Statistica descrittiva

indici

indici (o misure) di posizione

media campionaria di n osservazioni $x_1, x_2, ..., x_n$

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

per k campioni x_i ripetuti ciascuno con frequenza f_i

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k} x_i f_i$$

proprietà

Posto $y_i=a x_i+b$: $\bar{y}=a\bar{x}$

mediana m di n osservazioni $x_1 \leq x_2 \leq ... \leq x_n$

se n è dispari: $m=x_{(n+1)/2}$

se n è pari: $m=\frac{x_{n/2}+x_{(n/2)+1}}{2}$

moda

punto di massimo della distribuzione di frequenza
una distribuzione con un solo punto di massimo è detta unimodale
una distribuzione con più punti di massimo è detta plurimodale

indici di dispersione

varianza di n osservazioni $x_1, x_2, ..., x_n$

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

per k campioni x_i ripetuti ciascuno con frequenza f_i

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{k}(x_i-\bar{x})^2 f_i=(\frac{1}{n}\sum_{i=1}^{k}x_i^2 f_i)-(\bar{x})^2$$

proprietà

posto $y_i=a x_i+b$: $\sigma_y^2=a^2\sigma_x^2$

deviazione standard o scarto quadratico medio

$$\sigma=\sqrt{\sigma^2}$$

range di n osservazioni $x_1 \leq x_2 \leq ... \leq x_n$

differenza tra massima e minima osservazione

$$range=x_n-x_1$$

p-esimo quantile (o 100p-esimo percentile) di di n osservazioni $x_1 \leq x_2 \leq ... \leq x_n$

$p\in\mathbb{R}(0,1)$, si considera il numero np

se np non è intero: k è l'intero successivo , $Q_p=x_k$

se np è intero: $k = np$, $Q_p=\frac{x_k+x_{k+1}}{2}$

quartili

Q_1 primo quartile: quantile per $p = 0.25$
Q_2 secondo quartile: quantile per $p = 0.5$ (= mediana)
Q_3 terzo quartile: quantile per $p = 0.75$

differenza interquartile (IQR – InterQuartile Range)

$$IQR=Q_3-Q_1$$

indici di forma

coefficiente di asimmetria (skewness)

$$sk=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3$$

se vale zero indica che la distribuzione è simmetrica rispetto alla media
se positivo denota una coda verso destra
se negativo denota una coda verso sinistra

coefficiente di curtosi

$$curt=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4$$

misura quanto la distribuzione è appuntita

correlazioni

covarianza

di n osservazioni congiunte di 2 variabili $\{(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)\}$:

$$\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=(\frac{1}{n}\sum_{i=1}^{n}x_i y_i)-\bar{x}\,\bar{y}$$

se $\sigma_{xy}>0$ x e y sono direttamente correlate: a valori grandi (piccoli) di x corrispondo valori grandi (piccoli) di y;
se $\sigma_{xy}<0$ x e y sono inversamente correlate: a valori grandi (piccoli) di x corrispondo valori piccoli (grandi) di y;
se $\sigma_{xy}=0$ x e y sono incorrelate;

coefficiente di correlazione

$$\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \ ; \ -1\leqslant\rho_{xy}\leqslant 1$$

indice normalizzato, adimensionale ed invariante per trasformazioni lineari delle variabili

regressione lineare

retta $y=\hat{a}x+\hat{b}$ che meglio approssima la nuvola di punti (x_i, y_i)

$$\hat{a}=\frac{\sigma_{xy}}{\sigma_x^2} \ ; \ \hat{b}=\bar{y}-\bar{x}\frac{\sigma_{xy}}{\sigma_x^2}$$

valori stimati

$$\hat{y}_i=\hat{a}x_i+\hat{b}$$

rappresentano i valori stimati di y a partire dalla retta di regressione lineare

residui

$$r_i=y_i-\hat{y}_i$$

differenza tra i valori reali e stimati

valore previsto

$$\hat{y}_0=\hat{a}x_0+\hat{b}$$

x_0 è un valore diverso dai valori x_i già osservati

cambiamento di scala

$$\log(y)=\hat{a}\log(x)+\hat{b}$$
$$y=e^{\hat{b}}x^{\hat{a}}$$

devianza totale

$$DEV_{TOT}=DEV_{REG}+DEV_{RES}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$DEV_{REG}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \ ; \ DEV_{RES}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

coefficiente di determinazione

$$R^2=\frac{DEV_{REG}}{DEV_{TOT}}=1-\frac{DEV_{RES}}{DEV_{TOT}}=\frac{\sigma_{\hat{y}}^2}{\sigma_y^2} \ ; \ 0\leq R^2\leq 1$$

tanto più esso si avvicina ad uno tanto più la funzione di regressione trovata è buona.

Probabilità

definizioni

eventi elementari

tutti i possibili esiti di un esperimento aleatorio

evento

ogni sottoinsieme di uno spazio campionario discreto Ω

spazio campionario

insieme di tutti gli eventi elementari; può essere:

discreto

se gli elementi sono un numero finito o un'infinità numerabile

$P(\{\omega_k\})=p_k$

continuo

se è più numeroso (ad esempio: tutti i numeri reali in un certo intervallo)

linguaggio

insiemi	*eventi*
Ω , intero spazio campionario	evento certo
$\emptyset$, insieme vuoto	evento impossibile
insieme A	l'evento si verifica
insieme $\bar{A}$ complementare di A	l'evento non si verifica
$A\cup B$, (unione)	si verifica almeno uno dei due eventi
$A\cap B$, (intersezione)	gli eventi si verificano simultaneamente
$A\setminus B$, (sottrazione = $A\cap\bar{B}$)	si verifica A e non si verifica B
$A\cap B=\emptyset$, eventi disgiunti	gli eventi sono incompatibili
$B\subseteq A$ (B incluso in A)	B implica A

proprietà eventi A, B, C sottoinsiemi di Ω

$A\cup A=A$

$A\cap A=A$

$A\cup B=B\cup A$

$A\cap B=B\cap A$

$A\cup(B\cup C)=(A\cup C)\cup C$

$A\cap(B\cap C)=(A\cap C)\cap C$

$A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$

$A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$

$A\cup\emptyset=A$

$A\cap\emptyset=\emptyset$

$A\cup\Omega=\Omega$

$A\cap\Omega=A$

$A\cup\bar{A}=\Omega$

$A\cap\bar{A}=\emptyset$

$(\overline{A\cup B})=\bar{A}\cap\bar{B}$

$(\overline{A\cap B})=\bar{A}\cup\bar{B}$

$(\bar{\bar{A}})=A$

probabilità su Ω

$P:P(\Omega)\rightarrow[0,1]$

proprietà

$P(\Omega)=1$

$P(\emptyset)=0$

$P(\bar{A})=1-P(A)$

$P(A\cup B)=P(A)+P(B)-P(A\cap B)$

$P\left(\cup_{n=1}^{\infty}A_n\right)=\sum_{n=1}^{\infty}P(A_n)\quad,\ con\ A_i\cap A_j=\emptyset\ se\ i\neq j$

probabilità classica

la probabilità di un evento è il rapporto dei casi favorevoli ed il numero dei casi possibili

posto Ω di N elementi ω_k (k = 1, 2, .., N) e

$P(\{\omega_k\})=p\ ,\ \ (eventi\ elementari\ equiprobabili)$, A evento qualunque

$$P(A)=\sum_{\omega_k\in A}P(\{\omega_k\})=p|A|=\frac{|A|}{N}=\frac{|A|}{|\Omega|}$$

