

PARTE PRIMA

Elementi di statistica per la ricerca sociale

1. Introduzione al metodo statistico

Di cosa parleremo

In questo capitolo descriveremo in generale il metodo statistico, specificandone i maggiori campi di applicazione, gli obiettivi, la prassi metodologica e la terminologia di base. Saranno poi trattate sia le rappresentazioni tabellari, diffusissime per la presentazione dei risultati di ogni indagine statistica, che le rappresentazioni grafiche, strumenti descrittivi di elezione nei casi in cui non è necessaria una precisione elevata nella rappresentazione dei dati a vantaggio di un'immediata valutazione sintetica della distribuzione dei dati stessi; infine, verranno introdotti alcuni parametri di sintesi che permettono una valutazione dei risultati, quali le medie di posizione e di calcolo. Le formule descritte, determinanti per l'elaborazione dei dati osservati, sono state ridotte a quelle più utilizzate e permettono, al lettore con un minimo bagaglio matematico, di approfondire le tecniche di calcolo maggiormente usate.

1) Definizione del metodo

Cos'è il metodo statistico. Il metodo statistico è un insieme di teorie e di metodi che permette di raccogliere, descrivere e presentare insiemi di dati, allo scopo di ricavarne agevolmente informazioni che ne sintetizzino alcune caratteristiche e che forniscano previsioni affidabili su insiemi più grandi rispetto ai dati oggetto di osservazione.

Statistica: scienza, strumentale ad altre, concernente la determinazione dei metodi scientifici da seguire per raccogliere, elaborare e valutare i dati riguardanti l'essenza di particolari fenomeni di massa.

La parola **statistica** (dal latino *status*) deriva dal fatto che il suo utilizzo iniziale riguardava la raccolta, l'elaborazione e la diffusione di dati che descrivevano aspetti quantitativi delle caratteristiche specifiche di una nazione.

Obiettivo della statistica. L'obiettivo della statistica, individuabile dalla definizione, è di duplice natura: *sintetizzare* e *permettere estensioni e generalizzazioni*. *Sintetizzare* significa predisporre i dati raccolti in una forma tale da rappresentarli efficacemente, ovvero con relativamente pochi numeri e/o indici, in modo da comprendere le caratteristiche essenziali dei fenomeni analizzati attraverso i dati raccolti. La sintesi viene dunque incontro all'esigenza di semplificazione. I metodi e le tecniche sviluppati per soddisfare questa finalità appartengono alla statistica descrittiva. A volte, a causa soprattutto della limitata disposizione di risorse economiche e temporali, si pone l'esigenza di *estendere* il risultato delle analisi effettuate su gruppi limitati di unità statistiche (*il campione*) all'intero insieme di appartenenza (*universo statistico*). Le tecniche e i metodi che si usano in questo caso rappresentano il contenuto della statistica inferenziale.

I campi di applicazione. I campi di applicazione del metodo statistico sono tutti quelli in cui si presentano fenomeni ripetitivi su insiemi numerabili di elementi in cui occorre prendere decisioni in condizioni di incertezza (ricerca scientifica, farmacologia, interventi di natura economica e finanziaria ecc.). Anche se non ce ne accorgiamo, siamo bersagliati quotidianamente da dati di natura statistica (ad esempio i risultati di sondaggi) e inoltre facciamo uso della statistica frequentemente: ad esempio, quando facciamo previsioni sul tempo di permanenza di un autobus.

Esempi di applicazioni statistiche:

- valutazione dell'efficacia di un farmaco;
- valutazione della migliore produttività di sementi diverse di cereali;

- dati su nascite, matrimoni, morti su un certo territorio;
- dati sull'inflazione, sull'andamento dei prezzi.

2) Il metodo statistico

Affinché le informazioni deducibili dai dati rilevati possano essere correttamente utilizzate è necessario sapere:

- la terminologia;
- come organizzare un'indagine statistica;
- come rappresentare e analizzare i dati;
- quali sono gli indici che li caratterizzano;
- quali conclusioni trarre.

Un aspetto importante dell'indagine statistica è costituito dalle **fonti dei dati**, che devono essere scelte con criteri di affidabilità e veridicità indispensabili per raggiungere gli obiettivi del metodo. Per ottenere i dati necessari si può:

- predisporre ed eseguire esperimenti;
- condurre un sondaggio;
- effettuare uno studio sul campo;
- ricorrere a fonti accreditate di dati, pubbliche o private che hanno il compito di fornire i dati, reperiti con uno dei metodi precedenti.

