DISTRIBUZIONE CAMPIONARIA

Si abbia un carattere X osservabile sulle unità di una popolazione. La distribuzione di frequenza relativa (X discreto) o di densità di frequenza relativa (X continuo) è assimilabile ad una Variabile Casuale (VC) che descrive compattamente l'esperimento *"estrazione di una unità dalla popolazione e osservazione del carattere X sull'unità estratta"*.

X *(lettera maiuscola)* identifica l'insieme dei valori (modalità) che il carattere può assumere nella popolazione e f(x) è la distribuzione di probabilità (densità di probabilità) ovvero

$f(x) = P(X=x)$ se il carattere è discreto

$f(x)\,dx = P(x-\varepsilon \leq X=x \leq x+\varepsilon)$ con ε infinitesimo se il carattere è continuo

Se si estrae dalla popolazione un campione casuale semplice con replicazione di n unità, l'esperimento può essere descritto da una ennupla $(X_1, X_2, ..., X_n)$ di VC indipendenti e identicamente distribuite (i.i.d) che identificano il risultato della 1a, 2°, …, n-esima estrazione

Supponiamo di calcolare sulle osservazioni campionarie una "statistica", ovvero una funzione t dei valori osservati (ad esempio la media). La VC

$S=t(X_1, X_2,...,X_n)$ ***(con le X maiuscole)***

descrive l'insieme dei possibili risultati dell'esperimento, mentre

$s=t(x_1, x_2,...,x_n)$ ***(con le x minuscole)***

indica il valore di S che si ottiene calcolando t su una data realizzazione $(x_1, x_2,...,x_n)$ della VC $(X_1, X_2,...,X_n)$, ovvero su un campione estratto con la procedura descritta dalla VC ennupla.

La distribuzione di probabilità (densità di prob) della VC $t(X_1, X_2,...,X_n)$ si dice distribuzione campionaria della statistica S ed è

$$f(s) = f[t(x_1, x_2, \dots, x_n]$$
$$= P[(X_1 = x_1) \cap (X_2 = x_2) \cap \dots \cap (X_n = x_n)$$

Ma, essendo le VC $X_1, X_2,...,X_n$ indipendenti e avendo tutte la stessa distribuzione f(x) corrispondente alla distribuzione di frequenza del carattere X nella popolazione, risulta

$$f(s) = f[t(x_1, x_2, \dots, x_n] = f(x)f(x) \dots f(x)$$

La distribuzione campionaria è lo strumento che consente di valutare il livello di incertezza connesso all'uso di una statistica campionaria. La sua conoscenza permette di risolvere il problema diretto: partendo da una popolazione nota, estraendo un campione e calcolando sul campione una statistica, quali valori si possono presentare e con quale probabilità?

Se la dovessi costruire empiricamente, dovrei estrarre tutti i possibili campioni, calcolare la statistica su ciascuno di essi e derivare la probabilità di presentarsi di ogni risultato (i vari campioni sono eventi incompatibili e quindi la probabilità di un dato risultato è pari alla somma delle probabilità di tutti i possibili campioni che a quel risultato danno luogo)

Vedi esempio cartella excel distribuz campionaria

Non posso costruire empiricamente l'universo dei campioni (che ha dimensioni più grandi della popolazione stessa)

In taluni casi posso scegliere un modello che rappresenta la popolazione e derivare analiticamente la distribuzione campionaria.

DISTRIBUZIONE CAMPIONARIA DI UNA PROPORZIONE

Se il carattere è dicotomico la popolazione può essere rappresentata da una vc X di Bernoulli, che assume valore 1 (in caso di successo) e 0 (insuccesso), con media è p e varianza è p(1-p)

La distribuzione campionaria della proporzione di successi su n estratti è una binomiale Y/n i cui valori sono dati dal numero possibile Y di successi in n estrazioni con replicazione (Y=1, 2, …, n) diviso per il numero n di estrazioni fatte.

