

Statistica Campionaria

Vers. 1.1.1

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Outline

1 Campioni e Statistiche

- Campioni
- Statistiche e distribuzioni campionarie

2 Statistiche d'ordine

- Cenni su simulazione

Campioni I

In questa parte del corso ci occuperemo di statistica “inferenziale” e solo marginalmente di statistica descrittiva.

Statistica descrittiva

La statistica descrittiva è la branca della statistica che studia i criteri di rilevazione, classificazione, sintesi e rappresentazione dei dati appresi dallo studio di una popolazione o di una parte di essa (“campione”).

Statistica inferenziale

È il procedimento per cui si inducono le caratteristiche di una popolazione dall'osservazione di una parte di essa (“campione”), selezionata solitamente mediante un esperimento casuale (aleatorio). Da un punto di vista filosofico, si tratta di tecniche matematiche per quantificare il processo di apprendimento tramite l'esperienza.

Entrambe le definizioni utilizzano il concetto di “campione”, che è fondamentale per la statistica. L'idea alla base è che quando si vuole analizzare un fenomeno, e.g., la

Campioni II

temperatura nella città di Torino, l'efficacia di un nuovo farmaco, non si possono osservare tutti i possibili valori (la temperatura in tutti i punti di Torino o testare il farmaco sulla popolazione mondiale), ma lo si fa solo su un sottoinsieme, detto **campione**

Campione

Data una popolazione di unità, un campione $\mathbf{X} = (X_1, \dots, X_n)$ esprime una sua parte rappresentativa del tutto. Consente di indagare determinate caratteristiche e attribuirle, per estrapolazione, all'universo di riferimento. Lo spazio dei possibili risultati di un esperimento viene chiamato spazio campionario.

Andrebbe fatta una differenziazione tra campionamento da popolazioni finite o infinite, dove con infinita si intende anche “potenzialmente” infinita, come il numero di beni prodotti da un'azienda. Distingueremo tra i due campionamenti solo quando sarà necessario.

Campioni III

Attenzione: A meno che non venga specificato, assumeremo sempre che i campioni siano iid (o casuali)

Campione casuale (o campione iid)

Sia F_x una legge di probabilità, allora un vettore $\mathbf{X} = (X_1, \dots, X_n)$ si chiama **campione casuale** da F_x (di dimensione n) se le componenti di \mathbf{X} sono indipendenti e provengono tutte da F_x .

Notate che nella definizione le componenti di \mathbf{X} possono essere scalari o vettoriali (anche se in genere li indicheremo senza bold). la densità congiunta è

$$f(\mathbf{x}) = \prod_{i=1}^n f(x_i)$$

Quel che vogliamo fare con il campione è di “inferire” il valore dei parametri di F_x , che sono generalmente sconosciuti. Per esempio, se F_x è $N(\mu, \sigma^2)$ vogliamo usare il campione per poter dire qualcosa su μ e σ^2 , se è $P(\lambda)$, vogliamo dire qualcosa su λ

Campioni IV

Esempio 1

Un macchinario per la produzione di pezzi elettronici ha una probabilità p (generalmente sconosciuta) di produrre un oggetto difettoso. Per controllo di qualità si fa un'indagine, si raccolgono n pezzi e si controlla se sono difettosi ($X_i = 1$) o no ($X_i = 0$). Il vettore (X_1, \dots, X_n) è un campione da una $\text{Bern}(p)$

Molto volte non conosciamo la F_x , ma assumiamo che sia una forma che conosciamo, come nell'esempio seguente

Esempio 2

Vogliamo studiare il numero di figli delle famiglie italiane. Per questo prendiamo n famiglie scelte a caso e registriamo il numero di figli. Se assumiamo che il numero dei figli fatti da una famiglia si distribuisca come un Poisson (con lo stesso parametro), questo è un campione casuale da una Poisson.

Campioni V

Dall'esempio precedente è chiaro come la Poisson sia un'assunzione. E' difficile che siano Poisson, come è difficile che abbiano lo stesso parametro (potrebbe dipendere dall'entrata mensile della famiglia, lo stato sociale, l'età etc). Non ci occuperemo di come scegliere le assunzioni e come verificarle, ma daremo per buono che le assunzioni distributive siano corrette.

Nel “**Chunk datasets**” ci sono degli esempio di dataset che utilizzeremo nel corso, in cui dei campioni sono stati raccolti per analizzare problemi reali.