Nei casi in cui sia richiesta una tecnica più efficace, alternativa sia al questionario che all'intervista esterna, si fa ricorso alla cosiddetta *osservazione partecipante* che prevede – allo scopo di rendere il più possibile oggettive le osservazioni – l'inserimento e la parziale integrazione nei gruppi, costituiti ovviamente in questi casi esclusivamente da persone, da parte di chi deve condurre la raccolta dei dati.

Un posto importantissimo nel panorama dei fornitori di dati occupa in Italia l'**ISTAT**, Istituto Nazionale di Statistica (sito www.istat.it) che si occupa, tra l'altro, di organizzare i censimenti della popolazione, dell'industria e commercio e dell'agricoltura. Altre attività di rilevazione dell'ISTAT sono: i movimenti migratori della popolazione, la sa-

nità, l'istruzione, la cultura, il clima, il turismo, la pesca e la caccia, i prezzi, il lavoro, i bilanci familiari ecc.

La terminologia. Si definisce **carattere (o variabile)** un qualsiasi aspetto della realtà (ad esempio reddito, titolo di studio ecc.) suscettibile di assumere **valori** diversi, rilevabili attraverso l'osservazione. I modi con cui si presentano questi valori si chiamano **modalità**, mentre i soggetti nei quali sono stati osservati si definiscono **unità statistiche**, il cui insieme si indica con l'espressione **popolazione statistica** o **universo statistico**.

Le modalità di una variabile vengono definite a priori.

Esempi di carattere statistico:

- il carattere «sesso» si manifesta con le modalità Maschi e Femmine;
- il carattere «numero di figli» si manifesta con modalità espresse da numeri interi maggiori o uguali a zero;
- il carattere «titolo di studio» si esprime con modalità espresse da descrizioni, diverse da Paese a Paese, ma che individuano in maniera univoca il tipo di studi inerenti al titolo connesso; esempi: *licenza elementare, diploma di scuola superiore*.

Nel caso in cui si studi un carattere per volta si parlerà di statistica **univariata**, mentre nel caso si pongano in relazione tra loro due o più caratteri statistici sulla stessa popolazione si parlerà rispettivamente di statistica **bivariata** o **multivariata**.

Tipi di caratteri. I **caratteri statistici** possono essere suddivisibili sulla base di diversi punti di vista. Secondo la modalità di rappresentazione, i caratteri si distinguono in:

- **qualitativi** o **mutabili**: si esprimono attraverso attributi non numerici, ad esempio colore degli occhi, colore dei capelli ecc.; in tal caso, i dati statistici rilevati formano delle **serie** cosiddette **scnesse**;
- **quantitativi**: le cui modalità si esprimono attraverso numeri; ad esempio numero di figli, reddito annuo; in tal caso i dati statistici rilevati formano una **seriazione**.

I **caratteri quantitativi** possono a loro volta essere distinti, osservando l'insieme dei numeri con cui vengono rappresentati, in:

- **discreti**, quando le modalità sono solo alcuni numeri, in genere i numeri naturali o loro sottoinsiemi;
- **continui**, quando le modalità sono i numeri reali o loro sottoinsiemi.

Un'ulteriore classificazione dei caratteri si può effettuare in base alla scala di misurazione con cui si distinguono tra loro le modalità del carattere osservato. Secondo questo criterio, le scale di misura si classificano in:

- **scala nominale**; in questo tipo di scala il carattere, che è sempre qualitativo, si esprime secondo delle griglie di uso comune. Un esempio è il colore degli occhi: azzurro, marrone, verde, nero ecc.;
- **scala ordinale**; in questo caso il carattere è qualitativo ma le sue modalità sono suscettibili di essere poste in ordine (crescente o decrescente); ad esempio il titolo di studio: licenza elementare, licenza media, laurea;
- **scala ad intervalli**; in tal caso il carattere è quantitativo ma consente solo confronti per differenza tra le modalità espresse. Come esempio si può citare la temperatura. Infatti ha senso parlare solo di differenze di temperatura ma non, ad esempio, di rapporti tra i numeri che le esprimono.
- **scala di rapporti**; il carattere è quantitativo e sono permesse tutte le operazioni aritmetiche, incluso il rapporto, tra le modalità con cui viene espresso; ad esempio peso, altezza, reddito ecc.

