tau$: funzione integrale della funzione $x'(t)$ (pag. 5)

$\begin{bmatrix} C_j \\ n_j \end{bmatrix}$: distribuzione a frequenze assolute n_j per le categorie C_j (pag. 8)

$\begin{bmatrix} C_j \\ f_j \end{bmatrix}$: distribuzione a frequenze relative f_j per le categorie C_j (pag. 8)

Insiemi e successioni

I dati di origine sperimentale si presentano spesso non come singoli valori, ma come insiemi di valori. Richiamiamo al proposito alcune notazioni, mantenendo per ora l'ipotesi di prendere in considerazione solo insiemi costituiti da un numero finito di elementi:

- gli *insiemi* si indicano con una lettera maiuscola, e i loro elementi con una lettera minuscola, eventualmente seguita da un indice che distingue ogni elemento nell'insieme; per esempio l'insieme A è costituito dagli elementi a_i, cioè $\forall i = 1,...,n$, $a_i \in A$, essendo dunque n la cardinalità (cioè il numero degli elementi) dell'insieme A, a volte indicata anche come $n = \#A$; si può indicare anche $A = \{a_i\}$ o, più esplicitamente, $A = \{a_i\}_{i=1,\dots,n}$; per definizione, gli elementi di un insieme sono distinti, cioè $i \neq j$ implica $a_i \neq a_j$, e non ordinati (cioè per esempio {0, 1, 2} = {2, 1, 0});
- a partire da un insieme A possono essere costruite *successioni* di elementi di A, $x = \langle x_1, x_2, ... \rangle$, in cui $\forall i$, $x_i \in A$; una successione è dunque interpretabile come una funzione x: $N \to A$, che a ogni numero naturale $i = 1, 2, ...$, chiamato *indice* della successione (in certi casi è comodo far partire l'indice da 0), fa corrispondere un elemento $x(i) = x_i \in A$, dove in questo caso A è chiamato *insieme supporto* della successione; per costruzione, una successione può contenere elementi uguali (cioè per esempio $\langle 0,1,0,1,... \rangle$ è una successione ammessa), e l'ordine con cui gli elementi compaiono in una successione è significativo (cioè per esempio $\langle 0,1,2 \rangle$ e $\langle 2,1,0 \rangle$ sono successioni diverse).

Per esempio, sia $A = \{a_1 = 18, \dots, a_{13} = 30\}$ l'insieme dei voti che uno studente può prendere a un esame (tralasciando le insufficienze e la lode). Potrebbe allora essere costruita la successione dei voti presi da un certo studente in un certo intervallo di tempo, $x = \langle x_1, x_2, ... \rangle$ il cui generico termine x_i, dunque, rappresenta l'i-esimo voto rispetto all'ordine con cui i voti sono stati presi (si noti che in questo caso l'indice si riferisce all'ordine con cui i voti sono stati presi e non alla posizione del voto nell'insieme supporto A; e infatti, come detto, la successione può contenere due o più elementi uguali).

Variazioni assolute, relative, percentuali

Una prima informazione che si può cercare in una successione di dati sperimentali è relativa al suo *andamento*, e quindi in particolare alla variazione dei valori della successione al variare dell'indice. Dati due elementi di una successione, x_i e x_j, la *variazione assoluta* è semplicemente lo scarto del secondo dal primo:

$$v_{ass}(x_i,x_j) = x_j - x_i$$

così che, evidentemente, $v_{ass}(x_j,x_i) = -v_{ass}(x_i,x_j)$.

La *variazione relativa* è invece lo scarto del secondo valore dal primo rispetto al primo, trattato dunque come riferimento:

$$v_{rel}(x_i, x_j) = \frac{x_j - x_i}{x_i}$$

e la *variazione percentuale* è:

$$v_{rel\%}(x_i, x_j) = 100\frac{x_j - x_i}{x_i}$$

(sia la variazione relativa sia la variazione percentuale sono anche chiamate *tassi di variazione*) così che, in questo caso, $v_{rel}(x_j, x_i) \neq -v_{rel}(x_i, x_j)$.

