Ripasso

RIPASSO

- Se $X \sim N(\mu, \sigma^2)$ $\qquad \phi_X(t) = e^{t\mu + \frac{t^2}{2}\sigma^2}$

$$Z \sim N(0,1) \qquad \phi_Z(t) = e^{t^2/2}$$

$$\ln\left(\phi_Z(t)\right) = \frac{t^2}{2}$$

- Se $T = X_1 + X_2 \dots + X_n$ con $X_1, X_2 \dots X_n$ v.c. INDIPENDENTI

$$\phi_T(t) = \phi_{X_1}(t)\,\phi_{X_2}(t) \dots \phi_{X_n}(t) = \prod_{k=1}^{n} \phi_{X_k}(t)$$

- Se due v.c. presentano la stessa funzione generatrice dei momenti sono identicamente distribuite, ovvero presentano stessa funzione di ripartizione di probabilità, stessa funzione di massa di prob. se sono discrete stessa funzione di densità di prob. se sono continue.

Teorema del limite centrale

TEOREMA DEL LIMITE CENTRALE o CENTRALE DEL LIMITE (TLC)

Date $X_1, X_2 \dots X_n$ v.c. i.i.d. con valor medio $E[X_k]=\mu$ e varianza $Var(X_k)=\sigma^2 \ \forall k=1,\dots n$, allora $\forall a \in \mathbb{R}$

$$\lim_{n\to+\infty} P\left(\frac{\left(\sum_{k=1}^{n} X_k\right)-n\mu}{\sigma\sqrt{n}} \le a\right)=F_Z(a) \quad \text{se } Z\sim N(0,1)$$

Oss

$$Y_n=\sum_{k=1}^{n} X_k \qquad E[Y_n]=E\left[\sum_{k=1}^{n} X_k\right]=$$

$$=\sum_{k=1}^{n} E[X_k]=\sum_{k=1}^{n}\mu=n\mu$$

$$Var(Y_n)=Var\left(\sum_{k=1}^{n} X_k\right)\underset{\text{INDIP.}}{=}\sum_{k=1}^{n} Var(X_k)=\sum_{k=1}^{n}\sigma^2=n\sigma^2$$

$$\frac{Y_n-n\mu}{\sigma\sqrt{n}}=\frac{\left(\sum_{k=1}^{n} X_k\right)-n\mu}{\sigma\sqrt{n}}$$ è una "standardizzazione" della v.c.

$$E\left[\frac{Y_n-n\mu}{\sigma\sqrt{n}}\right]=\frac{E[Y_n]-E[n\mu]}{\sigma\sqrt{n}}=\frac{n\mu-n\mu}{\sigma\sqrt{n}}=0$$

$$Var\left(\frac{Y_n - n\mu}{\sigma\sqrt{n}}\right) = Var\left(\frac{1}{\sigma\sqrt{n}}Y_n - \frac{n\mu}{\sigma\sqrt{n}}\right) = \frac{1}{(\sigma\sqrt{n})^2}Var(Y_n) = \frac{n\sigma^2}{n\sigma^2} = 1$$

$\boxed{\text{DIM}}$

Se $X_1, X_2 .. X_n$ sono IDENTICAMENTE DISTRIBUITE

$$\phi_{X_1}(t) = \phi_{X_2}(t) = \dots = \phi_{X_n}(t)$$

IPOTESI PRELIMINARE: SUPPONIAMO $\boxed{\mu = 0 \text{ e } \sigma^2 = 1}$
(se così non fosse è possibile definire delle nuove variabili $\hat{X}_1, \hat{X}_2 .. \hat{X}_n$ tali che $\hat{X}_k = \frac{X_k - \mu}{\sigma}$)

$$\phi_{X_k}(t) = e^{\ln(\phi_{X_k}(t))} = e^{L_k(t)} \qquad L_k(t) = \ln \phi_{X_k}(t)$$

$$L(t) = L_1(t) = L_2(t) \dots = L_n(t)$$

$$L(0) = \ln(\phi_{X_1}(t))\Big|_{t=0} = \ln(\phi_{X_1}(0)) = \ln(1) = 0$$

$$\phi_Y(0) = 1 \quad \forall Y \text{ v.c.}$$

$$L'(t)\Big|_{t=0} = \frac{dL(t)}{dt}\Big|_{t=0} = \frac{d}{dt}\left(\ln(\phi_{X_1}(t))\right)\Big|_{t=0} =$$