Utilizzo di codici nella rappresentazione dei dati statistici. La raccolta ed elaborazione dei dati viene facilitata mediante l'utilizzo di codici.

In particolare si utilizzano opportuni codici per:

- le modalità delle variabili;
- identificare il/i questionario/i (progressivi);
- identificare la mancata informazione, il rifiuto di rispondere ecc., allo scopo di renderne più agevole la valorizzazione nei questionari.

Il campione. Quando l'universo statistico non può essere studiato completamente si ricorre ad un **campione** che sia in grado di fornire una rappresentazione di un insieme più vasto (sottoinsieme di una «popolazione» statistica). Per poter effettuare in modo efficiente la selezione dei dati oggetto del campione occorre tener conto di alcune caratteristiche che esso deve possedere, sia di tipo qualitativo (rappresentatività dell'intera popolazione, località, temporalità ecc.) che quantitativo (numerosità del campione). Infine, la pianificazione della raccolta dei dati deve tener presente fattori economici sia di tempo che di spesa, che determinano oggettivamente la confidenza delle analisi statistiche effettuate successivamente all'elaborazione dei dati stessi.

Le frequenze. Si definisce frequenza assoluta il numero di volte che una certa modalità si manifesta nella popolazione di riferimento. La distribuzione di frequenze nelle varie modalità descrive come il fenomeno in esame si manifesta nella popolazione o nel campione (di solito indicata con N).

Altre frequenze usate in statistica sono:

- **frequenza relativa**, definita come rapporto tra la frequenza assoluta di ciascuna modalità ed il numero di elementi costituenti la popolazione statistica;
- **frequenza percentuale, assoluta o relativa**, che esprime in termini percentuali rispettivamente la frequenza assoluta e la frequenza relativa di ciascuna modalità osservata;
- **frequenza cumulata**: può essere definita solo **per caratteri di tipo quantitativo** oppure per **caratteri qualitativi ordinabili**; per ogni modalità del carattere, essa si calcola sommando alla frequenza assoluta della modalità in esame, le frequenze assolute di tutte le modalità precedenti, già ordinate in senso crescente.

La successione delle frequenze che corrispondono alle modalità di un carattere qualitativo viene chiamata **serie statistica**. La distribuzione di caratteri di tipo **quantitativo** viene invece chiamata **seriazione**. Nel caso di *caratteri quantitativi continui*, le frequenze si riferiscono non ad una modalità espressa da un numero ma ad **intervalli** di valo-

ri, ognuno dei quali include l'infinità dei possibili valori corrispondenti ai numeri reali compresi in ciascun intervallo (almeno virtualmente). Ad esempio, la variabile «peso di un individuo» estrapolata dalla classe 50-60 kg comprende uno qualsiasi dei numeri reali esistenti tra 50 e 60. In questo caso, si procede ad un **raggruppamento in classi** corrispondenti agli intervalli di valori prescelti.

Ciascuna classe ha:

- un *limite inferiore* corrispondente al valore più piccolo che può effettivamente appartenere alla classe;
- un *limite superiore* corrispondente al valore più grande che può effettivamente appartenere alla classe;
- un *valore centrale* corrispondente al valore che è esattamente al centro tra il limite superiore e quello inferiore.

3) La rappresentazione dei dati: le tabelle e i grafici

Le tecniche di rappresentazione dei dati raccolti fanno uso principalmente di tabelle e di grafici.

Le due tecniche non sono necessariamente alternative; la prima privilegia l'esposizione numerica in forma sintetica, e la seconda la prospettiva visiva in forma d'immagine, meno accurata ma più immediata rispetto alle differenze tra i dati.

Le tabelle più usate sono le cosiddette tabelle di frequenza. Queste elencano le modalità (eventualmente raggruppate in classi o categorie di valori) insieme alle frequenze assolute (numerosità degli elementi per ciascuna modalità) e alle frequenze relative, ricavate dalle frequenze assolute divise per il totale delle osservazioni («N», che costituisce la popolazione statistica). I grafici sono, peraltro, molto utilizzati nei metodi di ricerca nel campo delle scienze sociali, dove l'effetto qualitativo visivo fornisce indicazioni efficaci, vista la numerosità delle osservazioni, evitando spesso il ricorso a rappresentazioni numeriche, la cui elaborazione è svolta, in generale, da comuni programmi standard per personal computer.