$|A|\ \grave{e}\ il\ numero\ di\ elementi\ di\ A$

permutazione di n oggetti

è ogni allineamento di n oggetti distinti in n caselle

$P_n=n!=n(n-1)(n-2)\cdots3\cdot2$

proprietà di n! (n fattoriale)

$0!=1$

$\frac{n!}{n}=(n-1)!$

$\frac{n!}{m!}=n(n-1)(n-2)\cdots(m+1)\ ,\ \ con\ m<n$

disposizione di n oggetti in k posti

è ogni allineamento di k oggetti scelti tra n oggetti distinti in k posti

$D_{n,k}=n(n-1)(n-2)\cdots(n-k+1)\ ,\ \ con\ 1\leq k\leq n$

$D_{n,n}=P_n=n!$

disposizione con ripetizione di n oggetti in k posti

è ogni allineamento di k oggetti scelti tra n oggetti e ripetibili, in k posti

$D_{n,k}^{*}=n^k\ ,\ \ con\ k\geq1$

combinazione di n oggetti di classe k

è ogni sottoinsieme di k elementi dell'insieme di n oggetti
(modi per scegliere k oggetti tra n)

$$C_{n,k}=\frac{D_{n,k}}{P_k}=\binom{n}{k}=\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\ ,$$

$$con\ n\geq1\ ;\ 0\leq k\leq n$$

coefficiente Binomiale

$$\binom{n}{k}=\binom{n}{n-k}=C_{n,k}\ ;\ \binom{n}{1}=n\ ;\ \binom{n}{0}=\binom{n}{n}=1$$

combinazione con ripetizione di k oggetti scelti fra n

ogni gruppo formato di k oggetti scelti fra n, che possono essere ripetuti
(modi per disporre k oggetti uguali in n posti)

$$C_{n,k}^{*}=\binom{n+k-1}{k}=\binom{n+k-1}{n-1}$$

permutazione con ripetizione di n oggetti uguali fra loro a gruppi

(allineamento in n posti di n oggetti)

$$P_{k_1,k_2,\ldots,k_r}^{*}=\frac{n!}{k_1!k_2!\ldots k_r!}$$

probabilità condizionata

probabilità dell'evento A, condizionata a B

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

proprietà

$$P(A\cap B)=P(B\cap A)=P(A|B)PB=P(B|A)P(A)$$
$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$
$$P(\bar{A}|B)=1-P(A|B)$$

probabilità totali

$$P(A)=\sum_{j=1}^{n} P(A|B_j)\cdot P(B_j)\ ,$$
$$con\ \cup_{j=1}^{n} B_j=\Omega\ ,\ B_i\cap B_j=\varnothing\ per\ i\neq j\ ,\ P(B_j)\neq 0\ per\ ogni\ j$$

caso notevole:

$$P(A)=P(A|B)P(B)+P(A|\bar{B})P(\bar{B})\ ,$$
$$con\{B,\bar{B}\}\ partizione\ di\ \Omega$$

formula di Bayes

$$P(B_k|A)=\frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{n} P(A|B_j)\cdot P(B_j)}\ ,\quad per\ ogni\ k$$

indipendenza di eventi

eventi A, B indipendenti

lo sono se soddisfano una delle seguenti condizioni

$$P(A\cap B)=P(A)\cdot P(B)$$
$$P(A|B)=P(A)$$
$$P(B|A)=P(B)$$

famiglia di eventi indipendenti

n eventi A_1, A_2, ..., A_n costituiscono una famiglia di eventi indipendenti se per ogni sottofamiglia di r eventi ($2\leq r\leq n$), la probabilità di intersezione di questi r eventi è uguale al prodotto delle probabilità di ciascuno di essi:

$$P(A_i\cap A_j)=P(A_i)P(A_j)\ ,\quad per\ ogni\ coppia\ di\ indici\ i\neq j$$
$$P(A_i\cap A_j\cap\ldots\cap A_n)=P(A_i)P(A_j)\ldots P(A_n)$$

data una famiglia di eventi indipendenti, anche sostituendo alcuni A_i con i complementari $\bar{A}_i$, rimane una famiglia di eventi indipendenti.

Affidabilità di un sistema

componenti in serie

il sistema funziona se e solo se funzionano tutti i componenti

affidabilità (probabilità che il sistema funzioni)

$$a=a_1\cdot a_2\cdots a_n$$

componenti in parallelo

il sistema funziona se e solo se funziona almeno un componente

affidabilità (probabilità che il sistema funzioni)

$$a=1-(1-a_1)\cdot(1-a_2)\cdots(1-a_n)$$

variabili aleatorie e modelli probabilistici

variabili aleatorie

variabile aleatoria (v.a.) discreta

è una qualunque funzione:

$$X:\Omega\to\mathbb{R}$$

$(X\in I)$, $con\ I\subseteq\mathbb{R}$ è un'abbreviazione di $\{\omega\in\Omega: X(\omega)\in I\}$

legge (o distribuzione) di una v.a.

applicazione che associa ad ogni intervallo $I\subseteq\mathbb{R}$ il numero:

$$P(X\in I)=P\{\omega\in\Omega: X(\omega)\in I\}$$

densità discreta di X

funzione che ad ogni valore assunto da X associa la probabilità che X assuma quel valore

$$p_X(x_k)=P(X=x_k)$$

proprietà

probabilità dell'evento $X\in I$:

$$P(X\in I)=\sum_{x_k\in I} p_X(x_k)\ ,\quad purché\ la\ serie\ converga$$

v.a. indipendenti

se scelti n intervalli $I_1, I_2, \ldots, I_n\subseteq\mathbb{R}$ si ha

$$P(X_1\in I_1, X_2\in I_2, \ldots, X_n\in I_n)=P(X_1\in I_1)\cdot P(X_2\in I_2)\cdots P(X_n\in I_n)$$

valore atteso, o media , o speranza matematica

$$\mu_X=EX=\sum_k x_k p_X(x_k)\ ,\quad per\ X\ discreta$$
$$\mu_X=EX=\int_{\mathbb{R}} t\cdot f_X(t)\,dt\ ,\quad per\ X\ continua$$

proprietà

$$E(aX+b)=a(EX)+b,\quad con\ a,b\in\mathbb{R}$$
$$E(X_1+X_2+\ldots+X_n)=EX_1+EX_2+\ldots+EX_n$$
$$E(X_1\cdot X_2\cdots X_n)=EX_1\cdot EX_2\cdots EX_n\ ,$$
$$con\ X_1, X_2, \ldots, X_n\ v.a.\ indipendenti$$
$$Ef(X)=\sum_k f(x_k)p_X(x_k)\ ,\quad purché\ la\ serie\ converga$$
$$E(aX_1+b)=aEX_1+b\ ,\ per\ ogni\ a,b\in\mathbb{R}\ (per\ v.a.\ continue)$$
$$E(g(X_1))=\int_{\mathbb{R}} g(t)f_{X_1}(t)\,dt\ ,\ per\ g:\mathbb{R}\to\mathbb{R}\ (per\ v.a.\ continue)$$

