STATISTICA INFERENZIALE: TRE FILE PDF

Il file PDF "Statistica Inferenziale (I)" è il primo dei seguenti tre file scaricabili dal Materiale Didattico (sotto la voce "Dispense") del Corso di Statistica (AA 2008-9)

Statistica Inferenziale (I): campionamento e stima puntuale	pag. 2-5	([1]-[8bis])
Statistica Inferenziale (II): stimatori e stima intervallare	pag. 6-9	([9]-[14bis])
Statistica Inferenziale (III): test statistici	pag. 10-12	([15]-[23bis])

I tre file sopra elencati riassumono in modo **schematico** ma **completo** i <u>concetti</u> e <u>risultati</u> principali della parte di Statistica Inferenziale del programma del Corso di Statistica già esposta agli studenti in aula a lezione dai docenti del Corso (escluso il modello di regressione lineare).

Contengono inoltre un <u>esercizio svolto</u> (vedi pag. 5) introduttivo all'applicazione di tali concetti e risultati (le dimostrazioni di tali risultati non sono contenute nei suddetti file, anch'esse comunque sono già state esposte a lezione dai docenti del Corso e, p. es., sono reperibili nelle copie dei lucidi delle lezioni corrispondenti presso Yellow Print).

I tre file sopra indicati, data la loro **schematicità**, possono essere particolarmente utili se vengono integrati dallo studente con **gli appunti presi alle lezioni dei docenti del Corso**, oppure, p. es., con le fotocopie dei lucidi delle lezioni reperibili presso Yellow Print, od anche, secondariamente, con i corrispondenti capitoli del libro di Ross. Un altro riferimento utile è il file PDF "Correlazione e Combinazioni Lineari" scaricabile dal Materiale Didattico del Corso. Anche questo file dà un riassunto schematico di concetti e risultati già esposti a lezione dai docenti del Corso, non contiene tuttavia le dimostrazioni di tali risultati (anch'esse comunque sono già state esposte a lezione e, p. es., sono reperibili nelle copie lucidi delle lezioni corrispondenti presso Yellow Print).

NB: Alcune parti di una prima versione parziale di questi file (qui rivista, migliorata e completata definitivamente) sono già state utilizzate da uno dei docenti del Corso come supporto didattico in aula in alcune lezioni in cui sono stati trattati i diversi argomenti di Statistica Inferenziale e quindi inserite nelle copie dei lucidi delle lezioni stesse. Un ringraziamento a chi segnalerà eventuali sviste, errori di stampa o battitura, ecc.

STATISTICA INFERENZIALE (I): campionamento e stima puntuale

[1]

Prima di tutto un breve ripasso sulle variabili aleatorie:

$$X = \begin{cases} x = 1, 2, \ldots 6 \\ p(x) = \dfrac{1}{6} = 0.1\overline{6} \end{cases} \Rightarrow \begin{cases} \mu_X = 1p(1) + 2p(2) + \ldots + 6p(6) = \boxed{3.5} \\ E(X^2) = 1^2 p(1) + 2^2 p(2) + \ldots + 6^2 p(6) = \boxed{15.1\overline{6}} \\ \sigma_X^2 = E(X^2) - (\mu_X)^2 = \boxed{2.9167} \end{cases}$$

$X =$ "**numero** dei componenti di **una qualsiasi famiglia** di un Comune di N famiglie", p. es. N = 1000, e di cui si sa che non vi sono famiglie con più di 6 componenti.

$X_t = X =$ "**numero** dei componenti della *t*-esima famiglia ($t = 1, 2, \ldots N = 1000$) del Comune"

Poiché **le probabilità** $p(x)$ **sono note** si può determinare il **valore esatto** di μ_X, $E(X^2)$ e σ_X^2 applicando le loro formule come fatto sopra (nel ns. caso, per semplicità, le probabilità sono costanti $\forall x$, ma in generale ovviamente non lo sono). Si ricordi inoltre che:

$$p(x) = \frac{1}{6} = 0.1\overline{6} \quad (x = 1, 2, \ldots 6)$$

è la probabilità che una qualsiasi famiglia del Comune abbia $x(= 1, 2, \ldots 6)$ componenti

è la probabilità di estrarre a caso dalle N famiglie del Comune una famiglia con $x(= 1, 2, \ldots 6)$ componenti (nel ns. caso è costante $\forall x$)

è frequenza relativa delle famiglie con $x(= 1, 2, \ldots 6)$ componenti (nel ns. caso è costante $\forall x$)

$p(x)$ dà, moltiplicata per 100, $16.\overline{6}\%$ cioè la percentuale delle famiglie con $x(= 1, 2, \ldots 6)$ componenti (nel ns. caso è costante $\forall x$)