Statistiche e Distribuzione Campionaria I

Statistica e Distribuzione Campionaria

Si definisce **statistica** una funzione $T(X_1, X_2, \dots, X_n)$ del campione (X_1, \dots, X_n) , e la sua distribuzione si chiama **distribuzione campionaria**.

Alcune delle statistiche più note sono

- Media campionaria $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$
- Varianza campionaria $S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$. Il perchè di $n - 1$ invece di n verrà chiarito quando introdurremo la **distorsione**.
- (\bar{X}, S^2) . Una statistica può avere anche valori vettoriali, matriciali etc
- Massimo campionario $X_{(n)} = \max(X_1, \dots, X_n)$;
- Vettore campionario ordinato $(X_1, X_2, \dots, X_n) = \text{sort}(X_1, X_2, \dots, X_n)$, che indica lo stesso vettore ma con gli elementi ordinati dal più piccolo $X_{(1)}$ al più grande $X_{(n)}$: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$;

Statistiche e Distribuzione Campionaria II

- conteggio valori minori o uguali di x : $C(x) = \sum_{i=1}^n I(X_i \leq x)$, dove $I()$ è la funzione che assume valore 1 se l'argomento è vero, e zero altrimenti.

Una statistica interessante è la **funzione di ripartizione empirica**, che è definita come

$$\hat{F}(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} = \frac{C(x)}{n}$$

che approssima la funzione di ripartizione

Statistiche e Distribuzione Campionaria III

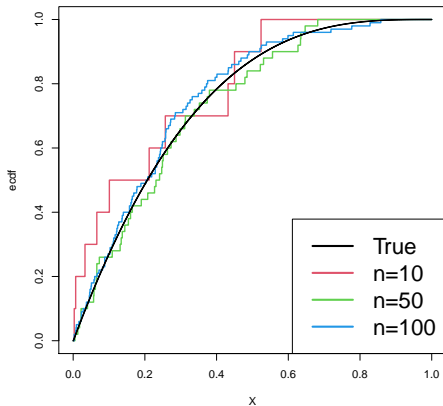


Figure: Alcuni esempi di funzioni di ripartizione empirica di una $B(1,3)$.

Statistiche e Distribuzione Campionaria IV

Altre statistiche importanti sono i quantili teorici (o campionari)

$$\hat{Q}(u) = \inf\{x : u \leq \hat{F}(x)\}$$

I quantili campionari più importanti sono $\hat{Q}(0.5)$ (mediana), $\hat{Q}(0.25)$ (primo quartile), $\hat{Q}(0.75)$ (terzo quartile). Ci sono altre definizioni dei quantili campionari, ma nel nostro caso possiamo dire che

$$\hat{Q}(u) = \begin{cases} X_{u \times n} & \text{se } u \text{ n è intero} \\ X_{u \times n + 1} & \text{altrimenti} \end{cases}$$

dove n è la dimensione campionaria

Statistiche e Distribuzione Campionaria V

Alcune distribuzioni campionarie ci saranno molto utili (soprattutto nei test d'ipotesi).

Teorema - Distribuzione di \bar{X} e S^2 nel caso normale

Siano X_1, \dots, X_n iid $N(\mu, \sigma^2)$, abbiamo i seguenti risultati:

- (i) \bar{X} e S^2 sono indipendenti
- (ii) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- (iii) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- (iii) $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1)$

Dimostrazione:

(i) Il punto 1 lo verifichiamo in 2 modi diversi:

Statistiche e Distribuzione Campionaria VI

- (i-1) Consideriamo il vettore di variabili aleatorie

$\mathbf{Q} = (\bar{X}, X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})'$ che è un vettore gaussiano perchè

$\mathbf{Q} = \mathbf{A}\mathbf{X}$ con $\mathbf{A} =$

$$\begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ 1 - \frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{1}{n} & 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ \dots & \dots & -\dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} \end{pmatrix}$$

(\mathbf{A} è di dimensione $(n+1) \times n$).

Visto che

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n Q_{i+1}^2}{n-1}$$

Statistiche e Distribuzione Campionaria VII

se io dimostro che $Q_1 = \bar{X}$ è indipendente da tutte le altre componenti di Q ,
dimostro che S^2 è indipendente da \bar{X} .

Questo corrisponde a vedere che la covarianza tra Q_1 e

$Q_{-1} = (X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})'$ è zero.

