NUMERO: 2238A

ANNO: 2017

A P P U N T I

STUDENTE: Tosti Michela

MATERIA: Statistica - Teoria - Prof. Vicario

INDIPENDENTI:
$P[A \cap B] = P(A) \cdot P(B)$
$P[A \cup B] = P[A] + P[B] + P[A \cap B]$
$P(A) \cdot P(B)$
ESCLUDENTESI:
$P(A \cap B) = \varnothing$
$P[A \cup B] = P[A] + P[B] - P[A \cap B]$
DIPENDENTI
$P[A \cap B] = \neq 0$

PER RICAVARE
$[A\,B] \mapsto$ USO P_{TOTALI}:
$P[A] = P[AB] + P[A\bar{B}]$
$[\bar{A}B]$
$\hookrightarrow P[B] = P[AB] + P[\bar{A}B]$

INDIPENDENTI: $P[A] = P[A/B]$
$P[B] = P[B/A]$

STATISTICA

CAP 1

Definizioni di Probabilità

- **classica** $P[E] = \frac{S}{n}$ — $\frac{n^\circ \text{ risultati favorevoli}}{n^\circ \text{ risultati possibili}}$

ATTENZIONE!
- risultati ugualmente possibili
- incompatibili

- $0 \leq P \leq 1$ (0 ↑ evento impossibile; 1 ↑ evento certo)

- **frequentista**

frequenza evento $f_E = \frac{n_E}{N}$ — $\frac{n^\circ \text{ volte in cui si presenta l'evento}}{n^\circ \text{ esperimenti}}$

↑ Relativa

assume come valore della P di un evento **E**, il valore limite a cui tende la frequenza relativa di quell'evento al tendere del n° di prove all'∞

$$\lim_{N \to \infty} P[|f_E - P[E]| < \varepsilon] = 1$$

- **soggettiva**

Verte sulla fiducia con cui il soggetto ritiene possibile il verificarsi dell'evento

- **assiomatica**

PREMESSE:
- Fenomeno Casuale (aleatorio): un fenomeno empirico il cui risultato non è prevedibile a priori ↦ non ho quindi REGOLARITÀ DETERMINISTICA ma STATISTICA.

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 3 di 58

detto E un qualsiasi evento costituito da $\#E$ punti campione si può definire:

$$\mathbb{P}[E] = \frac{\#E}{\#S}$$

↳ se ho: k componenti: $\rightarrow \frac{\#A_k}{\#S} = \frac{\binom{n}{k} k^k (M-k)^{n-k}}{\binom{M}{n}}$

da cui:

$$\mathbb{P}[A_k] = \frac{\binom{K}{k}\binom{M-K}{n-k}}{\binom{M}{n}}$$

diagramma ad albero

.... ecc...

Probabilità Condizionata

Probabilità che ha un evento A di verificarsi condizionatamente al verificarsi dell'evento B

$$\mathbb{P}[A|B]$$

e si calcola:

$$\mathbb{P}[A|B] = \frac{\#AB}{\#B} = \frac{\mathbb{P}[AB]}{\mathbb{P}[B]}$$

soddisfa tutti e 3 gli assiomi di K.

Probabilità totale

teorema:

Sia $E_1, E_2, \dots E_n$ una collezione di eventi incompatibili tali che $S = \bigcup_{i=1}^{n} E_i$ (eventi esaustivi) e $\mathbb{P}[E_i] \neq 0 \ \forall i = 1 \dots n$ Qualsiasi sia $F \subseteq S$ si ha:

$$\mathbb{P}[F] = \sum_{i=1}^{n} \mathbb{P}[F|E_i]\mathbb{P}[E_i]$$

Formule di Bayes

teorema:

Sia $E_1, E_2 \dots E_n$ una collezione di eventi incompatibili ed esaustivi e sia $\mathbb{P}[E_i] \neq 0$ per $\forall i = 1 \dots n$ Qualsiasi sia $F \subseteq S$ si ha:

$$\mathbb{P}[E_k|F] = \frac{\mathbb{P}[F|E_k]\mathbb{P}[E_k]}{\sum_{i=1}^{n} \mathbb{P}[F|E_i]\mathbb{P}[E_i]} \qquad \forall k = 1 \dots n$$

⇩ se $n = 2$

$$\mathbb{P}[E|F] = \frac{\mathbb{P}[F|E]\mathbb{P}[E]}{\mathbb{P}[F|E]\mathbb{P}[E] + \mathbb{P}[F|\bar{E}]\mathbb{P}[\bar{E}]}$$

Regole per il calcolo della $\mathbb{P}$

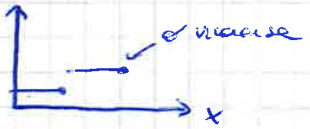
moltiplicativa: se $E_1, E_2 \dots E_n$ sono eventi appartenenti allo stesso spazio degli eventi e $\mathbb{P}[E_1, E_2 \dots E_n] \neq 0$, si ha:

$$\mathbb{P}[E_1 E_2 E_3 \dots E_m] = \mathbb{P}[E_1] \times \mathbb{P}[E_2|E_1] \times \mathbb{P}[E_3|E_2 E_1] \dots \times \mathbb{P}[E_n|E_1 E_2 \dots E_{n-1}]$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 5 di 58

La funzione di distribuzione, spiega come sono distribuiti i valori della variabile casuale, il termine cumulative, talvolte annesso, sta a ricordare che i valori di tale funzione vengono forniti in forma cumulata.
Ricorda: Tale funzione è univocamente definita per ogni variabile casuale.

