Riassunto Statistica

Dimix

2025-03-14

Indice

Capitolo 1 – Introduzione alla Statistica

Definizione

La **statistica** è l'arte di apprendere dei dati. Si occupa della raccolta, della descrizione e dell'analisi dei dati, possibilmente permettendo di trarne delle conclusioni.

La **statistica descrittiva** è quella parte della statistica che si occupa di descrivere e riassumere i dati

La **statistica inferenziale** è quella parte della statistica che si occupa di trarre conclusione dai dati

La statistica inferenziale si basa sul modello probabilistico che consiste nel fare un insieme di assunzioni sulle probabilità di ottenere un certo valore, per cui essa richiede la conoscenza della teoria della probabilità. L'inferenza statistica si basa sull'assunzione che importanti aspetti del fenomeno in analisi si possano rappresentare in termini di probabilità e giunge a conclusioni usando i dati per fare inferenza su queste probabilità.

Popolazione e campioni

Nella statistica è cruciale ottenere delle informazioni su tutto un insieme di elementi, che vengono definiti **popolazione**. Spesso la popolazione pero è troppo numerosa per poter analizzare ciascuno dei suoi membri, in questo caso si sceglie e si esamina un suo sottoinsieme, che viene definito **campione**

Affinché il campione ci dia informazioni su tutta la popolazione, esso deve essere scelto in modo da essere *rappresentativo* di tutta la popolazione. Rappresentativo significa che il campione deve essere scelto in modo che tutte le parti della popolazione abbiano uguale probabilità di fare parte del campione, quindi esso deve riflettere la variabilità reale della popolazione.

Un campione di k membri di una popolazione si dice **campione casuale**. o talvolta *campione casuale semplice* quando i membri sono scelti in modo che tutte le possibile scelte dei k membri siano ugualmente probabili.

Una volta scelto il campione casuale, si può utilizzare l'inferenza statistica per giungere a conclusioni sull'intera popolazione studiando gli elementi del campione.

Campione casuale stratificato

Un metodo più sofisticato del campionamento casuale semplice è il **campionamento casuale stratificato**. inizialmente si stratifica la popolazione in *sottopopolazioni*, ognuna dalle quali contiene unità simili secondo determinati criteri. in seguito da ogni strato si estrae casualmente un numero di unità proporzionale alla sua consistenza nella popolazione totale. in questo modo, le proporzioni di ciascun strato presenti nel campione rispecchiano esattamente quelle dell'intera popolazione.

La stratificazione è particolarmente efficace per conosce il membro *medio* della popolazione totale quando ci sono differenze tra le sottopopolazioni rispetto alla questione studiata.

La statistica descrittiva

Capitolo 2 - Descrivere insiemi di dati

Dati quantitativi e qualitativi

Una distinzione che si può fare sui dati osservabili riguarda il modo in cui questi sono misurati:

- **Dati quantitativi**: l'esito della misurazione è una quantità numerica.
- **Dati qualitativi**: l'esito della misurazione è un'etichetta appartenente a un insieme fissato di etichette, vengono anche detti categorici o nominali.

Classificazione dei dati qualitativi

I dati qualitativi si distinguono in dati binari, nominali e ordinali:

- Dati binari o booleani: possono assumere soltanto due valori tra loro non confrontabili, si utilizza *booleani* per evidenziare la presenza/assenza di una proprietà, mentre binari per indicare due etichette possibili.
- Dati nominali: non ammettono un confronto d'ordine tra i valori, ma è possibile stabilire una relazione di equivalenza.
- Dati ordinali: se abbiamo due valori diversi riusciamo a stabilire quale sia il più piccolo e quale il più grande, quindi esiste una relazione d'ordine tra i valori.

Classificazione dei dati quantitativi

I dati quantitativi si distinguono in discreti e continui a seconda dell'insieme di valori che possono assumere:

- Dati discreti: costituiscono variabili che possono assumere un insieme numerabile di valori distinti e separati, ad ogni valore corrisponde un significato specifico.
- Dati continui: in teoria possono assumere un qualsiasi valore all'interno di un intervallo, anche se nella pratica vengono appossimati a una precisione finita, per via della memorizzazione digitale.

Frequenze

La **frequenza assoluta** di un'osservazione x in un insieme di dati $A = \{x_1, \cdots, x_n\}$ è definita come il numero di volte in cui x compare in A.

In modo formale possiamo indicarla con f_x la frequenza assoluta di x, si ha che $f_x = \#\{j \in \{1, \cdots, n\} \mid x_j = x\}$

La **frequenza relativa** consente di esprimere la presenza di ogni valori in termini di proporzione rispetto all'intero campione. sia $A = \{x_1, \cdots, x_n\}$ un insieme di n dati e sia f_i la frequenza assoluta di un osservazione x_i in A, possiamo definire frequenza relativa di x_i il valore f_i/n.