$$= \left(\frac{1}{\phi_{X_1}(t)} \frac{d\phi_{X_1}(t)}{dt}\right)\Big|_{t=0} = \frac{1}{\phi_{X_1}(0)} E[X_1] = 1 \cdot \mu = 0$$

$$\frac{d\phi_Y(t)}{dt}\Big|_{t=0} = E[Y]$$

$$L''(t)\Big|_{t=0} = \frac{d^2L(t)}{dt}\Big|_{t=0} = \frac{d}{dt}\left(\frac{1}{\phi_{X_1}(t)} \frac{d\phi_{X_1}(t)}{dt}\right)\Big|_{t=0} =$$

$$= \left[\underbrace{-\frac{1}{(\phi_{X_1}(t))^2}}_{\frac{d\,\phi_{X_1}(t)}{dt}} \frac{d\phi_{X_1}}{dt} \cdot \frac{d\phi_{X_1}(t)}{dt} + \frac{1}{\phi_{X_1}(t)} \frac{d^2\phi_{X_1}(t)}{dt^2}\right]\Bigg|_{t=0} = \Rightarrow \frac{d^2\phi_Y(t)}{dt^2}\Big|_{t=0} = E[Y^2]$$

$$= \left[-\frac{1}{(\phi_{X_1}(0))^2} E^2[X_1] + \frac{1}{\phi_{X_1}(0)} E[X_1^2]\right] =$$

$$= \left[-1 \cdot \mu^2 + 1 \cdot (\sigma^2 + \mu^2) \right] \underset{\substack{\mu = 0 \\ \sigma^2 = 1}}{=} 1$$

$$E[X_1^2] = \sigma^2 + \mu^2$$

RIASSUMENDO

$$L(t)\big|_{t=0} = L(0) = 0$$

$$L'(t)\big|_{t=0} = L'(0) = 0$$

$$L''(t)\big|_{t=0} = L''(0) = 1$$

Questo è l'ingrediente principale per la dimostrazione. Usiamo la variabile casuale Sn.

$$S_n = \frac{\left(\sum_{k=1}^{n} X_k\right) - n\mu}{\sigma\sqrt{n}}$$

$E[S_n] = 0$, $Var(S_n) = 1$ (vedi oss. prima della dim)

Cosa dice il teorema?

Si vuole dimostrare che $\forall a \in \mathbb{R}$

$$\lim_{n \to +\infty} P(S_n \leq a) = F_Z(a) \qquad \text{se } Z \sim N(0,1)$$

ovvero

$$\lim_{n \to +\infty} F_{S_n}(a) = F_Z(a)$$

$$\Downarrow$$

$$\lim_{n \to +\infty} \ln\left(\phi_{S_n}(t)\right) = \ln\left(\phi_Z(t)\right) = \ln\left(e^{\frac{t^2}{2}}\right) = \frac{t^2}{2}$$

$$\lim_{n \to +\infty} \phi_{S_n}(t) = \phi_Z(t)$$

Spostiamo l’attenzione dal prodotto tra le funzioni di ripartizione ad una strada più semplice, le funzioni generatrici dei momenti.

$$\phi_{S_n}(t) = \phi_{\frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}}}(t) \underset{\substack{\mu=0 \\ \sigma^2=1}}{=} \phi_{\frac{\sum_{k=1}^{n} X_k}{\sqrt{n}}}(t) \underset{\text{INDIPENDENZA}}{=}$$

$$= \phi_{\frac{X_1}{\sqrt{n}}}(t)\, \phi_{\frac{X_2}{\sqrt{n}}}(t)\, \phi_{\frac{X_3}{\sqrt{n}}}(t) \cdots \phi_{\frac{X_n}{\sqrt{n}}}(t) = \prod_{k=1}^{n} \phi_{\frac{X_k}{\sqrt{n}}}(t) =$$

$$= \prod_{k=1}^{n} E\left[e^{t\frac{X_k}{\sqrt{n}}}\right] = \prod_{k=1}^{n} \phi_{X_k}\left(\frac{t}{\sqrt{n}}\right)$$

$$= \prod_{k=1}^{n} e^{L\left(\frac{t}{\sqrt{n}}\right)}$$

$$= e^{nL\left(\frac{t}{\sqrt{n}}\right)}$$

La dimostrazione è complicata perché pur utilizzando semplici operazioni di somma-prodotto, deve esserci un filo logico.