Si rappresenta:



Caratteristiche di tale funzione sono:

- non negatività
- non decrescenza → $F_X(x_1) \le F_X(x_2)$

$\lim_{x\to-\infty} F_X(x) = 0 \qquad \lim_{x\to+\infty} F_X(x) = 1$

- continua a dx → $\lim_{h\to 0^+} F_X(x+h) = F_X(x)$

Funzione di Densità

Funzione di densità discreta

definizione: Data una variabile casuale discreta X con codominio $\{x_1, x_2 \dots\} \subset \mathbb{R}$ è detta funzione $f_X(x)$ definita da:

$$f_X(x) = \begin{cases} \mathbb{P}[X = x_J] & x = x_J, \text{ con } J = 1 \dots n \\ 0 & x \neq x_J \end{cases}$$

è una funzione da $\mathbb{R}$ nell'intervallo chiuso $[0,1]$ che, per come è definita, gode delle seguenti proprietà:

i) $f_X(x) \ge 0 \quad \forall x \in \mathbb{R}$ (più precisamente $f_X(x) > 0 \ \forall x = x_J$, $f_X(x) = 0 \ x \neq x_J$)

ii) $\sum_{x_J} f_X(x_J) = 1$ dove la sommatoria è estesa a tutti i pi nonne

teorema: la funzione di densità discreta $f_X(x)$ è legata alla funzione di distribuzione cumulativa $F_X(x)$ della stessa variabile casuale X dalle seguenti relazioni:

$$F_X(x) = \sum_{J:\, x_J \le x} f_X(x_J)$$

$$f_X(x) = \begin{cases} F_X(x_J) - \lim_{h\to 0^+} F_X(x_J - h) & \text{se } x = x_J \\ 0 & \forall\, x \neq x_J \end{cases}$$

Funzione di Densità di Probabilità (continua)

Se la variabile X è continua la funzione $f_X(x)$ tale che:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$$

è detta funzione di densità di probabilità di X

Proprietà: (analoghe alla discreta:)

i) $f_X(x) \ge 0, \ \forall x \in \mathbb{R}$

ii) $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$

NOTA: se la funzione di densità discreta ha come codominio l'intervallo $[0,1]$ la funzione di densità di $\mathbb{P}$ $f_X(x)$ ha come codominio $[0,+\infty]$

NOTA 2:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 7 di 58

Disuguaglianza di Tchebycheff

Sia X una variabile casuale e $g(x)$ una funzione non negativa per $\forall x \in \mathbb{R}$, allora:

$$\mathbb{P}[g(x) \geq k] \leq \frac{\mathbb{E}[g(x)]}{k} \quad \forall k > 0$$

Corollario delle disuguaglianze di T:

$$\mathbb{P}[|X-\mu_X| \geq t\sigma_X] = \mathbb{P}[(X-\mu_X)^2 \geq t^2\sigma_X^2] \leq \frac{1}{t^2}$$

in forma equivalente:

$$\mathbb{P}[|X-\mu_X| < t\sigma_X] \geq 1 - \frac{1}{t^2}$$

e anche:

$$\mathbb{P}[\mu_X - t\sigma_X < X < \mu_X + t\sigma_X] \geq 1 - \frac{1}{t^2}$$

↑ è la probabilità che la massa di una qualsiasi variabile casuale X cada dentro le 2 deviazioni standard della sua media

Altre misure di dispersione; momenti

- **quantile**: si definisce quantile q-esimo di x_q di una variabile casuale continua X il più piccolo valore $X \in \mathbb{R}$ tale che $F_X(x_q) = q$

in modo equivalente, il quantile è il + piccolo valore $x \in \mathbb{R}$ tale che l'area sottesa alla curva di densità di $\mathbb{P}$ a sx di tale valore sia uguale a q

(per variabile casuale discrete siccome che il quantile x_q sia il più piccolo N° per cui si ha $F_X(x_q) \geq q$

Mediana: valore che soddisfa:

X: discreta $\quad \mathbb{P}[X \leq med(X)] \geq \frac{1}{2}$ e $\mathbb{P}[X \geq med(X)] \geq \frac{1}{2}$

X continua: $\int_{-\infty}^{med(x)} f_X(x)dx = \int_{med(x)}^{+\infty} f_X(x)dx$

Moda: valore in cui $f_X(x)$ ha il suo max (se lo possiede)

Escursione (range): Differenza tra il valore max e il valore min che X può assumere

Escursione interquantile: Differenza $Q_3 - Q_1$, tra il valore del 3° e del 1° quantile della v.c X

Momento di ordine n di X: valore atteso X^n se esiste; cioè:

$$\mu_n' = \mathbb{E}[X^n]$$

Momento centrale di ordine n rispetto a μ_X: il valore atteso di $(X-\mu_X)^n$ cioè:

$$\mu_n = \mathbb{E}[(X-\mu_X)^n]$$

NOTA:
- $\mu_1 = \mathbb{E}[(X-\mu_X)] = 0$
- $\mu_2 = \mathbb{E}[(X-\mu_X)^2] = \mathrm{var}[X]$
- se la funzione di densità $f_X(x)$ è simmetrica rispetto a μ_X, allora tutti i momenti centrali dispari sono nulli

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 9 di 58

cioè S è un insieme di n-uple in cui l'i-esimo elemento indica il risultato della i-esima prova

$(P[S]=p,\ P[\bar{E}]=1-p)$

La P che si verifichi una determinata n-upla costituita da x successi, e quindi n−x insuccessi, è data da $p^x q^{n-x}$ a causa dell'indipendenza delle prove e dal fatto che la stessa P da successo da prove

Teorema di Bernoulli

Si consideri una variabile casuale X Binomiale di parametri n e p, cioè X rappresenta il numero delle volte che in n prove indipendenti e ripetute si verifica un determinato evento E che ha probabilità p di presentarsi in ciascuna prova. La variabile casuale $V = X - np$ rappresenta lo scarto assoluto della v.c. X dal suo valor medio in quanto $E[x] = np$. Ne consegue che:

$$E[V] = E[X - np] = 0$$
$$\text{var}[V] = \text{var}[X - np] = \text{var}[x] = npq$$

Cioè V è una v.c. che ha la sua distribuzione con punti mass. $-np, -np+1, \dots nq$ con media = 0 e dispersione che cresce al crescere del N° delle prove n

↳ Il fatto che var[v] aumenti al crescere di n significa che all'aumentare del N° delle prove cresce la probabilità di trovare dei valori della variabile X sempre più distanti dal valor medio.