varianza

X v.a. discreta:

$$\sigma_X^2=VarX=E((X-EX)^2)=E(X^2)-(EX)^2$$

X v.a. continua:

$$\sigma_X^2=VarX=E(X^2)-(EX)^2=\int_{\mathbb{R}} t^2 f_X(t)\,dt-\left(\int_{\mathbb{R}} t\, f_X(t)\,dt\right)^2$$

proprietà

$$VarX\geq 0$$
$$VarX=E(X^2)-(EX)^2$$
$$Var(c)=0\ ,\quad per\ ogni\ costante\ c$$
$$Var(aX+b)=a^2 VarX\ ,\quad per\ ogni\ a,b\in\mathbb{R}$$
$$VarX=\sum_k (x_k-EX)^2 p_X(x_k)=\left(\sum_k x_k^2 p_X(x_k)\right)-(EX)^2$$
$$Var(X_1+X_2+\ldots+X_n)=VarX_1+VarX_2+\ldots+VarX_n\ ,$$
$$con\ X_i\ indipendenti$$

deviazione standard o scarto quadratico medio

$$\sigma_X=\sqrt{\sigma_X^2}=\sqrt{VarX}$$

covarianza

$$Cov(X,Y)=E((X-EX)\cdot(Y-EY))=E(XY)-EX\cdot EY\ ,$$
$$con\ X, Y\ v.a.\ con\ varianza\ finita$$

proprietà

$Cov(X,X)=VarX$
$Cov(X,c)=0 \;\; , \;\; per\,ogni\,costante\,c$
$Cov(X,Y)=Cov(Y,X)$
$Cov(X+Y,Z)=Cov(X,Z)+Cov(Y,Z)$
$Cov(Y,Y+Z)=Cov(X,Y)+Cov(X,Z)$
$Cov(aX,Y)=aCov(X,Y)$
$Cov(X,aY)=aCov(X,Y)$
$Var(X+Y)=VarX+VarY+2Cov(X,Y)$
$|Cov(X,Y)|\le\sqrt{VarX\cdot VarY} \quad (dis.\,Cauchy-Swartz)$

correlazione

due v.a. con varianza finita si dicono **incorrelate** se:
$Cov(X,Y)=0$
in tal caso:
$Var(X+Y)=Var(X)+Var(Y)$

coefficiente di correlazione di X, Y

$$\rho_{XY}\equiv\frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y}\equiv\frac{Cov(X,Y)}{\sqrt{VarX\cdot VarY}} \quad , \quad dove\,-1\le\rho_{XY}\le 1$$

se ρ_{XY} è vicino a zero: X e Y sono quasi indipendenti
se ρ_{XY} è positivo: ad X grande corrisponderà in genere una Y grande
se ρ_{XY} è negativo: ad X grande corrisponderà in genere una Y piccola
se $\rho_{XY}=\pm 1$ le v.a. sono una funzione lineare dell'altra: $Y=aX+b$

standardizzata di X

è una v.a. ottenuta da una v.a. X con media e varianza finite:

$$X^*=\frac{X-\mu_X}{\sigma_X}$$

$EX^*=0 \;\; ; \;\; Var\,X^*=1$

disuguaglianza di Cebicev

sia X una v.a. di valore atteso μ_X e varianza σ_X^2 finite, allora per ogni $\delta>0$:

$$P(|X-\mu_X|\ge\delta\sigma_X)\le\frac{1}{\delta^2} \quad \text{, ovvero}$$

$$P(|X-\mu_X|<\delta\sigma_X)=P(\mu_X-\delta\sigma_X<X<\mu_X+\delta\sigma_X)\ge 1-\frac{1}{\delta^2}$$

processo di Bernoulli

sequenza di esperimenti di Bernoulli indipendenti di uguale parametro p

esperimento bernulliano o prova di Bernoulli

è un esperimento aleatorio che può avere solo due esiti possibili:

- successo : con probabilità p
- insuccesso : con probabilità (1-p)

p è il parametro della prova di Bernoulli

processo di Bernoulli limitato

il numero di prove è finito

bernulliana di parametro p

$X\sim B(p)$
descrive l'esito di ogni prova di Bernoulli
$p_X(1)=p \;\; ; \;\; p_X(0)=1-p$
$EX=p \;\; ; \;\; VarX=p(1-p)$

la probabilità di ottenere, in n prove, una particolare sequenza di k successi e (n-k) insuccessi è:
$p^k(1-p)^{n-k}$
la probabilità di ottenere, in n prove, almeno un successo è:
$1-(1-p)^n$

Binomiale di parametri n e p

$X\sim B(n,p)$
conta il numero complessivo di successi ottenuti in n prove (estrazione con reimissione)

$$p_X(k)=\binom{n}{k}p^k(1-p)^{n-k} \;\; , \;\; k=0,1,2,\ldots,n$$

$EX=np \;\; ; \;\; VarX=np(1-p)$

$$sk(X)=\frac{1-2p}{\sqrt{np(1-np)}} \;\; ; \;\; curt(X)=3+\frac{1-6p(1-p)}{np(1-p)}$$

il numero di oggetti di tipo K che si trovano in un campione di n oggetti estratti con reimmissione da un insieme di N oggetti che contiene K oggetti di un tipo e (N-K) oggetti di un'altro è:

$$X\sim B(n,\frac{K}{N})$$

processo di Bernoulli illimitato

sequenza infinita di prove

Binomiale negativa di parametri -n e p

$X\sim B(-n,p)$
conta il numero di insuccessi che si ottengono prima di ottenere n successi

$$p_X(k)=\binom{n+k-1}{k}p^n(1-p)^k \;\; , \;\; k=0,1,2,\ldots$$

$$EX=n\frac{1-p}{p} \;\; ; \;\; VarX=n\frac{1-p}{p^2}$$

il numero Y di prove necessarie per ottenere n successi:

$$P(Y=k)=P(X+n=k)=P(X=k-n)=\binom{k-1}{k-n}p^n(1-p)^{k-n} \;,\;\; per\,k=n,n+1,n+2,\ldots$$

Geometrica di parametro p

$X\sim G(p)$
conta il numero di prove necessarie per ottenere il primo successo
$p_X(k)=p(1-p)^{k-1} \;\; , \;\; per\,k=1,2,3,\ldots$

$$EX=\frac{1}{p} \;\; ; \;\; VarX=\frac{1-p}{p^2}$$

Geometrica traslata di parametro p

$X\sim G'(p)$
conta il numero di insuccessi prima del primo successo
$p_X(k)=p(1-p)^k \;\; , \;\; per\,k=0,1,2,\ldots$

$$EX=\frac{1-p}{p} \;\; ; \;\; VarX=\frac{1-p}{p^2}$$

Ipergeometrica di parametri (N, K, n)

$X\sim G(N,K,n) \;\; , \;\; con\,N\ge k\,;\,N\ge n$
conta il numero di oggetti di tipo K che si trovano in un campione di n oggetti estratti senza reimmissione da un insieme di N oggetti che contiene K oggetti di un tipo e (N-K) oggetti di un altro.

