



DIPARTIMENTO DI STORIA, PATRIMONIO CULTURALE, FORMAZIONE E SOCIETÀ

Corso di Laurea Triennale
in
Scienze del Turismo (SciTUR)

SECS-S/01 – Statistica

Unità 1

A cura di
Andreina Anna D'Arpino, Carla Rossi

Indice

LA STATISTICA ATTRAVERSO LE SUE APPICAZIONI.....	1
Dalla formulazione del problema alla presentazione dei dati.....	1
Introduzione.....	1
I termini della Statistica.....	5
Fenomeno collettivo ed unità statistica.....	5
Carattere e modalità.....	6
Variabili di stato e di flusso.....	9
Popolazione e campione.....	10
Statistica descrittiva ed inferenziale.....	12
Studi osservazionali e studi sperimentali.....	13
Le fasi di un'indagine statistica.....	13
Le fasi di un'indagine statistica.....	14
La pianificazione della ricerca.....	14
La rilevazione dei dati.....	16
L'elaborazione dei dati.....	17
La presentazione dei risultati.....	18
L'interpretazione dei risultati.....	18
Principali modalità di rilevazione dei dati statistici.....	18
Le principali fonti di dati sul turismo.....	20
Le distribuzioni statistiche.....	21
Elaborazione dei dati e presentazione dei risultati.....	21
Rappresentazioni grafiche.....	28
Il diagrammi a barre.....	28
L'ortogramma.....	29
Diagrammi a settori circolari o a torta.....	30
L'istogramma.....	30
Il cartogramma.....	32
L'ideogramma.....	33
Bibliografia.....	34

LA STATISTICA ATTRAVERSO LE SUE APPICAZIONI

Dalla formulazione del problema alla presentazione dei dati

Introduzione

La Statistica, oggi più che mai, è diventata una scienza indispensabile per la lettura e interpretazione della realtà. Fornisce la metodologia e gli strumenti per una corretta impostazione di una ricerca o un'indagine, ma anche per leggere criticamente e interpretare correttamente le informazioni fornite dai media e dalle pubblicazioni specializzate. E' necessario conoscere e comprendere il percorso di una ricerca o di un'indagine per acquisire queste competenze.

Condurre uno studio non è mai un percorso lineare e privo di ostacoli quanto, piuttosto, una salita con false partenze, vicoli ciechi, svolte nella direzione sbagliata. Riprendendo uno schema illustrato da Keppel¹ e collaboratori, possiamo articolare in **6 passi** il processo:

- 1. Scelta del problema o del quesito di ricerca**
- 2. Rassegna della letteratura teorica ed empirica, eventualmente disponibile, sull'argomento**
- 3. Definizione delle ipotesi di ricerca**
- 4. Identificazione del Piano Sperimentale o di rilevazione e Realizzazione dell'esperimento**
- 5. Descrizione dei dati mediante statistiche descrittive e verifica delle ipotesi mediante statistiche inferenziali**
- 6. Preparazione di un rapporto finale.**

Sei fasi che non si sviluppano in modo sequenziale ma, riprendendo un'idea di John Dewey², secondo un modello circolare che parte dalla percezione dell'esistenza di un problema e dalla scelta di una strategia per affrontarlo, attraversa poi le varie fasi, fino ad arrivare alla valutazione finale che sarà origine o elemento per eventuali indagini successive.

¹ G. Keppel, W. J. Saufley & H. Tokunaga, *Disegno sperimentale e analisi dei dati in psicologia*, Napoli, 2001.

² John Dewey (1859-1952).

Scelta del Problema.

Affrontare uno studio significa, essenzialmente, porsi quesiti e cercarne le risposte. La scelta del problema e degli obiettivi dello studio è, tra le fasi, quella che concede più spazi di libertà.

Rassegna della letteratura.

Dopo aver scelto il tema e gli obiettivi dello studio, è fondamentale svolgere una rassegna bibliografica degli studi che hanno affrontato lo stesso problema o analoghe situazioni. Questi studi potranno essere di tipo teorico o sperimentale. Da entrambi i tipi si possono ricavare informazioni utili per lo studio che si sta affrontando. Una buona rassegna bibliografica è un passo tanto importante quanto spesso disatteso. Se in passato condurre una ricerca bibliografica comportava parecchi sacrifici quasi sempre, oggi, nell'era del digitale, le cose vanno sostanzialmente meglio e si può condurre una ricerca e "scaricare" gli articoli da Internet attraverso l'utilizzo di banche dati elettroniche.

Formulazione delle Ipotesi.

Dopo la scelta dell'argomento e l'esame della letteratura, diviene fondamentale formulare le ipotesi da mettere alla prova nello studio. E' importante sottolineare che quando un ricercatore conduce una ricerca molto raramente, o quasi mai, egli rileva i dati su tutti i soggetti della **popolazione di riferimento** ma rileva i dati su un gruppo più ristretto numericamente, un **campione**. Spieghiamo con un esempio il significato di questi due termini. Ammettiamo, ad esempio, di voler valutare l'utilizzo e il gradimento dei principali mezzi di comunicazione dei quindicenni italiani. La popolazione di riferimento è rappresentata dai quindicenni italiani, ma intervistare tutti i giovani interessati sarebbe molto dispendioso e difficilmente realizzabile, si preferisce svolgere l'indagine, somministrando un questionario ad un gruppo di qualche migliaio di soggetti, per esempio 5000 (anziché i circa 500.000 che compongono la popolazione), ovvero un campione, scelti in modo tale da essere un campione rappresentativo della popolazione (ogni soggetto della popolazione deve avere la stessa possibilità di essere incluso nel campione) ed estendere poi il risultato ottenuto all'intera popolazione. E' importante, quindi, comprendere che gli studi vengono spesso condotti sui campioni, ma servono per trarre conclusioni sulle popolazioni! Gli exit poll elettorali sono un esempio ben noto di questa procedura.

Il numero di elementi che compone un campione è detto numerosità campionaria.

A volte, lo scopo stesso della rilevazione obbliga a limitare il numero dei casi osservati: il collaudo di una partita di borse non giungerà mai a tagliarle tutte per verificare la qualità del cuoio. In altri casi, è necessario restringere l'osservazione ad una parte soltanto dei casi esistenti perché non tutte le

unità sono disponibili, o per difficoltà tecniche, o per limitatezza delle risorse finanziarie ed umane utilizzabili per l'analisi. Per esempio, i meteorologi "conoscono" la temperatura o la quantità di pioggia che cade in un paese, supponendo rappresentative per tutto il territorio le registrazioni di alcuni termometri e di alcuni pluviometri disposti in poche località opportunamente prescelte.

Quando un ricercatore formula un'ipotesi, egli la formula pensando alla popolazione e la verifica sul campione.

Identificazione del Piano Sperimentale o di rilevazione e realizzazione dello studio. Una volta stabilite le ipotesi di lavoro, occorre scegliere quali siano le condizioni più adeguate per condurre lo studio e come misurare le variabili³. Ciò implica la definizione delle variabili d'interesse e la scelta delle modalità di misurazione delle variabili stesse.

Inoltre occorre valutare in modo accurato il possibile intervento di variabili di disturbo che possono compromettere la **validità dello studio**. Lo sperimentatore, sia durante la fase di progettazione sia durante la fase di realizzazione dello studio, dovrà cercare di minimizzare l'effetto di tutte le possibili fonti di errore, cercando di salvaguardare la validità delle sue conclusioni. Una volta definito il Piano Sperimentale, il ricercatore provvederà alla raccolta dati, alla loro codifica e tabulazione in fogli elettronici che costituiranno la base dati. E' in questa fase che si entra nel vivo del metodo statistico.

Descrizione dei dati mediante statistiche descrittive e verifica delle ipotesi mediante statistiche inferenziali.

Dopo l'immissione dati, è buona regola procedere con il controllo della loro qualità onde evitare di condurre analisi statistiche su una base dati (o matrice dei dati) con errori di immissione. Questa raccomandazione, sempre presente nei libri di metodologia, è spesso disattesa e non è insolito che ricercatori, pur esperti, si ritrovino a dover compiere delle analisi più volte, per la presenza di errori nei dati che vengono "scoperti" troppo tardi. Dopo aver controllato la qualità dei dati, si procede, con *software* più o meno evoluti, al computo delle statistiche e al loro utilizzo per sintetizzare e descrivere i dati, interpretarli e verificare le ipotesi di lavoro.