$$(*) \quad E\left[e^{t\frac{X_k}{\sqrt{n}}}\right] = E\left[e^{\left(\frac{t}{\sqrt{n}}\right)X_k}\right] = \phi_{X_k}\left(\frac{t}{\sqrt{n}}\right)$$

$$\phi_{S_n}(t) = e^{\overset{(1)}{n L\left(\frac{t}{\sqrt{n}}\right)}}$$

$$\ln \phi_{S_n}(t) = n L\left(\frac{t}{\sqrt{n}}\right)$$

e dalle considerazioni precedenti si vuole dimostrare che

$$\lim_{n \to +\infty} n L\left(\frac{t}{\sqrt{n}}\right) = \lim_{n \to +\infty} \ln \phi_{S_n}(t) = \frac{t^2}{2}$$

La potenza della funzione generatrice dei momenti è che abbiamo fatto queste considerazioni solo con l'ipotesi che tutte abbiano la stessa distribuzione di probabilità e che siano indipendenti.

$$\lim_{n \to \infty} n L\left(\frac{t}{\sqrt{n}}\right) = \lim_{n \to +\infty} \frac{L\left(\frac{t}{\sqrt{n}}\right)}{n^{-1}}$$

Viene una forma indeterminata, quindi applichiamo il teorema di De L'Hospital.

$$= \lim_{n \to \infty} \frac{L'\left(\frac{t}{\sqrt{n}}\right)\left(-\frac{1}{2} t\, n^{-\frac{3}{2}}\right)}{-n^{-2}}$$

$$= \lim_{n \to \infty} \frac{L'\left(\frac{t}{\sqrt{n}}\right) \frac{1}{2} t}{n^{-\frac{1}{2}}}$$

$$\underset{\text{TEOREMA DI DE L'H.}}{=} \lim_{n \to +\infty} \frac{\frac{1}{2} t\, L''\left(\frac{t}{\sqrt{n}}\right)\left(\cancel{-\frac{1}{2}} t\, \cancel{n^{-\frac{3}{2}}}\right)}{\cancel{-\frac{1}{2} n^{-\frac{3}{2}}}} = \lim_{n \to +\infty} \frac{1}{2} t^2 L''\left(\frac{t}{\sqrt{n}}\right) =$$

La dimostrazione è conclusa, dimostrando che Sn tendente a infinito si distribuisce come una normale standard.

$$= \frac{1}{2} t^2 L''(0) \underset{L''(0) = 1}{=} \frac{1}{2} t^2$$

APPLICAZIONI DEL TLC

La prima è della binomiale.

- $X \sim B(n, p) \Rightarrow X = Y_1 + Y_2 \ldots + Y_n$
  con $Y_k \sim Be(p)$ INDIPENDENTI

Se $n \geq 30$ $X \dot{\sim} N(np, npq)$

Poi per una somma generale.

- Per qualsiasi $Y_1, Y_2 \dots Y_n$ v.c. i.i.d. con $n \geq 30$

$$X = Y_1 + Y_2 + \dots + Y_n \overset{\circ}{\sim} N(n\mu, n\sigma^2)$$

se $E[Y_k] = \mu$, $Var(Y_k) = \sigma^2$

Non sto standardizzando perché sto considerando un numero finito di variabili. Ad esempio una proprietà come questa viene usata per il calcolo della portata di un ponte, sapendo peso medio e varianza dei veicoli, si può dire complessivamente quanta portata serve, ponendo che i veicoli occupino spazio.

