Analisi dei dati

Un goliardico riassunto

Ollari Dmitri

27 dicembre 2022

Indice

Elenco delle figure

Capitolo 1

Introduzione alla statistica

1.1 Raccolta dati e statistica descittiva

1. Procedimento di raccolta dati
2. Attenzione a come si compone il campione
3. illustrare e sintetizzare i dati(statistica descittiva)

1.2 Inferenza statistica

Mediante l'Inferenza statistica si possono predirre risultati mediante l'analisi dei dati.

1. Tenere in conto il caso
2. modello probabilistico(assunzioni sulla probabilità)

1.3 Popolazione e campioni

Si vogliono risultati su gruppi estesi di persone, dove i vari sottogruppi devono essere rappresentati.
Il campione è rappresentativo solo se è casuale.

Capitolo 2

Statistica descrittiva

Permette la rappresentazione dei dati in maniera chiara, precisa, concisa e sintetica.
Si utilizzano diversi tipi di grafici in base al contesto.

2.1 Grandezze per sintetizzare i dati

Si hanno n dati.

$$n : x_1, x_2, \ldots, x_n \tag{2.1}$$

L'ampiezza(numerosità) dei dati sarà n.

2.1.1 Centro dei dati

Media campionaria

La media campionaria è una media pesata!

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

Dove la n rappresenta il numero di elementi presi in considerazione per la media.

Mediana campionaria

Avendo un insieme di dati n **ordinato**:

- Se n è dispari, la mediana campionaria è il valore in posizione $(n+1)/2$
- Se n è pari, la mediana campionaria è la media dei valori in posizione $n/2$ e $n/2+1$

Varianza

Avendo un'insieme di dati $x_1, x_2, \ldots, x_n$, la varianza campionaria è:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.3}$$

Devizione standard

Avendo un'insieme di dati $x_1, x_2, \ldots, x_n$, la deviazione standard è:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.4}$$

Weight-for-age GIRLS

Birth to 13 weeks (percentiles)

World Health Organization

				Percentiles (weight in kg)										
Week	L	M	S	1st	3rd	5th	15th	25th	50th	75th	85th	95th	97th	99th
0	0.3809	3.2322	0.14171	2.3	2.4	2.5	2.8	2.9	3.2	3.6	3.7	4.0	4.2	4.4
1	0.2671	3.3388	0.14600	2.3	2.5	2.6	2.9	3.0	3.3	3.7	3.9	4.2	4.4	4.6
2	0.2304	3.5693	0.14339	2.5	2.7	2.8	3.1	3.2	3.6	3.9	4.1	4.5	4.6	4.9
3	0.2024	3.8352	0.14060	2.7	2.9	3.0	3.3	3.5	3.8	4.2	4.4	4.8	5.0	5.3
4	0.1789	4.0987	0.13805	2.9	3.1	3.3	3.5	3.7	4.1	4.5	4.7	5.1	5.3	5.6
5	0.1582	4.3476	0.13583	3.1	3.3	3.5	3.8	4.0	4.3	4.8	5.0	5.4	5.6	5.9
6	0.1395	4.5793	0.13392	3.3	3.5	3.7	4.0	4.2	4.6	5.0	5.3	5.7	5.9	6.2
7	0.1224	4.7950	0.13228	3.5	3.7	3.8	4.2	4.4	4.8	5.2	5.5	5.9	6.1	6.5
8	0.1065	4.9959	0.13087	3.7	3.9	4.0	4.4	4.6	5.0	5.5	5.7	6.2	6.4	6.7
9	0.0918	5.1842	0.12966	3.8	4.1	4.2	4.5	4.7	5.2	5.7	5.9	6.4	6.6	7.0
10	0.0779	5.3618	0.12861	4.0	4.2	4.3	4.7	4.9	5.4	5.8	6.1	6.6	6.8	7.2
11	0.0648	5.5295	0.12770	4.1	4.3	4.5	4.8	5.1	5.5	6.0	6.3	6.8	7.0	7.4
12	0.0525	5.6883	0.12691	4.2	4.5	4.6	5.0	5.2	5.7	6.2	6.5	7.0	7.2	7.6
13	0.0407	5.8393	0.12622	4.3	4.6	4.7	5.1	5.4	5.8	6.4	6.7	7.2	7.4	7.8

WHO Child Growth Standards

Figura 2.1: Esempio percentili

2.1.2 Percentili campionari

Da questo graafico dell'OMS si capisce bene coasa sono i percentili.

