NUMERO: 2140A- ANNO: 2017

A P P U N T I

STUDENTE: Pinna Eleonora

MATERIA: Statistica - Prof. Pellerey e Vicario

METODO DEI MINIMI QUADRATI

DOMANDE D'ESAME ORALE

- Il metodo dei quadrati è una tecnica di ottimizzazione che permette di trovare una funzione rappresentata da una curva di regressione che si avvicini ad un insieme di dati

la funzione trovata deve essere quella che minimizza la somma dei quadrati delle distanze tra i dati osservati e quelli della curva che rappresenta la funzione stessa.

oppure

- È una tecnica di ottimizzazione grazie alla quale è possibile determinare una stima $\hat{y} = b_0 + b_1 x$ della retta di regressione $Y = \beta_0 + \beta_1 x + \varepsilon$, minimizzando la somma dei quadrati degli errori, ciascuno $\varepsilon_i = y_i - \hat{y}_i$, tra i valori della variabile ciascuna Y e i valori previsti con la retta in corrispondenza del valore della variabile indipendente x_i

$$\begin{cases} b_0 = \bar{Y} - b_1 \bar{x} \\ b_1 = \dfrac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \end{cases}$$

RESIDUI

Sono la realizzazione della nostra variabile casuale, errore casuale, se è davvero casuale

errore casuale = ε Si cerca di tener effetti di errore, di tutte le variabili non considerate nel modello quindi la componente casuale arriva a rappresentare altri errori di variabile

residui: $\sum_{i=1}^{n} (y_i - \hat{y}) = 0$

Se elevo al quadrato:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SS_{tot} — var. dovuta alla regressione → SS_{reg} — variabile residua → SS_{res}

Differenza tra valore osservato e il valore predetto dal modello

Formula di Bayes

Si usa tutte le volte in cui abbiamo un esperimento da fasi successive. Viene usata per calcolare la P di una causa che ha scatenato l'evento verificato.

Sia $E_1, E_2 \dots E_n$ una collezione di eventi incompatibili ed esaustivi e sia $P[E_i] \neq 0$ per ogni i.

Qualsiasi sia $F \subseteq S$ si ha:

$$P[A_i|B] = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{i \in I} P(A_i) \cdot P(B|A_i)}$$

Definizione Probabilità classica

La probabilità di un evento E è definita come il rapporto tra il numero n_A di risultati favorevoli (cioè il numero dei risultati che determinano E) ed il numero dei risultati possibili:

$$P[E] = \frac{n_A}{n}$$ → purché tutti i risultati siano ugualmente possibili e tra loro incompatibili.

Definizione Probabilità frequentista

Si definisce come frequenza assoluta n_E di un evento E il numero delle volte in cui si è ripetuto l'esperimento nelle medesime condizioni, il rapporto

$$f_E = \frac{n_E}{N}$$

presentato l'evento favorevole detto N il numero delle volte in cui si è ripetuto ...

$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$ → valida nell'ipot. di ripetere esp. all'infinito

Probabilità assiomatica e Kolmogoroff

Una funzione di probabilità P è una funzione di insieme che ha come dominio lo spazio degli eventi e come codominio l'intervallo $[0,1]$ e che soddisfa i seguenti assiomi:

- $P(A) > 0$ sempre → $0 \leq P(A) \leq 1$
- $P(S) = 1$, S = spazio campione
- Presa la successione di eventi $A_1, A_2 \dots A_n$, tra loro mutuamente esclusivi (incompatibili), se l'unione infinita è lo spazio di eventi A, allora:

$$P[\bigcup_i^\infty E_i] = \sum_i^\infty P[E_i]$$

CONSEGUENTE

- $P[\bar{E}] = 1 - P[E]$ → $P[\emptyset] = 0$
- $P[E_1 \cup E_2 \dots \cup E_n] = \sum_i^n P[E_i]$
- $P[E \cup F] = P[E] + P[F] - P[E \cap F]$

ERRORI PRIMA E SECONDA SPECIE

Si commette errore di prima quando rifiuto l'ipotesi nulla quando quella è invece corretta.

Si commette errore di seconda specie quando si accetta l'ipotesi nulla, nel caso in cui questa sia invece falsa

L'errore di prima specie è identificato con α, quello di seconda specie con β.

PROBABILITÀ CONDIZIONATA

Sia dato lo spazio di probabilità $(S, \mathcal{A}, P)$. Sia $B \in \mathcal{A}$, tale per cui $P(B) > 0$

Dato l'evento $A \subset \mathcal{A}$, è detta "probabilità di A condizionata a B":

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

→ esprime la probabilità che si verifichi A, condizionata dal fatto che si verifichi B.

- Non sempre una prob. condiz. diminuisce.

DISTRIBUZIONE POISSON (APPROSS. BINOMIALE)

È il modello adatto per modellare il numero di difetti o non conformità che si trovano in una unità di prodotto.

È una distribuzione discreta che esprime la probabilità per il n° di eventi che si verificano successivamente ed indipendentemente in un dato intervallo di tempo, sapendo che mediamente se ne verificano λ.

$$f_X(x) = f_X(x;\lambda) = \begin{cases} e^{(-\lambda)} \cdot \frac{\lambda^x}{x!} & \text{per } x = 0, \dots n \\ 0 & \text{altrove} \end{cases}$$

È una funzione sempre > 0

$E[X] = \lambda$ e $V[X] = \lambda$, significa che la dist. di Poisson è tanto più dispersa quanto più è grande la sua media

$\sim$ Bin → Si considera una successione di prove aventi parametri n e p; se il n° delle prove tende all'infinito e la prob. di successo tende a 0, in modo che np rimanga costante, si ha:

$$\binom{n}{x} p^x (1-p)^{n-x} \to \exp(-\lambda) \cdot \frac{\lambda^x}{x!}$$

IPERGEO $\sim$ BIN

Quando abbiamo un campionamento con reimmissione, la distribuzione che modellizza il n° di successi è la Bin

Se il campionamento è senza reimmissione usiamo la dist. ipergeo.

→ se però ho N e K grandi e se $\frac{K}{N}$ è costante, posto $\frac{K}{N} = p$

Ipergeo $\sim$ Bin

MASSIMI E MINIMI

Supponiamo che X sia una variabile casuale relativa all'esperimento con funzione di ripartizione F e densità f

Ripetiamo n repliche indipendenti dell'esperimento semplice e otteniamo un campione $(X_1 \dots X_n)$, di v.c indipendenti.

Utilizzo le statistiche d'ordine per mettere in ordine i dati

$$Z = \min\{X_1 \dots X_n\}$$
$$Y = \max\{X_1 \dots X_n\}$$

- $F_Y(t) = P[Y \le t] = P[X_1 \le t \dots X_n \le t] = \prod_{i=1}^{n} F_i(t)$

  - in particolare se $F_i = F \rightarrow F_Y(t) = (F(t))^n$

    $f_Y(t) = n F^{n}(t) \cdot f(t)$

- $F_Z(t) = P[\min\{X_1 \dots X_n\} \le t] = \dots = 1 - \prod^{n} (1 - F_i(t))$

  - in particolare se $F_i = F \rightarrow 1 - F_Z(t) = (1 - F(t))^n$

CAMPIONE

Insieme di n variabili casuali indipendenti, con stessa distribuzione

$$\{X_1 \dots X_n\}$$

CONTRASTO

Si definisce contrasto una combinazione lineare delle medie dei trattamenti, in modo che i coefficienti sommati diano 0.

$C \rightarrow$ la uso come media camp.

$V[C] \rightarrow$ faccio tabella con repliche e poi moltiplico

Poi standardizzo

Legge dei grandi numeri:

Se la numerosità n delle prove tende all'infinito, la probabilità che $\bar{x}_n$ assuma valori al di fuori di $(\mu-\varepsilon, \mu+\varepsilon)$ tende a 0, qualunque sia l'ampiezza ε dell'intervallo. In formula

$$\lim_{n\to+\infty} P[|\bar{x}_n - \mu| > \varepsilon] = 0$$

Equivalentemente:

Se n tende ad infinito, la probabilità che la media campionaria $\bar{x}_n$ converga alla media comune μ delle x_i è uguale a 1, ossia per ogni $\varepsilon > 0$,

$$\lim_{n\to+\infty} P[|\bar{x}_n - \mu| < \varepsilon] = 1$$

Formula probabilità totali:

Siano $A_1, A_2 \dots A_n$ eventi incompatibili e tali che $\bigcup_{i=1}^{n} A_i = S$ e $A_i \neq 0$ $\forall$ e sia B un altro evento qualsiasi, $B \subseteq S$

Vale $P[B] = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)$

Ipotesi di Poisson:

Affinché un fenomeno in cui è in atto un conteggio possa essere modellizzato mediante la d. di Poisson, devono essere rispettate 3 ipotesi:

1) Esiste quantità λ positiva tale che la probabilità che si verifichi esattamente un evento in un processo in intervallo di ampiezza Δt sia approssimativamente uguale a $\lambda \Delta t$:

$$P[\text{esattamente 1 evento nell'intervallo di tempo } \Delta t] = \lambda \Delta t + o(\Delta t)$$

2) La probabilità che nell'intervallo Δt si verifichino più di un evento è trascurabile (è più alta) rispetto alla probabilità che se ne verifichi esattamente 1:

$$P[2 \text{ o più eventi in } \Delta t] = o(\Delta t)$$

$$P[\text{n° eventi in } \Delta t] = 1 - P[1 \text{ evento}] - P[2 \text{ o più eventi}] =$$
$$= 1 - (\lambda \cdot \Delta t + o(\Delta t) - o(\Delta t))$$

3) L'evento E_1 rappresenta il verificarsi di n_1 eventi in qualsiasi intervallo (t_1, t_2), è indipendente dall'evento E_2 rappresentante il verificarsi di n_2 eventi in un intervallo (t_3, t_4) non contenuto in (t_2, t_3):

$$P[n_1 \text{ eventi in } (t_1, t_2) \text{ e } n_2 \text{ eventi in } (t_3, t_4)] = P[E_1] \cdot P[E_2]$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 11 di 386

R^2 FORMULA E SPIEGAZIONE

$SS_{TC} = SS_{reg} + \text{varianza residua}$

$R^2 = \frac{SS_{reg}}{SS_{TC}}$ → indice di determinazione multipla

È sempre compreso $0 \le R^2 \le 1$, esprime la proporzione della variabilità totale dei dati sperimentali rispetto al loro valore medio spiegata dal modello di regressione

- $R^2 = 0$ → la variabilità dovuta alla retta di regressione è nulla, perciò la retta di regressione è parallela all'asse delle ascisse. Non ha dunque nessuna capacità interpretativa.
- $R^2 = 1$ → la variabilità dovuta alla regressione coincide con quella totale, mentre quella residua è nulla. In questo caso tutti i punti sperimentali sono collocati sulla retta di regressione e quindi tutti allineati.