Se per esempio si passa nella successione da $x_1 = 40$ a $x_2 = 60$, allora $v_{ass}(x_1,x_2) = 60-40 = 20$ mentre $v_{rel}(x_1,x_2) = (60-40)/40 = 0{,}5$ e $v_{rel\%}(x_1,x_2) = 100 \times 0{,}5 = 50\%$. D'altra parte, la variazione da 60 a 40 corrisponde a un decremento del 33% (lo si verifichi!). Dunque ci si potrebbe chiedere come è possibile che

nel passaggio 40, 60, 40, corrispondente a una variazione assoluta nulla, la variazione percentuale sia invece 50–33 = 17%: la ragione è evidentemente che il primo 50% è calcolato su 40, mentre il secondo –33% su 60. Ciò suggerisce una considerazione generale:

per evitare questo genere di errori, mentre può essere efficace presentare i risultati di un'elaborazione nella forma di variazioni relative, *le elaborazioni dovrebbero essere effettuate in termini di variazioni assolute*.

Un caso specificamente interessante è quello in cui l'indice j del secondo termine della variazione dipende dall'indice i del primo termine, e in particolare $j = i + 1$, cioè si considerano variazioni di un termine rispetto al successivo della successione. Una volta che la successione $\langle x_i \rangle$ sia data, le variazioni possono allora essere espresse come funzioni di un solo indice, per esempio $v_{ass}(x_i) = x_{i+1} - x_i$. D'altra parte, al variare dell'indice i, i valori $v_{ass}(x_i)$ costituiscono una nuova successione, dipendente da quella di partenza attraverso la trasformazione v_{ass}. Schematicamente:

$$\langle x_i \rangle - v_{ass} \rightarrow \langle v_{ass}(x_i) \rangle$$

(e naturalmente la stessa cosa vale per v_{rel} e $v_{rel\%}$). Ciò mostra che *le variazioni possono essere intese come funzioni di secondo ordine*, che sono applicate a successioni e generano nuove successioni.

Calcolo sulle variazioni

Supponiamo che la variazione relativa tra elementi successivi di una successione sia costante, pari a k, cioè:

$$k = \frac{x_{i+1} - x_i}{x_i}$$

per esempio con $k = 0{,}1$, cioè 10%. Partendo da un valore x_0, ci si chiede quale sarà il valore della successione x_n dopo n passi (per esempio: se un prezzo, inizialmente pari a x_0, cresce del 10% all'anno, quale valore x_n avrà dopo n anni?). Il valore x_1 si ottiene incrementando x_0 del tasso k, cioè $x_1 = x_0 + kx_0$. Se definiamo $k' = 1 + k$ (nell'esempio, dunque, $k' = 1{,}1$), si ha che $x_1 = k'x_0$, $x_2 = k'x_1 = k'2x_0$ e quindi per iterazione:

$$x_n = k'^n x_0$$

(è dunque immediato verificare che se il primo valore della successione è x_1, come è abitudine in matematica, la formula diventa $x_n = k'^{(n-1)} x_1$).

Se, per esempio, $k' = 1{,}1$, $x_0 = 40$, $n = 5$:

x_0	40
x_1	44
x_2	48.4
x_3	53.24
x_4	58.56
x_5	64.42

Ci si può porre utilmente anche il problema inverso: quale tasso di variazione costante k' consente di far passare una successione da un valore iniziale x_0 a un valore finale x_n in n passi? La risposta si ottiene invertendo la formula precedente:

$$k' = \left(\frac{x_n}{x_0}\right)^{1/n}$$

(o anche, nel caso in cui il primo valore della successione sia x_1: $k' = (x_n/x_1)^{1/(n-1)}$).

Per esempio, supponiamo $x_0 = 40$, $x_n = 60$, $n = 5$; allora:

$$\bar{k} = \left(\frac{60}{40}\right)^{1/5} = 1{,}08$$

Per passare da 40 a 60 in 5 passi la successione deve crescere con un tasso costante dell'8% per passo.