Esempio di tabella statistica semplice

Voto (modalità)	Allievi (frequenza)
4	3
5	5
6	8
7	5
8	3

Nella prima colonna si pone la variabile osservata (in questo caso *voto*) nelle sue modalità (le righe della tabella), mentre nella seconda colonna si pone la frequenza assoluta con cui ciascuna modalità è stata rilevata. Talvolta nella seconda colonna di una tabella statistica viene riportata direttamente la frequenza relativa o, in taluni casi, quando queste sono molto piccole e impongono l'uso dei decimali, si presenta direttamente nella seconda colonna la frequenza relativa percentuale.

Esempio di tabella statistica con esposizione di vari tipi di frequenze rilevate

Voti	Allievi (frequenza)	Frequenza relativa	Frequenza relativa %
4	2	0,09	9
5	4	0,18	18
6	8	0,36	36
7	5	0,23	23
8	3	0,14	14
Totale	22	1	100

La somma delle frequenze relative è sempre uguale a 1, mentre la somma delle frequenze relative percentuali è sempre uguale a 100.

Esempio di tabella con raggruppamento in classi della variabile

Variabile statistica «classi di età»	Frequenza «numero di operai»
20 - 30	220
31 - 40	185
41 - 50	120
51 - 60	25

Particolarmente efficace risulta la rappresentazione tabellare quando occorre mettere a confronto i risultati di due indagini statistiche sullo stesso carattere eseguite in tempi o luoghi diversi oppure, come nell'esempio seguente, su campioni diversi:

Confronto tra distribuzioni

Voti 1° A	Allievi (frequenza)	Frequenza relativa	Frequenza relativa %
4	2	0,09	9
5	4	0,18	18
6	8	0,36	36
7	5	0,23	23
8	3	0,14	14
Totale	22	1	100

Voti 1° B	Allievi (frequenza)	Frequenza relativa	Frequenza relativa %
4	4	0,15	15
5	5	0,19	19
6	9	0,33	33
7	5	0,18	18
8	4	0,15	15
Totale	27	1	100

Naturalmente il confronto, nel caso di popolazioni statistiche di numerosità diverse, è possibile solo attraverso le frequenze relative (semplici o percentuali).

I grafici o diagrammi. Spesso in ambito statistico si fa uso di *grafici o diagrammi*. Questi possono essere di vario tipo e, in generale, offrono una percezione immediata della distribuzione di frequenze rappresentata. Naturalmente essi vengono presentati sempre a partire da una rappresentazione tabellare, anche se questa, per comodità, talvolta non viene mostrata.

Rappresenteremo con degli esempi i seguenti tipi di grafici:

- a barre orizzontali o verticali;
- a torta;
- a pile;
- a bastoncini;
- istogrammi;
- cartogrammi.

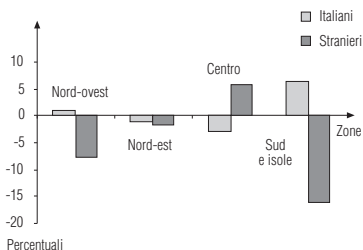
I diagrammi a barre sono particolarmente utili quando si devono esprimere frequenze negative, ottenute per differenze con dati precedenti il cui risultato è appunto negativo, come nell'esempio che segue.

Esempio di rappresentazione grafica di una tabella mediante un diagramma a barre

Tabella 1

Zona geografica	Italiani	Stranieri
Nord - ovest	1,1	-7,6
Nord - est	-1,0	-1,6
Centro	-2,9	5,9
Sud e isole	6,5	-16,1

Figura 1



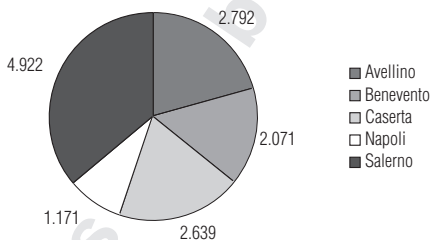
Arrivi (espressi sotto forma di variazioni percentuali rispetto allo stesso periodo dell'anno precedente) negli esercizi alberghieri per ripartizione geografica, durante il periodo di ferragosto 2001 (Fonte: ISTAT).

