Statistica e Analisi dei dati

Matteo Mascherpa

a.a. 2024/2025

Indice

Lezione 1

Leggere la prima dispensa.

Lezione 2

Leggi la seconda dispensa.

Lezione 3

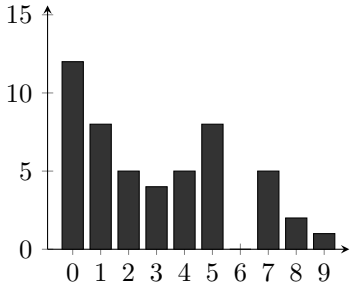
Tabelle e grafici di frequenza

Dato un campione di dati bisogna trovare un modo per disporre i dati in modo che sia di più facile lettura.

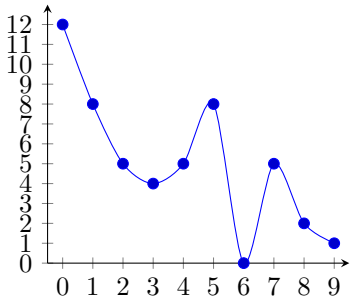
Grafi a bastoncini, a barre e poligonali

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.



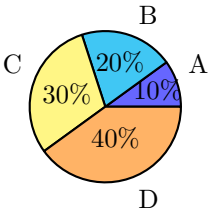
Il grafico poligonale infine serve a rendere visibile l'andamento di un dato unendo ogni punto con il successivo tramite un segmento creando quandi una sorta di funzione che indica l'andamento del valore sull'asse delle ordinate.

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 ossservazioni del dato. In caso la somma dei valori delle colonne farà 1 cioé tutte tutte le osservazioni.

Grafici a torta



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 $\frac{360 * f}{n}$.

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, 890
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 suffisso e i valori a destra sono i valori che hanno quel suffisso. Questo serve a riassumere i dati. In caso un ramo abbia troppe foglie si può dividere.

Lezione 4

Lezione 5

Lezione 6

Mediana Campionaria

Sia il campione: $23, 04, 02, 2, 110, 5, 7, 6, 7, 3$ la *media campionaria* è $\overline{x} = 140/7 = 20$. Se si vuole avere un valore che identifichi il centro del campione serve la *mediana campionaria* per indicarla uso m.

> Dati i valori di un campione ordinati in ordine crescente. Se la cardinalità del campione è **dispari** allora m è il valore intermedio della lista altrimenti, se la cardinalità è **pari** allora m è la media dei due valori intermedi.

Al contrario della *media campionaria* che prende in considerazione tutti i valori degli insiemi dati la *mediana campionaria* non è influenzata dai valori estremi.

Percentili campionari

La media campionaria è una caso di statistica nota come: *100p-esimo percentile campionario* con $p \in [0,1]$. Esso è un valore che è maggiore del(di almeno) $100p\%$ dei valori del campione e minore del(di almeno) $100(1-p)\%$ dei valori. Nel capo della mediana $p = 0,5$.

Come trovare il $100p - esimo$ percentile

1. Ordina i dati in ordine crescente
2. Se np non è intero, trova il più piccolo $\geq$ di np. Il valore è quello nella posizione trovata.
3. Se np è intero, allora il valore è la media tra i valori in posizione np e $np+1$

Quartili

I quartili suddividono il campionario dei dati in quattro parti 25% l'una.

1. Primo quartile: Il 25-esimo percentile.
2. Secondo quartile: Il 50-esimo percentile.
3. Terzo quartile: Il 75-esimo percentile.

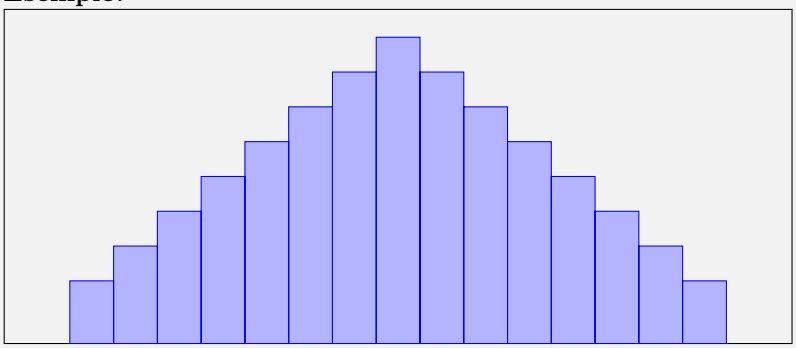
Gli insiemi di dati normali e la regola empirica

La maggior parte degli istogrammi hanno un simile aspetto. Sono spesso simmetrici sulla frequenza massima e assumono una forma a campana. L'insieme di questi istogrammi si dice*istogrammi normali*.

Un insieme si dice *normale* se il rispettivo istogramma:

1. Ha punto di massima in corrispondenza dell'intervallo centrale.
2. Spostandosi dal centro in una qualsiasi direzione l'altezza cala in modo da creare una forma a campana.
3. L'istogramma è simmetrico a rispetto all'intervallo centrale.

Esempio:



Regola empirica

Se un insieme è approssimativamente normale con media $\overline{x}$ con una devizione standard s.