Preparazione del rapporto finale di ricerca.

La fase conclusiva di uno studio è la stesura di un resoconto scritto nella forma di un testo da comunicare e rendere pubblico. Un buon resoconto deve sempre riportare tutte quelle informazioni che siano sufficienti a consentire al lettore di condurre a sua volta uno studio identico, dovrebbe quindi:

³ Nel prossimo paragrafo sarà ampiamente illustrato il significato.

1. rendere trasparente per altri ricercatori il percorso d'indagine compiuto;
2. comunicare i risultati più importanti ottenuti attraverso l'analisi dei dati;
3. effettuare un confronto con i risultati ottenuti da altri studi noti sullo stesso argomento.

■ Esempio

Una delle raccomandazioni mediche che appaiono sui pacchetti di sigarette dice che il fumo in gravidanza può provocare danni al feto, nascita prematura, e peso, alla nascita, inferiore alla norma.

Su che cosa si basano tali raccomandazioni?

Il monossido di carbonio, agente chimico presente nelle sigarette, riduce la quantità di ossigeno che arriva al feto e gli effetti fisiologici non sono ancora completamente chiari. Per studiare gli effetti del fumo, tra il 1960 e il 1967, fu condotto uno studio sulle donne in gravidanza della città di San Francisco. Allo studio parteciparono 15.000 famiglie.

Insieme all'informazione sulle abitudini al fumo della madre, alla nascita del bambino, venne misurato il peso di 1236 maschi, nati tra il 1960 e il 1961, e che erano sopravvissuti almeno 28 giorni (oltre ad una serie di altre variabili). Da questa ricerca si sono ottenuti interessanti risultati sugli effetti del fumo, confrontando, per esempio, il peso di neonati provenienti da madri fumatrici e non fumatrici. I risultati ottenuti saranno oggetto di riflessione, più avanti, quando analizzeremo i più importanti strumenti di indagine per l'analisi statistica dei dati.

E' un utile esercizio riconoscere, nella descrizione dello studio riportata, le diverse fasi dello studio.

I termini della Statistica

Introduciamo, ora, alcuni termini della statistica.

Fenomeno collettivo ed unità statistica

Definizione

I fenomeni collettivi sono quei fenomeni riferibili ad una moltitudine di oggetti in cui interessi studiare l'insieme degli oggetti nel suo complesso e non i singoli individui.

Esempi

Un'analogia suggestiva per comprendere meglio può riferirsi alle scienze naturali dove interessi studiare il comportamento di una specie e non quello di ogni singolo individuo della specie, così come lo studio del comportamento di un gas nella fisica è rivolto all'insieme delle sue molecole e non al loro singolo comportamento.

Tali fenomeni sono investigabili solo tramite una pluralità di osservazioni. Per essere ancora più precisi, i fenomeni analizzabili con metodi statistici sono quelli che si manifestano in modo variabile nei singoli individui e per cui sono disponibili più osservazioni di analoga natura. Sono fenomeni collettivi le vendite di un determinato bene in un periodo fissato, il reddito dei lavoratori di un certo settore industriale in un certo anno, la condizione di salute dei bambini della scuola materna in una data provincia in un periodo fissato e così via.

La statistica cerca di individuare delle leggi e delle regolarità che sono alla base di un fenomeno collettivo.

Nel corso di uno studio statistico, si parla di **popolazione** riferendosi ad un insieme di elementi ben definiti, che chiameremo **individui** o **unità statistiche**, e che sono considerati omogenei rispetto ad una o più caratteristiche legate al fenomeno che si intende studiare: gli "anziani" sono una popolazione omogenea rispetto all'età, definita generalmente da un limite inferiore, per esempio età \geq 60 anni, i bambini della scuola materna sono omogenei rispetto alla fascia di età e al

fatto di frequentare un certo tipo di scuola, gli occupati in una certa attività, per esempio, gli addetti alla sala, sono omogenei rispetto al tipo di lavoro svolto e così via. Popolazione statistica ha diversi sinonimi come **universo** o **collettivo**.

Per poter studiare un fenomeno collettivo è necessario, in prima analisi, stabilire quali siano gli elementi che costituiscono la popolazione di interesse.

■ Esempio

Se vogliamo studiare il fenomeno collettivo “vendite di iPod” in un certo periodo ed in un certo luogo, esamineremo come popolazione statistica il collettivo degli iPod in quel periodo, nel luogo prescelto: ogni iPod venduto rappresenta un’unità statistica.

Carattere e modalità

Definizione

Ogni fenomeno collettivo viene studiato mediante l’osservazione e la misurazione di una o più caratteristiche delle unità statistiche della popolazione di interesse, ognuna delle quali è denotata carattere o variabile, per sottolineare il fatto che può presentarsi in modo diverso nelle diverse unità statistiche. Ciascun carattere è presente in ogni unità statistica con una ben determinata modalità.

■ Esempi

Si consideri un gruppo di atleti iscritti ad una federazione calcistica. Per ogni calciatore è possibile rilevare una serie di informazioni come:

- l’età;
- la data di nascita;
- il luogo di nascita;
- la nazionalità;
- l’altezza;
- il peso;
- la pressione arteriosa;
- la frequenza cardiaca;

- il flusso espiratorio forzato;
- ecc.

Tali informazioni costituiscono i **caratteri** di ogni **unità statistica** (*calciatore*) considerata.

Se un calciatore proviene dal Brasile ed un altro dall'Estonia, si dice che il **carattere luogo di nascita** presenta nel primo calciatore **modalità** "Brasile", nel secondo "Estonia". Se due calciatori sono alti entrambi 182 cm allora il **carattere altezza** si presenta con la stessa **modalità** "182".

I caratteri rilevati in ciascuna unità statistica sono generalmente di natura assai diversa. Alcune volte esprimono delle informazioni qualitative relative ad attributi o qualifiche (come il sesso, il colore dei capelli, il luogo di nascita) e sono detti **qualitativi**.

Altre volte esprimono la misura di qualche grandezza, come il reddito, il peso, l'altezza, ecc., e sono, pertanto, caratteri **quantitativi**. In questo caso i caratteri hanno la peculiarità di essere espressi da numeri, che risultano da un processo di misurazione, che è possibile sommare e sottrarre fra loro, una volta che sia stata definita un'unità di misura.

Le modalità dei caratteri qualitativi vengono in genere contraddistinte mediante attributi che specificano le proprietà dell'unità a cui si riferiscono.

■ Esempio

Possiamo distinguere i dvd venduti secondo il **genere** (giallo, fantascienza, avventura, ...), il gruppo degli studenti universitari secondo il **sesso** (maschile, femminile), secondo **l'anno di corso** (I, II, III, IV, V).

Avremo:

	Carattere	Modalità
Genere	Qualitativo	Giallo, fantascienza, avventura, ecc.
Sesso	qualitativo	Maschio, Femmina
Anno di corso	qualitativo	I, II, III, IV e V

Per tradurre in cifre i caratteri quantitativi occorre preliminarmente definire un processo di misurazione, cioè occorre stabilire una scala e un'unità di misura da utilizzare per i confronti sulle unità statistiche di interesse.

Se si vuole misurare il peso dei neonati di un certo ospedale, si dovrà decidere se esprimerlo in chilogrammi, in ettogrammi, con quante cifre decimali.. È importante sottolineare che ogni misura non è mai esatta, ma è sempre affetta da un certo margine di errore. Se, nel misurare il peso, si ottiene il valore 2,52, vuol dire semplicemente che il rilevatore ha osservato una misura che è più vicina a 2,52 anziché a 2,51 o 2,53: il peso x è un qualche valore compreso in un intervallo centrato su 2,52: $2,52 - a < x < 2,52 + a$ dove a è appunto il “**margine di errore**” della misurazione.

A volte le misure consistono in semplici conteggi. Si procede in tal modo per la misurazione relativa al numero delle foglie di una pianta, al numero dei denti, dei vani degli alloggi, degli aerei arrivati in ritardo un determinato giorno in un certo aeroporto, ecc.