Teniamo presente che la somma di tutte le frequenze relative in un campione è sempre uguale ad 1.

Le **frequenze cumulate** si ottengono quando i valori di una variabile possono essere ordinati. il procedimento consiste nel disporre i valori in ordine crescente, calcolare le loro frequenze individuali e poi sommarle progressivamente: al primo valore si associa la sua frequenza, al secondo la somma della frequenza del primo e del secondo, al termo la somma delle frequenze dei primi tre e cosi via.

l'ultima frequenza cumulata rappresenta il totale dei casi osservati. inoltre possiamo applicare il concetto di frequenza cumulata sia alle frequenze assolute che a quelle relative, in caso di frequenze relative i valori cumulati variano da 0 a 1.

Quando i dati sono numerici o comunque ordinabili, un concetto affine alle frequenze cumulate è quello della **funzione cumulativa empirica**, nota anche come funzione di ripartizione empirica. Data una serie di osservazioni $x_1, \cdots, x_n$, la funzione empirica $\hat{F} : \mathbb{R} \to [0, 1]$ è definita in modo che per ogni $x \in \mathbb{R}$ essa assume il valore pari alla frequenza relativa delle osservazioni minori o uguali a x. in altre parole:

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(-\infty,x]}(x_i)$$

dove $\mathrm{I}_A : \mathbb{R} \to 0, 1$ è la funzione indicatrice dell'insieme A, che restituisce 1 se l'argomento appartiene ad A o 0 altrimenti: di conseguenza l'intervallo $(-\infty, x]$ include tutti i valori minori o uguali a x. Pertanto, per ogni x, $\hat{F}(x)$ rappresenta la frequenza relativa cumulata del massimo valore osservato che non supera x, e il grafo di questa funzione sarà a tratti costanti.

$$I_A(x) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases} \quad \Rightarrow \quad I_{(-\infty,x]}(x_i) = \begin{cases} 1 & \text{se } x_i \in (-\infty, x] \\ 0 & \text{se } x_i \notin (-\infty, x] \end{cases} \quad = \quad \begin{cases} 1 & \text{se } x_i \leq x \\ 0 & \text{se } x_i > x \end{cases}$$

in pratica rappresenta il numero di osservazioni dei miei campioni che sono minori o uguali di una certa x, diviso per il numero totale di campioni. La divisione per n è fatta per avere dei valori relativi

Le **frequenze congiunte e marginali**, quando si analizza un insieme di osservazioni, può essere utile considerare due caratteri contemporaneamente, in modo da verificare se esiste una relazione tra i valori dei due attributi. In questo caso, il concetto di frequenza si adatta contando il numero di occorenze in cui i due caratteri assumono contemporaneamente determinati valori. Questo conteggio porta alla definizione di *frequenza congiunta assoluta*, se invece si considera una frazione delle osservazioni, si parla di *frequenza congiunta relativa*.

Se il numero dei possibili valori osservabili per i caratteri non è elevato, possiamo rappresentare visivamente queste frequenze tramite una *tabella delle frequenze congiunte* o *tabella di contigenza*. in questa tabella, le righe sono associate ai valori di uno dei caratteri, mentre le colonne rappresentano i valori del secondo carattere. Gli elementi all'interno della tabella indicazione le frequenze congiunte per le coppie di valori.

Per facilitare ulteriori analisi, si riportano spesso nelle ultime colonne e nelle ultime righe della tabella le *frequenze marginali*, ottenute sommando rispettivamente i valori per ogni riga e per ogni colonna. Se si desiderano valori relativi, questi totali devono essere normalizzati rispetto al numero complessivo delle osservazioni.

Grafici

La **simmetria**, un insieme di dati è detto simmetrico attorno a un valore x_0 se, per ogni scostamento c da x_0, la frequenza dei valori $(x_0 - c)$ è uguale a quella dei valori $(x_0 + c)$. In tal caso, il valore x_0 viene detto **centro di simmetria** della distribuzione.

La **quasi simmetria** si presenta quando i dati non sono permettamente simmetrici, ma la distribuzione rimane quasi specuplare rispetto a un punto centrale.

Un modo semplice per rendersi conto se una distribuzione è (quasi) simmetrica consiste nel rappresentarla graficamente e osservare la sua forma.