Nel caso di distribuzioni di frequenze relative a variabili qualitative (serie sconnesse o ordinate), si fa spesso ricorso a una rappresentazione grafica con diagrammi distanziati, oppure ad un diagramma *circolare*, detto anche *a torta*, nel quale l'ampiezza dell'angolo al centro relativo a ciascuna frequenza è proporzionale ad essa, come nell'esempio riportato qui sotto:

Tabella 2

Province	Superficie (kmq)
Avellino	2.792
Benevento	2.071
Caserta	2.639
Napoli	1.171
Salerno	4.922
Totale	13.595

Figura 2



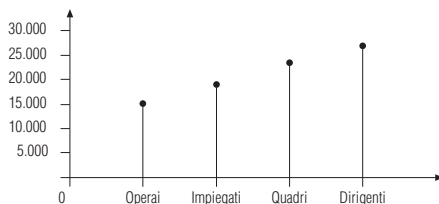
Superficie (in kmq) delle province della Campania.

Quando si devono rappresentare seriazioni, ovvero distribuzioni quantitative di un carattere discreto, si preferisce usare un diagramma del tipo a bastoncini la cui altezza è proporzionale alla frequenza, come nell'esempio seguente:

Tabella 3

Dipendenti	Redditi annui (in euro)
Operai	15.500
Impiegati	18.000
Quadri	23.500
Dirigenti	26.000

Figura 3



Distribuzione dei redditi medi annui lordi di 4 categorie di dipendenti di un'azienda.

Gli **istogrammi** sono dei grafici a forma di rettangoli, utilizzati in molte rappresentazioni di funzioni nelle quali sulle ascisse sono posti gli intervalli della variabile indipendente mentre i corrispondenti valori assunti dalla variabile dipendente sono posti sull'asse delle ordinate.

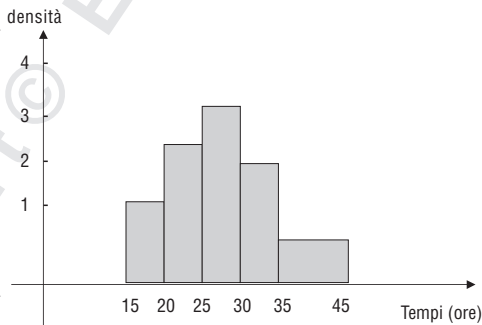
Tuttavia, in statistica, gli **istogrammi** privilegiano un'interpretazione intuitiva immediata, secondo la quale l'area dei rettangoli del grafico deve essere proporzionale alla **frequenza (assoluta o relativa)**. Per ottenere ciò, si ricorre alla definizione di una grandezza, chiamata **densità di frequenza**, priva di significato fisico, ottenuta dividendo ciascuna frequenza assoluta per la relativa ampiezza di classe; la densità viene normalmente posta sull'asse delle ordinate. Sull'asse delle ascisse è rappresentata l'ampiezza delle classi. Gli istogrammi vengono utilizzati, ovviamente, soltanto in presenza di *raggruppamenti in classi* di variabili quantitative.

Esempio di un istogramma statistico

Tabella 4

Tempi (ore)	Num. pezzi	Densità=Num. pezzi/tempi
15-20	6	1,2
20-25	12	2,4
25-30	16	3,2
30-35	10	2
35-45	6	0,6

Figura 4



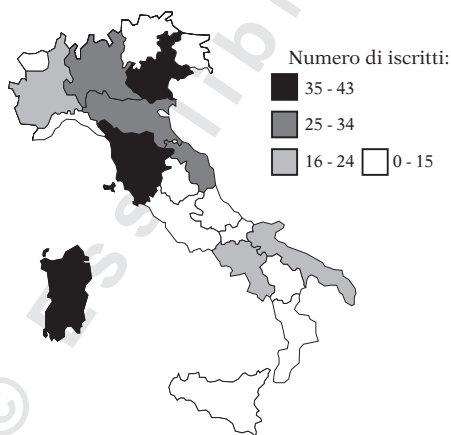
Distribuzione dei tempi di funzionamento di 50 pezzi prodotti da una macchina.

