Statistica

Esplicazione integrato con Appunti

Marco D'Epifano

Premessa

Chi siamo

Appunti Luiss è un progetto nato per rendere meno difficoltosa e più soddisfacente la vita universitaria.

Questo è stato possibile perché il team di appunti Luiss ha fatto una scoperta tanto banale quanto geniale: la collaborazione tra studenti tramite la condivisione di esperienze universitarie facilita il superamento degli esami. Tale collaborazione e condivisione, molto spesso, si concretizza nella produzione, anche involontaria, di lavori come appunti, compendi o esplicazioni.

Ora, dato che la diffusione di questo tipo di lavori aiuta lo studio e il superamento degli esami, il **favorire** tale diffusione è il primo obbiettivo che Appunti Luiss si propone.

Il secondo obbiettivo che ci proponiamo è quello di **valorizzare** questo tipo di lavori. Tale valorizzazione, per natura, produce un doppio effetto: favorisce la **diffusione,** incentivando gli studenti a produrne sempre di più, e costituisce la giusta **ricompensa** per gli studenti che li hanno prodotti agevolando anche il sostentamento dello studente stesso.

Insomma, quello che Appunti Luiss vuole fare è **aiutare** gli studenti e **premiare** coloro che hanno reso questo possibile.

Appunti Luiss Team

STATISTICA

Indice

STATISTICA DESCRITTIVA

La statistica è l'insieme delle tecniche con lo scopo di *raccogliere*, *elaborare* ed *interpretare* informazioni che riguardano fenomeni del mondo reale di qualsiasi tipo. È dunque l'analisi e la sintesi dei dati grazie al quale è possibile prendere decisioni.
La raccolta o rilevazione può avvenire secondo *censimento*, su tutte le unità statistiche, o secondo *campionamento*, su alcune unità statistiche. L'obiettivo della statistica è lo studio della popolazione, intesa come somma delle modalità corrispondenti a tutte le unità statistiche interessate dal fenomeno: in caso di campionamento si potrà generalizzare sull'intera popolazione utilizzando le metodologie dell'*inferenza statistica*.

CLASSIFICAZIONE DEI CARATTERI

Carattere: ossia una determinata informazione di interesse della statistica (es. *genere sessuale*), rilevata su **unità statistiche** (es. *studenti canale D Luiss*) appartenenti ad una **popolazione** (es. *studenti italiani*), che si può presentare con determinate **modalità** (es. *maschio/femmina*).

- **Qualitativi**: hanno per modalità delle denominazioni.
  - **Nominali**: se le modalità si possono ordinare solo con = oppure ≠; (es. *maschio/femmina*)
  - **Ordinali**: si è tra le modalità si può stabilire la relazione di ≤; (es. *titolo di studio*)
- **Quantitativi**: hanno per modalità dei numeri.
  - **Discreti**: modalità ∈ ℕ; (es. *numero addetti*)
  - **Continui**: modalità ∈ ℝ; (es. *statura*)

DISTRIBUZIONI

Unitaria:
lista delle unità con associata a ognuna di esse la modalità del carattere.

1 → A
2 → C
3 → B
4 → A

Supponendo che A > B > C

Di frequenza (ordinata):
lista delle modalità distinte (disposte in ordine crescente, il carattere deve essere almeno qualitativo ordinale) con associata frequenza **assoluta** (n di volte che tale modalità compare, ∑ = n unità), **relativa** (frequenza assoluta diviso n delle unità, ∑ = 1) e **percentuale** (frequenza relativa · 100, ∑ = 100).

x_i	n_i	N_i	f_i	F_i	p_i	P_i
A	2	2	2/4	2/4	50%	50%
B	1	3	1/4	3/4	25%	75%
C	1	4	1/4	4/4	25%	100%

Le lettere maiuscole riprendono i valori di quelle minuscole e sommano ad esso il valore precedente.

Approssimare le frequenze relative: per togliere di mezzo l'ultima cifra decimale e fare in modo che la somma delle frequenze dia 1 si deve prima troncare l'ultima cifra decimale, poi si sommano le frequenze e si vede quanto manca. Questo si assegna ai numeri con la

cifra decimale troncata più alta.