Calcoliamo il valore generico

$$\begin{aligned} Cov(\bar{X}, X_j - \bar{X}) &= Cov\left(\frac{\sum_{i=1}^n X_i}{n}, X_j - \frac{\sum_{i=1}^n X_i}{n}\right) = \\ &Cov\left(\frac{\sum_{i=1}^n X_i}{n}, X_j\right) - Cov\left(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n X_i}{n}\right) \end{aligned}$$

se considero che le X_i sono indipendenti, devo prendere solo le covarianza tra gli stessi elementi

$$Cov(\bar{X}, X_j - \bar{X}) = Cov\left(\frac{X_j}{n}, X_j\right) - \sum_{i=1}^n Cov\left(\frac{X_i}{n}, \frac{X_i}{n}\right) = \frac{\sigma^2}{n} - \frac{\sum_{i=1}^n \sigma^2}{n^2} = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$$

Statistiche e Distribuzione Campionaria VIII

Allora visto che S^2 è composto da una trasformazione di variabili aleatori indipendenti da \bar{X} , è anch'esso indipendente da \bar{X} .

- (i-2) Per questo secondo modo, notiamo prima che il numeratore della varianza campionaria può essere scritto come

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2\bar{X} \sum_{i=1}^n X_i = \sum_{i=1}^n X_i^2 - n\bar{X}^2$$

Definiamo poi una nuova variabile aleatoria $\mathbf{W} = \mathbf{A}\mathbf{X}$ dove \mathbf{A} è

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2 \cdot 1}} & -\frac{1}{\sqrt{2 \cdot 1}} & 0 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3 \cdot 2}} & \frac{1}{\sqrt{3 \cdot 2}} & -\frac{2}{\sqrt{3 \cdot 2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{4 \cdot 3}} & \frac{1}{\sqrt{4 \cdot 3}} & \frac{1}{\sqrt{4 \cdot 3}} & -\frac{3}{\sqrt{4 \cdot 3}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}$$

Statistiche e Distribuzione Campionaria IX

- \mathbf{A} (di dimensione $n \times n$) è ortogonale visto che $\mathbf{A}\mathbf{A}^T = \mathbf{I}_n$
- poichè \mathbf{W} è una trasformazione lineare di \mathbf{X} abbiamo che \mathbf{W} è normale con
 - valore atteso $\mathbf{A}\mu\mathbf{1}_n = (\mu\sqrt{n}, 0, 0, \dots, 0)^T$
 - matrice di covarianza $\mathbf{A}\sigma^2\mathbf{I}_n\mathbf{A}^T = \sigma^2\mathbf{I}_n$
- il primo elemento di \mathbf{W} è $W_1 = \sqrt{n}\bar{X}$

Visto che le componenti di \mathbf{W} sono indipendenti, se io dimostro che S^2 è funzione solo di (W_2, W_3, \dots, W_n) (quindi non del primo), dimostro che non dipende da \bar{X} . Consideriamo la seguente somma

$$W_2^2 + W_3^2 + \dots + W_n^2 = \mathbf{W}^T \mathbf{W} - W_1^2 = \mathbf{X}^T \mathbf{A}^T \mathbf{A} \mathbf{X} - \left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \right)^2 =$$
$$\sum_{i=1}^n X_i^2 - n\bar{X}^2 = (n-1)S^2$$

che dimostra la tesi

Statistiche e Distribuzione Campionaria \bar{X}

(ii) Il secondo punto è facile da dimostrare visto che \bar{X} è combinazione lineare di normali indipendenti, quindi normale, con

$$\begin{aligned}E(\bar{X}) &= E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n E(X_i)}{n} = \mu \\ \text{Var}(\bar{X}) &= \text{Var}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n \text{Var}(X_i)}{n^2} = \frac{\sigma^2}{n}\end{aligned}$$

(iii) Abbiamo che

$$\frac{\sum_{i=2}^n W_i^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

e $\frac{\sum_{i=2}^n W_i^2}{\sigma^2}$ è una somma di $(n-1)$ normali standard al quadrato $\left(\sum_{i=2}^n \left(\frac{W_i}{\sigma}\right)^2\right)$, e quindi $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

Statistiche e Distribuzione Campionaria XI


(iiii) Per i punti precedenti abbiamo che

$$(\bar{X} - \mu)/\sqrt{\sigma^2/n} \sim N(0, 1)$$

allora

$$\frac{(\bar{X} - \mu)/\sqrt{\sigma^2/n}}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$$

visto che il rapporto tra una normale standard $((\bar{X} - \mu)/\sqrt{\sigma^2/n})$, e la radice quadrata di un chi-quadrato $((n-1)S^2/\sigma^2)$ diviso i gradi di libertà $(n-1)$, è una t di student. \square