Variazioni relative: il problema della classe di riferimento

Consideriamo il seguente problema.

AFFERMAZIONE 1: l'esperienza mostra che il numero di macchine di tipo X che si guastano nel corso di un anno aumenta del 50% se le macchine stesse non sono sottoposte a manutenzione regolare.
AFFERMAZIONE 2: l'esperienza mostra che il 4% delle macchine di tipo X, se opportunamente sottoposte a manutenzione regolare, si guasta nel corso di un anno.
DOMANDA: nell'ipotesi, realistica, che il programma di manutenzione regolare abbia un costo significativo, è opportuno sottoporre le macchine a tale programma o no?
L'AFFERMAZIONE 1 sembra suggerire una risposta affermativa, ma consideriamo i numeri: su 100 macchine con manutenzione, 4 si guastano; senza manutenzione, l'aumento del 50% corrisponde dunque a un totale di 6 macchine che si guastano (infatti $(6-4)/4 = 0{,}5$). Se però teniamo conto anche dell'AFFERMAZIONE 2, le cose assumono una rilevanza un po' diversa: con il programma di manutenzione 96 su 100 macchine funzioneranno, senza il programma, le macchine funzionanti saranno invece 94; il peggioramento è dunque $(96-94)/96 \approx 0{,}02$: vale la pena di sostenere i costi del programma di manutenzione per evitare un peggioramento del 2%?
Naturalmente il discorso sarebbe diverso se, a parità di quanto asserito nell'AFFERMAZIONE 1, la percentuale di guasto per macchine sottoposte a manutenzione regolare fosse per esempio del 20%, corrispondente dunque a una percentuale di 80 macchine su 100 funzionanti con manutenzione e 70 macchine su 100 funzionanti senza, con un peggioramento di $(80-70)/80 \approx 0{,}12$; un peggioramento del 12% giustifica ben di più i costi in questione.

Variazioni in successioni temporali

L'andamento di una successione è caratterizzabile in termini di variazioni non solo attraverso le variazioni assolute e le variazioni relative, ma anche mediante le *variazioni relative in funzione del parametro* della successione stessa, che per dati di origine sperimentale è spesso il tempo, x: $T \to A$, con $T = \{t_1, t_2, ...\}$. In questi casi si valuta dunque la variazione dei valori della successione *in funzione del tempo*, sulla base dell'idea che, per esempio, se $x(t_1) = 2$ e $x(t_2) = 4$, la variazione ha un significato ben diverso se l'intervallo $[t_1,t_2]$ corrisponde a 1 secondo oppure a 1 giorno, benché la variazione assoluta $x(t_2)-x(t_1)$ sia la stessa. Per ogni coppia di valori contigui della successione, $x(t_i)$ e $x(t_{i+1})$, si può considerare perciò la variazione:

$$\frac{x(t_{i+1})-x(t_i)}{t_{i+1}-t_i}$$

Nel caso in cui gli intervalli $[t_i,t_{i+1}]$ abbiano ampiezza costante, $\forall i,\ t_{i+1}-t_i = \Delta t$, l'espressione precedente diventa:

$$\frac{x(t_i+\Delta t)-x(t_i)}{\Delta t}$$

chiamata, come è noto dall'analisi matematica, *rapporto incrementale* della funzione x.
Al variare dell'indice i, l'insieme ordinato di tali variazioni costituisce una nuova successione:

$$x'(t_1)=\frac{x(t_1+\Delta t)-x(t_1)}{\Delta t},\ x'(t_2)=\frac{x(t_2+\Delta t)-x(t_2)}{\Delta t},\ \ldots$$