$p(x)$ dà, moltiplicata per N = 1000, $166.\overline{6}$ cioè il numero delle famiglie con $x(= 1, 2, \ldots 6)$ componenti (nel ns. caso è costante $\forall x$)

$p(x)$ dà, con la opportuna formula, $\mu_X = 3.5$ cioè il numero medio di componenti delle famiglie del Comune

[2]

In generale nelle
APPLICAZIONI ECONOMICO-AZIENDALI, INGEGNERISTICHE e FISICHE

le **PROBABILITA'** (o DENSITA')**, e quindi** μ_X**,** $E(X^2)$ **e** σ_X^2

NON HANNO VALORI NOTI

$$X = \begin{cases} x = 1, 2, \ldots 6 \\ p(x) = \boxed{?} \end{cases} \Rightarrow \begin{cases} \mu_X = 1p(1) + 2p(2) + \ldots + 6p(6) = \boxed{?} \\ E(X^2) = 1^2 p(1) + 2^2 p(2) + \ldots + 6^2 p(6) = \boxed{?} \\ \sigma_X^2 = E(X^2) - (\mu_X)^2 = \boxed{?} \end{cases}$$

COME SI DETERMINANO TALI VALORI $\boxed{?}$

DETERMINAZIONE DEI VALORI di $p(x)$, μ_X, $E(X^2)$ e σ_X^2 [3]

Determinazione di $p(x), \mu_X, E(X^2)$, e σ_X^2 con

- CENSIMENTO (N)
  - dà il valore esatto
  - non è tempestivo ai fini decisionali
  - è troppo costoso
- **CAMPIONAMENTO** $(n \ll N)$
  - dà una **STIMA** che approssima bene il valore esatto con probabilità crescente al crescre di *n*
  - è tempestivo ai fini decisionali
  - è meno costoso

LA **STATISTICA INFERENZIALE** DA' I **METODI DI CAMPIONAMENTO** PER **STIMARE** I VALORI **NON NOTI** DI [4]

$p(x)$, μ_X, $E(X^2)$, σ_X^2, ecc.

Metodo di CAMPIONAMENTO CASUALE SEMPLICE (CCS) [5]

di dimensione $n \ (\ll N)$

$$\begin{array}{cccccccc}
\multicolumn{8}{c}{X = X_1, X = X_2, X = X_3, X = X_4, X = X_5, X = X_6, \ \dots \ X = X_n} \\
X_1, & X_2, & X_3, & X_4, & X_5 & X_6 & \dots & X_n \\
\Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \dots & \Downarrow \\
x_1 = 5, & x_2 = 3, & x_3 = 2, & x_4 = 3, & x_5 = 6 & x_6 = 2 & \dots & x_n = 1
\end{array}$$

X v.a. di riferimento = **Popolazione statistica** X

$(X_1, X_2, \dots X_n)$ = **Campione statistico**, $(x_1 = 5, x_2 = 3, \dots x_n = 1)$ = **realizzazione campionaria**

DEFINIZIONE FORMALE DI **CAMPIONE STATISTICO CASUALE SEMPLICE**

Le *n* v.a. X_t $(t = 1, 2, \dots n)$ che costituiscono il campione statistico $(X_1, X_2, \dots X_n)$ sono **tutte uguali** (ovvero, **tutte distribuite identicamente**) **alla v.a.** X **di riferimento** cioè alla **popolazione statistica** X (vedi anche il riquadro [1]). Inoltre il CSC richiede l'**indipendenza stocastica** di tali *n* v.a. Si ha dunque che:

un Campione Statistico Casuale Semplice è dato da

***n* v.a.** $(X_1, X_2, \dots X_n)$ **I.I.D.**

dove la prima "**I.**" è l'iniziale di "**I**ndipendenti" (o "**I**ndependent"), la seconda "**I.**" è l'iniziale di "**I**denticamente" (o "**I**dentically") e "**D.**" è l'iniziale di "**D**istribuite" (o "**D**istributed"). **L'ipotesi di indipendenza stocastica delle *n* v.a.** X_t **assicura che** la probabilità (o densità) dei valori possibili della v.a. X non cambi passando da una osservazione alla successiva. **In pratica l'indipendenza stocastica si realizza (a)** rappresentando l'insieme dei valori possibili della v.a. X di riferimento come un'urna da cui si estraggono casualmente *n* dei valori possibili, e **(b)** re-immettendo nell'urna ciascun valore osservato prima di effettuare l'estrazione successiva.