Consideriamo ora la variabile casuale X/n cioè il rapporto tra il n° di successi X in n prove ed il N° stesso delle prove (frequenza relativa)
Applicando le proprietà dei valori attesi:

$$E[X/n] = \frac{1}{n} E[x] = p$$

$$\text{var}[X/n] = \frac{1}{n^2} \text{var}[x] = \frac{pq}{n} = \frac{p(1-p)}{n}$$

Perciò la Frequenza ha una distribuzione con pt nome $0, 1/n, 2/n \dots 1$ valor medio p e dispersione che tende a 0 per $n \to \infty$

Introduciamo ora la v.c. $V/n = (X - np)/n$, essa rappresenta lo scarto relativo della v.c. X rispetto al suo valor medio

$$E[V/n] = E[X/n - p] = \frac{1}{n} E[x] - p = 0$$

$$\text{var}[V/n] = \frac{1}{n^2} \text{var}[v] = \frac{pq}{n}$$

cioè al crescere del numero delle prove diventano sempre più probabili i valori di piccoli scarti relativi, in quanto il valor medio dello scarto relativo è 0 e la sua dispersione decresce al crescere di n.

Si consideri ora il corollario della Disuguaglianza di Tchebycheff, assumendo come v.c. X la Frequenza X/n ricordando che $E[X/n] = p$ e $\sqrt{\text{var}[X/n]} = \sqrt{pq/n}$ si ha:

$$P\left[|X/n - p| < t\sqrt{pq/n}\right] \geq 1 - \frac{1}{t^2}$$

da cui, ponendo $\varepsilon = t\sqrt{pq/n}$ si ha:

$$P\left[|X/n - p| < \varepsilon\right] \geq 1 - \frac{pq}{n\varepsilon^2}$$

che vuol dire che, fissato una quantità piccola a piacere $\varepsilon > 0$, se il N° delle prove tende ad ∞ $P[|X/n - p| < \varepsilon]$ tende all'unità.

↳ Il risultato è il Teorema di Bernoulli ↦ enunciato:

In una serie di n prove, in ciascuna delle quali un evento ha probabilità p di manifestarsi, la probabilità che la differenza tra la frequenza con cui l'evento si è manifestato e p sia, in valore assoluto, inferiore ad una quantità assegnata è infinitamente limitata e tende alla certezza col crescere del numero n delle prove.

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 11 di 58

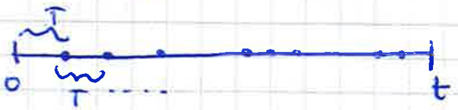
Processi stocastici

es: - code
- temperature e richiesta energia elettrica
- sistemi riconoscimento voce

def: schema astratto di un fenomeno empirico che si sviluppa secondo le leggi probabilistiche

Processi di Poisson

sono per esempio ① n° di difetti per unità di superficie e per unità di lunghezza di un certo prodotto proveniente da una produzione continua
② n° radiazioni emesse per unità di tempo da sostanze radioattive ......



lunghezza = t
T = variabile casuale che identifica ampiezza intervallo tra 2 eventi poissoniani successivi

Hp:
1) esiste quantità $\alpha > 0$ tale che la $\mathbb{P}$ che si verifichi esattamente un evento in un piccolo intervallo di lunghezza Δt sia $\simeq$ uguale a $\alpha \Delta t$ cioè:

$$\mathbb{P}[\text{esattamente 1 evento in } \Delta t] = \alpha \Delta t + o(\Delta t)$$

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

2) la $\mathbb{P}$ che nell'intervallo Δt si verifichi più di un evento è trascurabile rispetto alla $\mathbb{P}$ che se ne verifichi esattamente 1:

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

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

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

se tali Hp sono soddisfatte si ha il teorema:

la variabile casuale $N(t)$ che indica il n° di volte che un evento (per il quale sono soddisfatte le 3 Hp) si verifica in un intervallo qualsiasi di lunghezza t segue una distribuzione di Poisson con parametro $\lambda = \alpha t$ cioè:

$$\mathbb{P}[N(t) = x] = e^{-\alpha t} \frac{(\alpha t)^x}{x!} \quad \text{per } x = 0 \dots n \dots$$

% difetti
x nella binomiale

Distribuzione geometrica

Def: la variabile casuale X ha una distribuzione geometrica se la densità discreta di X è data da:

$$f_X(x) = f_X(x;p) = \begin{cases} p(1-p)^x & x = 0 \dots n \dots \\ 0 & \text{altrove} \end{cases}$$

con $0 < p \leq 1$

Come sempre: è una funzione di densità perché:

1) $f_X(x,p) > 0$
2) $\sum_{x=0}^{\infty} f_X(x,p) = 1$

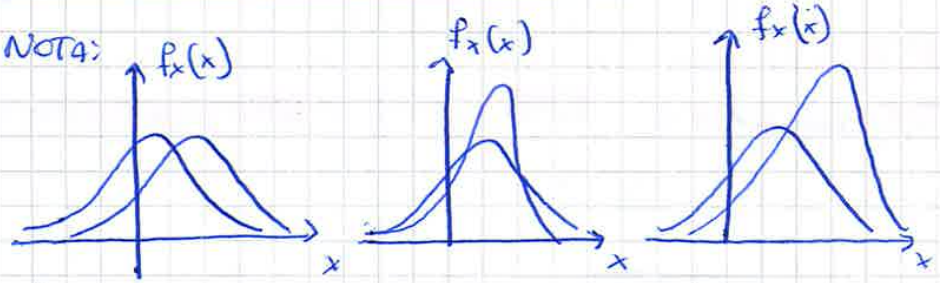
Si ha: $\mathbb{E}[X] = \frac{1-p}{p}$ $\quad var[X] = \frac{1-p}{p^2}$