Sempre l'analisi matematica insegna che, via via che l'intervallo Δt si riduce, la successione $\langle x'(t_i)\rangle$ così ottenuta approssima sempre meglio la *derivata* $x'(t)$ della funzione $x(t)$, che infatti è definita come:

$$x'(t)=\lim_{\Delta t \to 0}\frac{x(t+\Delta t)-x(t)}{\Delta t}$$

Per un certo valore $\Delta t > 0$ fissato, la successione $\langle x'(t_i)\rangle$ può dunque essere intesa come la "versione discreta" della derivata, calcolata sulla successione $\langle x(t_i)\rangle$, e come tale descrive la *velocità* della successione $\langle x(t_i)\rangle$ stessa (dovrebbe essere chiaro che su una successione, che è una funzione a dominio discreto, la derivata come tale non è definita). E' importante notare che i valori della successione $\langle x'(t_i)\rangle$ sono calcolabili numericamente (e in modo molto semplice, in effetti) a partire dai valori della successione $\langle x(t_i)\rangle$, anche in situazioni in cui non sia nota l'espressione (se pure esiste) che definisce analiticamente $\langle x(t_i)\rangle$: ciò mette in evidenza la notevole generalità di questa *strategia di soluzioni numeriche* per l'analisi di successioni.
E infatti la logica che consente di generare $\langle x'(t_i)\rangle$ a partire da $\langle x(t_i)\rangle$ è replicabile, e consente per esempio ottenere la velocità della velocità (cioè l'accelerazione) di $\langle x(t_i)\rangle$, $x''(t_i)=[x'(t_i+\Delta t)-x'(t_i)]/\Delta t$, e così via.

In certi casi potrebbe poi essere nota non direttamente la successione $\langle x(t_i)\rangle$ ma $\langle x'(t_i)\rangle$, successione che dunque descrive la variazione di $\langle x(t_i)\rangle$ nel tempo, cioè appunto la sua velocità. Si può comunque ricavare semplicemente $\langle x(t_i)\rangle$ da $\langle x'(t_i)\rangle$, risolvendo perciò un problema inverso. Da:

$$x'(t_1)=\frac{x(t_1+\Delta t)-x(t_1)}{\Delta t}=\frac{x(t_2)-x(t_1)}{\Delta t}$$

si ottiene iterativamente:

$$x(t_2)=x(t_1)+x'(t_1)\Delta t\ ,\ \ x(t_3)=x(t_2)+x'(t_2)\Delta t\ ,\ ...$$

Per sostituzione:

$$x(t_3)=x(t_1)+x'(t_1)\Delta t+x'(t_2)\Delta t=x(t_1)+[x'(t_1)+x'(t_2)]\Delta t$$

e così via, e quindi in generale:

$$x(t_{n+1})=x(t_1)+\sum_{i=1}^{n}x'(t_i)\Delta t$$

La successione $\langle x(t_i)\rangle$ così calcolata sulla successione $\langle x'(t_i)\rangle$ può dunque essere intesa come la "versione discreta" della funzione integrale:

$$x(t)=\int_{t_1}^{t}x'(\tau)d\tau$$

della funzione $x'(t)$.

La sintesi è interessante nella sua semplicità: nel caso discreto, la derivata corrisponde a una differenza e l'integrale a una somma, e come differenza e somma sono funzioni inverse l'una dell'altra, così derivata e integrale sono funzioni (di funzioni, dunque funzioni di secondo ordine) inverse l'una dell'altra.

Sintesi di successioni di dati sperimentali

In molti casi, non è necessario (o non è possibile) mantenere l'informazione completa sulla composizione di una successione, ed è invece appropriato (o è necessario) esprimere l'informazione disponibile in forma sintetica. A questo proposito si presentano due possibilità:

- in certi casi è nota un'espressione analitica che descrive tutti gli elementi della successione; per esempio, se la successione è $\langle x_i\rangle = \langle 0, 2, 4, 6, 8, ...\rangle$ allora si può scrivere $\forall i,\ x_i = 2i$; poiché da tale espressione è ricostruibile l'intera successione, in questa sintesi non si è persa alcuna informazione: si tratta di una *compressione lossless*;
- quando invece un'espressione analitica con queste caratteristiche non è nota, la sintesi implica una perdita di informazione; si pone allora il problema di scegliere un criterio di sintesi appropriato, in grado di ridurre la quantità di informazione sulla successione mantenendo nello stesso tempo l'informazione considerata rilevante: si cerca un criterio di *compressione lossy*.