Tali caratteri vengono detti **quantitativi discreti**

.Altre volte i caratteri **quantitativi** sono **continui**, come l'estensione di un certo territorio, la massa di un grave, la densità di un gas, la velocità di un veicolo e la loro misura può essere espressa con diversi gradi di precisione, una volta fissata un'unità di misura, tenendo conto dei sottomultipli, come nel caso citato dei neonati.

A loro volta i caratteri **qualitativi** si distinguono in **caratteri sconnessi** e **caratteri ordinati**.

Il sesso di un neonato è un carattere sconnesso in quanto non esiste alcun ordine tra le due modalità “maschio” e “femmina”. In tal caso le modalità del carattere possono soltanto essere confrontate per vedere se sono uguali o diverse. Si può, invece, considerare come carattere qualitativo ordinato l'ordine di nascita, ovvero “primogenito”, “secondogenito” e così via o il titolo di studio di uno dei genitori.

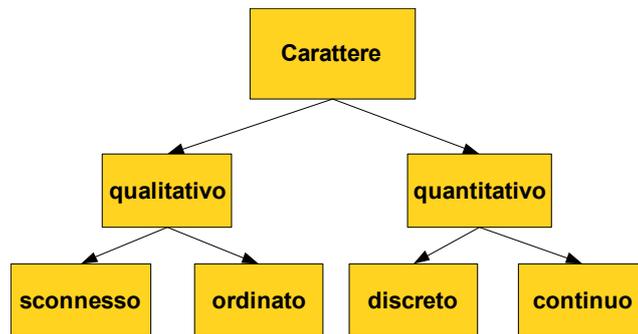


Figura 1

■ Esempio

Riprendendo l'esempio dei calciatori, prova ad collocare ciascun carattere nella colonna appropriata.

Carattere	qualitativo		quantitativo	
	sconnesso	ordinato	discreto	continuo
Età(in anni compiuti)				
Data di nascita				
Luogo di nascita				
Nazionalità				
Altezza				
Peso				
Pressione arteriosa				
Frequenza cardiaca				
Flusso espiratorio				

Variabili di stato e di flusso

E' opportuno distinguere ancora tra *variabili di stato* e *variabili di flusso*.

Definizione

Si definisce di stato una variabile rilevabile in un certo istante di tempo prefissato.

Una variabile è di flusso se per la sua rilevazione è necessario prendere in considerazione due istanti di tempo per poter valutare le unità statistiche entrate nella popolazione di interesse nell'intervallo temporale così determinato.

■ Esempio

Il totale delle camere di albergo di una certa regione è una **variabile di stato**, perché è, in un certo senso, una "fotografia" della situazione alberghiera nel territorio interessato. Il numero dei nuovi posti letto in una certa regione si può,

invece, rilevare solo rapportandolo al periodo di tempo in cui tali posti sono stati creati. I nuovi posti letto sono una **variabile di flusso**. L'informazione che resta associata ad una variabile di flusso è relativa agli aspetti dinamici del fenomeno di interesse. Analogamente la popolazione presente sul territorio italiano in un certo giorno è un **collettivo di stato**, le presenze alberghiere in un certo anno (o mese, o giorno) sono un **collettivo di flusso**.

Popolazione e campione

Definizione

Lo studio statistico può riguardare il fenomeno nella sua globalità, cioè considerare l'insieme di tutti i suoi componenti, ovvero l'universo o popolazione, oppure solo una sua parte che è detta campione.

La **popolazione** può essere **limitata** o **illimitata** a seconda che sia costituita da un numero finito o teoricamente infinito di osservazioni. Nel primo caso è possibile (ma spesso non fattibile) rilevare i dati su ogni singola osservazione (censimento), ma nella maggior parte delle indagini, esperimenti e prove scientifiche si rilevano i dati su un campione.

Una **popolazione illimitata**, invece, deve essere necessariamente studiata mediante un campione.

■ Esempi

1. Popolazione “limitata”- Una compagnia di assicurazione vuole studiare il fenomeno collettivo “danni subiti dalle imbarcazioni di carico assicurate”.

La definizione dell'oggetto dello studio identifica la popolazione statistica di interesse che è costituita dalle tutte le imbarcazioni assicurate (popolazione limitata): ognuna di queste costituisce un'unità statistica. Dato che si ipotizza che il numero di danni possa dipendere dal tipo di imbarcazione (5 possibili tipi), dall'anno di costruzione e dal tempo di servizio svolto dall'imbarcazione, occorre rilevare tutti questi caratteri per ogni unità statistica.

I caratteri considerati sono di tipologie diverse:

- Il **tipo di imbarcazione** è un carattere qualitativo sconnesso; le possibili modalità possono essere identificate dalle lettere A, B, C, D, E;

- L'**anno di costruzione** è un carattere qualitativo ordinato; si distinguono 4 diversi periodi (1990-94, 1995-99, 2000-04, 2005-09);
- La **durata del servizio** è un carattere quantitativo continuo e può essere misurata in anni, mesi, giorni, tenendo conto, almeno in teoria, anche dei sottomultipli.
- Il **numero di danni** è un carattere quantitativo discreto, trattandosi di un conteggio.

2. Popolazione “illimitata” - Una casa produttrice di insetticidi vuole valutarne l'efficacia su alcuni insetti.

Si prende in considerazione un campione che fa parte del numero infinito di osservazioni possibili (popolazione illimitata). Il protocollo di sperimentazione che viene utilizzato è molto complesso, noi proponiamo, per ovvi motivi, una versione semplificata.

Il campione è costituito dal lotto di terreno su cui si effettua la sperimentazione. L'unità statistica è rappresentata dalle foglie di ciascuna pianta che vengono analizzate prima e dopo il trattamento con l'insetticida.

I caratteri osservati sono:

- Le **varie specie di insetti** che si possono trovare sul lotto (mosca del melo, minatrice, ecc) che rappresenta un carattere qualitativo sconnesso.
 - La **quantità di sostanza**, espressa in litri/ha, carattere quantitativo continuo reso discreto.
 - I **tempi di osservazione** è una variabile discreta che assume valore 0 prima della somministrazione, 1 (dopo 3 giorni), 2 (dopo 6 giorni).
 - L'**effetto** è una variabile discreta e rappresenta la differenza tra il numero di insetti presenti sotto ciascuna foglia selezionata ai vari tempi di osservazione.
-

Statistica descrittiva ed inferenziale

Definizione

La statistica descrittiva è quella parte della statistica che fornisce criteri, metodi, indici e tecniche grafiche per sintetizzare e rendere interpretabile e divulgabile il dato raccolto.

La statistica inferenziale è quella branca della statistica che interviene quando dall'osservazione di un fenomeno sul campione (e dalle ipotesi fatte su un piccolo numero di dati) si vuole risalire alla conoscenza del fenomeno sull'intera popolazione e alla formulazione di una teoria da utilizzare nelle applicazioni "fino a prova contraria".

Le tecniche di statistica descrittiva trovano applicazione sia nel caso in cui vengano raccolti dati riguardo a tutti gli individui costituenti la popolazione oggetto di studio (come nel caso di un censimento), sia nel caso in cui si raccolgano informazioni solo su un campione estratto dalla popolazione.

Nel caso della statistica inferenziale, come è facile intuire, non è il campione in sé che interessa bensì la sua capacità di fornire informazioni sulla popolazione che rappresenta.

In altre parole, la Statistica inferenziale è quella parte della Statistica tendente ad interpretare le osservazioni in termini di modelli teorici che spiegano, nelle linee generali, il meccanismo secondo il quale si producono i dati esaminati. Sono proprio tali modelli matematici che permettono di generalizzare in modo induttivo i risultati, ottenuti da campionamento o da osservazioni indirette, dall'insieme dei dati osservati alla popolazione di riferimento. Tali modelli permettono anche, in molti casi, di valutare, probabilisticamente, eventuali margini di incertezza nella generalizzazione dei risultati.

In realtà la distinzione tra statistica descrittiva e inferenza statistica non ha molto senso perché la ricerca statistica su qualsiasi fenomeno collettivo è sempre di natura induttiva avendo sempre come scopo lo studio di un fenomeno complesso a partire da osservazioni comunque particolari o parziali.