$$p_X(k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \;\; , \;\; con\,0\le k\le n \;;\; k\le K \;;\; (n-k)\le(N-K)$$

$$EX=n\frac{K}{N} \;\; ; \;\; VarX=n\frac{K}{N}\left(1-\frac{K}{N}\right)\left(\frac{N-n}{N-1}\right)$$

approssimazione Binomiale

per N (e quindi K) molto grandi (N > 10n) è come se estraessimo con reimissione:

$$X\sim G(N,K,n)\rightarrow X\sim B(n,\frac{K}{N}) \;\; , \;\; per\,N\rightarrow\infty$$

$$p_X(k) \rightarrow \binom{n}{k} p^k (1-p)^{n-k} \quad , \quad per\ N \rightarrow \infty \quad , \quad p = \frac{K}{N}$$

$$EX = np \quad ; \quad VarX = np(1-p)\left(\frac{N-n}{N-1}\right)$$

$\left(\frac{N-n}{N-1}\right)$ **(fattore di correzione per la popolazione finita** (< 1))

Poisson di parametro λ > 0

$$Y \sim P_0(\lambda) \quad , \quad con\ \lambda > 0$$

permette di descrivere quantitativamente situazioni in cui non abbiamo accesso ai valori di N e p, ma possediamo un unica informazione numerica: il parametro λ (numero medio di arrivi)

$$p_Y(k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad , \quad per\ k = 0,1,2,\ldots$$

$$EY = \lambda \quad ; \quad VarY = \lambda$$

$$sk(X) = \frac{1}{\sqrt{\lambda}} \quad ; \quad curt(X) = 3 + \frac{1}{\lambda}$$

proprietà

se $X_i \sim P_0(\lambda_i)$ allora:

$$X_1 + X_2 + \ldots + X_n \sim P_0(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$$

approssimazione della Binomiale

per N molto grande e p molto piccolo:

$$X \sim B(N, p) \rightarrow Y \sim P_0(Np) \quad , \quad P(X=k) \rightarrow P(Y=k)$$

processo Poisson di intensità ν

permette di calcolare probabilità di eventi che accadono in un certo intervallo di tempo diverso da quello su cui abbiamo informazioni di partenza;

posto $\lambda = \nu t$ con ν numero medio di arrivi nell'unità di tempo, il numero X_t di arrivi nell'intervallo di tempo $[0, t]$ è dato da

$$X_t \sim P_0(\nu t)$$

$$p_{X_t}(k) = e^{-\nu t} \frac{(\nu t)^k}{k!} \quad , \quad per\ k = 0,1,2,\ldots$$

$$EX_t = \nu t \quad ; \quad VarX_t = \nu t$$

variabili aleatorie continue

densità continua f_x

determina la legge della v.a. continua X;
è una densità di probabilità

$$P(X \in I) \equiv \int_I f_x(t)\,dt \quad , \quad con\ I \subseteq \mathbb{R}$$

$$f_x : \mathbb{R} \rightarrow \mathbb{R} \quad ; \quad f_x(t) \geq 0 \quad , \quad per\ ogni\ t \in \mathbb{R} \quad ; \quad \int_{\mathbb{R}} f_x(t)\,dt = 1$$

proprietà

$P(X=t) = 0$, $per\ ogni\ t \in \mathbb{R}$ (la probabilità che assuma un valore fissato è nulla (integrale di un punto))

$$P(X \leq a) = P(X < a)$$

$$P(a \leq X < b) = P(a < X < b)$$

esempi di densità continue

densità uniforme

$$f_X(t) = \frac{1}{b-a} I_{(a,b)}(t) \quad , \quad a, b \in \mathbb{R}, a < b$$

con $\begin{matrix} I_{(a,b)}(t) = 1 \quad , \quad per\ t \in (a,b) \\ I_{(a,b)}(t) = 0 \quad , \quad per\ t \notin (a,b) \end{matrix}$ (funzione indicatrice)

$$P(X \in J) = \int_J \frac{1}{b-a} I_{(a,b)}(t)\,dt = \frac{1}{b-a} |(a,b) \cap J|$$

densità di Cauchy

$$f_X(t) = \frac{1}{\pi(1+t^2)}$$

$$P(a < X < b) = \int_a^b \frac{1}{\pi(1+t^2)} = 1/\pi(\arctan(b) - \arctan(a))$$

densità Normale Standard

"curva a campana" di Gauss, o curva degli errori

$$f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt$$

funzione di ripartizione di X (f.d.r.)

equivale alla densità discreta nel caso continuo

$$F_X(t) : \mathbb{R} \rightarrow [0,1]$$

$$F_X(t) = P(X \leq t) \quad , \quad per\ ogni\ t \in \mathbb{R}$$

$$F_X(t) = \int_{-\infty}^{t} f_X(y)\,dy \quad , \quad per\ X\ continua$$

$$F_X(t) = \sum_{x_k \leq t} p_X(x_k) \quad , \quad per\ X\ discreta$$

proprietà

$$se\ t_1 < t_2\ ,\ (X \leq t_1) \subseteq (X \leq t_2)\ ,\ P(X \leq t_1) \leq P(X \leq t2)\ ,\ (F_X(t)\ è\ monotona\ crescente)$$

$$F_X(t) \rightarrow 1 \quad per\ t \rightarrow +\infty$$

$$F_X(t) \rightarrow 0 \quad per\ t \rightarrow -\infty$$

$$F_X(b) - F_X(a) = P(X \leq b) - P(X \leq a) = P(a < X \leq b) \quad , \quad con\ a, b \in \mathbb{R} \quad , \quad a < b$$

la f.d.r. di una v.a. continua è sempre una funzione continua nei punti in cui la densità è continua; in questi punti è derivabile:

$$F'_X(t) = f_X(t)$$

quantile α-esimo (q_α)

$$P(X \leq q_\alpha) = \alpha \quad , \quad con\ q_\alpha \in (a,b) \subseteq \mathbb{R} \quad , \quad \alpha \in (0,1)$$

variabili aleatorie legate al processo di Poisson

legge Esponenziale di parametro ν

$$Y \sim Esp(\nu) \quad , \quad con\ \nu > 0$$

misura l'istante del primo arrivo in un processo di Poisson X_t di intensità ν, o il tempo di attesa tra due arrivi successivi;
è l'unico modello adeguato a rappresentare il tempo di vita di un apparecchio non soggetto ad usura

$$F_Y(t) = 1 - e^{-\nu t} \quad , \quad per\ t > 0$$

$$F_Y(t) = 0 \quad , \quad per\ t \leq 0$$

$$f_Y(t) = \nu e^{-\nu t} \quad , \quad per\ t > 0$$

$$f_Y(t) = 0 \quad , \quad per\ t < 0$$

$$E(Y) = \frac{1}{\nu} \quad ; \quad Var\ Y = \frac{1}{\nu^2}$$

$$sk(X) = 2 \quad ; \quad curt(X) = 9$$

stimatore non distorto per legge Esponenziale

$$U = T\frac{n-1}{n} = \frac{n-1}{\sum_{i=1}^{n} X_i}$$

$$\hat{\nu}=\frac{n-1}{\sum_{i=1}^{n} X_i}=\frac{n-1}{n}\frac{1}{\bar{X}_n} \quad , \quad (stima\,di\,\nu)$$

legge Gamma di parametri n (intero positivo) e ν (intero positivo)