La VC Binomiale è di fatto una media di n bernoulliane indipendenti risultando

$$\frac{Y}{n} = \frac{1}{n}\sum X$$ **da cui**

$$E\left(\frac{Y}{n}\right) = E\left(\frac{1}{n}\sum X\right) = \frac{1}{n}\sum E(X) = \frac{1}{n}np = p$$

$$VAR\left(\frac{Y}{n}\right) = \frac{1}{n^2}\sum VAR(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

E la sua distribuz di probabilità è

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

DISTRIBUZIONE CAMPIONARIA DI UNA MEDIA
campioni tratti da popolazione normale

Se il carattere è continuo e la popolazione può essere rappresentata da una vc X Normale con media μ e s.q.m. σ , la media campionaria

$$\overline{X} = \frac{1}{n}\sum_{i} X_i$$

ha distribuzione campionaria anch'essa normale con media pari a

$$E(\overline{X}) = \mu$$

e varianza pari a

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

Se la varianza è nota, si può costruire la VC standardizzata

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Che ha distribuzione normale con media 0 e varianza 1.

Se la varianza non è nota, si può trasformare la VC $\overline{X}$ nel modo seguente

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

dove

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \mu)^2$$

È la varianza calcolata sul campione estratto (ma divisa per n-1 e non per n)

La VC T_n ha una distribuzione t di Student che risulta

$$f(t) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\sqrt{\pi g}\cdot\Gamma\left(\frac{g}{2}\right)}\left(1 + \frac{t^2}{g}\right)^{-\frac{g+1}{2}}$$

con

$$\Gamma(g) = \int_0^{+\infty} e^{-x}x^{g-1}dx$$

Dove g è detto gradi di libertà e, nel caso in esame, è pari ad n-1
Se g è ≥3, la vc T ha media pari a 0 e varianza pari a g/(g-2).

STIMA PUNTUALE

$(X_1, X_2, \dots, X_n)$**è la VC ennupla che descrive l'esperimento campionario**

Supponiamo di dover stimare il valore che assume un parametro di interesse della popolazione (per es la media). Chiamiamo questo parametro θ**.**

Dare un valore come stima di θ **a partire dalle n osservazioni campionarie significa applicare alle VC che definiscono il campione una funzione che le sintetizzi. In generale**

$$T_n = f(X_1, X_2, \dots, X_n)$$

T_n **è detto stimatore e definisce la "regola" con la quale sintetizzare le osservazioni campionarie. Quando lo si calcola su un particolare campione estratto** $(x_1, x_2, \dots, x_n)$ **fornisce una stima**

$$t_n = f(x_1, x_2, \dots, x_n)$$

Non si può valutare la stima sulla base della differenza $|t_n - \theta|$ **essendo** θ **incognito.**

Si valuta il processo di stima, ovvero lo stimatore

Corretto (non distorto)
$E(T_n) = \theta$ **distorsione** $B(T_n) = E(T_n) - \theta$

Dovrebbe avere un piccolo Mean Square Error $MSE(T_n) = E[(T_n - \theta)^2]$**, ma la grandezza non è calcolabile essendo θ incognito. Se lo stimatore è corretto, tuttavia,** $E(T_n) = \theta$ **e, quindi, MSE è pari alla varianza, risultando**

$$MSE(T_n) = E\left[(T_n - \theta)^2\right] = E\left[(T_n - E(T_n))^2\right]$$

Pertanto, se due stimatori T_{n1} **e** T_{n2} **sono entrambi corretti, possono essere valutati in base alla loro varianza e si dirà che lo stimatore** T_{n1} **è più efficiente di** T_{n2} **se**

$$Var(T_{n1}) < Var(T_{n2})$$

Se lo stimatore è corretto e vale

$$\lim_{n\to\infty} E\left[(T_n - E(T_n))^2\right] = 0$$

Si dice che è consistente in media quadratica