  ↓ situazione ideale
- $R^2 = \frac{0}{0}$ → tutti i punti osservati sono allineati perpendicolarmente all'asse x e la retta di regressione è $y = y_i = \bar{y}$

INDICE DI CORRELAZIONE LINEARE

Si usa per quantizzare la relazione tra due variabili.

$$r_{xy} = \frac{Cov[x,y]}{\sqrt{Var[x] \cdot Var[y]}} \quad \rightarrow \quad -1 \le r_{xy} \le 1$$

Se:

- $r_{xy} > 0$ → x e y sono correlate positivamente.
- $r_{xy} = 0$ → x e y sono incorrelate. (sono indipendenti)
- $r_{xy} < 0$ → x e y sono correlate negativamente.

COVARIANZA:

Indice che permette di vedere come una relazione lineare tra due variabili statistiche, x e Y.

$$Cov[x,Y] = \begin{cases} \text{continue} \rightarrow \iint (x - E[x])(Y - E[Y]) \cdot f_{xy}(x,y)\, dx\, dy \\ \text{discrete} \rightarrow \sum_{x_i} (x_i - E[x])(Y_i - E[Y]) \cdot P_{xy}(x,y) \end{cases}$$

- $Cov > 0$ = correlate positivamente
- $Cov < 0$ = correlate negativamente
- $Cov = 0$ → incorrelate

• Def. ASSIOMATICA → $\mathcal{A}$ = Spazio degli eventi = $\{A_i, A_i \subseteq S\}$
o algebra

La terna $(S, \mathcal{A}, P)$ crea lo SPAZIO DI PROBABILITÀ se:

1) $S \in \mathcal{A}$
2) $A \in \mathcal{A} \rightarrow \bar{A} \in \mathcal{A}$
3) $\{A_n, n \in \mathbb{N}, A_n \in \mathcal{A}\} = \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

dove S = insieme di risultati possibili
o, meglio, spazio campione

Legge di Morgan: $\overline{A \cap B} = \bar{A} \cup \bar{B}$
$\overline{A \cup B} = \bar{A} \cap \bar{B}$

DEFINIZIONE PROBABILITÀ, Kolmogorov

Dato un esperimento, dato S e data $\mathcal{A}$, chiamiamo probabilità una funzione

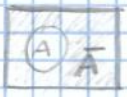
$P: \mathcal{A} \rightarrow \mathbb{R}$
$P: A \in \mathcal{A} \mapsto P(A) \in \mathbb{R}$

tale che:

i) $0 \leq P(A) \leq 1$
ii) $P(S) = 1$
iii) Presa la successione di eventi $A_1, A_2 \ldots A_n$, tra di loro mutuamente esclusivi, vale $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$

PROPRIETÀ ELEMENTARI DELLA PROBABILITÀ:

• $A \in \mathcal{A}$ $P(\bar{A}) = 1 - P(A)$ → infatti $A \cup \bar{A} = S$ e $P(A \cup \bar{A}) = P(S) = 1$



• $P(\emptyset) = 0$ perché $P(\bar{A}) = 1 - P(A)$
$P(\bar{S}) = 1 - P(S) = 1 - 1 = 0$
$\emptyset$ = insieme vuoto

• $A, B \in \mathcal{A}$ $A \subseteq B \rightarrow P(A) \leq P(B)$

Dimostro: $P(B) = P(A \cup [B - A]) = P(A) + P(B - A) \geq P(A)$
perché $P(B - A) > 0$

• Presi $A, B \in \mathcal{A}$ vale: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

va tolto perché altrimenti andrebbe calcolato due volte

ELEMENTI STOCASTICAMENTE INDIPENDENTI

Quando la probabilità $P(A|B) = P(A)$ allora A e B sono detti stocasticamente indipendenti.

- Vera solo se $P(B) \neq 0$
- Lo sono quando B in realtà non condiziona A

Per verificare che sono stoc. indipendenti:

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

PER ESERCIZI: se sono stocas. indipendenti, posso ricavare l'intersezione da formula sopra. Se non lo sono lui deve essere dato nel testo.

NOTA: indipendenza $\not\Longleftrightarrow$ incompatibilità

esempio → $A = \{\text{pari}\}$ $B = \{\leq 4\}$

sono indipendenti ma non incompatibili

↓

$A \cap B = \{2, 4\}$

Generalmente se sono incompatibili non sono indipendenti, sono sempre entrambi quando $P(A) \circ P(B) = 0$ e quindi $A \circ B$ è $\emptyset$

INDIPENDENZA PER 3 O PIÙ ELEMENTI

A, B, C

i) $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

ii) $\begin{cases} P(A \cap B) = P(A) \cdot P(B) \\ P(A \cap C) = P(A) \cdot P(C) \\ P(B \cap C) = P(B) \cdot P(C) \end{cases}$

$S = \{a, b, c, d\}$

$A = \{a, b\}$ $B = \{a, b\}$

$C = \emptyset$

i $\not\Longleftrightarrow$ ii → Dimostrazione:

$P(A \cap B \cap C) = P(\emptyset) = 0 = P(A) \cdot P(B) \cdot P(C)$

Però: $P(A \cap B) = \frac{1}{2} \neq P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

oppure

$P(A \cap B) = P(\{b\}) = \frac{1}{4}$ e $P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$P(A \cap C) = P(\{a\}) = \frac{1}{4}$ e $P(A) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$S = \{a, b, c,$

$A = \{a, b\}$

$B = \{b, c\}$

$C = \{a, c\}$

ma: $P(A \cap B \cap C) = P(\emptyset) = 0 \not\Longleftrightarrow P(A) \cdot P(C) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$

FORMULA DI BAYES

Sia dato $(S, \mathcal{F}, P)$. Sia data la partizione di S

$$P(A_j \mid B) = \frac{P(A_j) \cdot P(B \mid A_j)}{\sum_{j \in I} P(A_j) \cdot P(B \mid A_j)}$$

Si usa tutte le volte in cui abbiamo un esperimento a fasi successive. (es. cosa è successo a metà?)

Dimo:

$P(E_1 \cap E_2) = P(E_1) \cdot P[E_2 \mid E_1]$ oppure $P[E_1 \wedge E_2] = P[E_2] \cdot P[E_1 \mid E_2]$

osservo che entrambe hanno primo membro uguale, posso quindi eguagliare 1° e 2°

$$P(E_2) \cdot P[E_1 \mid E_2] = P(E_1) \cdot P(E_2 \mid E_1)$$

da cui ottengo

$$P(E_1 \mid E_2) = \frac{P(E_1) \cdot P(E_2 \mid E_1)}{P(E_2)} \quad \text{oppure} \quad P(E_2 \mid E_1) = \frac{P(E_2) \cdot P(E_1 \mid E_2)}{P(E_1)}$$

- Usato per calcolare la P di una causa che ha scatenato l'evento verificato.

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 19 di 386

FUNZIONI MISURABILI

$(S_1, \mathcal{A}_1) \xrightarrow{X} (S_2, \mathcal{A}_2)$ si hanno 2 esperimenti

$X : S_1 \to S_2$ X grande è una funzione definita tra S_1 e S_2

Diciamo che X sia misurabile se preso un qualsiasi evento $A_2 \in \mathcal{A}_2$ la sua controimmagine $X^{-1}(A_2)$ è un evento nello spazio $\mathcal{A}_1$, cioè $X^{-1}(\mathcal{A}_2) \in \mathcal{A}_1$

esempio sul quad

Definizione:

Dato $(S, \mathcal{A}, P)$ e dato $(R, \mathcal{B} \dots)$ è detta variabile casuale una funzione $X : S \to R$ $\quad X : S \to X(S) \in R$ se tale funzione è misurabile.

$\mathcal{B}$ → algebra di Borel (sottinsieme dei reali) è il più piccolo spazio degli eventi contenente tutti gli intervalli $(a,b]$, $[a,b]$, $[a,b)$ ... e tutte le loro unioni ed intersezioni infinite.

Dato la variabile casuale X è poi possibile definire una probabilità P_X così:

$\forall B \in \mathcal{B} \qquad P_X(B) = P[X^{-1}(B)] = P[X \in B]$

$\in \mathcal{A}$ → ben definita grazie alla misurabilità

È detta variabile casuale:

$X : S \to R$

$X : s \in S \to X(s) \in R$ tale che sia misurabile.

Come si descrivono le variabili casuali?

Notazione con lettere maiuscole sono insiemi numerici di esperimenti PRIMA di essere effettuati.
es. x = esperimento lancio del dado
y = tempo vita auto dopo l'esperimento. Diventano consistenti due tale con le maiuscole X, Y, Z.
$X = x \quad Y = y$)

Ha come dominio lo spazio dei campioni S e come codominio la retta reale

- $\lim_{t \to -\infty} F_x(t) = 0$

Infatti:

$$\lim_{t \to -\infty} F_x(t) = \lim_{t \to -\infty} P[x \in (-\infty, t]] = P[x \in (-\infty, -\infty)] = F_x(-\infty) = 0$$

- F_x sono continue da destra, ma non da sinistra o almeno non per forza

$$\lim_{t \to t_0^+} F_x(t) = F_x(t_0)$$

$$\lim_{t \to t_0^+} F_x(t) = \lim_{t \to t_0^+} P[x \in (-\infty, t)] = P\left[\lim_{t \to t_0^+} x \in (-\infty, t)\right] =$$

$$= P[x \in (-\infty, t_0)] = F_x(t_0)$$

VARIABILI DISCRETE O CONTINUE

due variabili possono essere continue o discrete:

- x è detta continua se F_x è derivabile (tranne al più in un numero finito di punti
- x è detta discreta se può assumere valori in un insieme di cardinalità finita o al più numerabile.

ASSOLUTAMENTE CONTINUE

è così definita se $\exists \frac{dF(t)}{dt} = f_x(t)$

$f_x(t)$ = DENSITÀ

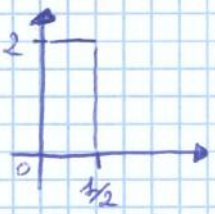
$$P[x \in A] = \int_A f_x(t)\,dt \quad \rightarrow F_x(t) = P[x \le t] = P[x \in (-\infty, t]] = \int_{-\infty}^{t} f_x(s)\,ds$$

Proprietà delle f_x:

- $f_x(t) \ge 0$
- $\int_{-\infty}^{\infty} f_x(t)\,dt = 1$

Nota: $0 \le P_x(t) \le 1$

$$f_x(t) = \begin{cases} 0 & t \in [0, \frac{1}{2}] \\ 2 & t \in (0, \frac{1}{2}] \end{cases}$$



Descrizione congiunta

(x_1, x_2) è discreta se i valori assumibili sono di cardinalità finita o numerabile

→ $P_x(t_1, t_2) = P_{(x_1,x_2)}(t_1, t_2) = P[x_1 = t_1, x_2 = t_2]$

- (x_1, x_2) è ass. continua se $\exists f_{(x_1,x_2)}(t_1, t_2)$ tale che $P[(x_1, x_2) \in A] = \iint_A f_{(x_1,x_2)}\, t_1 t_2 \; dt_1 dt_2$, $A \subseteq \mathbb{R}^2$

DISTRIBUZIONI MARGINALI

Fanno riferimento alla sola x_1 o alla sola x_2

$P_{x_1}(t_1) = \sum_{t_2} P_{(x_1,x_2)}(t_1, t_2)$ (in colonna nella tabella)