Le bambine che fanno parte del 25 esimo percentile hanno un peso alla nascita di 2.9 Kg, il 25% della bambine pesano meno e il 75% delle bambine pesano di più.

Il percentile divide i dati in due parti ben distinte, se il dato non è unico(due dati), si fa la media aritmetica.

Procedura

Ho n dati e voglio il valore del percentile k-esimo:

$$p = \frac{k}{100} \tag{2.5}$$

$$np = n\frac{k}{100} \tag{2.6}$$

Se np non è intero, arrotondo al numero per eccesso successivo.

Nota bene

Sono quartili campionari:

- 25-esimo percentile
- 50-esimo percentile
- 75-esimo percentile

Notare che il 50-esimo percentile è la media campionaria.

2.1.3 Box plot

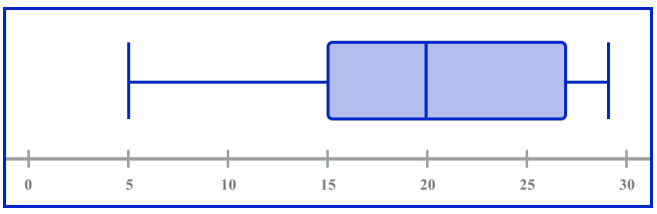

Figura 2.2: Esempio box plot

Il disegno del box plot segue le seguenti regole:

- linea da dato minore a dato maggiore
- box dal primo al terzo quartile

La lunghezza della linea rappresenta il campo di variazione(range dei valori), mentre la lunghezza del box rappresenta lo scarto interquartile.

2.2 Disuguaglianza di Chebyshev

Avendo un'insieme di dati $x_1, x_2, \ldots, x_n$ con media campionaria $\bar{x}$ e deviazione standard $s > 0$.
Denoto con S_k l'insieme degli indici dei dati compresi tra $\bar{x} - ks$ e $\bar{x} + ks$.
Essendo $\#S_k$ il numero di elementi dell'insieme S_k.(AKA cardinalità)
Ne deriva la seguante equazione se $k \geq 0$:

$$\frac{\#S_k}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2} \tag{2.7}$$

2.3 Campioni normali

Forma della distribuzione **caratteristica**:

- a campana
- massimo sulla mediana
- simmetrica

Può essere approssimativamente normale o sbilanciata(skewed).

2.3.1 Regola empirica

- Il $\sim 68\%$ dei dati si trova nella spazio $\bar{x} \pm s$
- Il $\sim 95\%$ dei dati si trova nella spazio $\bar{x} \pm 2s$
- Il $\sim 99.7\%$ dei dati si trova nella spazio $\bar{x} \pm 3s$

2.4 Dati bivarianti

Sono serie di dati diverrsi che hanno una correlazione fra di loro.
La formula della correlazione è:

$$(x_i - \bar{x})(y_i - \bar{y}) \tag{2.8}$$

Se il risultato è:

- $> 0 \Rightarrow$ correlazione positiva
- $< 0 \Rightarrow$ correlazione negativa

2.4.1 Coefficiente di correlazione campionaria

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{2.9}$$

- $r > 0 \Rightarrow$ correlazione positiva
- $r < 0 \Rightarrow$ correlazione negativa

La correlazione è diversa dalla relazione di causa effetto.