- $$S_n = \frac{\left(\sum_{k=1}^{n} X_k\right) - n\mu}{\sigma\sqrt{n}} \overset{\circ}{\sim} N(0,1) \quad \text{se } n \geq 30$$

$\Downarrow$ divido numeratore e denominatore per n

$$\frac{\left(\frac{\sum_{k=1}^{n} X_k}{n} - \mu\right)}{\frac{\sigma}{\sqrt{n}}} \overset{\circ}{\sim} N(0,1)$$

Ma:

$\Downarrow$ e denominatore

$$\frac{\left(\overbrace{\frac{\sum_{k=1}^{n} X_k}{n}}^{\bar{X}} - \underbrace{\mu}_{= E[\bar{X}]}\right)}{\underbrace{\frac{\sigma}{\sqrt{n}}}_{= \sqrt{Var(\bar{X})}}} \overset{\circ}{\sim}$$

per $n \geq 30$ $\quad \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

es. 1

Individui di una popolazione hanno peso di media 167 e scarto quadratico medio 27 in opportune unità di misura.

1) Per 36 individui calcolare

$$P(163 < \overline{X} < 171)$$

$$E[\overline{X}] = 167, \quad Var(\overline{X}) = \frac{(27)^2}{36} \quad \left(\frac{\sigma^2}{n}\right)$$

36 è più grande di 30, quindi si può applicare in maniera approssimata il teorema del limite centrale.

Per TLC $\quad \overline{X} \sim N\left(167, \frac{(27)^2}{36}\right)$

$$P(163 < \overline{X} < 171) = P(\overline{X} \leq 171) - P(\overline{X} \leq 163)$$

Scrivere minore, minore o uguale è la stessa cosa. Standardizziamo.

$$= P\left(\frac{\bar{X}-167}{\frac{27}{\sqrt{36}}} \le \frac{171-167}{\frac{27}{\sqrt{36}}}\right) - P\left(\frac{\bar{X}-167}{\frac{27}{\sqrt{36}}} \le \frac{163-167}{\frac{27}{\sqrt{36}}}\right) =$$

$$= P\left(\underbrace{\frac{\bar{X}-167}{\frac{27}{\sqrt{36}}}}_{\sim N(0,1)} \le \frac{171-167}{\frac{27}{\sqrt{36}}}\right) - P\left(\underbrace{\frac{\bar{X}-167}{\frac{27}{\sqrt{36}}}}_{\sim N(0,1)} \le \frac{163-167}{\frac{27}{\sqrt{36}}}\right) =$$

Quelle evidenziate diventano circa delle normali standard.

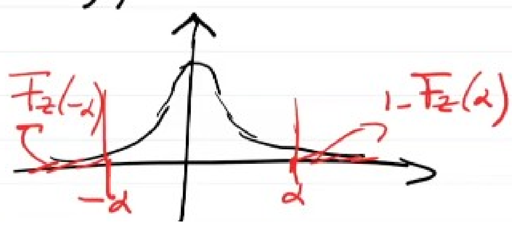
$$\underset{Z\sim N(0,1)}{=} P\left(Z \le \frac{4\cdot 6}{27}\right) - P\left(Z \le -\frac{4\cdot 6}{27}\right) =$$

$$= P\left(Z \le \frac{8}{9}\right) - P\left(Z \le -\frac{8}{9}\right) = F_Z\left(\frac{8}{9}\right) - F_Z\left(-\frac{8}{9}\right)$$

Si va di calcolatrice e tavole.

$$\underset{(*)}{=} F_Z\left(\frac{8}{9}\right) - \left(1 - F_Z\left(\frac{8}{9}\right)\right) = 2F_Z\left(\frac{8}{9}\right) - 1$$

$$(*) \quad F_Z(-\alpha) = 1 - F_Z(\alpha)$$



Si può ridurre il calcolo a un'unica funzione di ripartizione con la simmetria della normale standard.

2) Cosa cambia se il campione è composto da 144 individui $\Rightarrow$ per casa

Cambia la deviazione standard.

es. 2

AULA DI CAPIENZA 150 STUDENTI, SI SA CHE SOLO IL 30% DEGLI STUDENTI SEGUE IN PRESENZA. L'AULA È SUFFICIENTE PER 450 ISCRITTI?

Non applichiamo il teorema del limite centrale subito, capiamo che variabile serve.