ESAME : **Definizione** : x_1 e x_2 sono della sfocastrocca. in dipendenza se $P_{(x_1,x_2)}(t_1, t_2) = P_{x_1}(t_1) \cdot P_{x_2}(t_2)$

$\forall t_1, t_2 \in \mathbb{R}$

INDIPENDENZA:

Nota: $P_{(x_1,x_2)}(t_1, t_2) \not\longrightarrow P_{x_1}(t_1), P_{x_2}(t_2)$

NO (a meno che non siano indipendenti)

Nel caso delle ass. continue

$f_{x_1}(t_1) = \int_{\mathbb{R}} f_{(x_1,x_2)}(t_1, t_2)\, dt_2$

INDIPENDENTI se $f_{(x_1,x_2)}(t_1, t_2) = f_{x_1}(t_1) \cdot f_{x_2}(t_2)$ $\forall t_1, t_2$

Inoltre:

$P[x \in A, Y \in B] = P[x \in A] \cdot P[Y \in B]$

$\forall B, A \in \mathbb{R}$

$F_{xy}(a, b) = F_x(a) \cdot F_y(b)$

Definizione:

Dati $A_1, A_2, A_3 \dots A_n$ diciamo che sono indipendenti se preso un loro sottinsieme qualsiasi,

$\{A_2, A_5, A_{13}\}$ vale $P(A_2 \cap A_5 \cap A_{13}) = P(A_2) \cdot P(A_5) \cdot P(A_{13})$

Nota

Se A, B sono indipendenti, allora lo sono anche i complementari $\bar{A}, \bar{B}$

Ip: $P(A \cap B) = P(A) \cdot P(B)$

Ts: $P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot P(\bar{B})$

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(\overline{\overline{A \cup B}}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)] =$$
$$= 1 - P(A) - P(B) + P(A) \cdot P(B) = 1 \cdot (1 - P(A)) - P(B) \cdot (1 - P(A)) =$$
$$= (1 - P(B)) \cdot (1 - P(A)) = P(\bar{B}) \cdot P(\bar{A})$$

Formula prodotto:

(S, a, P) sia data una famiglia di eventi $A_1, A_2 \dots A_n$

vale $P(A_1 \cap A_2 \dots \cap A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \dots P(A_n | A_1 \cap A_2 \dots A_{n-1})$

VARIANZA

$Var[x] = E[(x - E[x])^2]$ = varianza di x, cresce al crescere della dispersione della v.c. x

Il posto della varianza può essere usata la DEVIAZIONE STANDARD $\sigma = \sqrt{Var[x]}$

OSSERVAZIONE:

$$V[x] = \begin{cases} \sum_{x_i} (x_i - E[x])^2 \cdot p_x(x_i) & \text{DISCRETA} \\ \int_R (x - E[x])^2 \cdot f_x(x)\, dx & \text{ASS. CONTINUA} \end{cases}$$

Nota: togliendo il quadrato

$$\int_R (x - E[x])\, f_x(x)\, dx = \int_R x \cdot f_x(x)\, dx - E[x] \int_R f_x(x)\, dx = E[x] - E[x] \cdot 1 = 0$$

Perché no con il valore assoluto?

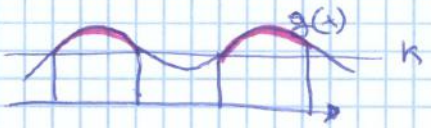
$E[|x - E[x]|]$ = scarto medio assoluto

DISUGUAGLIANZA DI TCHEBYCHEFF

Teorema

Sia x v.c. qualsiasi, sia $g: \mathbb{R} \to \mathbb{R}^+$

Allora: $P[g(x) \geq k] \leq \frac{E[g(x)]}{k}$ $\forall k$



Dimostrazione (ass. continua)

$$E[g(x)] = \int g_x \cdot f_x(x)\, dx = \int_{\{x: g(x)<k\}} g(x) f_x(x)\, dx + \int_{\{x: g(x)\geq k\}} g(x) f_x(x)\, dx \geq \int_{\{x: g(x)\geq k\}} g(x) \cdot f_x(x)\, dx \geq$$

$$\geq \int_{\{x: g(x)\geq k\}} k\, f_x(x)\, dx$$

$$E[g(x)] \geq \ldots \geq \int_{\{x: g(x)\geq k\}} k \cdot f_x(x)\, dx = k \int_{x: g(x)\geq k} f_x(x)\, dx = k \cdot P[g(x) \geq k]$$

$$P[g(x) \geq k] \leq \frac{E[g(x)]}{k} \rightsquigarrow g(x) = (x - E[x])^2$$

Corollario di Tchebycheff

$$P[|x - \mu_x| < t\sigma_x] \geq 1 - \frac{1}{t^2}$$

$\frac{1}{t^2}$ → minorante o limite inferiore

Dato y, $P[|y - E[y]| \geq a \cdot \sigma_y] \leq \frac{1}{a^2}$

$y = \bar{x}_n \qquad \varepsilon = a\sigma_y \leftrightarrow a = \frac{\varepsilon}{\sigma_y} \leftrightarrow a^2 = \frac{\varepsilon^2}{\sigma_y^2} = \frac{\varepsilon^2}{\frac{p(1-p)}{n}} \rightarrow \frac{1}{a^2} = \frac{p(1-p)}{n\varepsilon^2}$

$$P[|\bar{x}_n - p| \geq \varepsilon] \leq \frac{p(1-p)}{n\varepsilon^2} \underset{n \to \infty}{\to} 0$$

LEGGE DEI GRANDI NUMERI

x_i indipendenti, stessa distribuzione

$E[x] = \mu < \infty$

$V[x] = \sigma^2 < \infty$ (deviazione standard)

$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}$

Man mano che aumento gli esperimenti ($\bar{x}$), la media diventa sempre più stretta.

DISTRIBUZIONE DI POISSON

È il modello adatto per modellare il numero di difetti o non conformità che si trovano in una unità di prodotto.

$x \sim Poiss(\lambda)$ con $\lambda > 0$ e $\lambda \in \mathbb{R}^+$, $\lambda = n \cdot p$

se può assumere valori in $\mathbb{N} = \{0, 1 \dots\}$ con probabilità

$$P[x=k] = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \qquad \forall k \in \{0, 1 \dots\}$$

- Nota: $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1$ → quindi questa è una distribuzione.

- È un caso particolare della binomiale:

Sia $x \sim Bin(n, p)$ considerando $n \to \infty$ $p \to 0$ sotto il vincolo che $n \cdot p = \lambda$ è resta costante

$$P[x=k] = \binom{n}{k} p^k (1-p)^{n-k} \to \frac{n!}{k!(n-k)!} \cdot \frac{\lambda^k}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} =$$

$$= \frac{\lambda^k}{k!} \cdot \frac{n(n-1)\cdots(n-k+1)\cdot(n-k)!}{n^k \cdot (n-k)!} \cdot \left(1 - \frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} = \frac{\lambda^k}{k!} \cdot 1$$

($\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$, $n \to \infty$; $\left(1 - \frac{\lambda}{n}\right)^k \to 1$, $n \to \infty$; $\left(1 - \frac{\lambda}{n}\right)^{-k} \to e^{-\lambda}$, $n \to \infty$)

ovvero $\frac{\lambda^k}{k!} \cdot e^{-\lambda}$ → distribuzione di Poisson quando n è molto grande e p molto piccolo.

$n \gg 100$ e $p \leq 0{,}05$

MEDIA: $E[x] = \lambda = n \cdot p$ VARIANZA $= n \cdot p(1-p) = np - p(np) = \lambda - p\lambda$ tendono a λ, $p \to 0$

Proprietà geometrica:

Sia $x \sim Geom(p)$, allora $P[x \geq k] = (1-p)^k$

Dimostrazione:

$$P[x \geq k] = \sum_{i=0}^{\infty} P[x=i] = \sum_{i=k}^{\infty} (1-p)^i \cdot p = p \sum_{i=k}^{\infty} (1-p)^i = p \sum_{i=k}^{\infty} (1-p)^{i+k} =$$

$$= p \sum_{0}^{\infty} (1-p)^i \cdot (1-p)^k = (1-p)^k \cdot \underbrace{p \cdot \sum_{i=0}^{k} (1-p)^i}_{1} = (1-p)^k$$

perché $p \cdot \sum_{i=0}^{\infty} (1-p)^i = p \cdot \frac{1}{1-(1-p)} = 1$

PROPRIETÀ DI NON-ITERAZIONE

Facendo la ripetizione della stessa cosa mille volte, questo non comporta la probabilità.

Sia $x \sim Geo(p)$. Siano $n, k \in \mathbb{N}$, allora

$$P[x = k+n \mid x \geq k] = P[x=n]$$

Dimostrazione:

come mai non è ripetizione?

$$P[x = k+n \mid x \geq k] = \frac{P[x=k+n, x \geq k]}{P[x \geq k]} = \frac{P[x=k+n]}{P[x \geq k]} =$$

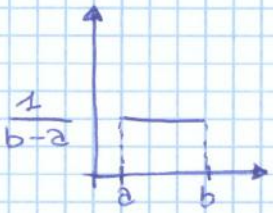
$$= \frac{(1-p)^{k+n} \cdot p}{(1-p)^k} = (1-p)^n \cdot p = P[x=n]$$

Somma di 2 geometriche indipendenti non è più indipendente

DISTRIBUZIONI PER FUNZIONI CONTINUE

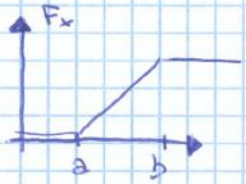
DISTRIBUZIONE UNIFORME

$X \sim U[a,b]$ $\quad a < b$ $\quad$ se $f_X(t) = \begin{cases} \frac{1}{b-a} & t \in [a,b] \\ 0 & \text{altrimenti} \end{cases}$



Funzione di distribuzione cumulativa:

$$F_X(t) = \begin{cases} 0 & t < a \\ \frac{t-a}{b-a} & t \in [a,b] \\ 1 & t > b \end{cases}$$



MEDIA: $E[X] = \frac{a+b}{2}$

Dimostrazione:

$$E[X] = \int t \cdot f_X(t)\,dt = \int_a^b t \cdot \frac{1}{b-a}\,dt = \frac{1}{b-a}\left[\frac{t^2}{2}\right]_a^b = \frac{1}{b-a}\cdot\frac{1}{2}\left(b^2-a^2\right) =$$

$$= \frac{(a+b)(b-a)}{2(b-a)} = \frac{a+b}{2}$$

VARIANZA: $V[X] = \frac{(b-a)^2}{12}$

Dimostrazione:

$$V[X] = E[X^2] - (E[X])^2 = \int_a^b t^2 \cdot \frac{1}{b-a}\,dt - \frac{(a+b)^2}{4} = \frac{1}{b-a}\left[\frac{t^3}{3}\right]_a^b - \frac{(a+b)^2}{4}$$

$$= \dots = \frac{(b-a)^2}{12}$$

Proprietà esponenziale

- Se $x \sim Exp(\lambda)$, allora $r_x(t) = \lambda$ costante

  Infatti,

$$r_x(t) = \frac{f_x(t)}{\bar{F}_x(t)} = \frac{\lambda \cdot e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