Capitolo 3

Elementi di probabilità

Ci sono due interpretazioni, **frequentista** o **soggettivista**, nel'interpretazione frequentista si ha che la probabilità è una proprietà dell'esito stesso dell'esperimento e si ricava ripetendo l'esperimento.(scienza)

Nel caso di interpretazione soggettivista la probabilità viene precisata in base alla fiduci dell'esito.(filosofia e finanza)

3.1 Spazio degli esiti ed eventi

- **Spazio degli esiti(S oppure Ω)** = insieme degli esiti possibili
- **Eventi(E)** = insieme i cui elementi sono esiti possibili
- **Unione eventi** $A \cup B$ = OR logico
- **Intersezione eventi** $A \cap B$ = AND logico, si indica molto spesso mediante una virgola
- $A \cap B = \emptyset$ significa che sono due insiemi disgiunti o mutualemnte esclusivi
- A^B significa complementari
- $A \subset B$ significa che A è un sottoinsieme di B

3.2 Assiomi della probabilità

1. La probabilità di un evento E è: $0 \leq P(E) \leq 1$
2. La probabilità dello spazio degli esiti $P(S)$ è 1
3. La probabilità che si verifichi almeno un eventi di un insieme di eventi mutualemnte esclusivi è uguale alla somma delle loro probabilità $P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

3.3 Spazio esiti equiprobabili

Si capisce molto facilmete guardando la probabilità di lanciare un dato e le facce hanno tutte la stesa probabilità di uscire.

$S = \{1, 2, \ldots, n\}$

La probabilità $P = \frac{1}{n}$

3.3.1 Coefficiente binomiale

Voglio determinare il numero di diversi gruppi di r oggetti che si possono formare scegliendoli da un'insieme di n.

$$\binom{n}{k} = \frac{n!}{r!(n-r)!} \tag{3.1}$$

3.4 Probabilità condizionata

La probabilità condizionata di E dato F è:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{3.2}$$

3.5 Fattorizzazione di un evento e formula di Bayes

$$E = (E \cap F) \cup (E \cap F^C) \tag{3.3}$$

$$\begin{aligned} P(E) &= P(E \cap F) + (E \cap F^C) \\ &= P(E|F)P(F) + P(E|F^C)P(F^C) \\ &= P(E|F)P(F) + P(E|F^C)[1 - P(F)] \end{aligned} \tag{3.4}$$

- E intersezione F sarebbe un and logico
- E intersezione F^C risulta nell'insieme E togliendo le parti in comune con F

3.6 Eventi indipendenti

Se due eventi E ed F sono indipendenti, allora

$$P(E \cap F) = P(E)P(F) \tag{3.5}$$

Sono indipendenti e significa che la riuscita o meno di un evento non condiziona la riuscita dell'altro.
La formula generale è:

$$P(\cap_{i=1}^{r} E_{ai} = \prod_{i=1}^{r} P(E_{ai})) \tag{3.6}$$

Capitolo 4

Variabili aleatorie e valore atteso

4.1 Variabili aleatorie

Con le variabili aleatorie si assegna la probabilità ai valori possibili.

Esempio dei due dadi: La probabilità che il risultato della somma sia 3 si scrive così

$$P(\{X = 3\}) = P\{(2, 1), (1, 2)\} \tag{4.1}$$

Quindi si hanno 2 casi favorevoli su 36 totali, che porta la probabilità al valore di 0.055.

4.2 Variabili aleatorie discrete e continue

Le variabili aleatorie che prendono valori da insiemi finiti o numerabili prendono il nome di **discrete**.

In alternativa le variabili non numerabili prendono il nome di **continue** e sono tutti quei valori che fanno parte dei numeri reali.

4.2.1 Funzione di ripartizione

$$F(x) = P(X \leq x) \tag{4.2}$$

La funzione $F(x)$ rappresenta la probabilità con la quale la variabile aleatorie X sia $\leq$ di x.

La notazione $X \sim F$ indica che F è la funzione di ripartizione di X.

4.2.2 Variabili aleatorie discrete e continue

Se X è una variabile aleatoria discreta, la sua **funzione id massa di probabilità(PMF)** è:

$$p(a) = P(X = a) \tag{4.3}$$

4.2.3 Funzioni di densità di probabilità di X(PDF)

$$1 = P(X \in \Re) = \int_{-\infty}^{+\infty} f(x)dx \tag{4.4}$$

dove il limiti di integrazione sono dati dal range dei dati che stiamo considerando.