X = 'n° studenti che decidono di frequentare in presenza'

X è una binomiale. 450 studenti, con probabilità di frequenza 3/10.

$$X \sim B\left(450, \frac{3}{10}\right)$$

3/10 non è molto grande, potrei approssimare a una poissoniana. Però la consegna dice:

$$P(X \leq 150)$$

Si può usare la poissoniana, ma diventa molto faticoso fare le somme, quindi conviene approssimarla a una gaussiana (teorema del limite centrale).

$$X \sim B\left(450, \frac{3}{10}\right) \overset{\circ}{\sim} N\left(450 \cdot \frac{3}{10}, 450 \cdot \frac{3}{10} \cdot \frac{7}{10}\right) =$$

$$= N\left(135, \frac{945}{10}\right)$$

$$P(X \le 150) = P\Big(\overbrace{\frac{X-135}{\sqrt{94.5}}}^{\dot\sim N(0,1)} \le \frac{150-135}{\sqrt{94.5}}\Big) \underset{TLC}{=} F_Z\left(\frac{15}{\sqrt{94.5}}\right)$$

Un miglioramento del calcolo si può fare con l’approssimazione alla continuità.

Approssimazione alla continuità

APPROSSIMAZIONE ALLA CONTINUITÀ

$n \ge 30$

Se $X \sim B(n, p) \;\dot\sim\; N(np, npq)$

$k \in \{0, \dots n\}$

$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ se considero $X \sim B(n, p)$

$P(X=k) = 0$ se $X \;\dot\sim\; N(np, npq)$

Questa differenza viene bypassata nel caso in cui si consideri la variabile casuale gaussiana, non consideri il singolo valore ma un intervallo di lunghezza unitaria centrato in k.

per superare questo problema considero

$$P(X=k) = P(k-0.5 < X \le k+0.5)$$

Ha senso sia nel caso discreto che nel caso continuo. Nel caso discreto equivale alla probabilità X = k, nel caso continuo diventerà un probabilità diversa da 0.

$$\Downarrow$$

$$P(k-0.5<X\le k+0.5)=\binom{n}{k}p^k(1-p)^{n-k}$$ se X è vista come binomiale

$$P(k-0.5<X\le k+0.5)=P(X\le k+0.5)-P(X\le k-0.5)\neq 0$$ se X è vista come v.c. gaussiana

Come al solito, bisogna standardizzare.

Regola generale che vale se $X \sim B(n,p) \approx N(np, npq)$

$$P(X=k)=P(k-0.5<X\le k+0.5)$$

$$P(X<k)=P\left(X \underset{(<)}{\le} k-\tfrac{1}{2}\right)$$

$$P(X\le k)=P\left(X\le k+\tfrac{1}{2}\right)$$

$$P(X>k)=P\left(X \underset{(>)}{\ge} k+\tfrac{1}{2}\right)$$

$$P(X\ge k)=P\left(X\ge k-\tfrac{1}{2}\right)$$

Queste regole derivano dall'idea di considerare i numeri naturali tutti racchiusi in degli intervalli.

Ogni volta che ho una probabilità che X sia uguale, maggiore, minore a uno di questi valori interi, la approssimo all'intervallo di lunghezza unitaria centrato in k.

TORNANDO ALL'ESEMPIO PRECEDENTE

$$P(X \leq 150) = P(X \leq 150{,}5)$$

Usiamo ½ per via di un motivo sperimentale, l'intervallo unitario approssima meglio e non lascia scoperto alcun valore della variabile casuale.

Inferenza statistica

Faccio una differenza tra statistica descrittiva e inferenza statistica, comunemente chiamate statistica, ma con un significato diverso.

STATISTICA (DESCRITTIVA) = ha come scopo la descrizione della o delle caratteristiche (misurabili) degli individui di una popolazione

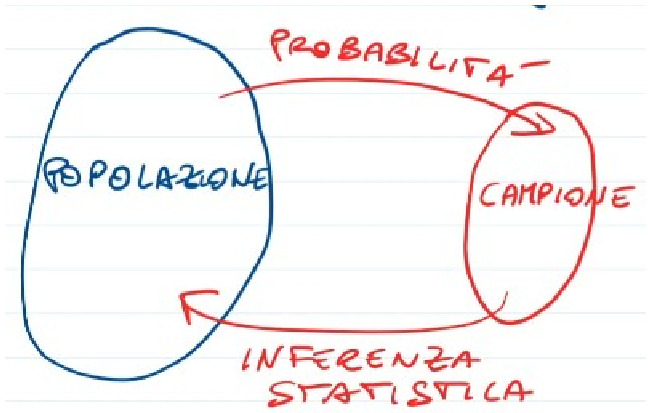
Studia gli individui di una popolazione che sono conosciuti, cioè rispetto a una caratteristica che si vuole studiare (come l'Istat, lo scopo fino al 2011 era di descrivere le caratteristiche di tutta la popolazione). Gli individui possono anche essere oggetti inanimati.