Infine, per serie di tipo *territoriali*, cioè per frequenze di una variabile osservata, su territori omogenei, si fa ricorso ai cosiddetti *cartogrammi*; nel caso seguente, viene presentato il cartogramma relativo alla distribuzione del numero di italiani (distinti per regione) appartenenti a una data associazione culturale.

Tabella 5

Regioni	Iscritti
Piemonte	16
Lombardia	34
Veneto	43
Emilia Romagna	33
Toscana	42
Marche	32
Puglia	24
Campania	23
Sardegna	35
Altre regioni	0

Figura 5



Le regioni in bianco sono quelle in cui il numero degli iscritti non supera i 15;
 le regioni in nero sono quelle in cui il numero degli iscritti è compreso tra 35 e 43.

Test di verifica

1. Nel caso di distribuzioni di variabili qualitative, il grafico più efficace per rappresentarle è:

- a) Il cartogramma.
- b) Il diagramma a bastoncini.
- c) Il diagramma con rettangoli distanziati.
- d) L'istogramma.
- e) Tutti i grafici sopra indicati.

2. Individuare l'opzione esatta:

- a) Il colore degli occhi e l'età sono variabili di tipo qualitativo.
- b) Il colore degli occhi è una variabile qualitativa mentre l'età è quantitativa.
- c) Il colore degli occhi è una variabile quantitativa mentre l'età è qualitativa.
- d) Il colore degli occhi e l'età sono entrambe variabili di tipo quantitativo.
- e) Tutte le risposte precedenti sono errate.

3. Individuare la sequenza corretta delle fasi di un'indagine statistica:

- a) Rilevazione, analisi, spoglio, rappresentazione dei dati.
- b) Spoglio, rilevazione, analisi, rappresentazione dei dati.
- c) Rilevazione dei dati, spoglio, rappresentazione, analisi dei dati.
- d) Analisi dei dati, rappresentazione, spoglio, rilevazione dei dati.
- e) Analisi dei dati, spoglio, rappresentazione, rilevazione dei dati.

4. In una scala ordinale:

- a) Il carattere è sempre qualitativo.
- b) Il carattere è sempre quantitativo.

- c) Il carattere può essere qualitativo oppure quantitativo.
- d) Tutte le risposte precedenti sono errate.
- e) Tutte le risposte precedenti sono corrette.

5. Gli istogrammi vengono utilizzati:

- a) Per rappresentare variabili qualitative.
- b) Per rappresentare variabili quantitative a valori discreti.
- c) Per rappresentare variabili quantitative raggruppate in classi.
- d) Per rappresentare variabili di tipo territoriale.
- e) Tutti le risposte fornite sono corrette.

Soluzioni e commenti

1. Risposta corretta **c)**. Il diagramma con rettangoli distanziati, unitamente a quello circolare, permette una visione immediata per serie di tipo *qualitativo*. Il cartogramma viene utilizzato per serie di tipo *territoriale*, il diagramma a bastoncini è molto utile per variabili *quantitative discrete* mentre l'istogramma viene utilizzato per rappresentare variabili *quantitative raggruppate in classi*.
2. Risposta corretta **b)**. Il colore degli occhi è una variabile qualitativa poiché viene espresso con una descrizione (blu, neri ecc.) mentre l'età è quantitativa poiché si esprime con numeri (20,30,40 ecc.).
3. Risposta corretta **c)**. Infatti si parte dalla rilevazione dei dati che può essere svolta attraverso dei questionari/interviste; successivamente si procede allo spoglio che ne permette una migliore organizzazione da cui si ricava la rappresentazione più efficace per gli stessi. Infine si procede all'analisi dei dati per ricavarne parametri utili ad esprimere in modo sintetico le caratteristiche più salienti della distribuzione analizzata.

4. Risposta corretta **a)**. Infatti una scala si dice *ordinale* quando esprime un carattere qualitativo, le cui modalità possono essere messe in ordine crescente o decrescente.
5. Risposta corretta **c)**. L'istogramma infatti fornisce una rappresentazione a forma di rettangoli in cui la base è rappresentata dall'ampiezza delle classi in cui viene suddivisa la variabile, necessariamente di tipo quantitativo.