

CI SONO NUMERI E NUMERI

Tutti i giorni e tutto il giorno siamo circondati da numeri. Così si esprime chi sostiene che il caffè abbassa la pressione e chi ci dice le attese di vita che possiamo avere. Tutto questo è frutto delle scienze statistiche. Un frutto che non di rado è semplificato, a volte frainteso e in certi casi anche banalizzato dai comunicatori. Può quindi essere opportuno conoscere un po' meglio la statistica. Lo facciamo con il professor Eugenio Brentari che non solo è uno statistico, ma anche uno dei grandi divulgatori della materia

Non tutti i numeri sono uguali e non tutti i numeri possono essere trattati nello stesso modo.

Quando si analizza una caratteristica di un certo fenomeno, questa si può presentare in diverse forme solitamente indicate con il termine *modalità*. Se, per esempio, si stesce analizzando la provenienza di un vino italiano, le modalità potrebbero essere le venti regioni italiane.

Considerato un soggetto, un prodotto, un vino ovvero un'unità *statistica*, su di essa possono essere analizzate una o più caratteristiche, rilevate secondo criteri di classificazione ben precisi.

Queste caratteristiche (*caratteri*) non sono sempre uguali e nell'analizzare questi dati sono necessari approcci diversi.

Un qualsiasi manuale di statistica di base, nella prima parte del volume fa una distinzione di caratteri e di scale di misurazione: si suggerisce, per esempio, il volume di Michele Zenga, *Introduzione alla Statistica Descrittiva*, Vita e Pensiero, Milano, 1998 oppure quello di Domenico Piccolo, *Statistica*, Il Mulino, Bologna, 2010 che entrambi trattano questo argomento nel secondo capitolo.

CARATTERI E SCALE DI MISURAZIONE

La prima distinzione è sui caratteri: *qualitativi* e *quantitativi*. I primi hanno modalità che possono essere espresse con vocaboli (maschio o femmina; nubile, sposata, divorziata, ...) mentre i caratteri quantitativi hanno modalità che sono espressione di una quantità, di una misura (numero di figli, reddito, tempi di percorrenza ecc.).

La seconda distinzione è sulle scale di misurazione delle modalità di un carattere.

Quando non è possibile stabilire nessun ordinamento tra le modalità del carattere si parla di *scale nominali*. Le regioni di provenienza di un vino possono essere elencate in diversi modi (alfabetico, per estensione, per numero di bottiglie prodotte, ...) ma sono elenchi convenzionali.

Quando invece è possibile stabilire un ordinamento tra le modalità del carattere allora si parla di *scale ordinali*. L'intensità del colore di un vino rosso è una scala tipicamente ordinale, dal colore meno intenso a quello più carico.

Il livello di misurazione più elevato si raggiunge con le *scale a intervalli* e le *scale di rapporti*. Si tratta di scale che utilizzano dei numeri per indicare delle quantità





reali. In queste scale è possibile non solo ordinare le modalità, ma anche sommare, sottrarre, dividere, ... La differenza sostanziale tra le due scale è che la scala a intervalli ha uno zero arbitrario mentre nella scala di rapporti lo zero è una quantità reale. Solo in quest'ultimo caso ha senso fare dei rapporti tra modalità, mentre non ha senso, per esempio, dire che l'anno 2000 è il doppio dell'anno 1000, proprio perché in questo caso lo zero è convenzionalmente legato alla nascita di Gesù Cristo.

LE FREQUENZE

A seconda del carattere e della scala di misurazione esistono strumenti diversi per sintetizzare i dati rilevati su una certa unità statistica. Per una scala nominale sarà, per esempio, possibile determinare se una modalità si presenta più frequentemente delle altre.

Considerando le venti regioni italiane e un lotto di 200 bottiglie di vino prodotte in Italia, si può determinare quante bottiglie siano state imbottigliate in una certa regione. Questo valore prende il nome di *frequenza assoluta* e può essere indicato con n_i . Se, per esempio, 24 bottiglie provengono dal Friuli-Venezia Giulia allora $n_i = 24$.

È chiaro che se si sommano le frequenze assolute di tutte le regioni allora il totale delle bottiglie sarà ancora pari a 200.

REGIONI	N_i	F_i	%
Abruzzo	4	0,02	2
Emilia-Romagna	22	0,11	11
Friuli-Venezia Giulia	24	0,12	12
Basilicata	2	0,01	1
...
...
Totale	200	1	100

Tabella 1: esempio di tabella riassuntiva per le varie frequenze

Indicato con n il numero totale delle unità statistiche considerate (nell'esempio considerato le 200 bottiglie) e utilizzando il simbolo S (che si chiama *sommatoria* e indica appunto fare la somma dei valori), allora

$$\sum_{i=1}^k n_i = n$$

La precedente formula può essere letta nel seguente modo: sommando dalla prima alla k -ma modalità (nell'esempio delle regioni italiane $k = 20$) le relative *frequenze assolute* si ottiene il totale delle unità osservate n (nel caso delle bottiglie $n = 200$). Dividendo ciascuna frequenza assoluta per il totale n si ottengono le *frequenze relative* f_i

$$f_i = \frac{n_i}{n}$$

In questo caso la somma delle *frequenze relative* sarà pari a 1

$$\sum_{i=1}^k f_i = 1$$

Moltiplicando per 100 le frequenze relative si ottengono le *percentuali* (%) a tutti note. Nell'esempio si può affermare che il 12% delle bottiglie considerate proviene dal Friuli-Venezia Giulia

$$f_i \cdot 100 = \frac{24}{200} \cdot 100 = 0,12 \cdot 100 = 12\%$$

Una volta determinate le varie frequenze è possibile costruire una tabella che riassume quanto fatto (tabella 1).

Per chi è pratico di *Excel* e ha a disposizione tutti i singoli valori può determinare le frequenze assolute con l'istruzione = *Frequenza* (non semplicissima) oppure con le Tabelle Pivot (molto semplice, una volta capito come funziona). Ma questo potrà essere materiale per una prossima puntata.

EUGENIO BRENTARI

Università di Brescia