ESEMPIO: Variabile aleatoria X con PDF:

$$f(x) = \begin{cases} c(4x - 2x^2) & 0 < x < 2 \\ 0 & else \end{cases} \tag{4.5}$$

Per ottenere c:

$$1 = c\int_0^2 (4x - 2x^2)dx = c\left[2x^2 - 2\frac{x^3}{3}\right]_{x=0}^{x=2} = c\frac{8}{3} \tag{4.6}$$

$c = \frac{3}{8}$

4.3 Coppie di vettori di variabili aleatorie

Utili a valutare la relazione tra variabili aleatorie.

4.3.1 Distribuzione congiunta per variabili aleatorie continue

X e Y sono congiuntamente continue se esiste $f(x,y) > 0$ definita per tutti x, y tale che ogni sottoinsieme C del piano cartesiano sia:

$$P((x,y) \in C) = \iint_{x,y \in C} f(x,y)dxdy \tag{4.7}$$

Dove con $f(x,y)$ si intende la densità congiunta.
Facendo vari giri matematici si ottiene:

$$\int_b^{b+db} \int_a^{a+da} f(x,y)dxdy \simeq f(a,b)dadb \tag{4.8}$$

Vale solo se da **e** db **sono molto piccoli e** $f(a,b)$ **è continua in** (a,b).

4.3.2 Variabili aleatorie indipendenti

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{4.9}$$

4.3.3 Generalizzazione a più variabili aleatorie

Ripartizione:

$$F(a_1, a_2, \ldots, a_n) = P(x_1 \leq a_1, \ldots, x_n \leq a_n) \tag{4.10}$$

Se sono variabili discrete:

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{4.11}$$

Se sono continua devo usare le funzioni con gli integrali mega in sbatti.

4.3.4 Indipendenza

Infinite variabili aleatorie sono indipendenti se ogni loro sottogruppo finito è formato da variabili aleatorie indipendenti.

mancano la densita condizionale e la distribuzioen condizionale

4.4 Valore atteso

X variabili aleatoria con valori $x_1, \ldots,$ il valore atteso di X:

$$E[X] = \sum_i x_i P(X = x_i) \tag{4.12}$$

Il valore della sommatoria deve converge ad un numero inferiore a infinito.
Praticamente la media pesata dei possibili valori di X, con pesi dati dalle probabilità.
$E[X]$ viene chaimata media di X oppure aspettazione (expectation).

4.5 Proprietà del valore atteso

Se X è una variabile aleatoria discreta con funzione di massa di probabilità p, allora, per ogni funzione reale g:

$$E[g(X)] = \sum_x g(x)p(x) \tag{4.13}$$

Se X è una variabile aleatoria continua con funzione di densità di probabilità f, allora, per ogni funzione reale g:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx \tag{4.14}$$

4.6 Variana

Misura della variabilità della variabile aleatoria, la media $E[X]$ rappresenta il baricentro ma non è sufficiente.

Sia X una variabile aleatoria con media μ, la varianza di X si denota con $Var(X)$:

$$Var(X) = E[(X - \mu)^2] \tag{4.15}$$

Però spesso viene più comoda la seguente formula:

$$Var(X) = E[X^2] - E[X]^2 \tag{4.16}$$

4.6.1 Deviazione standard

per calcolare la deviazioen standard basta:

$$\sqrt{Var(X)} \tag{4.17}$$

4.7 La covarianza e la varianza della somma di variabili aleatorie

Avendo due variabili aleatorie X e Y con media μ_X e μ_Y, la loro covarianza se esiste è:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{4.18}$$

Proprietà

- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, X) = Var(X)$
- $Cov(aX, Y) = aCov(X, Y) = Cov(X, aY)$

4.7.1 Coefficiente di correlazione lineare

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \tag{4.19}$$

4.8 La funzione generatrice dei momenti