- Siano x_1, x_2, x_4

  $x_i \sim Exp(\lambda_i)$ indipendenti. Allora $Y = \min\{x_1, x_2 \dots x_4\}$ è ancora esponenziale di parametro $\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_4$

MEDIA: $E[x] = \frac{1}{\lambda}$

Dimostrazione:

$$E[x] = \int_0^\infty \lambda t \cdot f_x(t)\, dt = \int_0^\infty \lambda t \cdot e^{-\lambda t}\, dt = \dots = \frac{1}{\lambda}$$

VARIANZA: $V[x] = \frac{1}{\lambda^2}$

Dimostrazione:

$$V[x] = \int_0^\infty \lambda t^2 \cdot e^{-\lambda t}\, dt - (E[x])^2 = \dots = \frac{1}{\lambda^2}$$

WEIBULL

$x \sim Exp(\lambda)$ $\quad r_x(t) = \lambda$

Consideriamo un tasso

$$r_x(t) = \beta \cdot t^{\beta-1} \cdot \alpha^\beta \qquad \alpha, \beta > 0$$

$$\bar{F}_x(t) = e^{-(\alpha t)^\beta} \rightarrow x \sim Weibull(\alpha, \beta)$$

N.B.: Se $\beta = 1$ → distribuzione esponenziale

Varianza della popolazione non nota:

σ^2 sconosciuta $\rightarrow S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$

$T = \frac{\bar{x}_n - \mu}{S_n/\sqrt{n}}$ $\quad T \sim$ variabile casuale di Student con $n-1$ gradi di libertà

$P\left[-t_{n-1,1-\frac{\alpha}{2}} < T \leq t_{n-1,1-\frac{\alpha}{2}}\right] = 1-\alpha$

Intervallo di fiducia:

$I = (L_i, L_s] = \left(\bar{x}_n - t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x}_n + t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right)$

utilizzabile solo con popolazioni distribuite solo con una normale $\sim N(\mu, \sigma^2)$

DETERMINAZIONE DELLA NUMEROSITÀ CAMPIONARIA:

I_α = semiampiezza

$I_\alpha = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ oppure $I_\alpha = t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$

$n \geq \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{I_\alpha}\right)$

TEST IPOTESI

$H_0: \mu = \mu_0$
$H_A: \mu \neq \mu_0$

livello di fiducia $1-\alpha$, $x \sim N(\mu, \sigma^2)$
n = num. campioni , σ^2 nota

$$t = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \rightarrow \begin{cases} \text{se } t_{calc} \text{ è "vicina" a zero, accetto } H_0 \\ \text{se } t_{calc} \text{ è "lontana" da zero, rifiuto } H_0 \end{cases}$$

Regione accettazione: $R_{acc} \left(-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}\right)$

Regione rifiuto: $R_{rif} \left(-\infty, -t_{1-\frac{\alpha}{2}}\right) \cup \left(t_{1-\frac{\alpha}{2}}, +\infty\right)$

α = prob. errore del primo tipo

Tre tipi di test:

TEST BILATERALE:

$H_0: \mu = \mu_0$ $\quad t_{calc} > t_{1-\frac{\alpha}{2}}$
$H_A: \mu \neq \mu_0$ $\quad t_{calc} < -t_{1-\frac{\alpha}{2}}$

TEST UNILATERALE: $(\bar{x}_{max})$

$H_0: \mu = \mu_0$
$H_A: \mu > \mu_0$ $\quad t_{calc} > t_{1-\alpha}$

TEST UNILATERALE: $(\bar{x}_{min})$

$H_0: \mu = \mu_0$
$H_A: \mu < \mu_0$ $\quad t_{calc} < -t_{1-\alpha}$

TEOREMA LIMITE CENTRALE

Siano $x_1, x_2 \dots x_n$ delle variabili casuali indipendenti con stessa distribuzione, con $E[x_i] = \mu$ e $V[x_i] = \sigma^2$

Sia $Y_n = \frac{x_1 + x_2 \dots x_n - n\mu}{\sqrt{n} \cdot \sigma}$, allora per "n grande" vale (perché la media è uguale)

l'approssimazione $Y_n \sim N(0,1)$

Sia data una popolazione distribuita con densità $f(x)$ avente media μ e varianza σ^2 finite. Detta $\bar{x}_n$ la media di un campione casuale di dimensione n estratto da essa, posso allora la media campionaria segue una dist. normale normale al tendere di n ad infinito.

Per grande si im… anche $n \sim 10$ o $n \sim 20$ simmetrica

Se però partissi … anche avere n…

RICORDA INFATTI

Siano $x_1 \dots x_n$ indipendenti e $\sim N(\mu, \sigma^2)$

$\bar{x}_n = \frac{S}{n}$ dove $S = x_1 + x_2 \dots + x_n \sim N(n \cdot \mu, n \cdot \sigma^2)$

da cui ottengo

$$\bar{x}_n = \frac{x_1 + \dots x_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \begin{cases} E[\bar{x}_n] = \frac{n \cdot \mu}{n} = \mu \\ V[\bar{x}_n] = V\left[\frac{x_1 + \dots x_n}{n}\right] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \end{cases}$$

↓ Conferma la legge dei grandi numeri

- Se $n \to \infty$, $\bar{x}_n \sim N(\mu, 0)$

Standardizzo $\bar{x}_n$!

$$\frac{\bar{x}_n - \mu}{\sqrt{\sigma^2/n}} = \frac{\frac{x_1 + x_2 \dots + x_n}{n} - \frac{n\mu}{n}}{\sqrt{\frac{\sigma^2}{n}} \cdot \frac{n}{n}} = \frac{x_1 + x_2 \dots + x_n - n\mu}{\sqrt{n\sigma^2}} =$$

$$= \frac{x_1 + x_2 \dots x_n - n\mu}{\sqrt{n} \cdot \sigma} \sim N(0,1)$$ vale per $x_i \sim N(\mu, \sigma^2)$

Sia x una v.c., sia $Y = \phi(x)$, dove $\phi : \mathbb{R} \to \mathbb{R}$. Conosciamo F_x e f_x

Come sono fatte F_y e f_y?

CASO DISCRETO

Sia ϕ invertibile (e sia ϕ^{-1} la sua inversa), conviene lavorare con la definizione.

$$P_y(t) = P[Y=t] = P[\phi(x)=t] = P[x=\phi^{-1}(t)] = P_x[\phi^{-1}(t)]$$

- Se invece ϕ non fosse invertibile

$$P_y(t) = P[y=t] = P[\phi(x)=t] = P[\{x_i : \phi(x_i)=t\}]$$

CASO CONTINUO

Conviene usare la def. cumulata, $x, Y = \phi(x)$

Se ϕ invertibile $\rightsquigarrow \phi^{-1}$ è l'inversa

[ϕ monotona crescente]

$F_x \rightsquigarrow F_y$: $F_y(t) = P[Y \le t] = P[\phi(x) \le t] = P[x \le \phi^{-1}(t)] = F_x(\phi^{-1}(t))$

$$f_y(t) = \frac{d F_y(t)}{dt} = \frac{dF_x(\phi^{-1}(t))}{dt} = f_x(\phi^{-1}(t)) \cdot \frac{1}{\phi^{-1}(\phi^{-1}(t))}$$

per definizione

RELAZIONI TRA INDIPENDENZA E INCORRELAZIONE

(x,y) indipendenti $\rightarrow$ (x,y) sono anche incorrelate

$\downarrow$

non è sempre vero il contrario

Problema

Analizzo la relazione tra x ed y.

$Cov[x,y] = 5$ 5 è "grande" oppure è "vicino" a zero

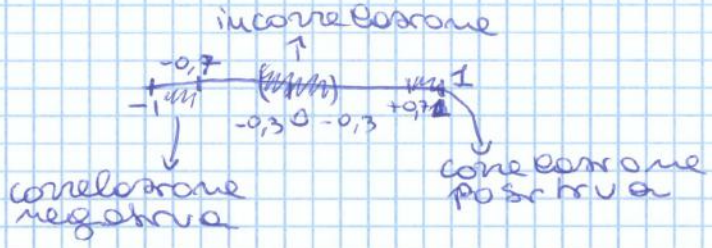
$\rightarrow$ indice adimensionale $\rightarrow$ INDICE DI CORRELAZIONE LINEARE

$$r_{x,y} = \frac{Cov[x,y]}{\sqrt{V[x]-V[y]}}$$

$x \sim$ anni
$y \sim$ euro

Teorema

$-1 \leq r_{x,y} \leq 1$



PROCESSI DI POISSON

Sono una particolare classe dei processi di conteggio. Si applicano quando valgono le ipotesi seguenti:

Ipotesi 1:

Esiste quantità α positiva tale che la probabilità che si verifichi esattamente un evento in un piccolo intervallo di lunghezza Δt sia approssimativamente uguale a $\alpha \Delta t$, cioè:

$$P[\text{esattamente un evento nell'intervallo di tempo } \Delta t] = \alpha \Delta t + o(\Delta t)$$

$$\text{con } \lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

Ipotesi 2:

La probabilità che nell'intervallo Δt si verifichi più di un evento è trascurabile (e più alta) rispetto alla probabilità che se ne verifichi esattamente uno, cioè:

$$P[\text{due o + eventi nell'Intervallo } \Delta t] = o(\Delta t)$$

$$P[\text{n° eventi in I } \Delta t] = 1 - P[\text{1 evento}] - P[\text{2 o più eventi}] =$$
$$= 1 - (\alpha \cdot \Delta t + o(\Delta t) - o(\Delta t))$$

Ipotesi 3:

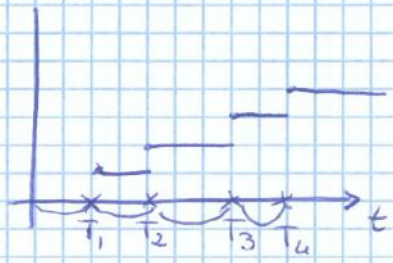
L'evento E_1 rappresenta il verificarsi di n_1 eventi in qualsiasi intervallo di tempo (t_1, t_2) è indipendente dall'evento E_2 rappresentante il verificarsi di n_2 eventi in un intervallo (t_3, t_4) non contenuto (parzialmente o del tutto) in (t_1, t_2), cioè:

$$P[n_1 \text{ eventi in } (t_1, t_2) \text{ e } n_2 \text{ eventi in } (t_3, t_4)] = P[E_1] \cdot P[E_2]$$

Un processo di conteggio che soddisfa queste 3 ipotesi è detto

PROCESSO DI POISSON D'INTENSITÀ α

TEOREMA



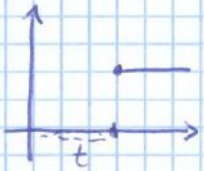
T_i = intervalli tra eventi

Sia $N = \{N(t), t \geq 0\}$ un processo di Poisson d'intensità α ($\alpha > 0$).

Allora le variabili T_i sono variabili con distribuzione esponenziale di parametro α, indipendenti. E viceversa.

