Statistica & Analisi dei dati

Mascherpa Matteo

a.a. 2024-2025

Indice

1 Introduzione alla Statistica

1.1 Definizioni:

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

Statistica descrittiva: La parte della statistica che si occupa di descrivere e riassumere i dati.

Statistica inferenziale: La parte della statistica che si occupa di trarre conclusioni dai dati. (Richiede quindi di conoscere i concetti di probabilità)

Popolazione: L'insieme della totalità dei dati.

Campione: Sottoinsieme omogeneo delle popolazione.

Campione Casuale: Il campione è un sottoinsieme della popolazione dove ogni elemento ha la stessa probabilità di essere scelto.

Frequenza assoluta: Numero di volte che un dato compare in un campione. Indicata con f_j dove j è un elemento del campione.

Frequenza relativa: Frazione di volte che un dato compare nel campione. Indicata con $f'_j (= \frac{f_j}{n})$.

1.2 Popolazione e Campioni

In statistica si vuole ottenere informazioni da un insieme di dati(*popolazione*. Ma dato la natura delle popolazione, spesso enorme, se ne prende un sottoinsieme di distribuzione omogenea(*Campione*) dove ogni elemento ha una stessa probabilità di essere scelto. In caso il campione non abbia questa qualità le informazioni da esso estratto sono inconcludenti poiché non *rappresentativo*. Una volta ottenuto il *campione casuale* si può usare tecniche di *statistica inferenziale* per ottenere da esso le informazioni volute(dato che dal campione si possano ricavare, da un insieme di dati sul colore dei vestiti non si potrà in ogni caso trovare alcun dato sui gusti dei gelati).

Campione casuale stratificato

Prendere valori a caso non è spesso la migliore delle scelte, è una scelta buona, ma esistono opzioni migliori. Per questo esiste il concetto di ***campione casuale stratificato***. Prima di prendere gli elementi casuali del campione si divide la popolazione in *classi* secondo un qualche criterio. Una volta si prende da ogni classe un numero di elementi proporzionale alla cardianlità della classe. Se una classe è composta dal 15% della popolazione allora il 15% del campione dovrà essere preso casualmente dalla quella classe, e così via.

2 Rappresentazione dei dati

Serve un metodo per rendere i dati leggibili e facilmente interpretabili, a questo scopo ci vengono in soccorso un potente strumento grafico, i *grafici*.

2.1 Tabelle e Grafici

Avendo un insieme di dati di dati disposti nel modo più elementare possibile la lettura diventa complessa, invece, gli stessi dati, una volta riarrangiati vengono letti con semplicità.

2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4,
0, 4, 0, 1, 8, 9, 7, 0, 1, 7, 2, 5, 5, 4, 3, 3, 0,
0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1

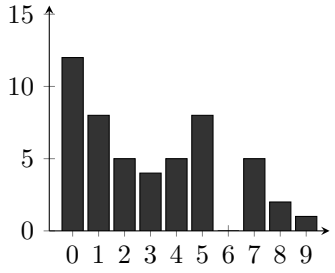
Valore	**Frequenza**
0	12
1	8
2	5
3	4
4	5
5	8
7	6
8	2
9	1

2.2 Grafici a bastoncini, a barre & poligonali

Varie sono le tipologie di grafici, tra le opzioni ci sono grafici a bastoncini, a barre e poligonali.

DRAW HERE A STICK CHART.

Grafico dalla lettura intuitiva dove ogni dato è rappresentato da un semplice segmento.



Spesso utilizzati per la rappresentazione di dati, per ogni dato sull'asse delle ascisse esiste un parallepipedo dall'altezza equivalente al valore del dato che si vuole sappresentare. Comodo per la visualizzazione dei valori di varie categorie diverse in un unico grafico.

Un insieme di dati si dice simmetrico al valore x_n se le frequenze dei valori $x_n - c$ e $x_n + c$ sono le stesse per ogni c. Si dice *quasi simmetrico* se i valori sono precisamente uguali ma sono solamente simili, è quindi una proprietà meno restrittiva.