RAGGRUPPAMENTO IN CLASSI

nella costruzione delle classi è opportuno seguire alcuni criteri:

- *Classi disgiunte*: in modo da evitare ambiguità nell'attribuzione delle osservazioni alle classi;
- *Classi esaustive*: ossia a tutte le osservazioni devono poter essere collocate in una delle classi;
- *Numero di classi*: deve garantire un compromesso fra sintesi e dettaglio;
- *Ampiezza delle classi*: dove possibile sempre la stessa oppure può essere opportuno costruire classi di maggior ampiezza per evitare classi vuote e classi di minore ampiezza per non disperdere informazioni;

N persone	n_i	f_i	A_i	H_i	h_i
0 |–3	5	0,25	3 (3 - 0)	1,7	0,08
3 |–6	7	0,35	3 (6 - 3)	2,3	0,12
6 |–11	6	0,30	5 (11 - 6)	1,2	0,05
≥ 11	2	0,10	5 (valore max registrato, 15, - 11)	0,4	0,02
	(frequenza assoluta della classe)	(frequenza relativa della classe)	(**Ampiezza** = $x_i - x_{i-1}$)	(**Densità assoluta** = n_i/A_i)	(**Densità relativa** = f_i/A_i)

Per fare un confronto tra classi bisogna depurare la frequenza dall'ampiezza della classe ottenendo la densità assoluta che va bene, come la densità relativa, per fare confronti **inter-distribuzione**.

Per confrontare **distribuzioni diverse** si ricorre alla densità relativa.

DISTRIBUZIONE IN 2 VARIABILI

Supponiamo di avere 14 prodotti appartenenti a tre diverse linee di produzione (A, B e C) e rileviamo quali sono difettosi e quali no (lo scopo è sapere quale linea produce più prodotti difettosi). Mostriamo direttamente la distribuzione di frequenza in due variabili nella seguente **tabella di contingenza**:

	D	ND		
A	4 (n_{11})	2 (n_{12})	4 + 2 ($n_{1\cdot}$)	Frequenze marginali (assolute) di riga
B	2 (n_{21})	1 (n_{22})	2 + 1 ($n_{2\cdot}$)	
C	1 (n_{31})	4 (n_{32})	1 + 4 ($n_{3\cdot}$)	
Frequenze (assolute) congiunte	4 + 2 + 1 ($n_{\cdot 1}$)	2 + 1 + 4 ($n_{\cdot 2}$)	Totale generale = 14	
	Frequenze marginali (assolute) di colonna			

Distribuzione relativa: dividere ogni frequenza (congiunta, quindi ogni singola cella, e marginale, quindi i totali di righe e colonne) per il totale generale (questo in distribuzione relativa deve essere ovviamente 1).

INDICI DI POSIZIONE

Si intende una sola unità rappresentativa dell'intera distribuzione di frequenza secondo un determinato carattere rilevato su unità statistiche (detto anche *valore medio*). Per tutti valgono le proprietà:

• *Consistenza*: se tutte le modalità statistiche sono uguali allora sono uguali all'indice di posizione;

• *Interno*: poiché compreso tra il valore minimo e quello massimo;

Moda

Modalità di massima *frequenza* assoluta, relativa o percentuale.

Classe modale

Classe di massima *densità* assoluta o relativa.

Unico indice per caratteri qualitativi non ordinabili (può non essere unica).

Media aritmetica

Data una distribuzione di frequenza secondo un determinato carattere quantitativo, essa è la modalità di ripartizione del carattere tra le unità.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{n} = \sum_{i=1}^{n} x_i \cdot f_i$$

Dato un **raggruppamento in classi** di un determinato carattere quantitativo, essa si calcola facendo la media ponderata per le frequenze corrispondenti dei *valori centrali* delle classi (valore centrale: somma degli stremi /2).