Teorema: se X è una v.c. con densità geometrica con parametro p si ha:

$$\mathbb{P}[X \geq i + j \mid X \geq i] = \mathbb{P}[X \geq j]$$

con $i, j = 0, 1 \dots n$

NOTA:



STESSA VARIANZA DIVERSE MEDIE — DIVERSA VARIANZA STESSE MEDIE — DIVERSA VARIANZA DIVERSE MEDIE

ATTENZIONE ↦ se $\mu = 0$ e $\text{Var} = 1$ (cioè $\sigma = 1$) DISTRIBUZIONE NORMALE STANDARDIZZATA

↳ SI TRASFORMA la V.C X in Z con $Z = (X - \mu)/\sigma$

→ LA NORMALE è SIMMETRICA rispetto a $x = \mu$ (facendo la STANDARDIZZATA è SIMMETRICA rispetto a $x = 0$)

→ SI HA che $x \to \pm\infty$ la $f_X(x)$ TENDE a 0

LA FUNZIONE DI DISTRIBUZIONE CUMULATIVA NORMALE è:

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(v-\mu)^2/(2\sigma^2)} dv$$

a partire di quanto la ORDINATA con cumulata è più ostica ed ha parametro >.

Si ha:

$$E[X] = \mu \quad \text{e} \quad \text{Var}[X] = \sigma^2$$

se NORMALE STANDARDIZZATA

$$E[Z] = 0 \qquad \text{Var}[Z] = 1$$

→ qualunque sia la V.C (qualunque) le $\mathbb{P}$ che essa appartenga ad un INTERVALLO SIMMETRICO rispetto al valor medio e di semiampiezza 2 volte la DEVIAZIONE STANDARD è circa 95%

↳ INOLTRE la $\mathbb{P}$ che una variabile casuale normale appartenga ad un INTERVALLO centrato sul valore medio e di semiampiezza tre volte la σ è 99,73%.

DISTRIBUZIONE Esponenziale

Def: la V.C. X ha una DISTRIBUZIONE esponenziale (negativa) se la sua funzione di densità di $\mathbb{P}$ è:

$$f_X(x) = f_X(x;\lambda) = \begin{cases} \lambda e^{(-\lambda x)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

con $\lambda \in \mathbb{R}^+$

– è FUNZIONE DI densità probabilità come prima, perché:

1) $f_X(x) > 0$
2) $\int_{-\infty}^{\infty} f_X(x;\lambda)\,dx = 1$

Si ha: $E[X] = \frac{1}{\lambda}$ $\quad \text{Var}[X] = \frac{1}{\lambda^2}$

TEOREMA: Sia X una variabile casuale esponenziale con parametro λ si ha:

$$\mathbb{P}[X > s+t \mid X > s] = \mathbb{P}[X > t] \quad \text{con } s,t \in \mathbb{R}$$

→ SI PUÒ anche DEFINIRE una V.C ↦ T che rappresenta la lunghezza dell'INTERVALLO temporale che INTERCORRE TRA 2 manifestazioni successive dell'evento

$$\mathbb{P}[T > t] = \mathbb{P}[\text{NESSUNA manifestazione IN UN INTERVALLO DI lunghezza } t]$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 15 di 58

Campione: (sottoinsieme della popolazione), è un insieme di n unità statistiche, scelte tra quelle che costituiscono la popolazione, in base a criteri oggettivi

campionamento sistematico: si tratta di scegliere come elemento che andrà a formare il campione, ad esempio ogni 30-esimo pezzo prodotto da una macchina, ogni 20-esimo nome di una lista di nomi, si ha che il **PASSO DI CAMPIONAMENTO** $K=[N/n]$, il più grande intero contenuto in N/n

campionamento stratificato: si divide la popolazione in un numero prestabilito di sottopopolazioni o strati, dalle quali si estraggono delle unità che andranno a costituire il campione totale. Ad esempio, dovendo estrarre un campione di un certo tipo di componenti per valutare le loro difettosità, una stratificazione fossibile potrebbe essere quella di dividere la popolazione costituita dai componenti, in 3 strati corrispondenti a 3 turni lavorativi. La variabilità interna allo strato sarà < della variabilità della caratteristica sotto indagine.

Vantaggio:
- aumenta efficienza perché si riduce l'errore di campionamento senza aumentare la numerosità
- possibilità di analizzare gli strati con diverse precisioni, aumentando la numerosità del campione negli strati in cui si vuole una conoscenza più approfondita **Oversampling** e diminuendo viceversa **Undersampling**

Campionamento con il metodo delle quote
Divido la popolazione in gruppi sulle base della caratteristica (oggetto di studio) per i quali sono noti i pesi % di ognuno nei confronti della popolazione a questo punto vengono definite le quote, cioè il n° di elementi da prelevare da ciascun gruppo. Il campione, sarà l'insieme costituito da tutte le unità estratte.

Campionamento a grappoli: la popolazione viene vista come un insieme di grappoli (clusters) non ricoprentesi, in quanto a si trova delle difficoltà di estrarre un campione dall'intera popolazione, a differenza dei precedenti metodi di campionamento, in questo caso sono i grappoli ad essere scelti in modo casuale tra quelli costituenti la popolazione. Tutte le unità appartenenti ai grappoli scelti vanno a far parte del campione. Va fatto osservare che in tal caso la numerosità del campione, elemento determinante per l'inferenza statistica, è nota solo al termine del campionamento, in quanto la numerosità dei grappoli, in genere non è uguale e non è nota a priori.