1. Approssimativamente 68% dei dati si trovano nell'intervallo: $[\overline{x} - s, \overline{x} + s]$
2. Approssimativamente 95% dei dati si trovano nell'intervallo: $[\overline{x} - 2s, \overline{x} + 2s]$
3. Approssimativamente $99,7\%$ dei dati si trovano nell'intervallo: $[\overline{x} - 3s, \overline{x} + 3s]$

Grafico ramo-foglia

Una versione di 'istogramma' sdraiato sul lato, utilizzato per notare velocemente se un grafico è normale.

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

Sulla sinistra si trova il *prefisso* del valore e sulla destra i valori con quel suffisso.

La scelta del suffisso deve essere scelto in modo da rendere comprensibile il grafico. Perde la sua utilità su campioni di grandi dimensioni.

Coefficenti di correlazione campionaria

Si consideri campione formato da dati accoppiati $(x_1, y_1)(x_2, y_2) \ldots, (x_n, y_n)$, il *coefficente di correlazione campionaria* quantifica in che misura i grandi valori di x corrispondono ai grandi valori di y. La correlazione può essere positiva, negativa, etc....
Si consideri un campione così composto: $(x_i, y_i)|i = 1, \ldots, n$ le medie campionarie sono $\overline{x}, \overline{y}$. Considero per la i-esima coppia considero lo scarto di x con la sua media e lo scarto di y con la sua media $(x_i - \overline{x}, y_i - \overline{y})$. Saprò quindi se x supera la sua media e se y supera la sua media, se $x - \overline{x} > 0$ allora x è maggiore viceversa, stesso vale per y. Si può facilmente notare che se lo scarto per x e per y hanno lo stesso segno il prodotto sarà positivo e in caso di disomogeneità avrà segno negativo. Si ottiene così il tipo di correlazione tra una coppia di dati, lo scarto del campione viene calcolato somma il valore degli scarti ($\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$). Si standardizza la somma divedendo per $n - 1$ e per le due deviazioni standard.

Dati S_x e S_y rispettivamente, le diviazione standard di x e y. Il *coefficente di correlazione campionaria* detta r con le coppie $(x_i, y_i), i = 1, \ldots, n$:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)\sqrt{\frac{\sum_{i=1}^{n}(x-\overline{x})^2 \sum_{i=1}^{n}(y-\overline{y})^2}{(n-1)^2}}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x - \overline{x})^2 \sum_{i=1}^{n}(y - \overline{y})^2}}$$

1. r è compreso $[-1, 1]$
2. r è uguale $+1$ se, per una costante a $y_i = a + bx$ dove b è costante positiva.
3. r è uguale -1 se, per una costante a $y_i = a + bx$ dove b è costante negativa.
4. Se r è il é *coefficente di correlazione campionaria* per i dati $x_i, y_i, i = 1, \ldots, n$ allora per qualunque costante a, b, c, d, r allora il *coefficente di correlazione campionaria* per i dati: $a + bx_i, c + dy_i, i = 1, \ldots, n$.

Si può dire anche che $|r|$ è la misura dell'intensità della *relazione lineare*. Un valore di 0.9 indica una forte relazione mentre un valore come 0.3 ne indica una debole.

Formula per il calcolo di r:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Lezione 7

Concentrazione

Può servire per valutare la dispersione di un carattere, dove minima significa che è poco distribiuta e massima significa che è molto distribiuta. Dato il campione $a_1 \leq a_2 \leq \ldots a_n$ allora il totale è: $\sum_{i=1}^{n} a_i$. La frazione di ricchezza dei primi i individui è: $\forall i = 1, \ldots, n_i = \frac{i}{n}\%$ individui. $Q_i = \frac{1}{\text{TOT}} \sum_{i=1}^{n} q_k$.

Dimostrazione:
... (Recupera)

Tipi di Concentrazione

Concentrazione massima:

I dati saranno così formati:

$$a_i \to 0, 0, 0, 0, \ldots, \text{TOT} q_i \to 0, 0, 0, 0, \ldots, 1.$$

Concentrazione minima:

I dati saranno così formati:

$$a_i \to \overline{a}, \overline{a}, \ldots, \overline{a} Q_i = \frac{1}{\text{TOT}}$$

.:

Trasformazione dei dati

A volte serve modificare un campione di dati per renderlo più espressivo, per farlo non si fa altro che aèèlicare una funziona ad ogni membro del campione.

Data una funzione **iniettiva lineare** allora una volta trasformato le frequenze nel campione rimangono invariate.

1. Traslazione: Utile quando il campione è molto compatto ma di valore molto alto allora lo si può traslare per una più semplice visualizzazione. $f(x) = x + k$.
   Questa trasformazione cambia:
   (a) Media Campionaria
   (b) Mediana Campionaria
   (c) Moda
   (d) Quantili.

   La trasformazione non cambia:
   (a) Varianza
   (b) Deviazione standard
   (c) Scarto/range-interquartile
   (d) range dei dati
2. Scalatura: Utile a espandere o contrarre i dati. Utile quando i dati sono molto dispersi o molto compressi. $f(x) = hx$. Con $h = \frac{1}{max_x}$ si può scalare i dati nell'intervallo $[0, 1]$. Si può anche arrivare ad avere un range di $[0, n]$ con $h = \frac{1}{max_x * n}$ Questa trasformazione cambia:
   (a) Media Campionaria
   (b) Mediana Campionaria
   (c) Moda
   (d) Quantili.
   (e) Varianza
   (f) Deviazione standard
   (g) Scarto/range-interquartile
   (h) range dei dati

   La trasformazione non cambia:
3. Standardizzazione

Lezione 8