Studi osservazionali e studi sperimentali

Definizione

I dati raccolti secondo uno schema sperimentale predisposto ad hoc si dicono dati sperimentali; i dati che sono generati secondo un processo di rilevazione non predisposto dallo sperimentatore si dicono dati osservazionali.

La differenza tra dati osservazionali e dati sperimentali sta, quindi, nel fatto che i primi sono prodotti attraverso un processo che non è stato pianificato da chi utilizza i dati; i dati sperimentali, invece, sono prodotti secondo uno schema pianificato da chi conduce l'indagine.

■ Esempi

Rientrano tra gli studi osservazionali quelli epidemiologici, nei quali vengono presi in considerazione dati aggregati (popolazioni, comunità, gruppi) derivati, generalmente, da statistiche correnti con i quali si ricavano informazioni generali sulla diffusione delle malattie e/o dei fattori di rischio in una popolazione.

Un'altra tipologia di studi osservazionali sono gli studi caso-controllo, con essi si ottengono, in tempi relativamente brevi, informazioni attendibili sui fattori di rischio di una malattia. Il disegno dello studio prevede che si considerino sempre due gruppi di soggetti: uno di malati che costituiscono i **casì**, ed uno di soggetti aventi le stesse caratteristiche dei primi, dai quali differiscono solo per la malattia (i **controlli**), dal confronto è possibile dedurre informazioni su caratteristiche che possono essere legate alla malattia.

Tra gli studi sperimentali troviamo, per esempio, le sperimentazioni cliniche (Clinical trials) e gli studi sul campo.

Con le prime si valuta, seguendo un protocollo ben preciso e regolamentato da norme, l'efficacia di due o più trattamenti farmacologici o di altro tipo.

Negli studi sul campo, invece, si eseguono interventi preventivi su soggetti a rischio di contrarre una determinata malattia e se ne verifica l'effetto. Sono inoltre studi sperimentali tutte le rilevazioni a campione per le indagini di mercato, gli exit poll e così via.

Spesso gli studi osservazionali permettono di formulare quelle ipotesi da sottoporre a verifica mediante opportune sperimentazioni. Basti pensare alle **osservazioni** sull'oscillazione dei lampadari del duomo di Pisa da parte di Galileo che suggerirono la legge di isocronia delle piccole oscillazioni, verificata poi da Galileo stesso mediante **esperimenti** pianificati e condotti in laboratorio.

Le fasi di un'indagine statistica



Fasi di un'indagine statistica

Entrando nello specifico dell'indagine statistica concentreremo l'attenzione su alcune delle fasi precedentemente elencate, operando una ulteriore specificazione del percorso, ovviamente si tratta di un'ipotesi di schematizzazione, nella realtà, i confini tra le varie fasi o sottofasi non sono sempre ben definiti. Esse sono:

- pianificazione;
- rilevazione;
- elaborazione;
- presentazione;
- interpretazione.

La pianificazione della ricerca

La prima fase di una qualunque ricerca relativa a un fenomeno collettivo consiste nella pianificazione.

Definizione

La pianificazione consiste nella definizione del fenomeno e degli obiettivi dell'indagine, nell'individuazione del collettivo, delle unità statistiche, nella scelta dei caratteri del collettivo e nella definizione delle relative modalità (per i caratteri qualitativi) o dei processi di misura (per i caratteri quantitativi).

■ Esempio

Si vuole studiare il fenomeno “*vendita di televisori nel 2014*” stabilendo l'obiettivo di individuare i gusti degli acquirenti in merito alla marca, la tipologia, la dimensione, altre caratteristiche tecniche, ecc. E' possibile, da queste premesse identificare il collettivo, che consiste nei televisori venduti nel 2014: ogni televisore è un'unità statistica. L'unità di rilevazione: il rivenditore. I caratteri di interesse sono: la marca del televisore (qualitativo sconnesso), la tipologia (qualitativo

sconnesso), la dimensione (quantitativo continuo convenzionalmente espresso in pollici).

Le modalità dei caratteri qualitativi sono identificate anch'esse dagli obiettivi:

• per la marca: Samsung, Sony, Rex, ...;

per la tipologia: CRT, LCD, plasma, ...;

.....

Occorre, sempre, pianificare la rilevazione nei minimi dettagli in modo tale che non vi siano ambiguità di nessun genere.

E' necessario tener conto dei seguenti punti:

- a) definire con precisione la popolazione, l'unità di rilevazione e l'unità statistica, in altre parole indicare con chiarezza una regola che consenta di stabilire se un'unità è o non è un elemento della popolazione di interesse;
- b) stabilire i caratteri quantitativi e qualitativi che interessa rilevare per il perseguimento dell'obiettivo dell'indagine e definire le relative modalità di interesse e scale di misura;
- c) indicare i mezzi tecnici per raccogliere le informazioni su tali caratteri (schede di rilevazione, questionari, ecc.);
- d) fissare l'estensione della rilevazione in ordine al tempo, allo spazio, alle disponibilità dei mezzi tecnici e finanziari.

Per quanto riguarda il punto a), la popolazione di riferimento viene determinata direttamente nella fase di definizione degli obiettivi. L'unità statistica e l'unità di rilevazione vengono analogamente definite dagli obiettivi e possono non coincidere. Questo avviene se l'unità di rilevazione è costituita da più unità elementari, su ciascuna delle quali interessa raccogliere informazioni.

Ad esempio, nel corso del censimento della popolazione residente in Italia, l'unità di rilevazione è la famiglia, infatti, in sede di raccolta delle informazioni, ad ogni famiglia viene fornito un modulo da compilare. In esso, però, si richiedono notizie su ciascun componente del nucleo familiare: ogni componente è un'unità statistica.

Anche nel caso dell'indagine sui televisori relativa all'esempio, è verosimile che l'unità di rilevazione sia diversa dall'unità statistica. L'unità di rilevazione è, infatti, il rivenditore, che, per ogni unità statistica (televisori venduti nel 2014) riempie un modulo di rilevazione o scheda contenente le informazioni sui caratteri di interesse. Dal punto di vista tecnico, lo strumento di rilevazione è ragionevolmente una semplice scheda (cartacea o elettronica se i dati sono raccolti via Internet), del tipo di quella riportata sotto.

Rivenditore.....	Data della vendita.....
n. d'ordine del televisore	Marca.....
Tipo.....	Dimensione (in pollici)
Angolo di visuale.....	

Osserviamo che le prime tre informazioni raccolte permettono di identificare univocamente l'unità statistica, in modo da non rischiare di contare più volte una stessa unità o, viceversa, di trascurarne qualcuna.

Finita la fase di pianificazione è possibile iniziare la fase di rilevazione.

La rilevazione dei dati

Definizione

La rilevazione è quel complesso di operazioni attraverso le quali si acquisiscono le informazioni sulle caratteristiche (o caratteri) di interesse per ciascuna unità statistica del collettivo considerato; da questa fase scaturiscono i dati statistici elementari o dati grezzi, che entrano a far parte della tabella unità/caratteri, che costituisce la matrice dei dati.

Supponiamo di aver riempito le schede relative ai televisori, secondo il modello riportato sopra, la matrice dei dati è una semplice tabella in cui a ogni riga corrisponde un'unità statistica e su ogni colonna è registrata la modalità o la misura di uno dei caratteri rilevati. La prima colonna contiene le informazioni che identificano l'unità statistica.

Tabella 1 – Tabella unità/caratteri o matrice dei dati relativa all'indagine sui televisori.

Unità statistica	Rivenditore	Marca	Tipo	Dimensione	Angolo di visuale
1	Mondonics	Sony	CRT	26"	170
2	Cimabue	Samsung	plasma	46"	178
3	Mediaeurope	Grundig	LCD	35"	176
4	Cimabue	Samsung	LCD	42"	170
.....

Per agevolare l'archiviazione e le elaborazioni i caratteri qualitativi, vengono generalmente *codificati* attraverso una relazione biunivoca tra le loro modalità e un opportuno insieme numerico.