Dimostrazione:

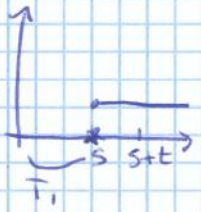
$$\bar{F}_{T_1}(t) = P[T_1 > t] = P[N(t) = 0] = \frac{(\alpha t)^0}{0!} \cdot e^{-\alpha t} = e^{-\alpha t}$$



e questa è proprio la

$$\bar{F}_{T_2}(t) = P[T_2 > t] = \int_0^\infty P[T_2 > t \mid T_1 = s] \cdot f_{T_1}(s)\, ds =$$

$$= \int_0^\infty P[N(t+s) - N(s) = 0 \mid T_1 = s] \cdot f_{T_1}(s)\, ds \cdot \bar{T}_1$$



per indipendenza

$$= \int_0^\infty P[N(t) = 0] \cdot f_{T_1}(s)\, ds = P[N(t) = 0] \cdot \int_0^\infty f_{T_1}(s)\, ds =$$

$$= P[N(t) = 0] = \frac{(\alpha t)^0 \cdot e^{-\alpha t}}{0!}$$

Campione:

Un insieme di n variabili casuali indipendenti $(x_1, x_2, x_3 \dots x_n)$ e ciascuna x_i ha lo stesso distribuzione: $x_i \sim f(\,)$

- Attenzione: sono variabili casuali, non numeri!

Dopo l'estrazione si osservano i dati → realizzazione del campione

STATISTICA:

$g(x_1, x_2 \dots x_n)$ = una funzione di variabili casuali (e quindi a sua volta una variabile casuale) che non contiene parametri da stimare.

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

$\max(x_1, x_2 \dots x_n) - \min(x_1, x_2 \dots x_n)$

$x_1 + 2x_2 - 4x_3 + \dots - x_n$

$\vdots$

$\bar{x}_n - \mu$ non è una stima

STIMATORI:

Statistiche che vengono usate per stimare un parametro ϑ o una qualche sua funzione

Esistono stimatori migliori di altri?

PROPRIETÀ STIMATORI

- correttezza:

$T = g(x_1, x_2 \dots x_n)$ stimatore di ϑ

$E[T] = \vartheta \rightarrow$ T stimatore corretto o non deviato

$E[T] - \vartheta = \delta$ è detta distorsione

È corretto se $\bar{x}_n = \frac{1x_1 + 1x_2 + 1x_3}{\text{numero}} = \mu$

- consistenza:

T stimatore corretto si dice uno stimatore consistente se $Var[T] \to 0$ quando la dim del campione $\to +\infty$:

consistente se $\lim_{n \to \infty} Var[T_n] = 0$

Dimostrazione:

$$E[S_n^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right] = \frac{1}{n-1}\cdot E\left[\sum_{i=1}^{n}\left\{x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2\right\}\right] =$$

$$= \frac{1}{n-1}\cdot E\left[\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}x_i\,\bar{x}_n + \sum_{i=1}^{n}\bar{x}_n^2\right] =$$

$$= \frac{1}{n-1}\cdot\left\{\sum_{i=1}^{n}E[x_i^2] - 2E\left[\sum_{i=1}^{n}x_i\cdot\bar{x}_n\right] + \sum_{i=1}^{n}E[\bar{x}_n^2]\right\} =$$

$$= \frac{1}{n-1}\cdot\left\{n\cdot(\sigma^2+\mu^2) - n\,E\left[\frac{(x_1+x_2+\dots x_n)}{n}\cdot\bar{x}_n\right] + n\,E[\bar{x}_n^2]\right\} =$$

$$= \frac{1}{n-1}\cdot\left\{n\sigma^2+n\mu^2 - n\left(\frac{\sigma^2}{n}+\mu\right)\right\} = \frac{1}{n-1}\left\{n\sigma^2+n\mu^2 - \sigma^2 - n\mu^2\right\} =$$

$$= \frac{\sigma^2(n-1)}{(n-1)} = \sigma^2$$

Invece la sua varianza: ???

$$V[S_n^2] = \frac{1}{n}\left(E[x_i^4] - \frac{n-3}{n-1}\cdot E\left[(x_i-E[x_i])^4\right]\right) \quad , \text{ per } n\to+\infty = 0$$

ERRORE QUADRATICO MEDIO (MSE)

Si usa per valutare l'efficienza di stimatori non corretti

$\delta = E[T] - \vartheta$ → distorsione

$$MSE_T = E[(T-\vartheta)^2] = E[((T-E[T])+(E[T]-\vartheta))^2] =$$

$$= E[(T-E[T])^2 + 2(T-E[T])\cdot(E[T]-\vartheta) + (E[T]-\vartheta)^2] =$$

$$= E[(T-E[T])^2] + 2E[(T-E[T])\cdot(E[T]-\vartheta)] + E[S^2] =$$

$$= V[T] + 2(E[T]-\vartheta)\cdot(E[T-E[T]]) + \sigma^2 =$$

$$= V[T] + 2(E[T]-\vartheta)\cdot\{E[T]-E[E[T]]\} + \sigma^2 =$$

$$= V[T] + \sigma^2$$

(il termine $\{E[T]-E[E[T]]\} = 0$)

poiché $E[E[T]] = E[T]$ perché il valore atteso è un n° e quindi il suo valore atteso è il n° stesso.

Da qui ottengo che quella parte della formula si annulla.

Teorema

$E[S_n^2] = \sigma^2$ $\quad V[S_n^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$

Distribuzione S_n^2?

STIME INTERVALLARI Stime rappresentate da un intervallo di valori (invece che un singolo valore) in cui cade il parametro da stimare con data probabilità.

Sia X popolazione, μ è da stimare, σ^2 nota e $\sigma^2 = 400$.
Per stimare μ uso $\bar{x}_n$, $n = 64$ → $\hat{\mu}$ [stima di μ attraverso $\bar{x}_{64}$]

$P[|\bar{x}_n - \mu| < 5]$

prob. che il valore compromesso si discosti dalla media di non più di 5

$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{64}\right)$

$= P[-5 < \bar{x}_n - \mu < 5] = P\left[\frac{-5}{\sigma/\sqrt{64}} < \frac{\bar{x}_n - \mu}{\sigma/\sqrt{64}} < \frac{5}{\sigma/\sqrt{64}}\right] = P\left[\frac{-5}{20/8} < Z < \frac{5}{20/8}\right] =$

$= P[-2 < Z < 2] = 0,96$

cercando da tavola

$$U = \sum_{i=1}^{2} \left(\frac{x_i - \bar{x}_m}{\sigma} \right)^2 = \left(x_1 - \frac{x_1+x_2}{2} \right)^2 + \left(x_2 - \frac{x_1+x_2}{2} \right)^2 = \left(\frac{x_1-x_2}{2} \right)^2 + \left(\frac{x_2-x_1}{2} \right)^2 =$$

$$= 2\left(\frac{x_1-x_2}{2} \right)^2 = \left(\frac{x_1-x_2}{\sqrt{2}} \right)^2 = t^2 \sim x_1^2$$

↓ Una normale standard al quadrato

Ci serve per:

Dato $(x_1, x_2 \dots x_m)$

$$U = \frac{(m-1)\,S^2}{\sigma^2} = \frac{(m-1)\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2}{\sigma^2} = \sum_{i=1}^{m}\left(\frac{x_i - \bar{x}_m}{\sigma} \right)^2 \sim \chi^2_{m-1}$$

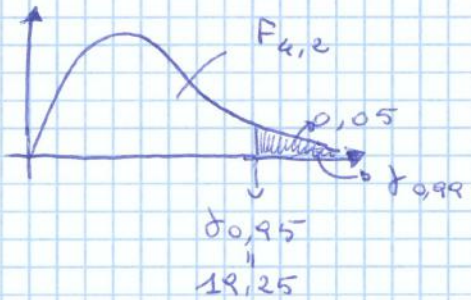
$$\rightarrow S^2 = \frac{\sigma^2}{m-1} \cdot U \quad \text{dove } U \sim \chi^2_{m-1}$$

DISTRIBUZIONE F (di Fisher)

U e V due variabili casuali indipendenti con distribuzione χ^2 rispettivamente con m e n gradi di libertà

$$F = \frac{U/m}{V/n} \sim F_{m,n} \qquad [F_{m,n} \neq F_{n,m}]$$

Si dice di avere una distribuzione F con m e n gradi di libertà



→ Esistono due tavole diverse di Fisher: - quantile 95 - quantile 99

$$f_{m,n,1-\alpha} = \frac{1}{f_{n,m,\alpha}} \rightarrow \text{inverti } m \text{ e } n$$

Quando si usa la t studentt?

$\{x_1, x_2 \dots x_n\}$ $\quad x_i \sim N(\mu, \sigma^2)$

$\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad$ e $\quad \frac{\bar{x}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

Supponiamo che σ^2 non sia nota → la sostituisco con una stima

$$\frac{\bar{x}_n - \mu}{\sqrt{\frac{s_n^2}{n}}} = \frac{\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{s_n^2 n}{n \cdot \sigma^2}}} = \frac{\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum (x_i - \bar{x}_n)^2}{(n-1)\sigma^2}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Nota! la distribuzione di t è più larga di quella standard

STIMA PUNTUALE (Stima per intervalli)

Ad ogni stima del rapporto σ si associa un intervallo di valori per il quale possiamo affermare che, con un certo grado di fiducia, contiene il valore vero del parametro.

$$P[L_i < \sigma < L_s] = 1 - \alpha$$

Stima per intervalli: $P[L_i \leq \sigma \leq L_s] = 1 - \alpha$

L_i: estremo inferiore di fiducia; L_s: estremo superiore di fiducia; $1-\alpha$: livello di fiducia

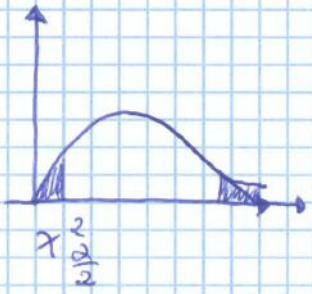
tipicamente α = 0,01 / 0,05 / 0,1 $\quad$ $1-\alpha$ = 0,99 / 0,95 / 0,9

(6) Test varianza

$H_0: \sigma^2 = \sigma_0^2$ $\quad \alpha, n$ fissati

$V = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$ (se è vero $\sigma^2 = \sigma_0^2$)

Lo calcolerò → se assume valori compatibili con una χ^2_{n-1} → accetto
altrimenti rifiuto



(7) Test uguaglianza tra 2 varianze

$X_1 \sim n_1$ $\quad X_2 \sim n_2$ $\quad H_0: \sigma_1^2 = \sigma_2^2 \rightarrow \frac{\sigma_1^2}{\sigma_2^2} = 1$

$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$

È vero se $\frac{\sigma_1^2}{\sigma_2^2} = 1 \rightarrow F = \frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1)$

PRIMA DELL'ANOVA – SCOMPOSIZIONE DELLA VARIANZA

Esempio:

2 popolazioni, 3 osservazioni per ogni popolazione. $n=6$

$A \to x_1, x_2, x_3 \to \mu_x, \sigma_x^2$
$B \to y_1, y_2, y_3 \to \mu_y, \sigma_y^2$
$\Big\} \to \mu = \sigma^2$ (totali)

Vale:

$$\mu = \frac{1}{6}\left[\sum_{i=1}^{3} x_i + \sum_{i=1}^{3} y_i\right] = \frac{1}{2}\left[\sum \frac{x_i}{3} + \sum \frac{y_i}{3}\right] = \frac{\mu_x + \mu_y}{3}$$