Si noti che nella distinzione precedente non ha alcun ruolo specifico il "caso": se la successione è casuale un'espressione analitica che ne consenta la ricostruzione non può essere nota e quindi rientra nel secondo caso; ma una successione può rientrare nel secondo caso pur non avendo nulla di casuale: è sufficiente che un'espressione analitica che descriva la successione non sia nota.

Un semplice esempio di sintesi del secondo tipo: data la successione $\langle x_i\rangle = \langle 26, 24, 30, 24, 21, 25\rangle$ dei voti che un certo studente ha preso nei suoi esami universitari, si potrebbe mantenere solo il voto minimo $x_{min} = 21$ e il voto massimo, $x_{max} = 30$; in questo modo si passa dai sei numeri della successione a due numeri, ma evidentemente si perde informazione: non si sa più né quanti né quali voti sono stati presi, e tantomeno si conosce l'ordine con cui sono stati presi.

La statistica fornisce degli strumenti per effettuare queste "sintesi del secondo tipo", spesso realizzate a partire non dall'intera successione, chiamata in questo contesto *popolazione*, ma da sue sotto-successioni, chiamate *campioni*.

Popolazioni, campioni e campionamento

Quando una popolazione è molto numerosa diventa praticamente necessario operare su un suo campione (in inglese *sample*, e non *standard*, che significa invece "campione di misura", oltre che "norma tecnica"), cioè effettuare un'operazione preliminare di *campionamento* (*sampling*). D'altra parte, un campione è correttamente utilizzabile al posto della popolazione da cui è estratto solo se è sufficientemente *rappresentativo* della composizione della popolazione stessa.

Per esempio, un'università vuole conoscere la situazione retributiva dei suoi laureati recenti, che supponiamo siano 1000. A questo scopo predispone un questionario che invia a 200 laureati, ottenendo 100 risposte.

- Con quale criterio dovrebbero essere scelti i 200 laureati a cui inviare il questionario?
- Le 100 risposte ottenute possono essere considerate rappresentative dell'intera popolazione?

Un campione non rappresentativo viene detto "distorto", o anche "polarizzato" (in inglese *biased*). Naturalmente un campione dovrebbe essere usato al posto della popolazione da cui è estratto solo quando si hanno buone ragioni per ritenerlo non distorto.
E così ci si dovrebbe chiedere come selezionare i 200 laureati destinatari del questionario in funzione del tipo di occupazione (se fosse vero che i consulenti guadagnano più degli impiegati, non sarebbe una buona idea interpellare solo consulenti o solo impiegati), dell'area geografica di lavoro (potrebbe essere che gli stipendi non siano ovunque gli stessi), e così via.
Dovrebbe essere perciò chiaro che per costruire un campione non polarizzato le competenze di statistica non bastano: occorre anche conoscere a fondo il fenomeno oggetto dello studio.
Popolazioni e campioni sono dunque ciò da cui la statistica trae origine. Questo chiarisce l'ipotesi alla base della possibilità stessa di sviluppare modelli statistici: *è necessario che le entità in considerazione, pur diverse tra loro, siano trattate come elementi di insiemi su cui sia significativo e utile produrre informazione sintetica*. Nelle situazioni in cui, al contrario, "ogni caso è unico", nessuna tecnica statistica può essere impiegata.

Schemi di campionamento

Data una popolazione, esistono due modalità fondamentali per ottenere dei campioni da essa. Immaginando che la popolazione sia costituita dalle palline contenute di un'urna, caratterizzate per esempio ognuna da un colore (non è evidentemente necessario che palline diverse abbiano colori diversi), si costruisce un campione estraendo via via palline dall'urna e trascrivendo di ognuna il colore, ripetendo l'operazione di estrazione per tante volte per quanti elementi dovrà avere il campione da costruire. Le due modalità sono distinte in base a ciò che si fa di ogni pallina estratta:

- in un caso la pallina viene reinserita nell'urna; chiamiamo *campionamento con ripetizione* (o "con reintroduzione") questo schema;
- nell'altro caso la pallina non viene reinserita nell'urna; chiamiamo *campionamento senza ripetizione* (o "senza reintroduzione") questo schema.