Invece l'inferenza statistica ha uno scopo diverso.

(INFERENZA) STATISTICA o STATISTICA INFERENZIALE = determinare le proprietà di una popolazione partendo da un sottoinsieme di individui scelti a caso (CAMPIONE)

Si parte da un sottoinsieme della popolazione per determinare le caratteristiche di una popolazione.

Confrontiamo statistica e probabilità: immaginiamo la popolazione, un insieme numeroso di individui accomunati da una certa somiglianza rispetto alle caratteristiche che vado a studiare. Immaginiamo un sottoinsieme della popolazione chiamato campione. La teoria della probabilità ci dice: se prendiamo un campione da una popolazione che caratteristiche avrà quel campione conoscendo la popolazione. L'inferenza statistica fa l'operazione contraria: partendo da un campione cerca di asserire qualcosa sulla popolazione.



L'inferenza statistica, si dice, è il problema inverso della probabilità. Per farla, serve conoscere la teoria della probabilità. È l'ultima teoria che nasce.

POPOLAZIONE = INSIEME MOLTO NUMEROSO (O INFINITO) DI INDIVIDUI CHE HANNO IN COMUNE UNA O PIÙ CARATTERISTICHE MISURABILI

CAMPIONE = SOTTOINSIEME DI UNA POPOLAZIONE FORMATO DA INDIVIDUI SCELTI A CASO.

L'ipotesi fondamentale dice che tutti gli individui di una stessa popolazione rispetto a una caratteristica da misurare abbiano tutti una distribuzione di probabilità uguale.

IPOTESI FONDAMENTALE (SOTTOINTESA) = LA CARATTERISTICA (O LE CARATTERISTICHE) DI CIASCUN INDIVIDUO DELLA POPOLAZIONE SI COMPORTANO IN MANIERA DESCRIVIBILE CON UNA (O PIÙ) V.C. CON DISTRIBUZIONE IDENTICA E INDIPENDENTE RISPETTO AGLI ALTRI INDIVIDUI DELLA STESSA POPOLAZIONE.

Vuol dire che il campione può essere descritto come un set di variabili casuali indipendenti identicamente distribuite.

⇓

IL CAMPIONE VIENE IDENTIFICATO COME UN INSIEME DI V.C. i.i.D CON FUNZIONE DI RIPARTIZIONE F CARATTERISTICA DELLA POPOLAZIONE

Di solito F non è nota oppure è nota a meno di qualche parametro e scopo dell'inferenza statistica è quello di stimare/determinare F e i suoi parametri.

Due tipi di inferenza statistica: F non conosciuta (per esempio l'altezza di una popolazione non so se si comporta come una gaussiana, una variabile uniforme, un'esponenziale…), oppure conosco F ma non conosco mu o sigma quadro o tutte e due.

STATISTICA = FUNZIONE DELLE V.C. DEL CAMPIONE

Una qualsiasi funzione.

es. già noto MEDIA CAMPIONARIA

$$\overline{X} = \frac{\sum_{k=1}^{n} X_k}{n}$$

Se la forma di F NON è nota si parla di INFERENZA NON PARAMETRICA

Se la forma di F è nota ma non sono noti alcuni o tutti i parametri si parla di INFERENZA PARAMETRICA

Nel caso della media campionaria sommo tutte le variabili del campione e divido per il numero degli elementi del campione. Essendo una funzione delle variabili del campione, la statistica è a sua volta una variabile casuale. Statistiche diverse possono servire per stimare parametri diversi (per esempio con una gaussiana, voglio stimare mu uso la media campionaria, se voglio stimare sigma quadro uso la varianza campionaria).

VARIANZA CAMPIONARIA

$$S^2 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})^2}{n-1}$$