La media totale è la media delle medie

$$\sigma^2 = \frac{1}{6}\left[\sum_{i=1}^{3}(x_i-\mu)^2 + \sum_{i=1}^{3}(y_i-\mu)^2\right] = \frac{1}{6}\left[\sum(x_i-\mu_x+\mu_x-\mu)^2 + \sum(y_i-\mu_y+\mu_y-\mu)\right] =$$

$$= \frac{1}{6}\left[\sum(x_i-\mu_x)^2 + \sum(\mu_x-\mu)^2 + 2\sum(x_i-\mu_x)(\mu_x-\mu) + 2\sum(x_i-\mu_y)(\mu_y-\mu)\right]$$

$$= \frac{1}{6}\left[3\sigma_x^2 + 3(\mu_x-\mu)^2 = \dots\right]$$

moltiplico e divido per 3 e mi viene il quadrato della varianza

$$2(\mu_x-\mu)\cdot\sum(x_i-\mu) = 2\cdot(\mu_x-\mu)\left[3\sum\frac{x_i}{3} - 3\mu\right] = 0$$

μ → stesso per μ_y

Da cui ottengo:

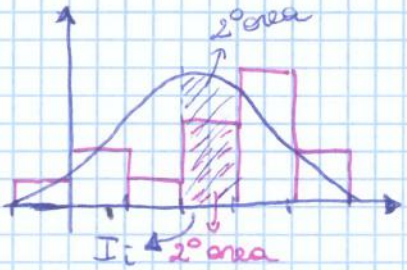
$$\sigma^2 = \frac{\sigma_x^2+\sigma_y^2}{2} + \frac{(\mu_x-\mu)^2+(\mu_y-\mu)^2}{2}$$

- $\frac{\sigma_x^2+\sigma_y^2}{2}$: Media delle varianze → VARIANZA INTERNA AI GRUPPI (WITHIN)
- $\frac{(\mu_x-\mu)^2+(\mu_y-\mu)^2}{2}$: Varianza delle medie → VARIANZA TRA GRUPPI (BETWEEN) → se molto alto indica che le medie sono molto diverse

TEST χ^2 PER BONTÀ ADATTAMENTO

$H_0: f_X = f_0$

$H_A: f_X \neq f_0$



- f_0
- f empirico (istogramma dati campione)

Devo misurare la distanza

= = frequenza di osservazione su I_i

= = frequenza "teorica di individui che avrei dovuto osservare su I_i" (se H_0 vera)

$$\sum_{i=1}^{N} \frac{(n_i - n p_i)^2}{n p_i} = W$$

n_i = n° osservazioni del campione

p_i = prob. di cadere I_i

n = tot. osservazioni campione

N = tot di sotto intervalli

Teorema:

Se vale H_0 $(f_X = f_0)$ allora $W \sim \chi^2$

g.l. = $N-1$ se nella f_0 teorica non ci sono parametri stimati

$N-K-1$ se nella f_0 teorica ci sono K parametri stimati col campione

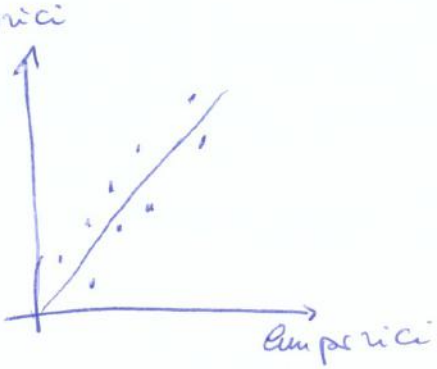
$$R_{Acc} = [0, \chi^2_{1-\alpha}]$$

Pregio: può essere utilizzato anche per discrete

Difetto: campioni molto numerosi (almeno 5 osservazioni in ogni sottinsieme)

QQ - PLOT

n° oss	osserv	quant. empirico	quant. teorico
1	x_1	x_1	$F_0^{-1}(x_1)$
2	x_2	x_2	$F_0^{-1}(x_2)$
3	x_3	x_3	$F_0^{-1}(x_3)$
⋮	⋮		⋮



↳ controimmagine di dist. empirica $(\frac{1}{n})$

Qui

CÈZANNE VINCENZO - SUDÈ

Rettangolare → n° righe > n° colonne

Se tutte le unità (variabili) sono quantitative, abbiamo una matrice di tipo matematico

unità statistica — variabile

Marca — Altezza ... Colore

↓ 1°3 sono di tipo quantitativo

↓ Ultime 2 sono di tipo qualitativo

Colore: qualitativa ordinale → ho un grado di giudizio

Riduzione dimensionale spesso tra 2 colonne

↓

REGRESSIONE

Var obs

Variabile stat. - continua: età: non è discreta perché in realtà se dico ho "20 anni" non sono non esatti, quindi è continua

correlazione (tabella)

p-value → è stato fatto il test di ipotesi

$H_0: \rho_{ij} = 0$ contro $H_a: \rho_{ij} \neq 0$

p-value accetto se ho dal 10% in poi,
se ↓
non posso rifiutare che
ci sia un legame
lineare

MATRICE BLOCK: vengono rappresentati a due a due i valori delle variabili

Come si può rappresentare:
- matrice completa
- triang. superiore
- triang. inferiore → lei preferisce questa

Se numero dei punti è lungo e sottile, ho relazione di tipo lineare.
p-value e grafici non soddisfano mai da soli,

Test:

Lett. greche = ? (variatre:)
maius = stimatori
minusc. = stime

$H_0: \rho_{x,y} = 0$

$H_A: \rho_{x,y} \neq 0$

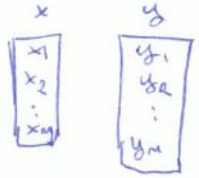
Si può dimostrare che se H_0 è vera, la dist. campionaria del coef. di correlazione (R) è tale che la statistica:

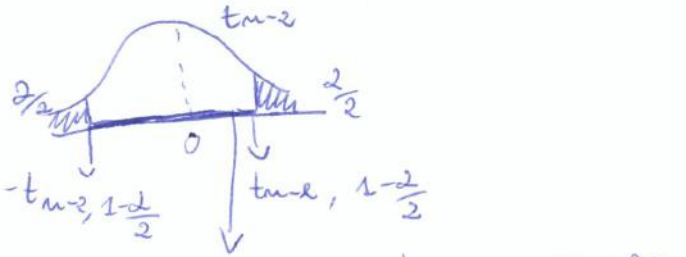
gradi di libertà

$$T_r = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \sim t_{n-2}$$

$\{(x_1,y_1),(x_2,y_2)\ldots(x_n,y_n)\}$ → r

$$t_{calc} = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$$





Es. su slide:

$\bar{x}_{min}, \bar{x}_{max}$

$\bar{x}_{s,min}$ $\bar{x}_{s,max}$

$-t\ldots$ $t_{n-2,1-\frac{\alpha}{2}}$

$$t_{n-2,1-\frac{\alpha}{2}} = \frac{r_{calc}}{\sqrt{1-r_{calc}^2}}\sqrt{n-2}$$

Un campione 0,80 non mi dice niente sul fatto che siamo correlate o no.
(rivedi)

$$t^2_{n-2,1-\frac{\alpha}{2}} = \frac{r_c^2}{1-r_c^2}(n-2) \rightarrow t^2_{n-2,1-\frac{\alpha}{2}} - r_c^2 t^2_{n-2,1-\alpha/2} = r_c^2(n-2) =$$

$$= r_c^2\left(t^2_{n-2,1-\frac{\alpha}{2}} + n - 2\right) = t^2_{n-2,1-\alpha/2}$$

$$|r_c| = \sqrt{\frac{t^2_{n-2,1-\alpha/2}}{t^2_{n-2,1-\alpha/2}+n-2}}$$

facendo limite → $\alpha = 5\%$ e $n = 5$

-0.878 $+0.878$

↑ decresce di n si avvicina

CONTRASTI (contrasts)

Tramite ANOVA (e relativi output) è possibile effettuare test relativi a combinazioni di medie (dello stesso, diversi livelli)

ES:

Anova $H_0: \mu_0 = \mu_1 = \dots \mu_k$

supponete test → rifiutare H_0

allora testate H_0: $c_0\mu_0 + c_1\mu_1 \dots c_k\mu_k = d$

[es: rifiuto $\mu_1 = \mu_2 = \mu_3$. Non sarà che $\mu_1 = \frac{\mu_2 + \mu_3}{2}$?

$2\mu_1 - \mu_2 - \mu_3 = 0$]

$$H_0: \sum_{i=1}^{k} c_i \mu_i = d$$

Possono essere tali che $\sum_{i=1}^{k} c_i = 0$

In questo caso si parla di contrasto

Esempio:

1) $\mu_1 = 2\mu_2 - \mu_3 \rightarrow \mu_1 - 2\mu_2 + \mu_3 = 0$

$\bar{c} = (1, -2, 1)$ → è un contrasto

2) $\mu_1 = \mu_2 + \mu_3 \rightarrow \mu_1 - \mu_2 - \mu_3 = 0$

$\bar{c} = (1, -1, -1) \Rightarrow \sum \neq 0$

non è un contrasto

$$C \sim N\left(\sum c_i\mu_i, \sum c_i^2 \frac{\sigma^2}{n}\right)$$

$H_0: \sum_{i=1}^{k} c_i \mu_i = d$ usare $C = \sum_{i=1}^{k} c_i \bar{y}_{\cdot i}$ come stimatore

$$E[C] = E\left[\sum_{i=1}^{k} c_i \bar{Y}_{\cdot i}\right] = \sum_{i=1}^{k} c_i \cdot E[\bar{Y}_{\cdot i}] = \sum_{i=1}^{k} c_i \cdot \mu_i \rightarrow \text{corretto}$$

$$V[C] = V\left[\sum_{i=1}^{k} c_i \cdot Y_{\cdot i}\right] = \sum_{i=1}^{k} c_i^2 \cdot V[\bar{Y}_{\cdot i}] = \sum_{i=1}^{k} c_i^2 \cdot \frac{\sigma^2}{n}$$

($V[\bar{Y}_{\cdot i}] = \frac{\sigma^2}{n}$)

dove n = n° misure su ogni livello

σ^2 = varianza errore sperimentale

2 fattori $\begin{cases} A \sim k \\ B \sim m \end{cases}$ più repliconcroni (p repliC per ogni coppia)

$$C = \begin{cases} \sum C_i \bar{y}_{\cdot i} & \text{per fatt A} \\ \sum C_j \bar{y}_{\cdot j} & \text{per fatt B} \end{cases}$$

$$\sigma^2_C = \frac{\sigma^2}{np} \sum_{i=1}^{k} C_i^2 \qquad = \frac{\sigma^2}{np} \sum_{j=1}^{m} C_j^2$$

per fatt A per fattore B

gradi libertà? → quelli dell'errore sperimentale

Pregi e difetti di questo test

difetto → utilizzabile solo per F_0 ass. continua

pregio → utilizzabile anche con campioni di numerosità molto piccola ($n = 5, 6 \dots$)