Grafici per le sequenze relative

A volte è conveniente, al posto di avere le frequenze assolute, visualizzazare le frequenze *relative*. Dato f la frequenza di x allora posso avere un grafico *frequenza relativa* $\frac{f}{n}$ dove n è il numero totale di osservazioni del dato. In caso la somma dei valori delle colonne farà 1 cioé tutte tutte le osservazioni.

Grafici a torta

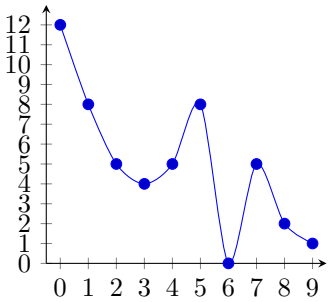
Raggrupamenti di dati e istogrammi

Serve quando la quantità dei dati è tale da rendere inutile la rappresentazione normale in un grafico. In tal caso conviene raggruppare in classi i dati. Trovare il numero perfetto di classi è spesso molto complesso e ci si può accontentare con un compromesso tra, scegliere poche classi(a costo di perdere molte informazioni) e scegliere molte classi(in modo da mantenere il significato dei dati nel campione). Di solito il valore è compreso in $[5, 10]$. Il libro usa la convenzione di includere in una classi il suo valore inferiore e non quello superiore: $[\lim_{inf}, \lim_{sup})$. Ogni valore è la media dei valori nella classe.

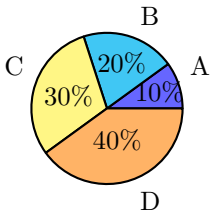
Diagrammi Ramo-Foglia

22	123
23	234, 567, 890
24	345
25	012, 034, 078, 234
26	123, 456, 789
27	234, 345, 456
28	567, 678, 789, 890

Comodo per rapprentare un piccolo campione di dati. Dove il valore a sinistra è il prefisso e i valori a destra sono tutti i valori che hanno quel prefisso. Questo serve a riassumere i dati. In caso un ramo abbia troppe foglie si può dividere su più linee per una maggiore leggibilità.



Il grafico poligonale infine serve a rendere visibile l'andamento di un dato unendo ogni punto con il successivo tramite un segmento creando quandi una sorta di funzione che indica l'andamento del valore sull'asse delle ordinate.



In caso che i dati non siano numerici una valida opzione è il grafico a torta che indica le frequenze relative di ogni dato. La percentuale di grafico da assegnare ad un solo valore si calcola $\frac{f}{n}$ dove n è il numero totale di osservazioni e per ottenere. Per invece ottenere l'angolo del grafico a torta la formula diventa $\frac{360*f}{n}$.

3 Statistica descrittiva

3.1 Tipi di dato

Prima di addentrarci nella statistica si deve introdurre il concetto di tipo di dato.

- **Dati quantitativi**: Dove il valore è effettivamente una quantità numerica
- **Dati qualitativi**: Dove il valore è un *tag* che appartiene ad un insieme di *tag*.

3.1.1 Dati quantitivi

Le principali tipologie di dati quantitativi dipendono dall'insieme di valori che il dato può valere. Esse si dividono in:

- **Discreti** dove il valore appartiene ad un insieme *discreto*, cioé che tra un valore e un altro può non esserci un valore intermedio. Un esempio di dati **quantitativi discreti** può essere l'insieme dei voti di un esame compresi nell'intervallo $[1; 30]$.
- **Continui** dove il valore appartiene ad un insieme continuo, cioé tra due valori ne esiste sempre un terzo intermedio. Un esempio di dati **quantitativi continui** può essere le percentuali di presenze di un gruppo di operai.

3.1.2 Dati qualitativi

Le principali categorie sono: *binari*, *nominali* e *ordinali*.

- Si definisce **binario** quel valore che ad ogni osservazione può ottenere un solo tra due arbitrari valori distinti e non confrontabili. Un esempio potrebbe essere un campione di animali composto da soli cani e gatti, ad ogni osservazione il valore può essere uno di due.