Esempio

Se consideriamo nuovamente la tabella dei dati sui televisori venduti, possiamo porre, a titolo di esempio, le seguenti corrispondenze:

Marca: Sony = 1, Samsung = 2, ...

Tipo: CRT = 1, LCD = 2, Plasma=3;

In questo modo la tabella unità/caratteri contiene solo valori numerici (anche il rivenditore può essere codificato in modo analogo).

Tabella 2: Tabella unità/caratteri o matrice dei dati relativa all'indagine sulle televisioni (in forma codificata)

Unità statistica	Rivenditore	Marca	Tipo	Dimensione	Angolo di visuale
1	1	1	1	26"	170
2	2	2	3	46"	178
3	3	3	2	35"	176
4	2	2	2	42"	170
.....

L'elaborazione dei dati

Definizione

L'elaborazione è quel complesso di operazioni attraverso le quali i dati grezzi vengono prima codificati e poi sintetizzati nei dati derivati più facilmente interpretabili. In alcuni casi è indispensabile l'elaborazione dei dati elementari, come, per esempio, nelle consultazioni elettorali.

Quando si svolge una consultazione elettorale, il risultato della medesima è ottenuto mediante lo spoglio di tutte le schede votate, ovvero attraverso il conteggio dei voti espressi e il successivo calcolo delle percentuali dei voti attribuiti a ciascun partito. Senza tale elaborazione non sarebbe possibile stabilire l'esito della consultazione elettorale.

La presentazione dei risultati

Definizione

La presentazione è l'esposizione dei grafici, delle tabelle, ecc., dove sono stati sintetizzati i dati rilevati.

Le tabelle e i grafici mostrati in TV con i risultati elettorali rappresentano un esempio di presentazione di dati statistici.

L'interpretazione dei risultati

Definizione

L'interpretazione è la spiegazione dei risultati, con le osservazioni finali e l'eventuale collegamento con altre indagini.

Riprendiamo in esame il caso dei dati relativi ad una consultazione elettorale; questi vengono semplicemente presentati in forma di tabelle e grafici e confrontati con i risultati ottenuti in consultazioni precedenti. Si cerca quindi di ottenere i commenti dei rappresentanti dei Partiti e degli studiosi di politica (politologi), che non sono altro che interpretazioni.

Principali modalità di rilevazione dei dati statistici

Le **rilevazioni statistiche** possono essere **continue**, quando si svolgono senza interruzioni (come quelle relative alle nascite, ai matrimoni, ai decessi, nonché quelle effettuate dai rilevatori sismici, meteorologici, osservatori epidemiologici, ecc.).

Sono **periodiche** le rilevazioni effettuate ad intervalli regolari di tempo, come il censimento della popolazione o delle imprese (ogni dieci anni), i prezzi al consumo nelle città campione per il calcolo dell'aumento del costo della vita (ogni mese), le indagini sulle Forze lavoro (trimestrale), ecc.

Vengono dette **occasional** le rilevazioni effettuate in circostanze particolari (sondaggi pre-elettorali, indagini riguardo alle opinioni su argomenti vari, ecc.).

Lo strumento di rilevazione varia a seconda dello scopo della rilevazione e del tipo di caratteri da rilevare.

Tra i metodi di acquisizione dei dati da popolazioni umane, o comunque da popolazioni le cui unità sono connesse all'organizzazione umana (per esempio: imprese, scuole, aziende agricole, ecc.), particolare rilievo assume **l'intervista**. Essa consiste nel rivolgere alcune domande ad alcune unità che compongono la popolazione di interesse (eventualmente tutte) e nel registrare le risposte a tali domande. Per porre un'intervista le domande vengono raccolte in un apposito modello detto questionario.

La struttura del **questionario** viene opportunamente progettata in relazione alla tecnica di intervista utilizzata: intervista diretta, autocompilazione, intervista telefonica, ecc.. La scelta della tecnica di intervista è legata agli obiettivi della ricerca, alle caratteristiche della popolazione di riferimento, ai tempi e alle risorse disponibili.

Un'altra tipologia di rilevazione è la **consultazione di registri, ruoli, archivi**, ecc., predisposti in precedenza, da cui le informazioni sono generalmente estratte attraverso schede di rilevazione. Attraverso di esse si rilevano alcuni caratteri predeterminati delle unità che formano il collettivo, per mezzo di "domande" prefissate e fornendo, in molti casi, per ogni risposta, la scelta tra una preindividuata rosa di risposte.

Le principali fonti di dati sul turismo

La gestione delle informazioni relative al turismo è affidata ad organismi istituzionali quali:

- l'ISTAT, Istituto Italiano di Statistica,
- l'EUROSTAT, Ufficio di Statistica dell'Unione Europea,
- il WTO, World Tourism Organization,

Essi hanno il compito di effettuare indagini e ricerche statistiche, insieme ad Enti, consorzi, imprese, Pubbliche Amministrazioni erogatori di servizi diversi che però raccolgono informazioni sul turismo, informazioni che possono essere utilizzate ed analizzate come statistiche.

Le definizioni principali utilizzate nelle statistiche del turismo sono stabilite e valide a livello mondiale

Esse fanno riferimento allo **spostamento sul territorio** che deve essere verso luoghi diversi da quelli abitualmente frequentati, alla **durata dello spostamento** che non deve superare un certo limite oltre il quale il visitatore stesso diventerebbe un residente del luogo. Tale limite è convenzionalmente fissato (in sede ONU) in un anno. La durata minima è invece di 24 ore ed un pernottamento nel luogo visitato è la discriminante che, all'interno del fenomeno, fa distinguere l'escursionismo (meno di 24 ore e nessun pernottamento) dal turismo.

Altra discriminante è il **motivo principale dello spostamento** che deve essere diverso dal trasferimento di residenza (definitivo o temporaneo) e dall'esercizio di un'attività lavorativa retribuita a carico dei fattori residenti nel luogo visitato. Dal turismo si escludono quindi i movimenti migratori, anche stagionali, legati al lavoro nonché i trasferimenti dei diplomatici e dei militari (e delle loro famiglie) di stanza in un Paese diverso dal proprio, nonché i rifugiati, i nomadi ecc.

Una delle principali indagini sull'attività alberghiera si svolge in tre periodi particolari dell'anno: Pasqua, la settimana di Ferragosto e il periodo Natale-Epifania. L'indagine effettuata su un campione di 2.000 alberghi estratti secondo un disegno campionario stratificato a due stadi, che prevede come unità di primo stadio le province, mentre gli esercizi (unità di secondo stadio) sono selezionati secondo una stratificazione per classe di tipologia di località e classe di posto-letto. L'indagine relativa alla "domanda" turistica è effettuata sulla base di un campione: 14.000 famiglie (3.500 per trimestre). I periodi di osservazione sono i trimestri gennaio-marzo, aprile-giugno, luglio-settembre e ottobre-dicembre. Per ciascun periodo vengono rilevate informazioni sui viaggi con almeno un pernottamento realizzati dai residenti e conclusi nel trimestre, indipendentemente dalla loro data di inizio.

Tra le principali Pubblicazioni dell'Istat/Sistan sul turismo ricordiamo:

- le Statistiche sull'offerta turistica italiana
- le Statistiche relative alla bilancia dei pagamenti turistica
- le Statistiche economiche sul settore del turismo
- le Statistiche sulla domanda turistica dei viaggi degli italiani

Le distribuzioni statistiche

Elaborazione dei dati e presentazione dei risultati

Una volta imparato ad impostare un'indagine statistica e a predisporre gli strumenti di rilevazione dei dati, occorre sintetizzare i dati raccolti, organizzati nella matrice dei dati grezzi o tavola unità/variabili, mediante opportune elaborazioni e presentarli in modo chiaro mediante tabelle, indici statistici e rappresentazioni grafiche.

■ Esempio

Ripartiamo dall'indagine sui televisori che abbiamo preso in considerazione nella fase di rilevazione statistica.

I caratteri, oggetto di indagine, sono:

Marca del televisore (carattere qualitativo sconnesso) con le seguenti codifiche:
Sony = 1 Samsung = 2 Grunding = 3 Rex = 4
Tipo di televisore (carattere qualitativo sconnesso):
CRT = 1 LCD = 2 Plasma = 3
Dimensione (espressa in pollici) carattere quantitativo continuo, reso discreto.
....
Angolo di visuale (espressa in gradi sessagesimali) carattere quantitativo continuo, reso discreto.
....

