Manuale dell'ingegnere intrippato con la statistica

What are the odds?

30 agosto 2019

1 Statistica descrittiva

1.1 Le grandezze che sintetizzano i dati

1.1.1 Media

Dato un insieme $x_1, x_2, ..., x_n$ di dati, si dice media campionaria la media aritmetica di questi valori.

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

1.1.2 Mediana

Dato un insieme di dati di ampiezza n, lo si ordini dal minore al maggiore. La mediana è il valore che occupa la posizioone $\frac{n+1}{2}$ in caso di un insieme dispari, o la media tra $\frac{n}{2}$ e $\frac{n}{2}+1$ se pari.

1.1.3 Moda

La moda campionaria di un insieme di dati, se esiste, è l'unico valore che ha frequenza massima.

1.1.4 Varianza e deviazione standard campionarie

Dato un insieme di dati $x_1, x_2, ..., x_n$, si dice varianza campionaria (s^2), la quantità

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Una comodità per il calcolo è che

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

Si dice **deviazione standard campionaria** e si denota con s, la quantità

$$s := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

(la radice quadrata di s^2)

1.1.5 Percentili campionari e box plot

Sia k un numero intero $0 \leq k \leq 100$. Dato un campione di dati, esiste sempre un dato che è contemporaneamente maggiore del k percento dei dati, e minore del $100 - k$ percento. Per trovare questo dato, dati n e $p = \frac{k}{100}$:

1. Disponiamo i dati in ordine crescente
2. Calcoliamo np
3. Il numero cercato è quello in posizione np, arrotondato per eccesso se non intero.

Il 25-esimo percentile si dice *primo quartile*, il 50-esimo *secondo* (ed è pari alla mediana), il 75-esimo *terzo*. Il box plot è un grafica con un quadrato sulla linea dei dati, con i lati sul primo e terzo quartile, e un segno sul secondo.

1.2 Disuguaglianza di Chebyshev

Siano $\overline{x}$ e s media e deviazione standard campionarie di un insieme di dati. Nell'ipotesi che $s > 0$, la disuguaglianza di Chebyshev afferma che per ogni reale $k \geq 1$, almeno una frazione $(1 - 1/k^2)$ dei dati cade nell'intervallo che va da $\overline{x} - ks$ a $\overline{x} + ks$. Usando il ~~pessimo~~ *fantastico* linguaggio da statista: sia assegnato un insieme di dati $x_1, ..., x_n$ con media campionaria $\overline{x}$ e deviazione standard campionaria $s > 0$. Denotiamo con S_k l'insieme degli indici corrispondenti a dati compresi tra $\overline{x} - ks$ e $\overline{x} + ks$. Sia $\#S_k$ il numero dei suddetti. Allora abbiamo che

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

1.3 Insiemi di dati bivariati e coefficiente di correlazione campionaria

A volte non abbiamo a che fare con dati singoli, ma con coppie di numeri, tra i quali sospettiamo l'esistenza di relazioni. Dati di questa forma prendono il nome di *campione bivariato*. Uno strumento utile è il diagramma di dispersione. Una questione interessante è capire se vi sia correlazione tra i dati accoppiati. Parleremo di correlazione positiva quando abbiamo una proporzionalità diretta tra i due, di correlazione negativa quando abbiamo una proporzionalità inversa.

1.3.1 Coefficiente di correlazione campionaria

Dato un campione bivariato (x_i, y_i), sono definite le medie $\overline{x}$ e $\overline{y}$. Possiamo senz'altro dire che se un valore x_i è grande rispetto alla media, la differenza $x_i - \overline{x}$ sarà positiva, mentre se x_i è piccolo, la differenza sarà negativa. Quindi, considerando il prodotto $(x_i - \overline{x})(y_i - \overline{y})$, sarà positivo per correlazioni positive, negativo per correlazioni negative. Se l'intero campione mostra quindi un'elevata correlazione, ci aspettiamo che la somma di tutti i prodotti $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ darà una buona stima della correlazione. Normalizziamola dividendo per $(n-1)$ e per il prodotto delle deviazione standard campionarie, e otteniamo il **coefficiente di correlazione campionaria**

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

con s_x e s_y deviazioni standard campionarie di x e y.

1.3.2 Proprietà del coefficiente di correlazione campionaria

Sebbene parleremo meglio di questo bastardo nella sezione sulla regressione, elenchiamo qui alcune proprietà:

1. $-1 \leq r \leq 1$
2. Se per opportune costanti a e b, con $b > 0$ sussiste la relazione lineare $y_i = a + b_x$, allora $r = 1$.
3. Se per opportune costanti a e b, con $b < 0$ sussiste la relazione lineare $y_i = a + b_x$, allora $r = -1$.
4. Se r è il coefficiente di correlazione del campione (x_i, y_i), $i = 1, ..., n$, allora lo è anche per il campione $(a + bx_i, c + dy_i)$, purché le costanti a e b abbiano lo stesso segno.