DETERMINAZIONE DELLE STIME DI [6]

$p(x)$, μ_X, $E\left(X^2\right)$, σ_X^2, ecc.

$$\begin{array}{cccccccc} X=X_1, & X=X_2, & X=X_3, & X=X_4, & X=X_5, & X=X_6, & \ldots & X=X_n \\ X_1, & X_2, & X_3, & X_4, & X_5 & X_6 & \ldots & X_n \\ \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \ldots & \Downarrow \\ x_1=5, & x_2=3, & x_3=2, & x_4=3, & x_5=6 & x_6=2 & \ldots & x_n=1 \end{array}$$

X v.a. di riferimento = **popolazione statistica** X

$\left(X_1, X_2, \ldots X_n\right)$ = **Campione statistico**, $\left(x_1=5, x_2=3, \ldots x_n=1\right)$ = **realizzazione campionaria**

stima di θ: $\hat{\theta}$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	n	$\rightarrow$	∞
$p(1)$: $\hat{p}(1)$	0	0	0	0	0	0	…	"$\rightarrow$"	$p(1)$
$p(2)$: $\hat{p}(2)$	0	0	1/3	1/4	1/5	2/6	…	"$\rightarrow$"	$p(2)$
$p(3)$: $\hat{p}(3)$	0	1/2	1/3	2/4	2/5	2/6	…	"$\rightarrow$"	$p(3)$
$p(4)$: $\hat{p}(4)$	0	0	0	0	0	0	…	"$\rightarrow$"	$p(4)$
$p(5)$: $\hat{p}(5)$	1	1/2	1/3	1/4	1/5	1/6	…	"$\rightarrow$"	$p(5)$
$p(6)$: $\hat{p}(6)$	0	0	0	0	1/5	1/6	…	"$\rightarrow$"	$p(6)$
μ_X: $\hat{\mu}_X=\bar{x}_n$	5	4	$3.\overline{3}$	3.25	3.8	…	…	"$\rightarrow$"	μ_X
$E(X^2)$: $\hat{E}(X^2)$	25	17	$12.\overline{6}$	11.45	16.6	…	…	"$\rightarrow$"	$E(X^2)$
σ_X^2: $\hat{\sigma}_X^2=s_n^2$	0	1	1.5	1.1875	2.16	…	…	"$\rightarrow$"	σ_X^2

QUALI SONO LE FORMULE USATE PER CALCOLARE LE STIME ? [7]

SONO QUELLE <u>GIA' USATE</u> PER I DATI OSSERVATI IN <u>STATISTICA DESCRITTIVA</u>

CON I SIMBOLI CAMBIATI

Ad esempio nella colonna n = 5 del riquadro [6] si ha

$$\boxed{M\left(X\right)}=5\frac{1}{5}+3\frac{2}{5}+2\frac{1}{5}+6\frac{1}{5}=3.8=\frac{5+3+2+3+6}{5}=\frac{x_1+\ldots+x_5}{5}=\frac{1}{5}\sum_{t=1}^{5}x_t=\boxed{\bar{x}_5=\hat{\mu}_X}$$

$$\boxed{M\left(X^2\right)=}5^2\frac{1}{5}+3^2\frac{2}{5}+2^2\frac{1}{5}+6^2\frac{1}{5}=16.6=\frac{5^2+3^2+2^2+3^2+6^2}{5}=\frac{x_1^2+\ldots+x_5^2}{5}=\frac{1}{5}\sum_{t=1}^{5}x_t^2=\boxed{\hat{E}\left(X^2\right)}$$

$$\boxed{V\left(X\right)}=M\left(X^2\right)-M\left(X\right)^2=2.16=\hat{E}\left(X^2\right)-\left(\bar{x}_5\right)^2=\boxed{s_5^2=\hat{\sigma}_X^2}$$

N.B. (I) in realtà per **piccoli campioni** (con $n<30$) la stima di σ_X^2 data da s_n^2 si corregge con il fattore di correzione $\frac{n}{n-1}$, cioè si utilizza $s_n^2\frac{n}{n-1}=s_c^2=\hat{\sigma}_X^2$, **vedi riquadro [11]**

N.B. (II): le stime si calcolano **SOLO e DIRETTAMENTE** per la dimensione n del campione prefissata, p. es. se è fissato n = 5 allora per ottenere le stime si calcola **solo e direttamente** la colonna con n = 5 del riquadro [6].

N.B. (III): rispondere alle domande (1), (2), (3a) e (4a) dell'Esercizio nel riquadro [8].