Statistiche e Distribuzione Campionaria XII

Teorema - Distribuzione delle statistiche media e varianza campionaria nel caso di due normali

Siano X_1, \dots, X_m iid $N(\mu_x, \sigma_x^2)$, e Y_1, \dots, Y_n iid $N(\mu_y, \sigma_y^2)$, con $X_i \perp Y_j, i = 1, \dots, m, j = 1, \dots, n$, allora

- (i) $\bar{X} - \bar{Y}$ e S_x^2/S_y^2 sono indipendenti
- (ii) $\bar{X} - \bar{Y}$ ha una distribuzione normale di media $\mu_x - \mu_y$ e varianza $\sigma_x^2/m + \sigma_y^2/n$
- (iii) $\sigma_y^2 S_x^2 / (\sigma_x^2 S_y^2)$ ha una distribuzione F con $m - 1$ e $n - 1$ gdl.

Dimostrazione:

- (i) Per il teorema precedente $(\bar{X}, \bar{Y})^T$ e $(S_x^2, S_y^2)^T$ sono composti da elementi indipendenti tra di loro
- (ii) Da dimostrare (da fare come esercizio)

Statistiche e Distribuzione Campionaria XIII

(iii)

$$\frac{\sigma_y^2 S_x^2}{\sigma_x^2 S_y^2} = \frac{((m-1)S_x^2/\sigma_x^2)/(m-1)}{((n-1)S_y^2/\sigma_y^2)/(n-1)}$$

e dal teorema precedente sappiamo che

$$((m-1)S_x^2/\sigma_x^2) \sim \chi^2(n-1)$$

e

$$((n-1)S_y^2/\sigma_y^2) \sim \chi^2(n-1)$$

e quindi

$$\frac{\sigma_y^2 S_x^2}{\sigma_x^2 S_y^2} \sim F(m-1, n-1)$$

visto che il rapporto di due chi quadrati divisi per i gradi di libertà è una F.



Statistiche d'ordine I

Assumiamo di avere un campione casuale $\mathbf{X} = (X_1, \dots, X_n)$ e il campione ordinato $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$, con $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. La variabile $X_{(j)}$ viene chiamata la j -esima statistica d'ordine. La distribuzione di $X_{(j)}$ è

$$F_{X_{(j)}}(x) = P(X_{(j)} \leq x)$$

che equivale alla probabilità j o più elementi di \mathbf{X} siano $\leq x$. Per ogni singola v.a. la probabilità di essere $\leq x$ è data da

$$F_X(x) = P(X \leq x)$$

e quindi

$$F_{X_{(j)}}(x) = \sum_{k=j}^n \binom{n}{k} F_X(x)^k (1 - F_X(x))^{n-k}$$

dove il coefficiente binomiale serve a tener conto delle diverse combinazioni.

Ci sono casi particolari:

Statistiche d'ordine II

Massimo

$$F_{X_{(n)}}(x) = P(\max(X_1, \dots, X_n) \leq x) = \prod_{i=1}^n F_x(x) = [F_x(x)]^n$$

Minimo

$$\begin{aligned} F_{X_{(1)}}(x) &= P(\min(X_1, \dots, X_n) \leq x) = 1 - P(\min(X_1, \dots, X_n) > x) = \\ &= 1 - \prod_{i=1}^n P(X_i > x) = 1 - \prod_{i=1}^n (1 - F_x(x)) = 1 - [1 - F_x(x)]^n \end{aligned}$$

Simulazione I

La simulazione di variabili casuali è molto utile quando si vogliono studiare caratteristiche che non si riescono a verificare matematicamente, o per verificare empiricamente risultati noti. Per esempio, nel **“chunk simulazioni”** ci sono dei esempi in cui si dimostra empiricamente che la media campionaria tende alla media vera. Nel corso utilizzeremo simulazioni per fare esempi e per mostrare situazioni “particolari” o casi specifici.

La simulazione è un argomento molto vasto che non tratteremo. Però va ricordato che i numeri ottenuti tramite simulazioni sono chiamati pseudo-random, perchè non sono veramente casuali, ma creati tramite un algoritmo. L'algoritmo assicura che la sequenza di numeri pseudo-casuali sia indistinguibile da “veri” numeri casuali. Ogni algoritmo inizia specificando un seme che cambia ogni volta che viene creato un numero casuale (vedere **“chunk simulazioni”**).