$Y \sim \Gamma(n,\nu)$

misura l'istante dell'ennesimo arrivo in un processo di Poisson X_t di intensità ν

$$F_Y(t)=1-\sum_{k=0}^{n-1} e^{-\nu t}\frac{(\nu t)^k}{k!} \quad , \quad per\ t>0$$

$$F_Y(t)=0 \quad , \quad per\ t\le 0$$

$$f_Y(t)=\nu e^{-\nu t}\frac{(\nu t)^{n-1}}{(n-1)!}=C_{n,\nu}t^{n-1}e^{-\nu t} \quad , \quad per\ t>0 \ ,$$

$$f_Y(t)=0 \quad , \quad per\ t<0$$

$$C_{n,\nu}=\frac{\nu^n}{(n-1)!}$$

$$E(Y)=\frac{n}{\nu} \ ; \ Var\,Y=\frac{n}{\nu^2}$$

legge Gamma di parametri r e ν (reali positivi)

$Y \sim \Gamma(r,\nu)$

descrive il tempo di vita di un apparecchio la cui propensione al guasto cresce col tempo, fino al limite ν

$$f_Y(t)=C_{r,\nu}t^{r-1}e^{-\nu t} \quad , \quad per\ t>0$$

$$f_Y(t)=0 \quad , \quad per\ t<0$$

$$E(Y)=\frac{r}{\nu} \ ; \ Var\,Y=\frac{r}{\nu^2}$$

assenza di memoria

$$P(Y\ge T-t|Y\ge T)=P(Y\ge t)$$

$$P(Y\ge T+t)=P(Y\ge t)\cdot P(Y\ge T)$$

se una v.a. continua soddisfa questa proprietà, allora ha legge Esponenziale se è continua e legge Geometrica traslata se discreta

istantaneous failure rate (propensione istantanea al guasto)

$$Z(t)=\frac{f_Y(t)}{1-F_Y(t)}$$

per la legge Esponenziale:

$$Z(t)=\nu \quad , \quad per\,t>0$$

per la legge Gamma:

$$Z(t)=C_n\frac{t^{n-1}}{\sum_{k=0}^{n-1}\frac{(\nu t)^k}{k!}}=\frac{\nu^n}{(n-1)!}\frac{t^{n-1}}{\sum_{k=0}^{n-1}\frac{(\nu t)^k}{k!}}$$

densità di Weibull

utile a rappresentare il tempo di vita di un apparecchio

posto $Z(t)=c\,t^{\beta}$ si trova:

$$F_Y(t)=1-e^{\frac{-c\,t^{\beta+1}}{\beta+1}} \quad , \quad con\,\beta>-1$$

$$f_Y(t)=c\,t^{\beta}e^{\frac{-c\,t^{\beta+1}}{\beta+1}}$$

se $\beta>0$ l'apparecchio invecchia

se $-1<\beta<0$ l'apparecchio migliora col tempo

se $\beta=0$ si ritrova la legge Esponenziale

modello Normale

legge Normale standard

$Z\sim N(0,1)$

$$F_Z(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{\frac{-y^2}{2}}dy\equiv\Phi(t)$$

$$f_Z(t)=\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}\equiv\varphi(t)$$

$$E(Z)=0 \ ; \ Var\,Z=1$$

proprietà

$\Phi(-t)=1-\Phi(t) \quad , \quad (simmetria)$

calcoli con i quantili

posto z_α quantile α-esimo della legge Normale standard:

$$z_\alpha=-z_{1-\alpha}$$

$$P(Z<z_\alpha)=\alpha$$

$$P(Z>z_{1-\alpha})=\alpha$$

$$P(|Z|>z_{1-\alpha/2})=\alpha$$

$$P(|(Z)|<z_{(1+\alpha)/2})=\alpha$$

legge Normale (o gaussiana) di media μ e varianza σ^2

$X\sim N(\mu,\sigma^2)$

rappresenta bene gli errori di approssimazione

$$F_X(t)=\Phi(\frac{t-\mu}{\sigma})$$

$$f_X(t)=\frac{1}{\sigma}\varphi(\frac{t-\mu}{\sigma})=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(t-\mu)^2}{2\sigma^2}}$$

$$EX=\mu \ ; \ Var\,X=\sigma^2$$

$$sk(X)=0 \ ; \ curt(X)=3$$

la v.a. $Z=\frac{X-\mu}{\sigma}$ ha legge Normale standard

proprietà

posto $X_1\sim N(\mu_1,\sigma_1^2)$, $X_2\sim N(\mu_2,\sigma_2^2)$ indipendenti:

$$X_1+X_2\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$$

posto $a,b\in\mathbb{R}$:

$$aX_1+b\sim N(a\mu_1+b,a^2\sigma_1^2)$$

relazione tra legge Normale e legge Normale standard

$$Z\sim N(0,1) \ \Rightarrow \ \sigma Z+\mu\sim N(\mu,\sigma^2)$$

$$X\sim N(\mu,\sigma^2) \ \Rightarrow \ \frac{X-\mu}{\sigma}\sim N(0,1)$$

errori

$Y=misura\,di\,una\,grandezza\,fisica$

$\nu=valore\,vero$

$X=errore\,di\,misura$

$\mu=errore\,sistematico$

$E_c=errore\,casuale$

$\sigma^2=inacuratezza\,della\,misura$

$X\sim N(\mu,\sigma^2)\ ,\ X=\mu+E_c$

$E_c\sim N(0,\sigma^2)$

$E(E_c)=0$

$EY=\nu+\mu$

media campionaria

se $X_i\sim N(\mu,\sigma^2)$ sono v.a. indipendenti ed identicamente distribuite (i.i.d.):

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$$

$$E\,\bar{X}_n = \mu \;\; ; \;\; Var\,\bar{X}_n = \frac{\sigma^2}{n}$$

media campionaria standardizzata

$$S_n^* = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \quad , \quad n = 1,2,3,\ldots$$

teorema del limite centrale

$$P(S_n^* \le t) \to \Phi(t) \quad per\; n \to \infty \quad , \quad t \in \mathbb{R}$$

approssimazione Normale

Date X_i $v.a.i.i.d.$, $EX_i = \mu$, $VarXi = \sigma^2$ con n abbastanza grande:

$$\bar{X}_n \simeq N(\mu, \frac{\sigma^2}{n}) \quad ossia \quad P(\bar{X}_n < t) \simeq \Phi\left(\sqrt{n}\left(\frac{t-\mu}{\sigma}\right)\right)$$

$$\sum_{i=1}^{n} X_i \simeq N(n\mu, n\sigma^2) \quad ossia \quad P(\sum_{i=1}^{n} X_i < t) \simeq \Phi\left(\frac{t - n\mu}{\sqrt{n}\,\sigma}\right)$$

approssimazione Normale di Gamma per n grande:

$$Y \sim \Gamma(n, \lambda)$$

$$Y \simeq N(\frac{n}{\lambda}, \frac{n}{\lambda^2})$$

$$F_Y(t) = P(Y < t) \simeq \Phi\left(\frac{\lambda t - n}{\sqrt{n}}\right)$$

approssimazione Normale della Binomiale:

approssimazione utile in problemi di campionamento

NOTA: vale se: $np > 5 \;\; ; \;\; n(1-p) > 5$

$$Y \sim B(n, p)$$

$$Y \simeq N(np, np(1-p))$$

$$F_Y(t) = P(Y \le t) \simeq \Phi\left(\frac{t - np}{\sqrt{np(1-p)}}\right) \quad , \quad (per\; v.a.\; continua)$$

$$F_Y(k) = P(Y \le k) \simeq \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right) \quad ,$$

$$k = 0,1,2,\ldots,n \quad , \quad (per\; v.a.\; discreta)$$

momenti ed indici di forma per v.a.