Campione fisico	Campione ontico
insieme di 'm' elementi per ognuno dei quali vengono rilevate determinate caratteristiche (individui, oggetti...)	insieme corrispondente delle modalità osservate nel campione fisico

Modalità termine usato per indicare uno dei possibili modi con cui le caratteristica viene rilevata sull'unità statistica

Spunta: modo sintetico di trascrivere una serie di dati consistente nell'inserire mediante un simbolo grafico il n° di ripetizioni osservate per ognuna delle modalità assunte

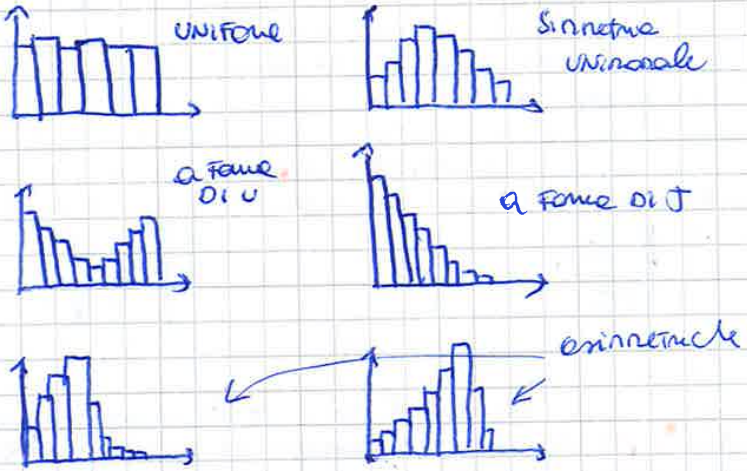
© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 17 di 58

classe corrispondente. INTERPRETAZIONE visiva più semplice, se le classi hanno AMPIEZZA UGUALE IN QUESTO IN TAL caso l'altezza è proporzionale alla FREQUENZA

Se combino 2 classi, la FREQUENZA diventa la somma delle frequenze, ed è uguale all'area del NUOVO rettangolo che quindi sarà più basso e largo.

Frequenza specifica: Si ottiene DIVIDENDO ogni frequenza di classe per la propria AMPIEZZA

Gli ISTOGRAMMI: delineano INDICATIVAMENTE la FORMA della DISTRIBUZIONE della popolazione della quale proviene il campione esaminato.



STEAM AND LEAF PLOT (DIAGRAMMI A FOGLIE E RAMO) rappresentano una variante rispetto agli ISTOGRAMMI, NON SI PERDE PERÒ INFO grazie all'ARTIFICIO di conservare numeri per costruire le barre la SUDDIVISIONE IN classi (di eguale ampiezza), elencate IN ORDINATE, è di regola su base decimale. (VEDI ESEMPIO).
L'aspetto corrisponde a quello di un ISTOGRAMMA, a DIFFERENZA di quello, però, consente di risalire ad ogni singolo valore dell'INSIEME esaminato.

→ la 1ª colonna riporta le FREQUENZE cumulate partendo dai 2 estremi e vi è INDICATA tra parentesi la FREQUENZA corrispondente al RAMO contenente la mediana
→ la 2ª colonna (RAMO) corrisponde alle decine
→ le righe (FOGLIE) ai decimali

La suddivisione IN rami o classi (di eguale ampiezza) elencate nelle II colonne è di regola su base decimale e il N° MAX di rami si determina sempre con la FORMULA del LOGARITMO

STEM-AND-LEAF Display: VOTI

STEM AND LEAF OF VOTI : N = 74
LEAF UNIT = 1.0

1	2	5
2	3	6
4	4	06
5	5	5
8	6	457
10	7	36
13	8	149
16	9	047
21	10	35568
24	11	148
30	12	013469
34	13	2477
(7)	14	3457889 → SIGNIFICATO: 143
33	15	011567899
23	16	0148
21	17	37
19	18	34567
14	19	46
12	20	0356
8	21	338
5	22	2
4	23	0678

quante volte compare il 25 ↑ Cumulate

quante volte compare il 25+36

quante volte compare 25+36+40+46

la cumulata inizia a decrescere dalla mediana 32-1

Ho 74 valori, la mediana è la media tra le 36ª e 37ª osservazione

144 145 147 147 148 149

35-8

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 19 di 58

Si usano solitamente quattro 'cause' principali corrispondenti alle 4 categorie: uomo, macchina, metodo, materiale ("4m"), in corrispondenza di ciascuna di esse si traccia una lisca (queste sono le cause principali), si individuano poi le cause secondarie che influenzano le principali e si disegnano come piccole lische attaccate alla grande lisca a cui si riferiscono. Si completa mettendo la testa del pesce con una caratteristica che possa essere utile.

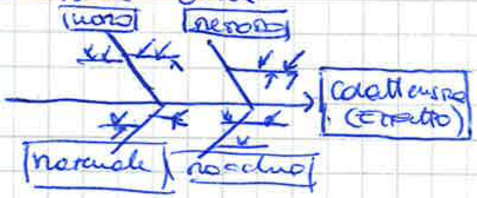


Diagramma a torta

Ogni fetta indica quale frazione del totale corrisponde ad ognuna delle categorie considerate
→ ampiezza settore più è la frequenza assoluta e'

$$\sum_{i=1} n_i = n$$

Pittogramma

Pittogramma: disegno stilizzato dell'elemento sotto indagine usato per costruire il grafo, che può assumere tipologie diverse. Facile lettura, imprecisi.