• *Trasformazione lineare*: la media aritmetica delle trasformazioni lineari delle modalità è uguale alla trasformazione lineare della media aritmetica;

• *Associativa*: la media è quel valore che sostituito a ogni osservazione lascia invariata la loro somma;

• *Somma degli scarti* $x_i - \bar{x} = 0$;

• *Somma dei quadrati degli scarti* dalla media viene minimizzata dalla media;

Mediana e Quartili

Modalità che divide l'elenco ordinato delle n modalità in 2 parti di uguale numerosità:

- *n dispari*: una mediana in posizione $\frac{n+1}{2}$ su N_i oppure 0,5 su F_i (vedere la modalità corrispondente);

- *n pari*: due mediane nelle posizioni $\frac{n}{2}$ e $\frac{n+1}{2}$ su N_i oppure 0,5 e 0,51 su F_i (vedere le modalità corrispondenti);

Quartili

Tre modalità che dividono la distribuzione ordinata di frequenza in quattro parti di uguale

numerosità:

- Q_1: modalità nella posizione $\frac{n}{4}$ su N_i oppure 0,25 su F_i;
- Q_2: modalità nella posizione $\frac{n}{2}$ su N_i (coincide con la mediana);
- Q_3: modalità nella posizione $\frac{3}{4}n$ su N_i oppure 0,75 su F_i;

Se la posizione è un numero decimale ci saranno due posizioni nelle quali può esserci la stessa modalità o due diverse, in questo caso si fa la semisomma.

Percentili

Modalità che divide la distribuzione ordinata di frequenza in una parte che comprende *a* modalità e in un'altra che ne comprende *a - 1* ($0 \leq a \leq 1$). (Detti anche *Quantili*).

• *Trasformazione lineare*; I quantili sono insensibili ai valori anomali dipendendo solo dalle posizioni.

• Il numero di *scarti positivi* dalla mediana è uguale al numero di *scarti negativi* dalla mediana;

• *Somma degli scarti dalla mediana in valore assoluto* è minimizzata dalla mediana;

DISTRIBUZIONI

Dati gli indici di posizione e disegnato il grafico a barre si può classificare la distribuzione di frequenze:

- **simmetrica**: se si verifica l'uguaglianza delle frequenze dei valori equidistanti dalla mediana che conseguentemente coinciderà con la moda e la media aritmetica;
- **asimmetrica**:
  - **negativa**: coda a sinistra nel grafico a barre, media aritmetica < mediana;
  - **positiva**: coda a destra nel grafico a barre, media aritmetica > mediana;

INDICI DI VARIABILITÀ

Variabilità: attitudine di un carattere a presentarsi con modalità diverse nelle diverse unità statistiche (se le modalità sono tutte uguali vuol dire che la variabilità è nulla). Si può calcolare la variabilità in due modi:

- **MUTUA VARIABILITÀ**

Si analizza la variazione delle modalità tra di loro. I seguenti indici si usano insieme:

- **Range**: modalità max – modalità minima (o x_n - x_1 nella distribuzione ordinata)
  Detto anche *Intervallo* o *campo di variazione*, è la massima differenza che si riscontra tra due qualsiasi modalità (R = 0 vuol dire che tutte le modalità sono uguali).
- **Scarto Interquartile**: SI = Q_3 - Q_1

- **INTORNO AD UN POLO**

Si analizza la variazione delle modalità rispetto a un centro.

- **Varianza**: $s^2 = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot n_i}{n}$ o $\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot f_i$

Con un'altra formula: $\frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{n} - \bar{x}^2 = \bar{x}_{(2)} - \bar{x}^2$.

La varianza è la media degli scarti (dalla media) al quadrato, ossia un indice di dispersione intorno alla media

($s^2 = 0$ vuol dire che non c'è variabilità). Poiché la somma degli scarti è nulla (scarti positivi e scarti negativi si compensano) per avere un indice che esprima l'ordine di grandezza degli scarti si neutralizza il segno elevando al quadrato (ha quindi un'unità di misura diversa da quella del fenomeno di studio). In caso di raggruppamento in classi la varianza si stima utilizzando il valore centrale quale valore rappresentativo della classe.

• *Trasformazione lineare*: ax + b = y allora $s^2_y = a^2 \cdot s^2_x$;

• *Disuguaglianza di Chebyshev*: $fr(|x - \bar{x}| < \varepsilon) \geq 1 - \frac{s^2}{\varepsilon^2}$

Ciò vuol dire che disponendo anche solo di varianza e media io posso sapere qual è la percentuale di osservazioni comprese nell'intervallo ($\bar{x}$ - ε, $\bar{x}$ + ε) scegliendo la ε in base al mio scopo.