momento r-esimo di X

$$\mu'_r = E(X^r)$$

$$\mu'_r = \sum_k x_k^r\, p_X(x_k) \quad , \quad per\; X\; discreta$$

$$\mu'_r = \int_{\mathbb{R}} x^r f_X(x)\, dx \quad , \quad per\; X\; continua$$

momento r-esimo centrato di X

$$\mu_r = E((X - EX)^r)$$

$$\mu_r = \sum_k (x_k - \mu)^r\, p_X(x_k) \quad , \quad con\;\; \mu = EX \quad , \quad per\; X\; discreta$$

$$\mu_r = \int_{\mathbb{R}} (x - \mu)^r f_X(x) dx \quad , \quad per\; X\; continua$$

coefficiente di asimmetria (skewness) di una v.a. X con μ'₃ finito

misura l'assimetria di X rispetto al valore atteso

$$sk(X) = \frac{\mu_3}{\mu_2^{3/2}} = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

coefficiente di curtosi di una v.a. X con μ'₄ finito

misura quanto la densità di X sia appuntita

$$curt(X) = \frac{\mu_4}{\mu_2^2} = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \quad , \quad \mu = EX \quad , \quad \sigma^2 = Var\,X$$

statistica inferenziale

campionamento e stime

definizioni

modello statistico

famiglia di leggi di v.a., dipendenti da uno o più parametri incogniti:

$$\{p_X(x; \vartheta) : \vartheta \in I\}$$

ϑ è un vettore di parametri

campione casuale di ampiezza n

estratto da una popolazione di densità $p_X(x;\vartheta)$ è una ennupla di v.a. indipendenti e identicamente distribuite (i.i.d) $(X_1, X_2, \ldots, X_n)$, ciascuna avente legge $p_X(x;\vartheta)$.

stima di parametri e stimatori

stima puntuale dei parametri

stimare il valore vero del parametro (o dei parametri) a partire dal campione casuale

stima del parametro p della popolazione bernulliana

$$\hat{p} = \bar{x}_n \quad , \quad con\; x_i\; valori\; effettivamente\; osservati$$

statistica T

è una qualsiasi v.a. T funzione del campione casuale $(X_1, X_2, \ldots, X_n)$ di ampiezza n estratto da una popolazione di legge $p_X(x, \vartheta)$:

$$T = f(X_1, X_2, \ldots, X_n) \quad , \quad con\; f : \mathbb{R}^n \to \mathbb{R}$$

stimatore del parametro ϑ

statistica che viene usata per stimare il valore del parametro ϑ

è **corretto (non distorto)** se $ET = \vartheta$ altrimenti è detto **distorto**

stima del parametro ϑ

$$\hat{\vartheta} = f(x_1, x_2, \ldots, x_n) \quad , \quad calcolato\; a\; campionamento\; eseguito$$

stimatore consistente

$$var\,T_n \to 0 \;\; per \;\; n \to \infty \quad , \quad con\, T_n\, stimatore\; corretto\; di\; \vartheta$$

valore atteso della media campionaria

$$E\,\bar{X}_n = \mu$$

varianza della media campionaria

$$Var\,\bar{X}_n = \frac{\sigma^2}{n}$$

legge dei grandi numeri

$$P\{|\bar{X}_n - \mu| > \epsilon\} \to 0 \quad , \quad per\, n \to \infty$$

stime

stima di $\sigma^2 = h(\vartheta)$

$$S_n^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \quad , \quad (varianza\; campionaria)$$

a campionamento effettuato:

$$s_n^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}(\bar{x}_n)^2$$

stima popolazione Normale

$\hat{\mu} = \bar{x}_n$

$\hat{\sigma}^2 = s_n^2$

se μ è nota:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

stima popolazione Gamma

$$\hat{\lambda} = \frac{\bar{x}_n}{s_n^2} \quad ; \quad \hat{r} = \frac{\bar{x}_n^2}{s_n^2}$$

leggi

legge Chi quadro con n gradi di libertà

$$Y \sim X^2(n) \quad \equiv \quad Y \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

X_i sono v.a. indipendenti, ciascuna di legge N(0,1)

$f_Y(t) = c_n t^{\frac{n}{2}-1} e^{-\frac{t}{2}} \quad , \quad per\ t>0$

$f_Y(t) = 0 \quad , \quad per\ t<0$

$EY = n \quad ; \quad Var\,Y = 2n$

proprietà

posto $Y_1 \sim X^2(n_1)$, $Y_2 \sim X^2(n_2)$ indipendenti:

$Y_1+Y_2 \sim X^2(n_1+n_2)$

intervallo a cui una v.a. di legge Chi quadro appartiene con probabilità α:

$$P\left(X^2_{\frac{1-\alpha}{2}}(n) < Y < X^2_{\frac{1+\alpha}{2}}(n)\right) = \alpha$$

approssimazione Normale di Chi quadro per n grande

$X^2(n) \simeq N(n, 2n) \quad , \quad per\ n\ grande$

$$P(Y<t) \simeq \Phi\left(\frac{t-n}{\sqrt{2n}}\right)$$

$X^2_\alpha(n) \simeq z_\alpha\sqrt{2n}+n$

approssimazioni

Sia $X_i, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione di legge $N(\mu, \sigma^2)$, allora:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right) \sim X^2(n)$$

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}_n}{\sigma}\right) \sim X^2(n-1)$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim X^2(n-1)$$

$S_n^2\ e\ \bar{X}_n\ sono\ tra\ loro\ indipendenti$

legge t di student a n gradi di libertà

$$T \sim t(n) \quad ; \quad con\ T = \frac{Z}{\sqrt{Y/n}} \quad , \quad Z \sim N(0,1) \quad , \quad Y \sim X^2(n)$$

$$f_T(t) = c_n\left(1+\frac{t^2}{n}\right)^{-\frac{(n+1)}{2}} \quad , \quad per\ t \in \mathbb{R}$$

$ET = 0 \quad , \quad (tranne\ per\ n=1\ per\ cui\ non\ esiste\ finito)$

per $t \to \infty$ la t di student tende alla Normale standard

approssimazioni

Sia $X_i, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione di legge $N(\mu, \sigma^2)$, allora:

$$\frac{\bar{X}_n-\mu}{\sqrt{S_n^2/n}} \sim t(n-1)$$

calcoli con i quantili

posto $t_\alpha(n)$ quantile α-esimo della legge t(n):