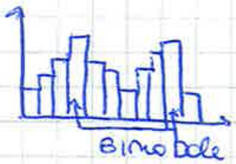
MISURE DESCRITTIVE

Cap 5

Misure di tendenza centrale

Moda: modalità con la frequenza maggiore tra le n osservazioni del fenomeno (può essere: unica, oppure multipla)

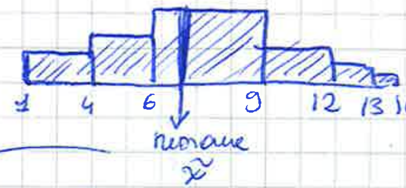
→ se i dati (quantitativi) sono distribuiti in classi si può riconoscere nel relativo istogramma una o più classi modali corrispondenti od altrettanti max locali, classi cioè con freq > rispetto alle altre



Mediana: ($\tilde{x}$ o Md), valore che occupa la posizione centrale (se dati ordinati); se n è dispari è proprio la centrale, se n è pari, è la media aritm. delle centrali

⇒ misure di tendenza di tipo centrale per dati sia qualitativi che quantitativi ma poiché i dati devono essere ordinati non può essere qualitativo sconnesso.

→ se dati raggruppati in classi, la mediana divide a metà l'area totale (frequenza) dell'istogramma



Così semplice se l'ampiezza delle classi è piccola, se così non fosse:

$$\tilde{x} = h_{\tilde{x}} + \frac{a_{\tilde{x}}}{n_{\tilde{x}}}\left(P_{\tilde{x}} - N_{\tilde{x}} - \frac{1}{2}\right)$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 21 di 58

MEDIA GEOMETRICA: $\bar{X}_g = \sqrt[n]{x_1 x_2 x_3 \dots}$

MEDIA ARMONICA: $\bar{X}_a = \dfrac{n}{\frac{1}{x_1} + \frac{1}{x_2} \dots}$

MEDIA QUADRATICA: $\bar{X}_q = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$

↳ media quadratica degli scarti o deviazioni dalla media aritmetica: $x_i - \bar{x}$:

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$ ↦ vedremo poi definizione più appropriata

MEDIA MOBILE: usata nelle analisi di serie temporali allo scopo di rimuovere effetto di fluttuazioni accidentali ed oscillazioni di breve periodo evidenziando invece le tendenze a carattere sistematico nel medio e lungo termine.

$$U_{t+h} = \sum_{j=0}^{k} p_j x_{t+j} \qquad t = 1, \dots, h-k$$

Ma, quale misura di tendenza usare?

⇨ DATI QUALITATIVI ↦ classe

⇨ QUANTITATIVI? _ media se pochi casi specialmente o se sono ammassati verso l'estremo o se vi è rischio di valori anomali

(in questi casi media aritmetica scarsamente significativa perché influenzata dagli estremi)

NOTA: MEDIA, MODA, MEDIANA rappresentano le principali misure di posizione

↳ se DISTRIBUZIONE SIMMETRICA e UNIMODALE coincidono

↳ " " FORTEMENTE ASIMMETRICA, la media può risultare troppo influenzata dai valori estremi per individuare opportunamente la tendenza centrale della distribuzione

MISURE DI DISPERSIONE (o variabilità)

deviazioni (o scarti) dalla media: $x_1 - \bar{x}$, $x_2 - \bar{x} \dots$ (dove $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$)

↳ non possiamo usare il loro valore come misura di variabilità in quanto $= \emptyset$

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

Possiamo però ricorrere alla: deviazione assoluta media

non comunemente matematicamente ↦ $\frac{1}{n} \sum_{i=1}^{k} n_i |x_i - \bar{x}|$

SCARTO QUADRATICO MEDIO (o DEVIAZIONE STANDARD)

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$

misuriamo la variabilità intorno al campione dei dati info sulla dispersione dei dati attorno alla media aritmetica

↳ varianza → è lo 's' al quadrato

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

Proprietà: se i dati campionari $\{x_1 \dots x_n\}$ vengono trasformati mediante trasf. lineari cioè, se $y_i = a + b x_i$ $(i = 1 \dots n)$ allora la varianza dei trasformati è uguale a quella originale moltiplicata per b^2

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 23 di 58

INDIPENDENZA E CONNESSIONE

Tabelle a doppia entrata

Cap 6

Frequenza congiunta (ASSOLUTA) di una combinazione (x_i, y_j) di due modalità, il n° di volte che compaiono le 2 caratteristiche X e Y secondo le modalità x_i, y_j e la si indicherà con n_{ij}.

$$\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = \sum_{j=1}^{k} n_{\bullet j} = \sum_{i=1}^{h} n_{i\bullet} = n$$

Frequenza congiunta relativa: f_{ij}, si ottiene dividendo la frequenza assoluta per il totale

$$f_{ij} = n_{ij}/n$$

Si ottiene sommando tutte le frequenze congiunte corrispon. a quelle coppie di modalità che contengono quella fissata.

$$n_{i\bullet} = \sum_{j=1}^{k} n_{ij} \quad e \quad n_{\bullet j} = \sum_{i=1}^{h} n_{ij}$$

la Freq Assoluta di una delle caratteristiche osservate in un fenomeno casuale nel quale vengono rilevate contemporaneamente 2 o + caratteristiche è detta: FREQUENZA MARGINALE (ASSOLUTA)

Freq. marginale relativa si ottiene dividendo quella assoluta x il tot.

$$f_{i\bullet} = n_{i\bullet}/n \quad f_{\bullet j} = n_{\bullet j}/n$$

↳ le frequenze marginali riguardano una sola caratteristica, ci sono anche le Frequenze condizionate di una variabile date le rimanenti, che riguardano più caratteristiche

OSSERVAZIONE: ↳ Siccome le frequenze marginali e le condizionate sono distribuzioni di frequenza, si potranno definire le solite misure descrittive (μ, σ) a seconda delle tipologie di caratteristiche rilevate

↦ se non vi è alcun legame si parla di INDIPENDENZA tra le variabili che rappresentano le caratteristiche sotto indagine

↳ Se si hanno 2 sole variabili, esse sono INDIPENDENTI se la distribuzione di frequenza della variabile Y condizionata a $X = x_i$ non cambia al variare di x_i con $i = 1 \dots h$ (idem X condiz a $Y = y_i$) data la simmetria della proprietà di indipendenza.