Nel caso di campionamento con ripetizione, il campione potrà contenere un numero arbitrario di elementi, anche maggiore del numero di elementi della popolazione, e potrà contenere elementi ripetuti; al contrario, nel caso di campionamento senza ripetizione, il campione potrà contenere al più tanti elementi quanti ne sono presenti nella popolazione, e non conterrà elementi ripetuti.

I fogli di calcolo mettono a disposizione strumenti espliciti per la costruzione di campioni solo attraverso estensioni / addon. Non è però difficile sviluppare quanto serve per tale costruzione.
Cominciamo dal più semplice caso del campionamento con ripetizione. Supponiamo che la popolazione contenga `num` elementi nel vettore `pop`. In ogni cella in cui dovrà essere contenuto un elemento del campione, dobbiamo generare un numero intero casuale tra `1` e `num`, da usare come puntatore per recuperare il contenuto della cella corrispondente nel vettore `pop`. La formula da usare è dunque:

```
=INDEX(pop,RANDBETWEEN(1,num))
```

Solo un poco più complesso è il caso del campionamento senza ripetizione. Generiamo, prima di tutto, un vettore `ran` di numeri casuali (con la solita funzione `RAND()`), tanti quanti elementi dovrà contenere il campione (e quindi al più `num`): possiamo certamente supporre che essi siano tutti diversi. Se per esempio un numero casuale è nella cella `E2`, con la formula `RANK(E2,ran)` otterremo l'indice ordinale, tra `1` e `num`, del valore in `E2` rispetto al vettore. Poiché gli indici ordinali ottenuti mediante i numeri casuali in `ran` saranno tutti diversi, possiamo usarli, analogamente a quanto fatto sopra, come puntatori, mediante la formula:

```
=INDEX(pop,RANK(E2,ran))
```

pop	*camp con rip*	ran	*camp senza rip*
blu	giallo	0.7401342811	blu
rosso	bianco	0.8588919064	rosso
verde	blu	0.5291483013	nero
blu	nero	0.2366397262	blu
giallo	bianco	0.3404931712	giallo
bianco	verde	0.8685777062	blu
verde	bianco	0.7146495625	giallo
nero	giallo	0.8158955504	verde
giallo	blu	0.6854898413	bianco
blu	giallo	0.6341535759	verde
num	blu		
10	blu	=RAND()	=INDEX(pop,RANK(E11,ran))
	giallo		
	rosso		
	blu		
	giallo		
	giallo		
	=INDEX(pop,RANDBETWEEN(1,num))		

Trasformazioni di campioni

Una volta che un campione è stato acquisito, può essere utile operare sui suoi elementi in modo che il campione trasformato soddisfi criteri dati. Vediamo qualche esempio, direttamente in riferimento agli strumenti dei fogli di calcolo.

Supponiamo che il campione sia nel vettore `camp`.
Traslare rigidamente (*shift*) gli elementi del campione in modo che il valore minimo sia `0`: ogni elemento `x` deve essere trasformato in `=x-min(camp)`.
Traslare e riscalare (*shift and scale*) gli elementi del campione in modo che il valore minimo sia `0` e il valore massimo `1`: ogni elemento `x` deve essere trasformato in `=(x-min(camp))/(max(camp)-min(camp))`.
Limitare il valore massimo degli elementi del campione a una soglia `y` data: ogni elemento `x` deve essere trasformato in `=min(x,y)`, espressione equivalente a `if(x<y,x,y)`.