- Si definisce **nominale**, un insime di dati non confrontabili ma che non sono limitati a soli due valori. Ampliando il precedente esempio se nel campione si aggiungono mucche e cavalli il campione sarà composto da *dati qualitativi nominali*.
- Si definisce **ordinale** quando nell'insieme di dati si possono mettere in relazione i valori del campione in modo da sapere quale valore sia inferiore o maggiore di un altro. Un esempio potrebbe essere quelle della misura di una cottura di bistecca, *al sangue* che è inferiore a *cottura media* che è inferiore a *ben cotta*. Nonostante non siano valori numerici si può comunque dedurre quale bistecca sia più o meno cotta.

3.2 Funzione cumulativa empirica

La **funzione cumulativa empirica**, (**ECDF**), strumento per stimare la distribuzione cumulativa di una variabile casuale in un campione di dati. Utile per scoprire la distribuzione dei dati nel campione.

Definizione: Dato un insieme di osservazioni $x_1, \ldots, x_n$ è definita come quella funzione $\hat{F} : \mathbb{R} \to [0; 1]$ tale che per ogni $x \in \mathbb{R}$ assume un valore pari alla *frequenza relativa* delle osservazioni minori o uguali a x. Si può vedere la formula come la stima della funzione di ripartizione.

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

3.3 Frequenze

Frequenza Assoluta

La **frequenza assoluta** di un'osservazione x in un campione di dati $A = \{x_1, \ldots, x_n$ è il numero di apparizioni di x in A. Formalmente indicata con f_x, la frequenza assoluta di x, ovvero $f_x = \#j \in \{1, \ldots, n\} : x_j = x$.

Frequenza Relativa

La **frequenza relativa** indica la presenza di ogni valore proporzionalmente sull'intero campione. Sia $A = x_1, \ldots, x_n$ il campione di dati e f_i la frequenza assoluta di $x_i \in A$ allora la *frequenza relativa* di x_i è $\frac{f_i}{n}$. Si può dimostrare quindi che la somma di tutte le frequenze relative è 1.

Sia f_i la frequenza di apparizione di x_i, $A = \{x_1, \ldots, x_n\}$ e n la cardinalità di A allora $\sum_{i=1}^{n} f_i = n$. La *frequenza relativa* è $f'_i = \frac{f_i}{n}$. La Somma di tutte le f' è $\sum_{i=1}^{n} f'_i = \sum_{i=1}^{n} \frac{f_i}{n}$. Ma essendo n sempre constante la posso portare fuori, $\frac{1}{n} \sum_{i=1}^{n} f_i$ ma sappiamo che la somma di tutte le frequenze è uguale a n arrivando a $\frac{n}{n} = 1$.

Frequenze cumulate

Le *frequenze cumulate* si hanno quando i valori sono ordinabili. Infatti per calcolare le *frequenze cumulate* serve ordinare i dati in ordine crescente. Si sommano poi progressivamente le varie

frequenze relaative partendo dal primo. Per il primo valore solo la prima frequenza e per il terzo valore la somma delle frequenze del primo, secondo e terzo valore e così via. Anche qua il valore deve appartenere a $[0;1]$ dove il valore 1 indica la *frequenza cumulata* dell'ultimo valore, visto che è la somma di tutte le frequenze. Per calcolarla ci avvaliamo della *funzione cumulativa empirica* $\hat{F} : \mathbb{R} \to [0;1]$, la formula è:

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

dove $I_A : \mathbb{R} \to [0,1]$ indica l'appartenenza a A dove 1 indica l'appartenenza all'insieme e 0 non appartiene all'insieme. Così si ha tutti i valori da $(-\infty, x)$ e si potrà sapere tutti i numeri minori di x.

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

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 per avere la *frequenza relativa*.

Frequenze congiuente e marginali

Le frequenze congiunte servono a considerare due caratteri contemporaneamente per verificare la relazione tra due attributi. Si può così contare il numero di occorrenze di due dati correlate. In casa ci sia una frazione si ottiene una *frequenza congiunta relativa*.

In caso il numero di correlazioni non fosse elevato allora si possono rappresentare visivamente su una tabella di congiunzione. Gli elementi all'interno della tabella indicano le *frequenze*(*assolute, relative*). Le *frequenze marginali* sono le frequenze ai margini del campione, in caso fossero relative i valori vanno normalizzati.