$P(T<t_\alpha(n)) = \alpha$

$P(T>t_{1-\alpha}(n)) = \alpha$

$P(|T|>t_{1-\alpha/2}(n)) = \alpha$

$P(|(T)|<t_{(1+\alpha)/2}(n)) = \alpha$

$t_{\frac{1-\alpha}{2}}(n-1) \simeq z_{\frac{1+\alpha}{2}} \quad , \quad approssimazione\ per\ n>120$

approssimazione di quantili tramite interpolazione lineare

$y = mx+q \quad ,$

$equazione\ della\ retta\ che\ passa\ per\ i\ punti\ \{q_1, t_\alpha(q_1)\}, \{q_2, t_\alpha(q_2)\}$

$$t_\alpha(x) = t_\alpha(q_1) - \frac{t_\alpha(q_2)-t_\alpha(q_1)}{q_2-q_1}(x-q_1) \quad , \quad con\ q_1<x<q_2$$

legge di fisher con m e n gradi di libertà

$$X \sim F(m,n) \quad ; \quad con\ X = \frac{U/m}{V/n} \quad , \quad U \sim X^2(m) \quad , \quad V \sim X^2(n)$$

proprietà

$\frac{1}{X} \sim F(n,m)$

$P(X<F_\alpha(m,n)) = \alpha$

$$P\left(\frac{1}{X} < \frac{1}{F_\alpha(m,n)}\right) = 1-\alpha$$

$$\frac{1}{F_\alpha(m,n)} = F_{1-\alpha}(n,m)$$

$$\frac{S_1^2}{S_2^2} = F(m-1, n-1)$$

intervallo di confidenza al livello del 100α% per h(ϑ)

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione di densità $f(x;\vartheta)$; siano $T_1 = t_1(X_1, X_2, \ldots, X_n)$, $T_2 = t_2(X_1, X_2, \ldots, X_n)$ due statistiche, e sia $h(\vartheta)$ una funzione del parametro che si vuole stimare; fissato un numero $\alpha \in (0,1)$, l'intervallo aleatorio (T_1, T_2) si dice intervallo di confidenza al 100α% per h(ϑ) se:

$P^\vartheta(T_1 < h(\vartheta) < T_2) = \alpha$

a campionamento eseguito l'intervallo (t_1,t_2) si dice "**calcolato al campione**";

h(ϑ) appartiene all'intervallo (t_1, t_2) con una confidenza del 100α%;

t_1 e t_2 sono detti **limiti di confidenza**

intervallo di confidenza per la media

(di una popolazione Normale o popolazione qualsiasi con n grande $(n \geq 30)$)

$$\hat{\mu} = \bar{X}_n \pm z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}} = \bar{X}_n \pm E \quad , \quad (con\ varianza\ nota)$$

$$\hat{\mu} = \bar{X}_n \pm t_{\frac{(1+\alpha)}{2}}(n-1)\sqrt{\frac{S_n^2}{n}} \quad , \quad (con\ varianza\ incognita)$$

stima dell'ampiezza per limitare l'errore E_0

$$n=t^2_{(1+\alpha)/2}(n-1)\frac{\sigma^2}{E_0^2} \quad , \quad (con\ varianza\ nota)$$

intervallo di confidenza per la frequenza p

valido per una popolazione bernoulliana e per grandi campioni $(n \geq 30)$

$$\hat{p}=\bar{X}_n \pm z_{(1+\alpha)/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \quad ; \quad se: n\,\bar{x}_n > 5 \quad , \quad n(1-\bar{x}_n) > 5$$

stima dell'ampiezza per limitare l'errore E_0

$$n=\left(\frac{z_{(1+\alpha)/2}}{2E_0}\right)^2$$

E_0 corrisponde a metà dell'intervallo di confidenza.

test di ipotesi

ipotesi statistica

è un'asserzione sul valore vero di un parametro incognito;
si dice **semplice** se specifica completamente il valore del parametro, altrimenti si dice **composta**

ipotesi nulla H_0

$$H_0 : \vartheta \in \Theta_0$$

ipotesi che si ritiene vera "fino a prova contraria";
rifiuteremo H_0 solo se i dati campionari forniranno una forte evidenza statistica contro di essa

ipotesi alternativa H_1

$$H_1 : \vartheta \notin \Theta_0$$

ipotesi vera solo se H_0 è falsa

errore di tipo I

rifiutiamo H0 quando è vera;
questo è considerato l'errore più grave

errore di tipo II

accettiamo H0 quando è falsa

regione critica o regione di rifiuto

è l'insieme R dei possibili risultati campionari che portano a rifiutare H_0 data la **regola di decisione**: $si\ rifiuti\ H_0\ se\ T(X_1, X_2, \ldots, X_n) \in I$:

$$R=\{(x_1, x_2, \ldots, x_n) : T(x_1, x_2, \ldots, x_n) \in I\}$$

la probabilità di rifiutare H_0 prima del campionamento:

$$P^{\vartheta}(T(X_1, X_2, \ldots, X_n) \in I)$$

ampiezza del test (o livello di significatività)

$$\alpha = sup_{\vartheta \in \Theta_0} P^{\vartheta}(T(X_1, X_2, \ldots, X_n) \in I)$$

rappresenta la massima probabilità di rifiutare l'ipotesi nulla quando questa è vera;
va stabilito piccolo a priori prima di eseguire il campionamento

p-value

numero pari al minimo livello di significatività a cui i dati campionari consentono di rifiutare l'ipotesi nulla;
se p-value = 0 siamo praticamente certi di non sbagliare

varianza campionaria pesata

media pesata delle varianze campionarie di due campioni n, m

$$S^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)+\sum_{i=1}^{n}(Y_i-\bar{Y}_n)}{n+m-2}$$

test sulla media di una popolazione Normale di varianza nota

$$z=\frac{\bar{x}_n-\mu_0}{\sigma/\sqrt{n}}$$

H_0	H_1	*rifiutare H_0 se*
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert z\rvert > z_{1-\alpha/2}$
$\mu\leq\mu_0$	$\mu>\mu_0$	$z>z_{1-\alpha}$
$\mu\geq\mu_0$	$\mu<\mu_0$	$z<-z_{1-\alpha}$

test sulla media di una popolazione Normale di varianza incognita

$$t=\frac{\bar{x}_n-\mu_0}{s_n/\sqrt{n}}$$

H_0	H_1	*rifiutare H_0 se*
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert t\rvert > t_{1-\alpha/2}(n-1)$
$\mu\leq\mu_0$	$\mu>\mu_0$	$t>t_{1-\alpha}(n-1)$
$\mu\geq\mu_0$	$\mu<\mu_0$	$t<-t_{1-\alpha}(n-1)$

test sulla frequenza p di una popolazione bernoulliana

$$z=\frac{\bar{x}_n-p_0}{\sqrt{p_0(1-p_0)/n}}$$

H_0	H_1	*rifiutare H_0 se*
$p=p_0$	$p\neq p_0$	$\lvert z\rvert > z_{1-\alpha/2}$
$p\leq p_0$	$p>p_0$	$z>z_{1-\alpha}$
$p\geq p_0$	$p<p_0$	$z<-z_{1-\alpha}$

test sulla differenza di due medie con varianze note

estraiamo due campioni n,m da due popolazioni normali indipendenti con varianze note;
questo test non va usato quando una varianza è almeno 4 volte l'altra

$$z=\frac{\bar{X}_n-\bar{Y}_m-\delta}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{n}}}$$