Ipotizziamo di aver rilevato i dati su 4 rivenditori di una certo comune nel corso del 2014, che hanno compilato in totale 30 schede, nella nostra matrice dei dati riportiamo i dati. Ogni riga rappresenta un televisore venduto (istanza), identificato da un numero progressivo da 1 a 30, nelle diverse colonne (campi) si riportano i

caratteri di interesse opportunamente codificati e la lunghezza di ciascun campo del record.

Nome campo	Descrizione	Lunghezza
NUM	Identificativo del televisore	2
RIV	Nome del rivenditore	1
MARCA	Marca del televisore	1
TIPO	Tipologia	1
DIM	Dimensione	2
AV	Angolo di visuale	3
MESE	Mese di vendita	2

Tabella 1: Matrice dei dati – indagine campionaria sulle vendite di televisori –
Anno 2014.

NUM	RIV	MARCA	TIPO	DIM	AV	MESE
1	1	1	1	26	170	4
2	1	2	3	46	178	5
3	1	3	2	35	176	6
4	1	2	2	42	170	4
5	1	4	1	37	175	5
6	1	3	3	47	176	4
7	2	2	2	42	176	6
8	2	1	2	47	178	6
9	2	4	2	27	178	7
10	2	4	2	52	176	4
11	2	3	1	46	176	5
12	2	2	3	19	160	12
13	3	1	3	37	172	9
14	3	1	1	14	170	6
15	3	2	2	37	176	11
16	3	3	3	42	178	2
17	3	3	1	26	170	4
18	3	4	2	32	174	5
19	4	1	2	57	176	6
20	4	4	3	43	178	6
21	4	3	3	46	170	7
22	4	1	2	40	178	9
23	4	4	2	46	178	12
24	4	4	2	52	176	12
25	4	2	3	50	174	5
26	4	2	3	42	174	6
27	4	2	1	14	174	6
28	4	2	2	57	178	7
29	4	3	2	42	178	12
30	4	3	3	46	176	5

Come si può osservare nella tab.1, l'analisi della matrice dei dati, pur limitata a solo 30 unità e 5 variabili, non rende agevole la comprensione del fenomeno. E' necessario organizzare gli stessi sotto forma di **tabelle**.

Fissiamo l'attenzione su uno dei **caratteri**, in altre parole su una delle colonne della matrice dei dati. Vogliamo vedere come si comporta il campione rispetto a tale carattere. Cominciamo con il considerare il carattere "**tipo**" (quarta colonna). Quando si considera l'analisi su una sola colonna della matrice dei dati si dice che stiamo effettuando un'**analisi statistica univariata**.

Ci chiediamo: **quanti televisori sono stati venduti nel 2014? Di questi, quanti LCD e quanti tradizionali?**

Considerando la quarta colonna della tabella e contando il numero di 1 (CRT), 2 (LCD) e 3 (al plasma) che compaiono, otteniamo la seguente tabella:

Tabella 2: Distribuzione di frequenza (assoluta) del carattere "tipo" nel campione di fornitori utilizzati.

Modalità	Frequenza assoluta n_i
CRT (tradizionale)	6
LCD	14
Plasma	10
Totale	30

In altri termini, abbiamo assunto che le unità statistiche della prima colonna, cui è associata la modalità 1 (rispettivamente 2 o 3) appartengano allo stesso insieme e in seguito abbiamo contato il numero di elementi di questo insieme. Nella costruzione di una tabella del tipo mostrato si definisce **frequenza assoluta** (si indica con n_i) corrispondente ad una certa modalità il numero di unità statistiche che presenta tale modalità. In tal modo nell'esempio si è proceduto per la modalità CRT, LCD e plasma. La tabella che mostra su due colonne affiancate, l'elencazione delle modalità di un carattere e le rispettive frequenze assolute, si chiama **distribuzione di frequenza (assoluta)** del carattere considerato, nel nostro esempio il carattere **tipo**.

Ad esempio, possiamo vedere che nel corso del 2014, sono stati acquistati 14 televisori LCD e solo sei televisori tradizionali.

Dividendo ogni frequenza assoluta per il totale delle unità statistiche considerate (nel nostro caso 30) si ottiene la **distribuzione di frequenza relativa** (si indica con f_i), che possiamo rappresentare in una tabella (cfr.tab.3).

Tabella 3: Distribuzione di frequenza (relativa) del carattere “tipo” nel campione di fornitori utilizzati.

Modalità	Frequenza relativa $f_i = \frac{n_i}{n}$
CRT (tradizionale)	6/30
LCD	14/30
Plasma	10/30
Totale	30/30

La distribuzione di frequenza relativa è spesso fornita in **forma percentuale**. Tale rappresentazione si ottiene moltiplicando per 100 tutti i valori delle frequenze relative, come mostrato nella tabella 4.

Tabella 4: Distribuzione di frequenza (percentuale) del carattere “tipo” nel campione di fornitori.

Modalità	Percentuale $f_i = \frac{n_i}{n} 100$
CRT (tradizionale)	20
LCD	46,67
Plasma	33,33
Totale	100

Riassumendo diamo la seguente definizione.

Definizione

-
- Si definisce *frequenza assoluta*, corrispondente a una certa modalità del carattere, il numero di unità statistiche che presenta tale modalità del carattere.
 - Si definisce *frequenza relativa*, corrispondente a una certa modalità del carattere, la proporzione tra il numero di unità statistiche che presenta tale modalità e il totale delle unità statistiche considerate.
 - Si definisce *frequenza percentuale*, corrispondente a una certa modalità del carattere, la proporzione tra il numero di unità statistiche che presenta tale modalità e il totale delle unità statistiche prese in esame moltiplicata per 100.

Prendendo in considerazione il carattere **marca**, allo stesso modo otteniamo la seguente tabella.

Tabella 5: Distribuzione di frequenza del carattere "marca".

Modalità	Frequenza assoluta n_i	Frequenza relativa $f_i = \frac{n_i}{n}$	Percentuale $f_i = \frac{n_i}{n} 100$
Sony	6	6/30	20
Samsung	9	9/30	30
Grundig	8	8/30	26,67
Rex	7	7/30	23,33
Totale	30	1	100

Nelle tabelle riportate sopra abbiamo costruito alcuni esempi di distribuzione *Statistica univariata*.

Possiamo ora dare una definizione formale rigorosa.

Definizione

Una distribuzione statistica univariata è una coppia di insiemi di cui il primo è l'insieme delle modalità o dei valori assumibili dal carattere considerato, il secondo, in corrispondenza con il primo, è l'insieme costituito dalle frequenze che ogni valore o modalità presenta nella matrice dei dati considerata. Le frequenze possono essere rappresentate in forma assoluta, relativa o percentuale. La distribuzione statistica fornisce un modello più compatto di rappresentazione dei dati che risultano, pertanto, più organizzati e quindi più leggibili.

Dimensione (pollici)	Frequenze assolute
14	2
19	1
26	2
27	1
32	1
35	1
37	3
40	1
42	5
43	1
46	5
47	2
50	1
52	2
57	2

Proviamo ora a considerare il carattere “*dimensione*” del televisore espressa in pollici. Si tratta, come abbiamo già visto, di un carattere quantitativo reso discreto. La tabella associata a tale carattere è riportata a fianco, come si può vedere, di non facile lettura, perché troppo analitica. Per renderla più leggibile è necessario procedere con il raggruppamento dei dati in classi.

Scegliamo di costruire le **classi**⁴ chiuse sia a sinistra sia a destra, vuol dire che gli estremi della distribuzione sono compresi nella classe. Ad esempio nella classe 43|–|46 sono state inserite le frequenze relative ai televisori che vanno da 43 pollici a 46 pollici (1+5).

Tabella 6: Distribuzione di frequenza del carattere “dimensione dei televisori venduti nel 2014”.