TEST χ^2 per BONTÀ ADATTAMENTO

$H_0: f_x = f_0$

$H_A: f_x \neq f_0$

f_0

- f empirica (istogramma dato campione)

Devo misurare la "distanza"

n_i = frequenza di osservazioni su I_i

np_i = frequenza "teorica" di individui che avrei dovuto osservare su I_i (se H_0 vera)

$$\sum_{i=1}^{N} \frac{(n_i - np_i)^2}{np_i} = W$$

n_i = n° osservazioni del campione

p_i = prob di cadere I_i

n = tot. oss. campione

N = tot di sottointervalli

Teorema

se vale H_0 $(f_x = f_0)$ allora $W \sim \chi^2$

g.l. = $N-1$ se nella f_0 teorica non ci sono parametri stimati con campione

$N-k-1$ se nella f_0 teorica ci sono k parametri stimati con campione

R. accettazione $[0, \chi^2_{1-\alpha}]$

ES (contrasti) Anova, 2 fattori sotto controllo (A - 4 livelli, B - 3 livelli)
2 repliche ciascuno

	DF	SS	MS	F	P-value
A	3	50	15,6		
B	2	210	105	~48	0,0001
Errore	18	40	2,2		
tot	23	300			

$H_0: \mu_1 = \mu_2 = \mu_3$ → rifiuto H_0

$\hat{\mu}_1 = 3.5 \quad \hat{\mu}_2 = 2.1 \quad \hat{\mu}_3 = 2.4$

$A_0: \mu_1 = \frac{1}{3}\mu_2 + \frac{2}{3}\mu_3 + 1$? $\rightarrow \mu_1 - \frac{1}{3}\mu_2 - \frac{2}{3}\mu_3 = 1$

$\bar{c}\left(1, -\frac{1}{3}, -\frac{2}{3}\right)$ stimatore corretto

$\sum c_i = 0$

$\hat{c} = \sum_{i=1}^{3} c_i \hat{\mu}_i = 1 \cdot \hat{\mu}_1 - \frac{1}{3}\hat{\mu}_2 - \frac{2}{3}\hat{\mu}_3 = 1 \cdot 3{,}5 - \frac{1}{3} \cdot 2{,}1 - \frac{2}{3} \cdot 2.4 = 1.2$

Se fosse vera H_0, $E[c] = 1$

$V[c] = \sigma^2 \cdot \sum c_i^2 = \frac{\sigma^2}{8}\left(1 + \frac{1}{9} + \frac{4}{9}\right) = \frac{\sigma^2}{8}\left(\frac{14}{9}\right) = \frac{2{,}2}{8}\left(\frac{14}{9}\right) = 0{,}4$

$2 \cdot 4$ → n° di osservazioni: 4 livelli del fattore A e ho preso 2 misurazioni

$V\left[\sum c_i \hat{\mu}_i\right] =$

$\sigma^2 \rightarrow 2.2$

$T_{calc} = \frac{c_{calc} - E[c]}{\sqrt{V[c]}} = \frac{1.2 - 1.0}{\sqrt{0{,}42}} = 0{,}31$ $\quad \alpha = 0{,}05$

Regione Accettazione? $= (t_{18,1-0,025}, t_{18,0.975})$ ← da output

$t_{18,1-0,025}$ → cerco in tabella

se poi $x_i \sim t$ $(F_i \equiv F)$ MAX + MIN

$F_z(t) = 1 - (1 - F(t))^n$

$f_z(t) = n(1-F(t))^{n-1} \cdot f(t)$ → inversa

$F(t)$ → dist cumulata $F(t) = P[x \le t]$

$\bar{F}(t)$ → funzione di sopravvivenza $\bar{F}(t) = P[x > t]$

$\bar{F}(t) = 1 - F(t)$

funzione di

→ $1 - F_z(t) = (1 - F(t))^n$

$\bar{F}_z(t) = (\bar{F}(t))^n$

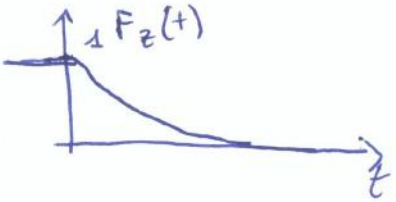
Es 1)

Siano $x_1, x_2, \dots x_n$ indipendenti e tali che $x_i \sim Exp(\lambda_i)$

$z = \min(x_i)$ → $E[z]$

$Y = \max(x_i)$ → $E[Y]$

$F_z(t) = \bar{F}_1(t) \cdot \bar{F}_2(t) \cdots \bar{F}_n(t) = e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} \cdots e^{-\lambda_n t} = e^{-(\lambda_1 + \lambda_2 + \dots \lambda_n)t}$



→ $z \sim Exp(\sum_{i=1}^{n} \lambda_i)$

$E[z] = \frac{1}{\lambda_1 + \lambda_2 + \dots \lambda_n}$

↓

valore atteso del minimo ha valore atteso inferiore a quello degli altri

$Y = \max(x \dots x_n)$

$F_Y(t) = F_1(t) \cdot F_2(t) \cdot \dots F_n(t) \qquad \forall t \geq 0$

$= (1-e^{-\lambda_1 t})(1-e^{-\lambda_2 t}) \dots (1-e^{-\lambda_n t}) = \prod_{i=1}^{n} (1-e^{-\lambda_i t})$

$E[Y] = ?$ → difficile?

Ci limitiamo al caso $n=2$, $\lambda_1 = \lambda_2 = \lambda_3$

$x_1, x_2 \sim Exp(\lambda)$ ind. $Y = \max(x_1, x_2)$

$F_Y = (1-e^{-\lambda t})^2 = 1 - 2e^{-\lambda t} + e^{-2\lambda t} \qquad (t \geq 0)$

$f_Y(t) = 2\lambda e^{-\lambda t} - 2e^{-2\lambda t} \cdot \lambda \qquad (t \geq 0)$

↓ derivo

$E[Y] = \int_0^\infty t \cdot (2\lambda e^{-\lambda t} - 2\lambda e^{-2\lambda t})\, dt = 2\int_0^\infty t \cdot \lambda e^{-\lambda t}\, dt - \int_0^\infty t 2\lambda e^{-2\lambda t}\, dt =$

$= 2 \cdot \frac{1}{\lambda} - \frac{1}{2\lambda} = \left(2 - \frac{1}{2}\right) \cdot \frac{1}{\lambda} = \frac{3}{2} \cdot \frac{1}{\lambda}$

Analisi della varianza

Un contadino vuole valutare l'efficacia di un fertilizzante per coltivazioni di legumi. Sul mercato sono disponibili 4 fertilizzanti prodotti da altrettante ditte. Allo stato attuale, il contadino è convinto che i 4 fertilizzanti sono sostanzialmente equivalenti. ***Come fare per accettare o confutare la sua convinzione??***

Si scelgono 5 appezzamenti di terreno diversi, vengono divisi in 4 parti uguali e ciascuna delle quattro parti viene *trattata* con un fertilizzante diverso. Si seminano poi i legumi ed alla fine della stagione i raccolti risultanti sono:

Campo	*Fertiliz. 1*	*Fertiliz. 2*	*Fertiliz. 3*	*Fertiliz. 4*	*Tot. riga*
I	12.6	13.1	12.7	12.9	51.3
II	6.2	7.0	6.7	7.0	26.9
III	4.8	5.1	4.8	5.2	19.9
IV	3.5	3.8	3.6	3.7	14.6
V	2.9	3.0	2.7	2.9	11.5
Tot. colon.	30	32	30.5	31.7	124.2

Rese (kg/m^2) ottenute negli appezzamenti trattati con fertilizzanti diversi, per ognuno di cinque campi (*)

(*) sono anche indicati i totali per riga, per colonna e generale

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 87 di 386

se mi concentro su una colonna, poiché ho lo stesso numero di dati la distribuzione è appross. ad una normale

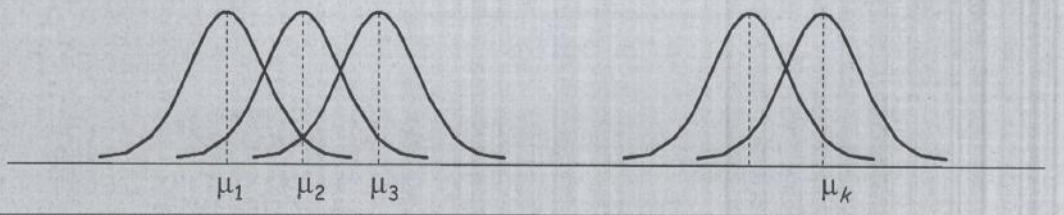
Requisiti:

ognuno dei k gruppi di osservazioni è un campione casuale che proviene da una popolazione ~ $N(\mu_i, \sigma^2)$, cioè la varianza è costante tra i gruppi

costante per tutti → detto ERRORE SPERIMENTALE

H_0: $\mu_1 = \mu_2 = \mu_i = \ldots = \mu_k$ + σ^2 costante equivale a:

I campioni provengono tutti dalla stessa popolazione



Come fare??? e proseguo → cerca meglio

Calcolo 2 varianze, se è soddisfatta la condizione che 1 sia nulla, allora è verificata

Per ogni colonna j, $j = 1, 2, \ldots, k$, si calcoli:

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2$$

Ogni s_j^2, $j = 1, 2, \ldots, k$, è una stima (corretta) della varianza comune σ^2. Ma allora: numerosità — varianza j-esima

$$\hat{\sigma}^2 = s_{pool}^2 \equiv \frac{1}{\sum_{j=1}^{k} (n_j - 1)} \sum_{j=1}^{k} (n_j - 1) \left(\frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \right)$$

stima varianza comune

$$\hat{\sigma}^2 = s_{pool}^2 \equiv \frac{1}{k} \sum_{j=1}^{k} s_j^2 = \frac{1}{k(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2$$ se i campioni hanno lo stesso numero di osservazioni

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 89 di 386

1° STIMA:

$$\hat{\sigma}^2 = \frac{1}{k(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \frac{SS_W}{k(n-1)} = MS_W$$

SS_W = Sum of Squares Within Samples

MS_W = Mean Squares Within Samples

$k(n-1)$ = gradi di libertà (degree of freedom)

I Mean Squares Within Samples (MS_W) sono una stima (sempre corretta, sia o non sia sotto H_0) della varianza σ^2 (incognita) comune alle k popolazioni

$$\frac{SS_W}{\sigma^2} \sim \chi^2_{k(n-1)}$$

2° STIMA:

Ma si può anche calcolare:

$$\hat{\sigma}^2_{\bar{y}} \equiv S^2_{\bar{y}} = \frac{\sum_{j=1}^{k} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}{k-1}$$