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n} \quad o \quad \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n}$$

quindi se le due variabili sono INDIPENDENTI, la frequenza relativa congiunta è uguale al prodotto delle 2 FREQUENZE RELATIVE marginali

↳ Anche in presenza di INDIPENDENZA si avranno scostamenti tra le frequenze congiunte osservate e quelle espresse dalle formule appena dette frequenze teoriche, tali scostamenti vengono chiamati CONTINGENZE (Assolute)

$$c_{ij} = n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n} \qquad \forall i = 1 \dots h \quad j = 1 \dots k$$

Queste definite come la differenza tra le frequenze assolute osservate e quelle teoriche.

↳ Si usano indici che sono MEDIE ASSOLUTE (INDICE DI CONNESSIONE DI MONTFORD) e QUADRATICHE (INDICE DI CONNESSIONE DI PEARSON) IN FORMA ASSOLUTA o NORMALIZZATI.

si chiamano

↳ INDICI DI CONNESSIONE IN QUANTO SI INDICA la presenza di un legame tra le 2 caratteristiche sotto indagine

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 25 di 58

Funzione interpolante ↦ spesso è una retta

$y = a + bx$, $\quad a, b \in \mathbb{R}$

Il metodo dei minimi quadrati richiede che sia

$$\min_{a,b} Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

⇩

Mediante una dimostrazione si ottiene:

$$a = \bar{y} - b\bar{x}$$

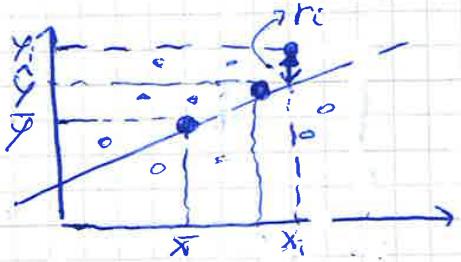
$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

Perciò l'equazione della retta interpolante la funzione di regressione è:

$$y = \bar{y} + \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

↦ Coeff angolare b strettamente legato alla covarianza tra le 2 variabili, perciò la retta è crescente se $cov > 0$, decrescente se $cov < 0$

RESIDUO $y_i - (a + bx_i)$ $\quad i = 1 \dots n$ è la differenza tra i valori osservati e i valori in ordine della funzione interpolante



↦ la somma dei residui è sempre $= 0$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \bar{y}) - b \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

con $r = \frac{S_{x,y}}{S_x S_y}$ COEFFICIENTE di correlazione lineare

La covarianza misura perciò una eventuale relaz. lineare tra le 2 variabili sotto indagine

$cov \simeq 0$ ↦ assenza di relazione lineare ↦ VARIABILI NON correlate

assenza di legame lineare tra le 2 variabili che NON implica l'assenza di ogni tipo di relazione

↦ INDIPENDENZA implica NON CORRELAZ.
↦ NON correlaz. NON implica INDIPENDENZA

$$r^2 = 1 - \frac{var[y - (a + bx)]}{var[y]} = 1 - \frac{\text{varianza residua di interpolazione}}{\text{varianza tot } y}$$

⇧

COEFFICIENTE DI DETERMINAZIONE lineare ↦ spiega la proporzione di riduzione della varianza tot, dovuta alla retta interpolante

$0 \le r^2 \le 1$ $\quad$ e $-1 \le r \le 1$

se $r^2 = r = S_{x,y} = 0$ la retta interpolante non apporta alcuna riduzione alla varianza residua ↦ Retta interp $y = \bar{y}$ e x e y non correlate

Se $r^2 = 1$ $(r = \pm 1)$ la varianza residua è nulla quindi tutti i punti (x_i, y_i) giacciono sulla retta interpolante ↦ tra x e y correlazione perfetta

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 27 di 58

DEFINIZIONE 3: Dato un campione $\{X_1, X_2 \dots X_n\}$ proveniente da una popolazione con densità $f(x)$ si definisce media campionaria la statistica

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

e coincide con la media aritmetica delle variabili che costituiscono il campione.

TEOREMA: Dato un campione $\{X_1 \dots X_n\}$ proveniente da una popolazione avente densità $f(x)$, posto $\bar{X}_n = \frac{1}{n}\sum_i X_i$ si ha:

$$\mathbb{E}[\bar{X}_n] = \mu \qquad \mathrm{Var}[\bar{X}_n] = \frac{\sigma^2}{n}$$

→ media e varianza di $f(x)$ stessa della

Si dice che, qualunque sia la distribuzione della media campionaria, essa ha μ = μ popolazione e σ = inversamente proporzionale allo $\sqrt{n}$. Questo significa che al crescere della numerosità del campione, i valori delle corrispondenti medie campionarie tendono a concentrarsi sempre di più attorno al loro valore medio che è la media della popolazione.

⇨ Da Tchebycheff si ha: **LEGGE dei GRANDI NUMERI**

$$\mathbb{P}\left[|\bar{x}_n - \mu| < t\sqrt{\frac{\sigma^2}{n}}\right] \geq 1 - \frac{1}{t^2} \Rightarrow \lim_{n\to\infty} \mathbb{P}[|\bar{X}_n - \mu| < \varepsilon] = 1$$

DEFINIZIONE 4 ↦ **VARIANZA CAMPIONARIA**

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

TEOREMA: Dato un campione $\{X_1 \dots X_n\}$ proveniente da una popolazione con $f(x)$, e posto S_n^2 si ha:

$$\mathbb{E}[S_n^2] = \sigma^2 \qquad \mathrm{Var}[S_n^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

con μ_4 e σ^2 rispettivamente il momento di ordine 4 e la varianza di $f(x)$

DISTRIBUZIONE MEDIA CAMPIONARIA

Teorema (del limite centrale) 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 esse, allora la media campionaria segue una distribuzione normale, al tendere di $n \to \infty$

Si può dire in modo equivalente che, detta Z la standardizzazione di $\bar{X}_n$, Z segue la distribuzione normale standardizzata:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

con $\mathbb{E}[\bar{X}_n] = \mu$
$\mathrm{Var}[\bar{X}_n] = \frac{\sigma^2}{n}$

Si vede graficamente che se il campione cresce ($n \to \infty$) la forma della distribuzione delle medie campionarie tende a quella di una normale.