Grafici per la frequenza

Se l'insieme di dati contiene un numero ridotto di valori distinti allora è rappresentabile con una tabella delle frequenze. Questa tabella associa a ciascun valore distinto osservato la sua frequenza assoluta. La somma di tutte le frequenze deve corrispondere al numero totale di osservazioni. Data una variabile statitisca $\boldsymbol{X}$ che può assumere vari valori, si elencano i valori distinti di $\boldsymbol{X}$ in una colonna e, a fianco di ognuno, la relativa frequenza di occorenza nel campione.

Per costruire la tabella delle frequenze relative da un insieme di dati, bisogna innanzitutto disporre i valori dei dati in ordine crescente. Si determinano i valori distinti e quante volte ciascuno di essi compaia. Si elencano questi valori distinti affiancati dalla loro frequenza f e dalla loro frequenza relativa f/n, dove n è il numero totalte di osservazioni nell'insieme di dati.

Grafici a bastoncini, a barre e poligonali

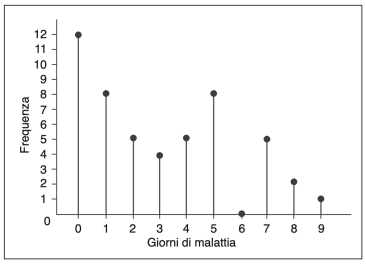

Figura 2.1 Un grafico a bastoncini.

I dati di una tabella di frequenza possono essere rappresentati graficamente in diversi modi. Uno dei più intuitivi è il *grafico a bastoncini*, in cui i valori della variabile statistica sono disposti lungo l'asse orizzontale, mentre le frequenze si riportano sull'asse verticale. Ogni valore viene quindi associato a un semplice segmento che parte dall'asse orizzontale e arriva all'altezza corrispondente alla relativa frequenza.

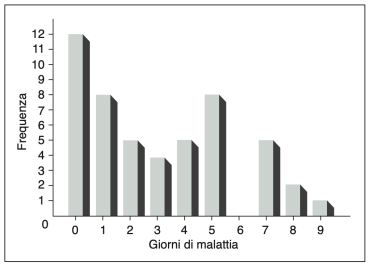

Figura 2.2 Un grafico a barre.

Un secondo tipo di rappresentazione, molto simile concettualmente, è il *grafico a barre*: anche in questo caso i valori si trovano sull'asse orizzontale e le frequenze su quello verticale, ma invece dei singoli segmenti si utilizzano barre di un certo spessore. Ciò permette di mettere in evidenza ciascuna categoria o classe di dati e risulta particolarmente efficace quando si vogliono confrontare categorie di grandezza diversa.

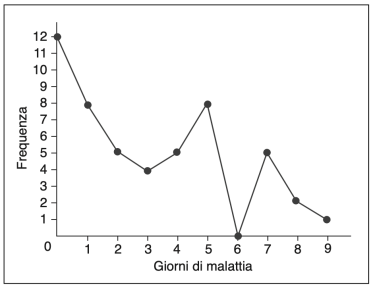

Figura 2.3 Un grafico poligonale.

Infine, esiste il *grafico poligonale*, in cui i valori (sempre disposti sull'asse orizzontale) vengono rappresentati da punti, collocati a un'altezza proporzionale alla loro frequenza, che vengono poi congiunti da segmenti. In questo modo si ottiene una linea spezzata che rende immediata la visualizzazione delle variazioni di frequenza da un valore all'altro, permettendo di apprezzare più facilmente tendenze o andamenti complessivi.

Diagramma a torta

In caso di dati non sono numerici si utilizza un diagramma a torta, esso consiste in un cerchio suddiviso in settori, uno per ogni valore distinto dei dati. Dato un valore con frequenza relativa f/n, allora l'area del settore corrisponde all'area del cerchio moltiplicata per f/n, ovvero un arco con un angolo di $360 \cdot (f/n)$ gradi.

Diagramma ramo-foglia

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575
26	183, 894, 982
27	671, 711, 744
28	345, 764, 913, 967

Tabella 1: Diagramma a stelo

Un modo efficiente di rappresentare un insieme di dati di dimensioni medie consiste nell'utizzare il *diagramma ramo-foglia* (o a stelo). Tale grafico si ottiene dividendo ciascun valore dei dati in due parti, chiamati appunto rami e foglie.

La scelta dei rami dovrebbe essere fatta in modo che il diagramma ramo-foglia che ne risulta sia informativo sui dati. Questi diagrammi sono particolarmente adatti a descrivere insiemi di dati dimensioni ridotte.

Questo grafico ha l'aspetto di un istogramma ruotato su un lato, con il vantaggio di contenere tutti i valori dei dati originali in ogni classe. Quando il grafico presenta troppe foglie per ogni riga, si può raddoppiare il numero di rami utilizzando due righe per ogni valore del ramo.