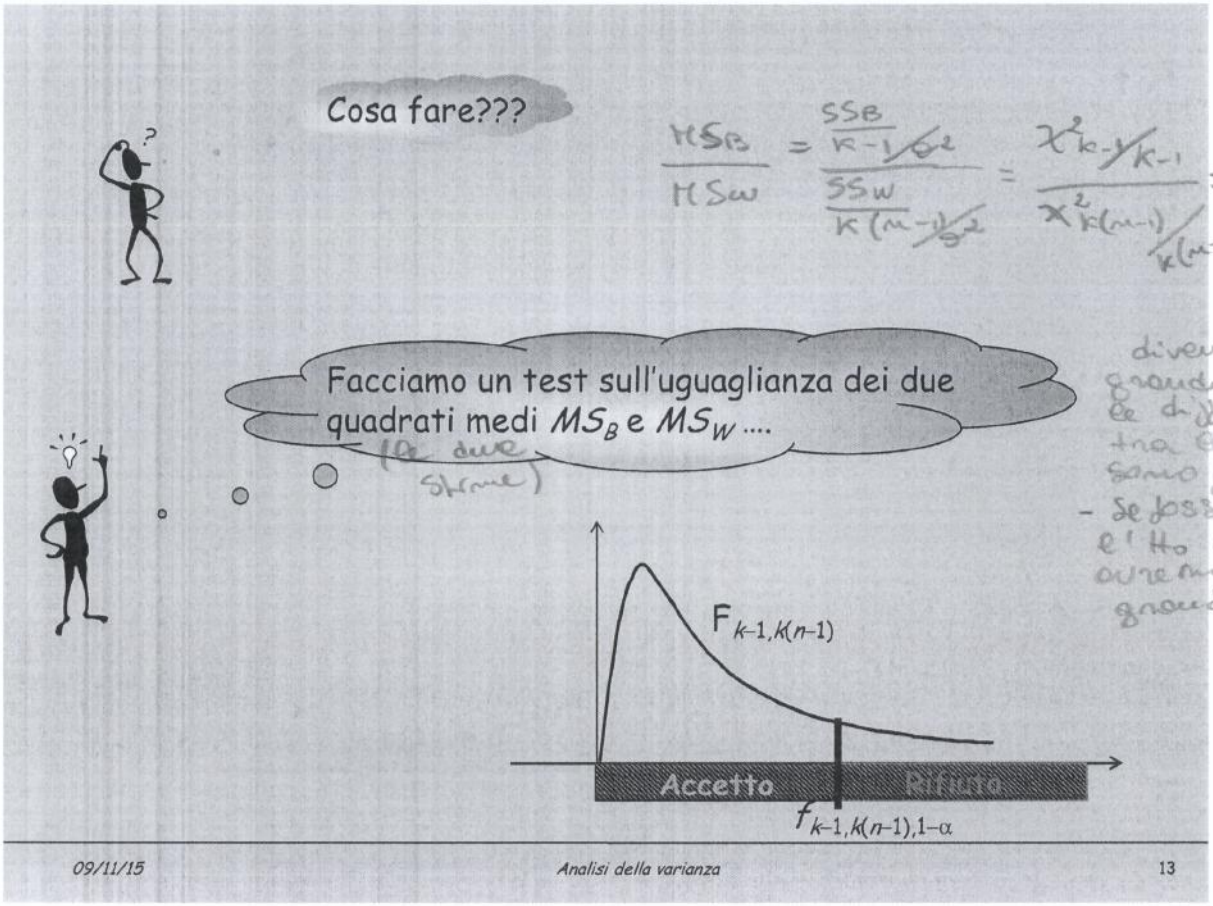
E' una stima corretta solo sotto H_0 della varianza $\sigma^2_{\bar{y}}$ delle k medie delle popolazioni rispetto alla media generale

Ma: $\sigma^2_{\bar{y}} = \frac{\sigma^2_y}{n}$ (nel caso dei fertilizzanti: $s^2_{\bar{y}} = 0.0364$)

$$S^2_y = nS^2_{\bar{y}} = \frac{n\sum_{j=1}^{k} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}{k-1} = \frac{SS_B}{k-1} = MS_B$$

$$\frac{SS_B}{\sigma^2} \sim \chi^2_{k-1}$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 91 di 386



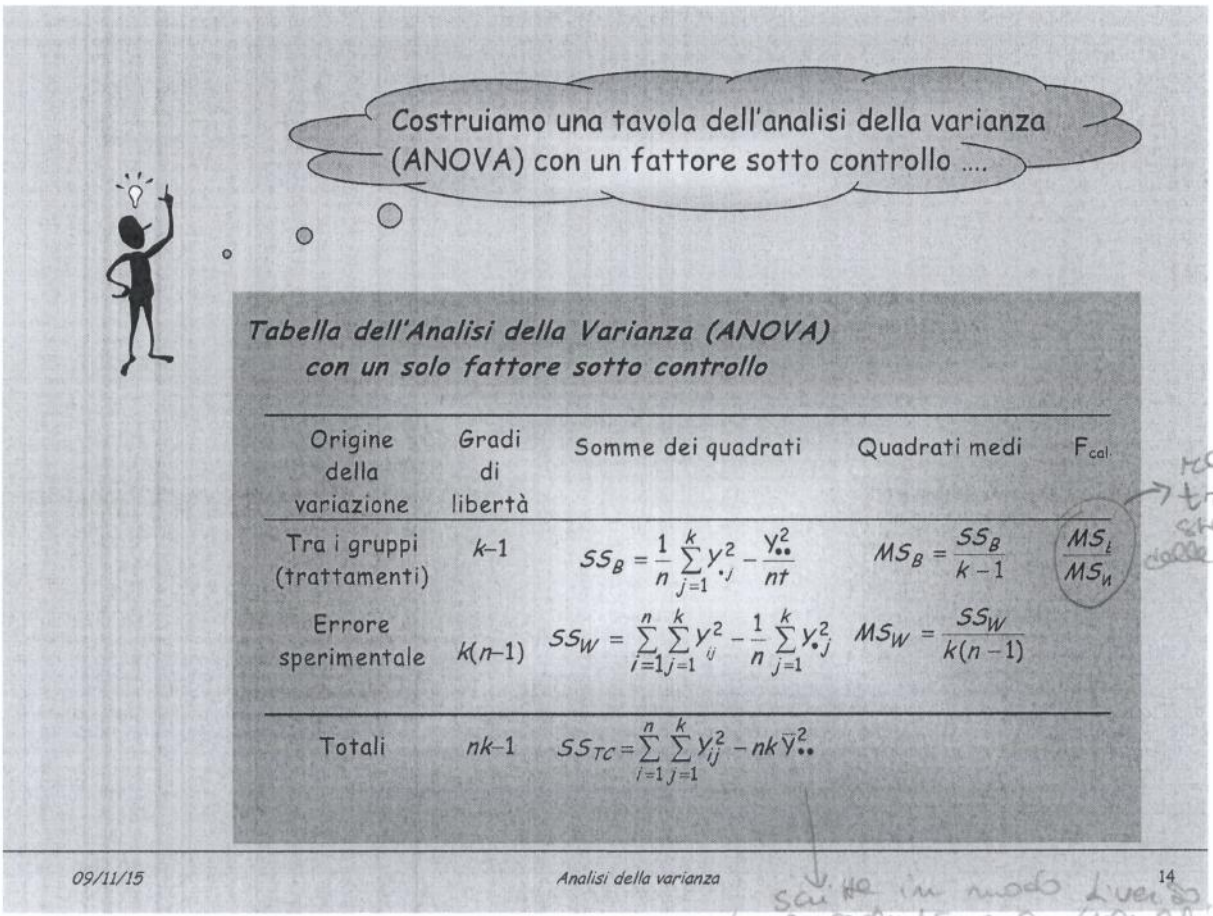


Tabella dell'Analisi della Varianza (ANOVA) con un solo fattore sotto controllo

Origine della variazione	Gradi di libertà	Somme dei quadrati	Quadrati medi	F_{cal}
Tra i gruppi (trattamenti)	$k-1$	$SS_B = \frac{1}{n}\sum_{j=1}^{k} Y_{\cdot j}^2 - \frac{Y_{\cdot\cdot}^2}{nt}$	$MS_B = \frac{SS_B}{k-1}$	$\frac{MS_t}{MS_u}$
Errore sperimentale	$k(n-1)$	$SS_W = \sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}^2 - \frac{1}{n}\sum_{j=1}^{k} Y_{\cdot j}^2$	$MS_W = \frac{SS_W}{k(n-1)}$	
Totali	$nk-1$	$SS_{TC} = \sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}^2 - nk\bar{Y}_{\cdot\cdot}^2$		

Dubbio!!!!

L'errore sperimentale (variazione dovuta al caso) è grande! Non è che è stata attribuita al caso una variazione che invece può essere stata originata (se non tutta, almeno in parte) a qualche altra causa?

	Fertiliz. 1	Fertiliz. 2	Fertiliz. 3	Fertiliz. 4	Tot. riga	Media riga
Varietà 1	12.6	13.1	12.7	12.9	51.3	12.825
Varietà 2	6.2	7.0	6.7	7.0	26.9	6.725
Varietà 3	4.8	5.1	4.8	5.2	19.9	4.975
Varietà 4	3.5	3.8	3.6	3.7	14.6	3.65
Varietà 5	2.9	3.0	2.7	2.9	11.5	2.875
Tot. colon.	30	32	30.5	31.7	124.2	
Media colonna	6	6.4	6.1	6.34		6.21

se sono sbagliato posso rifare il test

Obiettivo: Cercare quale parte di variazione totale è dovuta alla coltivazione di varietà diverse: presenza di un secondo fattore

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 95 di 386

Scomposizione della variabilità:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{\bullet\bullet})^2 = b\sum_{i=1}^{a}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 + a\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ij}-\bar{y}_{\bullet j}-\bar{y}_{i\bullet}+\bar{y}_{\bullet\bullet})^2$$

Somme dei quadrati totali corretti, con $ab-1$ g.d.l.	Somme dei quadrati dovuti al fattore A, con $a-1$ g.d.l.	Somme dei quadrati dovuti al fattore B, con $b-1$ g.d.l.	Somme dei quadrati dovuti ad errori sperimentali, con $(a-1)(b-1)$ g.d.l.

Tabella dell'Analisi della Varianza con due fattori sotto controllo

Origine della variazione	Gradi di libertà	Somme dei quadrati	Quadrati medi	F_{calc}
Tra colonne (fattore A)	$a-1$	$SS_A = \frac{1}{b}\sum_{i=1}^{a} Y_{i\bullet}^2 - \frac{Y_{\bullet\bullet}^2}{ab}$	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_{exp}}$
Tra righe (fattore B)	$b-1$	$SS_B = \frac{1}{a}\sum_{j=1}^{b} Y_{\bullet j}^2 - \frac{Y_{\bullet\bullet}^2}{ab}$	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_{exp}}$
Errore sperimentale	$(a-1)(b-1)$	$SS_{exp} = SS_{TC} - SS_A - SS_B$	$MS_{exp} = \frac{SS_{exp}}{(a-1)(b-1)}$	
Totali	$ab-1$	$SS_{TC} = \sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}^2 - ab\bar{Y}_{\bullet\bullet}^2$		

Se $F_{calc} = \frac{MS_A}{MS_{exp}} > F_{a-1,(a-1)(b-1),1-\alpha}$ si rifiuta l'ipotesi nulla che l'effetto del fattore A sia non significativo , al livello di significatività α

Se $F_{calc} = \frac{MS_B}{MS_{exp}} > F_{b-1,(a-1)(b-1),1-\alpha}$ si rifiuta l'ipotesi nulla che l'effetto del fattore B sia non significativo, al livello di significatività α.