Esercizio di Statistica inferenziale [8]

(Nota Bene: in quanto primo esercizio di questo tipo nel Corso di Statistica, il testo che segue è molto più dettagliato ed esplicito di quanto non saranno gli esercizi successivamente svolti in aula o che si possono trovare nei testi di esame)

E' noto che il reddito di esercizio di una qualsiasi azienda di un dato settore industriale (composto da *N* = 973 aziende) è una v.a. $X \sim N\left(\mu_X;\sigma_X^2\right)$ con reddito medio atteso μ_X e varianza σ_X^2 non noti. In realtà tali due parametri si potrebbero determinare in modo esatto facendo il censimento dei redditi delle aziende del settore, cioè: (a) rilevando tutti gli *N* = 973 redditi delle imprese del settore industriale e (b) calcolando con gli *N* = 973 redditi osservati le formule di media e varianza viste nella parte di Statistica descrittiva del corso. Tuttavia, poiché determinare in modo esatto reddito medio atteso μ_X e varianza σ_X^2 con il censimento richiede troppo tempo ed è troppo costoso, si preferisce stimare il valore di tali due parametri mediante un campione casuale semplice $(X_1, X_2, \ldots X_{80}, X_{81})$ con *n* = 81. La corrispondente realizzazione campionaria $(x_1, x_2, \ldots x_{80}, x_{81})$ degli *n* = 81 rediti osservati è stata riassunta nei due valori:

$$\sum_{t=1}^{81} x_t = 3321, \quad \sum_{t=1}^{81} x_t^2 = 180441$$

(1) Specificare quale è la popolazione statistica (ovvero la v.a. di riferimento)
(2) Specificare quali sono le formule che si usano per determinare μ_X e varianza σ_X^2 con il censimento delle *N* = 973 aziende del settore.
(3a) Determinare la stima del reddito medio atteso μ_X delle *N* = 973 aziende del settore sulla base della realizzazione campionaria di dimensione *n* = 81 osservata; (**3b)** si indichi inoltre lo stimatore utilizzato e le sue proprietà (la risposta è nel riquadro [14bis] e richiede la lettura dei riquadri [9] e [10]).
(4a) Determinare la stima della varianza σ_X^2 dei redditi delle *N* = 973 aziende del settore sulla base della realizzazione campionaria di dimensione *n* = 81 osservata. **(4b)** si indichi inoltre lo stimatore utilizzato e le sue proprietà (la risposta è nel riquadro [14bis] e richiede la lettura dei riquadri [9] e [11]).
(5) Determinare, sulla base della realizzazione campionaria osservata, l'intervallo di confidenza del reddito medio atteso μ_X con coefficiente di confidenza $1-\alpha = 0.95$ (la risposta è nel riquadro [14bis] e richiede la lettura del riquadro [13]).
(6) Infine, sulla base della realizzazione campionaria osservata, si esegua il test statistico bilaterale per il reddito medio atteso μ_X : $H_0 : \mu_X = 43.5$ contro $H_1 : \mu_X \neq 43.5$ al livello di significatività $\alpha = 0.05$ (la risposta è nel riquadro [23bis] e richiede la lettura dei riquadri da [15] a [21]).

[8bis]

Risposta alla Domanda (1). (vedi riquadri da [1], [3] e [5]). La v.a. di riferimento o popolazione statistica è la v.a. "reddito di esercizio" $X \sim N\left(\mu_X;\sigma_X^2\right)$ con reddito medio atteso μ_X e varianza σ_X^2 non noti.

Risposta alla Domanda (2). (vedi riquadri [1], [3], [6] e [7]).

$$M(X) = \frac{1}{N}\sum_{t=1}^{N} x_t = \frac{1}{973}\sum_{t=1}^{973} x_t = \mu_X$$

$$V(X) = M\left(X^2\right) - M(X)^2 = \frac{1}{N}\sum_{t=1}^{N} x_t^2 - M(X)^2 = \frac{1}{973}\sum_{t=1}^{973} x_t^2 - M(X)^2 = \sigma_X^2$$

Risposta alla Domanda (3a). (vedi riquadri [6] e [7]). La stima di μ_X è data da:

$$\bar{x}_{81} = \frac{1}{81}\sum_{t=1}^{81} x_t = \frac{1}{81}3321 = 41 = \hat{\mu}_X$$

Risposta alla Domanda (4a). (vedi riquadri [6] e [7]). E' $n = 81 > 30$, per stimare σ_X^2 va bene sia s_{81}^2 sia s_c^2 :

$$s_{81}^2 = \frac{1}{81}\sum_{t=1}^{81} x_t^2 - \left(\bar{x}_{81}\right)^2 = \frac{1}{81}\sum_{t=1}^{81}180441 - 41^2 = 2227.6667 - 1681 = 546.6667 = \hat{\sigma}_X^2$$

$$s_c^2 = \frac{81}{81-1} s_{81}^2 = 1.0125 \times 546.6667 = 553.5000 = \hat{\sigma}_X^2 \left(\cong 546.6667 = s_{81}^2\right)$$