H_0	H_1	*rifiutare H_0 se*
$\mu_X=\mu_Y+\delta$	$\mu_X\neq\mu_Y+\delta$	$\lvert z\rvert > z_{1-\alpha/2}$
$\mu_X\leq\mu_Y+\delta$	$\mu_X>\mu_Y+\delta$	$z>z_{1-\alpha}$
$\mu_X\geq\mu_Y+\delta$	$\mu_X<\mu_Y+\delta$	$z<-z_{1-\alpha}$

test sulla differenza di due medie con varianze incognite

estraiamo due campioni n, m da due popolazioni normali indipendenti con varianze incognite;
questo test non va usato quando una varianza è almeno 4 volte l'altra

$$t=\frac{\bar{X}_n-\bar{Y}_n-\delta}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}}$$

H_0	H_1	*rifiutare H_0 se*
$\mu_X=\mu_Y+\delta$	$\mu_X\neq\mu_Y+\delta$	$\lvert t\rvert > t_{1-\alpha/2}(n+m-2)$
$\mu_X\leq\mu_Y+\delta$	$\mu_X>\mu_Y+\delta$	$t>t_{1-\alpha}(n+m-2)$
$\mu_X\geq\mu_Y+\delta$	$\mu_X<\mu_Y+\delta$	$t<-t_{1-\alpha}(n+m-2)$

nel caso di campioni osservazioni accoppiate si considerano le differenze delle medie

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

test su due frequenze di due popolazioni bernoulliane indipendenti

estraiamo due campioni n, m da due popolazioni bernoulliane indipendenti $X \sim B(p_1)$, $Y \sim B(p_2)$;

questa procedura è valida se $\sum_{i=1}^{n} x_i > 5$; $\sum_{i=1}^{m} y_i > 5$

$$z = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}} \quad con \quad \hat{p} = \frac{n\bar{x}_n + m\bar{y}_m}{n+m}$$

H_0	H_1	rifiutare H_0 se
$p_1 = p_2$	$p_1 \neq p_2$	$\lvert z \rvert > z_{1-\alpha/2}$
$p_1 \leq p_2$	$p_1 > p_2$	$z > z_{1-\alpha}$
$p_1 \geq p_2$	$p_1 < p_2$	$z < -z_{1-\alpha}$

inferenze su una varianza

$$X^2 = \frac{(n-1)s_n^2}{\sigma_0^2}$$

H_0	H_1	rifiutare H_0 se
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$X^2 > X^2_{1-\alpha/2}(n-1)$ o $X^2 < X^2_{\alpha/2}(n-1)$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$X^2 > X^2_{1-\alpha}(n-1)$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$X^2 < X^2_{\alpha}(n-1)$

intervallo di confidenza

$$\left(\frac{(n-1)s_n^2}{X^2_{\frac{1+\alpha}{2}}(n-1)} \; , \; \frac{(n-1)s_n^2}{X^2_{\frac{1-\alpha}{2}}(n-1)} \right)$$

inferenze su due varianze

estraiamo due campioni n, m da due popolazioni normali indipendenti con medie incognite;

$$F = \frac{s_X^2}{s_Y^2}$$

H_0	H_1	rifiutare H_0 se
$\sigma_X^2 = \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$	$F > F_{1-\alpha/2}(n-1, m-1)$ $F < F_{\alpha/2}(n-1, m-1)$
$\sigma_X^2 \leq \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$F > F_{1-\alpha}(n-1, m-1)$
$\sigma_X^2 \geq \sigma_Y^2$	$\sigma_X^2 \leq \sigma_Y^2$	$F < F_{1-\alpha}(n-1, m-1)$

intervallo di confidenza

$$\left(\frac{1}{F_{\frac{1+\alpha}{2}}(n-1, m-1)} \frac{s_X^2}{s_Y^2} \; , \; \frac{1}{F_{\frac{1-\alpha}{2}}(n-1, m-1)} \frac{s_X^2}{s_Y^2} \right)$$

test Chi quadro di adattamento

ha lo scopo di verificare se certi dati empirici si adattino bene ad una distribuzione teorica assegnata;
si costruisce la seguente tabella:

classi	A_1	A_2	...	A_k	$\sum_{i=1}^{k}$
freq. rel. attese	p_1	p_2	...	p_k	1
freq. ass. attese	np_1	np_2	...	np_k	n
freq. ass. osservate	N_1	N_2	...	N_k	n
scarti quad. pesati	$\frac{(np_1-N_1)^2}{np_1}$	$\frac{(np_2-N_2)^2}{np_2}$	...	$\frac{(np_k-N_k)^2}{np_k}$	Q

le classi andranno accorpate in maniera tale che le frequenze assolute attese siano tutte maggiori o uguali a 5;
Chi quadro calcolato dal campione:

$$Q = \sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i}$$

$$Q \rightarrow X^2(k-1) \quad per \quad n \rightarrow \infty \quad , \quad con\ p_i\ assegnate\ a\ priori$$

$$Q \rightarrow X^2(k-1-r) \quad per \quad n \rightarrow \infty \ , \ con\ p_i\ calcolate\ dopo\ aver\ stimato\ r\ parametri\ incogniti$$

fissato α, si stabilisce la regola di decisione:

$si\ rifiuti\ H_0 \quad se \quad Q > X^2_{1-\alpha}(k-1-r)$ (si calcola tramite tabelle)

il p-value corrispondente al valore Q è:

$$\alpha = P(X > Q) \quad , \quad con\ X \sim X^2(k-1-r)$$

test Chi quadro di indipendenza

verifica l'indipendenza o meno di due variabili;
si costruisce una tabella di contingenza di rs classi:

	A_1	A_2	...	A_r	Tot.
B_1	n_{11}	n_{21}	...	n_{r1}	$n_{.1}$
B_2	n_{12}	n_{22}	...	n_{r1}	$n_{.2}$
...	...	...	...	...	...
B_s	n_{1s}	n_{2s}	...	n_{rs}	$n_{.s}$
Tot.	$n_{1.}$	$n_{2.}$	...	$n_{r.}$	n

si costruisce una tabella di rs classi:

	A_1	A_2	...	A_r
B_1	$\frac{n_1 \cdot n_1}{n}$	$\frac{n_2 \cdot n_1}{n}$	...	$\frac{n_r \cdot n_1}{n}$
B_2	$\frac{n_1 \cdot n_2}{n}$	$\frac{n_2 \cdot n_2}{n}$	...	$\frac{n_r \cdot n_2}{n}$
...	...	...	...	...
B_s	$\frac{n_1 \cdot n_s}{n}$	$\frac{n_2 \cdot n_s}{n}$	...	$\frac{n_r \cdot n_s}{n}$

ciascuna delle frequenze attese deve essere: $\frac{n_{i.} n_{.j}}{n} \geq 5$

si calcola il chi-quadro:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

fissato α, si stabilisce la regola di decisione:

$si\ rifiuti\ H_0 \quad se \quad Q > X^2_{1-\alpha}((r-1)(s-1))$

(si calcola tramite tabelle)

il p-value corrispondente al valore Q è:

$$\alpha = P(X > Q) \quad , \quad con\ X \sim X^2((r-1)(s-1))$$