Modalità —	Frequenza assoluta n_i	Frequenza relativa $f_i = \frac{n_i}{n}$	Percentuale $f_i = \frac{n_i}{n} 100$
14-26	5	5/30	16,67
27-32	2	2/30	6,67
33-37	4	4/30	13,33
38-42	6	6/30	20
43-46	6	6/30	20
47-51	3	3/30	10
52-57	4	4/30	13,33
Totale	30	1	100

⁴ Si possono costruire classi aperte o chiuse, nel primo caso gli estremi inferiore e superiore sono esclusi dalla classe, nel secondo caso sono compresi. Si possono costruire anche classi aperte (chiuse) a sinistra e chiuse (o aperte) a destra. Per indicare se la classe è chiusa si utilizza una barra verticale.

Osservazione

Raggruppare i dati provoca una perdita d'informazione, infatti non risultano più distinti tra loro i valori che cadono all'interno di una stessa classe. La perdita di informazione è tanto maggiore quanto più grande è l'ampiezza della classe. La tabella per dati raggruppati, però, è molto più leggibile ed è più facile darne una descrizione sintetica. Quest'ultima varia a seconda di come sono state scelte le ampiezze delle singole classi. E' sempre opportuno cercare un compromesso che consenta di ottenere una buona sintesi, senza troppa perdita di informazione.

Ci chiediamo ora: **quanti televisori sono stati venduti di dimensione minore di 38 pollici, oppure, maggiore di 51?**

Per rispondere alla prima domanda occorre sommare le frequenze minori di 38, nella tabella avremo:

$5 + 2 + 4 = 11$ televisori, nel secondo caso, avremo **4** televisori. Operando sulla tabella otteniamo la nuova tabella delle **frequenze assolute cumulate e percentuali** (si indicano, rispettivamente, con N_i e F_i).

Tabella 7: Distribuzione cumulata del carattere "dimensione dei televisori venduti nel 2014".

Modalità └─┘	Frequenza assoluta n_i	Frequenza assoluta cumulata N_i	Percentuale $F_i = \frac{N_i}{n} 100$
14-26	5	5	16,67
27-32	2	5+2=7	23,33
33-37	4	5+2+4=11	36,67
38-42	6	5+2+4+6=17	56,67
43-46	6	5+2+4+6+6=23	76,67
47-51	3	5+2+4+6+6+3=26	86,67
52-57	4	5+2+4+6+6+3+4=30	100
Totale	30		

Definizione

Si dice *distribuzione di frequenza cumulata di un carattere*, la distribuzione che associa ad ogni valore, la frequenza (assoluta, relativa o percentuale) dei valori osservati minori o uguali. Operativamente, la frequenza cumulata, indicata con N_i , è la somma delle frequenze corrispondenti a tutti i valori minori o uguali del valore considerato.

In modo analogo alle frequenze, anche le frequenze cumulate possono essere relative oltre che percentuali, la distribuzione che si ottiene si chiama **funzione di ripartizione empirica**.

Rappresentazioni grafiche

Una volta effettuata l'analisi di frequenza, è necessario rappresentare graficamente i risultati mediante opportuni diagrammi. La scelta del diagramma da utilizzare non è mai univoca, pertanto, ne analizzeremo alcuni.

Il diagrammi a barre

Un diagramma a barre consiste in una successione di colonne, segmenti verticali o rettangoli (a base uguale o arbitraria) per convenzione equidistanti, esse indicano le modalità del carattere, la cui altezza è uguale o proporzionale alla frequenza (assoluta, relativa o percentuale) della modalità corrispondente. Tale tipo di grafico è particolarmente adatto a rappresentare le distribuzioni di caratteri qualitativi. Si può usare, nello stesso tempo, per rappresentare la distribuzione di un carattere quantitativo discreto. Se il carattere è ordinato bisogna disporre le colonne seguendo lo stesso ordinamento delle modalità del carattere se, invece, il carattere è sconnesso è opportuno disporre le colonne a partire dalla più grande e finendo con la più piccola, o viceversa.

Il grafico, riportato come esempio, proviene dall'indagine effettuata dall'Istat⁵ sull'uso e l'abuso dell'alcol in Italia, in esso sono messi a confronto i consumi di bevande alcoliche

Fonte: ISTAT – Statistiche in breve - 2008

Figura 1

in giovani con età superiore a 14 anni, a distanza di 10 anni (1998 e 2008).

Spesso si confonde il diagramma a barre con l'istogramma⁶ che invece può essere utilizzato solo se si hanno caratteri quantitativi raggruppati in classi.

⁵ http://www.istat.it/salastampa/comunicati/non_calendario/20090423_00/

⁶ Anche nei fogli di calcolo più diffusi, spesso si confonde il diagramma a barre con l'istogramma, nel web si trovano innumerevoli esempi di uso scorretto del termine istogramma. La stessa

L'ortogramma

Nell'ortogramma un punto in un piano è individuato con un segmento e con l'ampiezza di un angolo. Dato un punto O (fig.2) chiamato origine delle coordinate polari, ed una semiretta di origine O, detta asse polare, ogni punto A del piano è definito dal valore del segmento OA, detto raggio vettore e dall'ampiezza dell'angolo α detta anomalia. Ogni dato statistico può, pertanto, essere rappresentato per mezzo di due elementi: lunghezza del raggio e ampiezza dell'angolo.

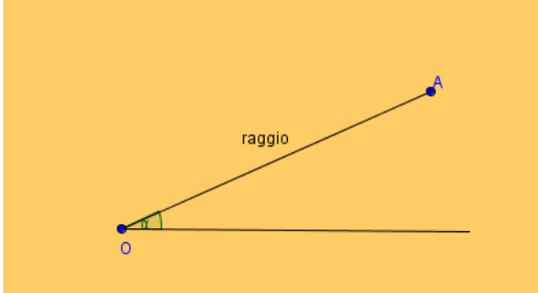


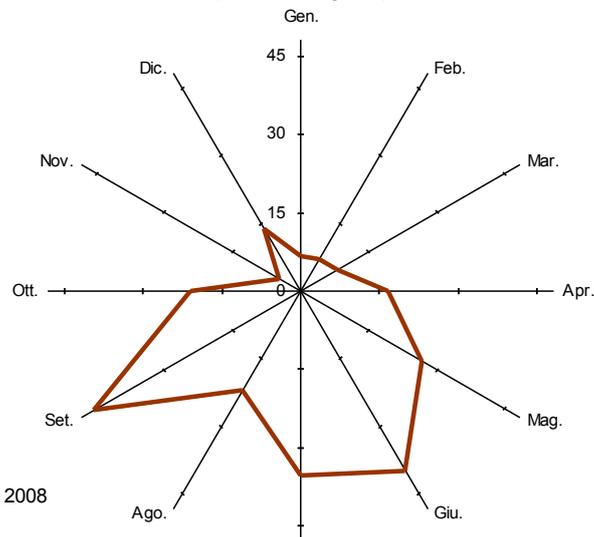
Figura 2

L'ortogramma è particolarmente indicato per rappresentare fenomeni ciclici. Si divide l'angolo giro nel numero delle modalità atte a rappresentare il fenomeno statistico (giorni della settimana, stagioni, mesi dell'anno, ecc.) si misurano sui raggi vettore le intensità del fenomeno o le frequenze corrispondenti alle modalità considerate. Si possono unire poi i punti su ciascun raggio con segmenti o curve chiuse.

Il grafico sottostante, è stato costruito con i dati mensili dei matrimoni celebrati in Italia nel corso del 2006.

Matrimoni per mese di celebrazione - Anno 2006

(valori in migliaia)



Fonte: ISTAT
L'Italia in cifre - 2008

Wikipedia, una delle più autorevoli e gettonate enciclopedie utilizzata da milioni di utenti, riporta una definizione corretta di istogramma ma un esempio grafico scorretto.