- **Scarto quadratico medio**: $s = \sigma = \sqrt{s^2}$

Detto anche *Scostamento q. m.* o *Deviazione standard*, è la radice quadrata della varianza, misura la distanza media delle osservazioni dalla media (le modalità oscillano tra $\bar{x} - s$ e $\bar{x} + s$, ossia i valori minimo e massimo).

- **Coefficiente di variazione**: $\frac{s}{\bar{x}}$

Permette di depurare lo scarto quadratico medio dalla media così da sapere quanto è rilevante s (un s di 1kg è diverso se si riferisce al peso medio di un adulto, 60kg, o a quello di un neonato, 3kg), è quindi un indice relativo di variabilità che permette il confronto tra distribuzioni di frequenza diverse.

- **Indice MAD**: 1,483 · *mediana*{$|x_i$ - mediana|}

Il Median Absolute Deviation è la mediana degli scarti (dalla mediana) in valore assoluto moltiplicato per una costante. Una sua caratteristica che lo contraddistingue dalla varianza è che risulta essere meno influenzato dalla presenza dei cosiddetti valori anomali (che possono derivare anche da errori di rilevazione).

NUMERI INDICI

Servono per studiare le serie storiche, ovvero le informazioni relative ad un carattere rilevato nel tempo.

Semplici

- **A base fissa**: il tempo b è fisso e il tempo t è mobile: ${}_bI_t = \frac{carattere\ al\ tempo\ t}{carattere\ al\ tempo\ b}$ e vuol dire quante unità di fenomeno ho al tempo t per ogni unità al tempo base.
- **A base mobile**: la base è il tempo precedente a t: $i_t = \frac{carattere\ al\ tempo\ t}{carattere\ al\ tempo\ t-1}$ e vuol dire quante unità di fenomeno ho al tempo t per ogni unità al tempo precedente.

t	X	${}_0I_t$	i_t
0	x_0	x_0/ x_0	-
1	x_1	x_1/ x_0	x_1/ x_0
2	x_2	x_2/ x_0	x_2/ x_1

Complessi

Sono la media di indici semplici (a base fissa nei due casi seguenti) ponderata per le unità di fenomeno al tempo base (Laspeyres) o al tempo corrente (Paasche).
Ipotizziamo due fenomeni A e B ognuno dei quali in ogni tempo può avere un valore x per q quantità (dove x solitamente è il prezzo) e due tempi b (base) e t (il tempo attuale):

- Indice di ${}_b\mathbf{Laspeyres}_t$: $\frac{\frac{x_t}{x_b}(x_b \cdot q_b)+\frac{x_t}{x_b}(x_b \cdot q_b)}{(x_b \cdot q_b)+(x_b \cdot q_b)}$ ovvero $\frac{(x_t \cdot q_b)+(x_t \cdot q_b)}{(x_b \cdot q_b)+(x_b \cdot q_b)}$ (le quantità sono fisse all'anno base);
- Indice di ${}_b\mathbf{Paasche}_t$: $\frac{\frac{x_t}{x_b}(x_b \cdot q_t)+\frac{x_t}{x_b}(x_b \cdot q_t)}{(x_b \cdot q_t)+(x_b \cdot q_t)}$ ovvero $\frac{(x_t \cdot q_t)+(x_t \cdot q_t)}{(x_b \cdot q_t)+(x_b \cdot q_t)}$ (le quantità sono di volta in volta quelle dell'anno corrente);

RAPPRESENTAZIONI GRAFICHE

Disegno della distribuzione di frequenza.

Caratteri qualitativi

- *Barre verticali/orizzontali* (anche suddivise o multiple);
- *Areali* (torta);

Caratteri quantitativi

Discreti:

- Non raggruppabili: *Aste o bastoncini*, *Grafico di dispersione* (per vedere la relazione tra due variabili);
- Raggruppabili: *Istogramma di frequenza* (barre, si deve rapportare la classe alla densità assoluta o relativa);

Non discreti:

- *Diagramma lineare* (cartesiano);