Nel caso di campioni di grandi dimensioni, problemi spesso critici sono relativi a dati mancanti (corrispondenti nel foglio di calcolo a celle vuote o il cui contenuto è per esempio `NA`, dall'inglese "*not available*") o a dati errati, di cui ci si può accorgere quando i valori sono palesemente incongruenti (un valore negativo per una lunghezza, un valore di qualche ordine di grandezza maggiore di tutti gli altri, ... In situazioni di questo genere è utile trasformare il campione in modo da "ripulirlo", ma le tecniche da adottare sono da scegliere caso per caso.

Distribuzioni

Un criterio molto generale di sintesi di una successione $\langle x_i \rangle$ con insieme supporto A si basa su una semplice procedura in due passi:

1. si costruisce una partizione di A, cioè un insieme di sottoinsiemi $\{C_j\}$ di A che siano mutuamente esclusivi (se $i \neq j$ allora $C_i \cap C_j = \varnothing$) ed esaustivi ($\cup_j C_j = A$); in questo modo, ogni elemento della successione è incluso in uno e un solo sottoinsieme C_j;
2. per ogni elemento C_j della partizione, si conta il numero n_j di elementi della successione $\langle x_i \rangle$ che appartengono al sottoinsieme C_j.

Per esempio, data ancora la successione dei voti $\langle x_i \rangle = \langle 26, 24, 30, 24, 21, 25 \rangle$ sull'insieme $A = \{18, ..., 30\}$:

- se la partizione è $\{C_1 = \{18, ..., 21\}, C_2 = \{22, ..., 25\}, C_3 = \{26, ..., 30\}\}$ allora $n_1 = 1, n_2 = 3, n_3 = 2$;
- se invece la partizione è $\{C_1 = \{18, ..., 24\}, C_2 = \{25, ..., 30\}\}$ allora $n_1 = 3, n_2 = 3$.

I valori n_j sotto detti *frequenze assolute*, e si chiamano *frequenze relative* i valori $f_j = n_j/n$, essendo n il numero degli elementi della successione (si noti che le frequenze relative sono tali per cui $f_j \in [0,1]$ e $\Sigma_j f_j = 1$). Dalle frequenze relative si ottengono poi le *frequenze percentuali* mediante $f_{\%j} = 100 n_j/n$.

Le successioni $\langle n_j \rangle$, $\langle f_j \rangle$ e $\langle f_{\%j} \rangle$ si chiamano *distribuzioni* sulla partizione $\{C_j\}$, rispettivamente a frequenze assolute, a frequenze relative e a frequenze percentuali, e gli elementi C_j si dicono *categorie* della distribuzione.
Una distribuzione può essere presentata come una tabella a due righe, categorie e frequenze, con un numero di colonne pari al numero di categorie impiegate nella partizione. Nell'esempio della prima partizione, a tre categorie, la distribuzione a frequenze assolute è:

$$\begin{bmatrix} C_j \\ n_j \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 \\ 1 & 3 & 2 \end{bmatrix}$$

e la distribuzione a frequenze relative è:

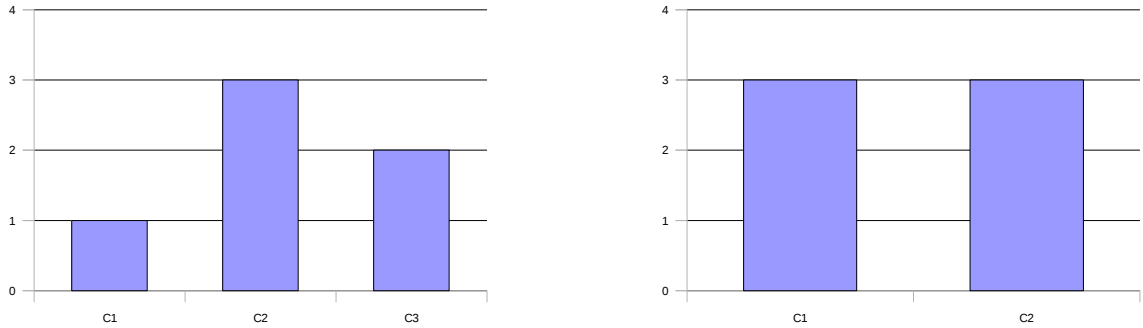
$$\begin{bmatrix} C_j \\ f_j \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 \\ 1/6 & 1/2 & 1/3 \end{bmatrix}$$