Figura 3

Diagrammi a settori circolari o a torta

I diagrammi circolari sono ideali per rappresentare sia fenomeni qualitativi sia fenomeni quantitativi se il numero delle modalità del carattere non è troppo elevato. Si deve dividere l'angolo giro, che si fa uguale all'intero fenomeno, nelle sue diverse modalità, in proporzione alle intensità o frequenze relative di ciascuna modalità del fenomeno. Ne deriva che i vari settori circolari ottenuti dalle ripartizioni sono proporzionali alle intensità o frequenze delle modalità del fenomeno considerato



Figura 4

L'istogramma

Nel caso di caratteri quantitativi continui non ha più senso parlare della frequenza di un singolo valore, poiché non è possibile osservare con esattezza quel dato valore. Il carattere viene raggruppato, pertanto, per classi di ampiezza e ne viene data una rappresentazione grafica attraverso l'istogramma: in ascissa si riportano le classi, adiacenti e rispettando la loro ampiezza, l'area di ogni rettangolo è rappresentata dalla frequenza assoluta o relativa. L'altezza di ogni istogramma deve essere uguale alla **densità di frequenza** (data

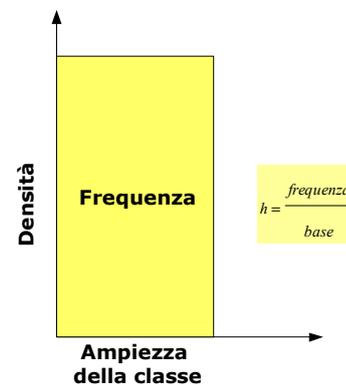


Figura 5

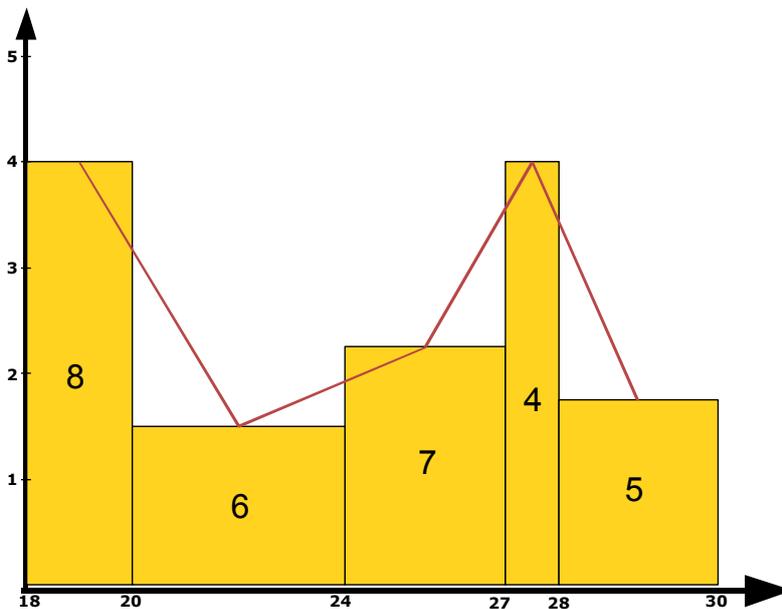
dal rapporto tra frequenza assoluta o relativa e ampiezza della classe)⁷. È ovvio che qualora le classi abbiano la stessa ampiezza (caso più comune), per l'altezza degli istogrammi si può usare indifferentemente la frequenza relativa o assoluta senza il rischio di errori di rappresentazione.

Consideriamo i dati riportati nella tabella riferiti al voto all'esame di statistica di 30 studenti. Le classi utilizzate sono aperte a destra e chiuse a sinistra, ne consegue che lo studente con un voto pari a 24 si troverà nella classe 24-27 e non 20-24.

Modalità ┌	Ampiezza della classe b_i	Frequenza n_i	Densità $\frac{n_i}{b_i}$
18-20	2	8	4
20-24	4	6	1,5
24-27	3	7	2,3
27-28	1	4	4
28 e oltre	3	5	1,67
Totale		30	

Come si può osservare, si tratta di classi di ampiezza diversa. La classe 28 e oltre viene riferita all'intervallo 28-30, essendo 30 il massimo voto conseguibile. Per prima cosa, sarà necessario calcolare l'ampiezza di ciascuna classe per poter individuare la

base di ciascun rettangolo. Una volta calcolata l'ampiezza si potrà misurare la densità (altezza) di ciascuna classe, dividendo la frequenza di ogni classe per l'ampiezza della classe stessa.



Dal momento che l'occhio percepisce meglio l'andamento di un fenomeno se è rappresentato con una linea, si possono considerare i punti medi della base superiore di ciascun rettangolo di un istogramma o di un diagramma a barre e unirli con una spezzata, ottenendo così la **spezzata delle frequenze** o **poligonale di frequenza**.

Figura 6

⁷ Tale quoziente è pari alla frequenza che compete ad una classe di ampiezza unitaria.

Il cartogramma

I cartogrammi sono utilizzati quando si deve rappresentare un fenomeno secondo una ripartizione territoriale. In questo caso si preferisce utilizzare una carta geografica. L'andamento del fenomeno è evidenziato mediante diverse colorazioni delle aree territoriali secondo una scala di graduazione opportunamente riportata sul cartogramma.

A titolo di esempio riportiamo il cartogramma relativo alla percentuale della popolazione italiana con età superiore a 64 anni nel 2011.

INDAGINE POPOLAZIONE OVER 65

Percentuale di popolazione 65 anni e oltre

Dati al 1° Gennaio
Gennaio 2011
FONTE ISTAT
Dati del 2011



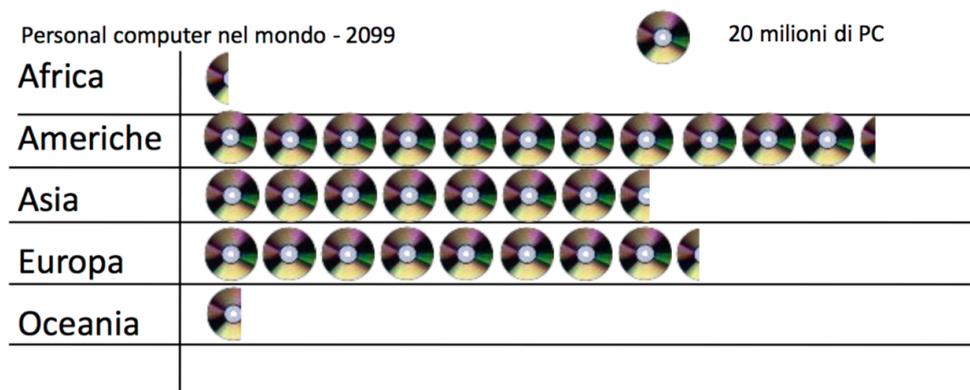
Figura 8

Osservando la carta geografica si può notare che il primato di invecchiamento della popolazione appartiene alla Liguria, con una percentuale maggiore del 24% di anziani. La regione più giovane è, invece, la Campania con una percentuale di over 65

L'ideogramma

Gli ideogrammi sono rappresentazioni grafiche effettuate, non con elementi geometrici, ma con figure reali schematizzate che traducono, in modo visivo, la natura del fenomeno considerato. Sono facili da comprendere e di notevole efficacia ma difettano in precisione. I simboli rappresentativi dell'intensità o frequenza delle modalità del fenomeno sono tali che la loro dimensione è proporzionale all'intensità o frequenza delle modalità stesse. Si possono, inoltre, avere più figure ripetute, tutte di uguale dimensione, alle quali attribuire un determinato valore. Si è preferito, per maggiore chiarezza, inserire in basso anche la tabella dei dati assoluti

Nel grafico è rappresentata la stima del numero di personal computer nel 2099 in ciascun continente.



Bibliografia

Calcolo delle probabilità ed elementi di statistica, L. Daboni – Collezione di matematica applicata, UTET, 1980
.....

Logic, the Theory of Inquiry; 1939 - tr. It. Logica, teoria dell'indagine, **J. Dewey** Giulio Einaudi Editore, Torino 1949; tr. di Aldo Visalberghi
.....

Metodologia della ricerca in educazione e formazione, P. Lucisano, A. Salerno, Carocci, Roma
.....

La matematica dell'incertezza, Didattica della probabilità e della statistica, C. Rossi – Zanichelli, 2000
.....

Disegno sperimentale e analisi dei dati in psicologia, G. Keppel, W. J. Saufley & H. Tokunaga, EdiSES, Napoli, 2001
.....

Statistica: metodi esplorativi e inferenziali, M. Fraire, A. Rizzi, Carocci Editore, Roma, 2005.