Diagramma di Pareto

I diagrammi di Pareto sono grafici a barre ordinate in ordine decrescente di frequenza, ai quali è spesso affianca una linea che rappresenta la frequenza cumulata. In questo modo, oltre a mostrare il numero di casi per ciascuna categoria, permettendo di evidenziare quali categorie contribuiscono maggiormente al totale, facilitando l'individuazione delle cause o delle categorie più rilevanti.

Istogrammi e raggruppamento dei dati

Utilizzare i grafici precedenti è un metodo efficace per descrivere un insieme di dati, ma alcuni di questi insiemi hanno troppi valori distinti per poter usare questo metodo. Perciò è necessario suddividere i valori in gruppi, o classi, e poi rappresentare con un grafico il numero di valori dei dati che cadono in ciascuna classe. Il numero di classi scelte è un compromesso tra:

- Scegliere poche classi al costo di perdere molte informazioni sui valori effettivi in una classe.
- Scegliere troppe classi, ottenendo frequenze troppo basse all'interno di ciascuna di esse.

I valori al bordo di una classe si chiamano **estremi** della classe. Si adotta la convenzione di inclusione a sinistra, cioè una classe include il suo estremo sinistro ma non quello destro.

Una volta suddivisi i dati in classi, si costruisce la tabella delle frequenze (e delle frequenze relative), e da questa si ottiene l'istogramma, un grafico a barre adiacenti che mostra la distribuzione dei dati in ciascuna classe. L'istogramma offre una visione immediata di come i valori si distribuiscono: per esempio, se sono concentrati in un certo intervallo, se ci sono vuoti senza osservazioni o se alcuni valori si distaccano notevolmente dagli altri. Pur non contenendo tutte le informazioni dell'insieme di dati originale, la tabella delle frequenze di classe e l'istogramma illustrano le caratteristiche fondamentali della distribuzione, come la simmetria, la dispersione e i possibili estremi isolati.

Diagramma di dispersione e insieme di dati a coppie

Un insieme di dati può consistere in coppie di valori che hanno una relazione di qualche tipo tra di loro. Ne viene che ogni elemento dell'insieme di dati sia costituito da un valore x e da uno y. Si indica con $(x_i, y_i),\ i = 1 \cdots n$ la i-esima coppia.

Un metodo per rappresentare un insieme di dati di questo tipo consiste nel considerare ogni elemento della coppia separatamente, producendo istogrammi (o diagrammi ramo-foglia) separati per ciascuno. Cosi facendo però, nonostante i due grafici ci diano molter informazioni sulle singole variabili (attributi), non si ha nessun tipo di informazione riguardo al rapporo tra queste due variabili.

Per capirne la relazione è necessario considerare i valori accoppiati di ciascun dato simultaneamente. Si possono allora rappresentare questi dati accopiati in un diagramma rettangolare e bidimensionale, in cui l'asse x rappresenta il valore x dei dati, e l'asse y il valore y. Cosi facendo si ottiene un **digramma di dispersione**.

Una delle ragioni per cui questo tipo di diagramma è utile consiste nella possibilità di fare previsioni sul valore y di una futura osservazione, noto il valore x. Per stimare il valore y a partire da x si cerca, in modo intuitivo, di trammare una **retta media** che approssimi l'andamento dei punti sul diagramma, ovvero una retta che passi *il più vicino possibile* a tutti i dati.

- in pratica, si ricorre a metodi di regressione lineare, come il *metodo dei minimi quadrati*, che permette di trovare l'equazione della retta (del tipo $y = a + bx$) minimizzando la somma delle distanze (al quadrato) tra i valori osservati (x_i, y_i) e i valori $\widehat{y}_i$ previsti dalla retta. Una volta trovata questa retta di miglior adattamento, per un qualunque valore x che possa presentarsi in futuro, si ottiene la stima di y semplicemente sostituiendo x nella equazione $y = a + bx$.

Il diagramma di dispersione, oltre a mostrare il comportamento relativo di due variabili e ad aiutarci nelle previsioni, è utile per riconoscere i valori anomali (**outlier**) che sono i punti che non sembrano seguire il comportamento degli altri. Una volta identificati questi valori, si può decidere quali di essi siano appropriati e quali invece siano causati da errori nella raccolta dati.

Capitolo 3 - Statistiche

Una **statistica** è una quantità numerica calcolata a partire da un insieme di dati.

Centralità

Media campionaria

Media pesata

Scarti

Mediana campionaria

Percentili campionari

Moda campionaria

Dispersione

Varianza campionaria

Deviazione standard campionaria

Scarto interquantile

Altri grafici

Box Plot

Q-Q Plot

Distribuzioni normali

Formulario

Frequenza cumulata:

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(-\infty,x]}(x_i)$$