Dato un insieme supporto A, la partizione più grezza è quella che contiene un'unica categoria $C_1 = A$, mentre la partizione più fine è quella che contiene tante categorie quanti elementi ci sono in A, e quindi tale che ogni categoria contiene uno e un solo elemento di A (stiamo naturalmente supponendo di trattare con un insieme supporto discreto).
La distinzione tra l'insieme supporto e la partizione più fine è dunque solo formale: d'ora in poi potremo quindi trattare sempre di categorie e distribuzioni su partizioni dell'insieme supporto.

Distribuzioni e diagrammi

Una distribuzione può essere visualizzata mediante un istogramma; nei due casi precedenti:

$$\begin{bmatrix} C_1 & C_2 & C_3 \\ 1 & 3 & 2 \end{bmatrix} \qquad \begin{bmatrix} C_1 & C_2 \\ 3 & 3 \end{bmatrix}$$

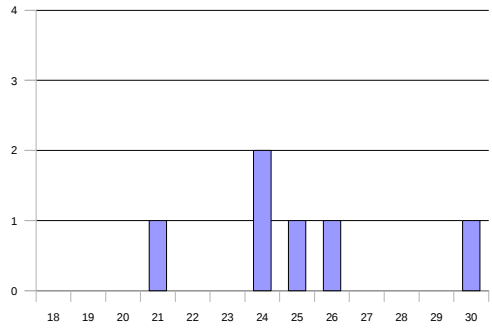


Nel caso in cui sia considerato significativo rappresentare i valori delle frequenze relative, eventualmente percentualizzati, l'informazione della distribuzione può essere visualizzata anche mediante un diagramma a torta (in inglese *pie chart*); in questi due casi:

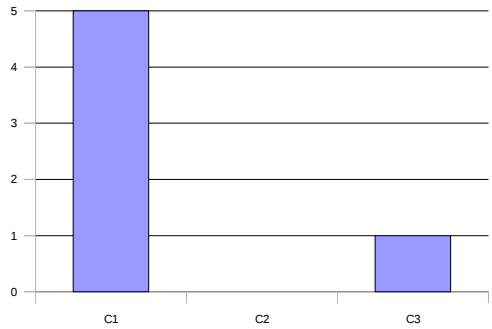
Distribuzioni: significatività empirica

La significatività empirica di una distribuzione dipende da un'appropriata scelta delle categorie:

- se il numero di categorie è elevato, analogo o addirittura superiore al numero degli elementi della successione, la distribuzione potrebbe risultare non significativa perché troppo "sparsa"; nell'esempio, l'istogramma corrispondente alla partizione più fine, in cui ogni categoria contiene un solo elemento, $\{C_1 = \{18\}, C_2 = \{19\}, \ldots\}$ è:



- se la partizione definita sull'insieme A è costituita da sottoinsiemi di cardinalità molto diversa, la distribuzione non è significativa perché "sbilanciata"; per esempio, se le categorie sono $\{C_1 = \{18,...,28\}, C_2 = \{29\}, C_3 = \{30\}\}$ l'istogramma è:



Si noti che una non appropriata categorizzazione impiegata nella raccolta dei dati può produrre degli effetti di distorsione nelle risposte; per esempio, data la domanda "quante ore al giorno lavori al calcolatore?" e date le due categorizzazioni: {fino a mezz'ora, da mezz'ora a un'ora, da un'ora a un'ora e mezzo, ..., più di due ore e mezzo} e {fino a due ore, da due ore a due ore e mezzo, ... più di quattro ore}, l'uso di una o dell'altra categorizzazione tipicamente modifica le risposte.
Un semplice criterio (teorico) di rappresentatività per un campione di una popolazione: un campione è rappresentativo se la sua distribuzione è almeno "abbastanza simile" a quella della popolazione da cui è estratto.