Osservazione
Se la distribuzione della popolazione è normale, la distribuzione delle medie campionarie è normale, con qualsiasi dimensione campionaria; infatti, la media campionaria è una particolare combo lineare e qualsiasi combinazione lineare di v.c normali, indipendenti, segue una normale

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 29 di 58

DISTRIBUZIONE DI FISHER

DEF: Siano U e V_2 VC INDIPENDENTI aventi entrambe DISTRIBUZIONE χ^2 con m e n GDL la V.C:

$$F = \frac{U/m}{V/n}$$

ha una DISTRIBUZIONE F DI FISHER con m e n gdl ovvero $F_{m,n}$

- m: GDL sopra nella frazione (il 1° che compare)
- n: GDL sotto nella frazione (il 2° che compare)

TEOREMA: Se $\{X_1 \dots X_m\}$ è un campione casuale DI DIMENSIONE m estratto da una popolazione DISTRIBUITA normalmente con MEDIA μ_X e varianza σ^2 e $\{Y_1 \dots Y_n\}$ è un campione casuale DI DIMENSIONE n estratto da una popolazione DISTRIBUITA normalmente con media μ_X e varianza σ^2 e se i 2 campioni sono INDIPENDENTI, allora la V.C:

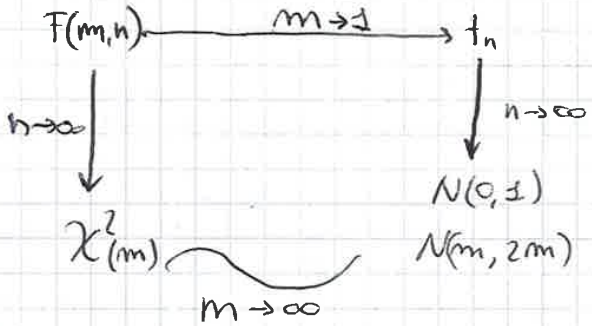
$$\frac{\sum_{i=1}^{m} (X_i - \bar{X}_m)^2 / (m-1)}{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 / (n-1)}$$

↳ segue una distribuzione $F_{m-1,\,n-1}$ (GDL, GDL)

↳ se voglio i percentili usati si applica

$$f_{m,n,1-\alpha} = \frac{1}{f_{n,m,\alpha}}$$

SCHEMA RIASSUNTIVO



STIMATORI PUNTUALI

METODO DEI MOMENTI Data una popolazione distribuito secondo la funzione di densità $f(\cdot, \theta_1 \dots \theta_k)$ se i momenti $\mu_r' = E[X^r] = \mu_r'(\cdot, \theta_1 \dots \theta_k)$ della variabile casuale $X \sim f(\cdot, \theta_1 \dots \theta_k)$, esistono finiti in n° uguale a quello dei parametri θ_i che si desiderano stimare, allora il METODO dei momenti consiste nel risolvere rispetto a $\theta_1 \dots \theta_k$ il sistema:

$$\begin{cases} \mu_1'(\theta_1 \dots \theta_k) = M_1' \\ \mu_2'(\theta_1 \dots \theta_k) = M_2' \\ \vdots \\ \mu_k'(\theta_1 \dots \theta_k) = M_k' \end{cases}$$

costituito delle prime k equazioni che si ottengono eguagliando i momenti campionari

$$M_r' = \frac{1}{n} \sum_{i=1}^{n} X_i^r$$ generati dal campione $\{x_1 \dots x_n\}$ provenienti da $f(\cdot, \theta_1 \dots \theta_n)$ ai momenti μ_r'

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 31 di 58

loro varianze, è quindi da preferirsi lo stimatore con varianza minore perciò:

$$eff(\bar{T}_1|T_2) = \frac{Var[T_2]}{Var[T_1]}$$ se < 1

Definizione: uno stimatore $T=(X_1 \dots X_n)$ del parametro θ si dice **consistente in media quadratica** per il parametro θ se:

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

Un esempio è la media campionaria $\bar{X}_n$ è uno stimatore corretto e consistente in media quadratica della media μ della popolazione da cui proviene il campione poiché $Var[\bar{X}_n] = \sigma^2/n \to 0$ per $n\to\infty$

Definizione: Uno stimatore $T=t(X_1 \dots X_n)$ del parametro θ si dice **consistente in probabilità** se converge in probabilità al parametro θ, cioè se per $\forall \varepsilon > 0$ si ha

$$\lim_{n\to\infty} P[|T-\theta| < \varepsilon] = 1$$

Dove la dipendenza dell'argomento del limite da n deriva dalla dipendenza dello stimatore dalla numerosità del campione

Osservazione 1 la consistenza in media quadratica implica la consistenza in probabilità

Osservazione 2 la media campionaria $\bar{X}_n$ nella classe degli stimatori lineari è il migliore stimatore per la media della popolazione

Osservazione 3 la varianza campionaria S_n^2, oltre ad essere uno stimatore corretto della varianza σ^2 della popolazione da cui proviene il campione è anche consistente

Stima per intervalli

Definizione: Si definisce intervallo di fiducia per il parametro θ un intervallo casuale che include il vero valore del parametro con una prefissata P chiamato livello di fiducia

In altre parole si tratta di associare un prefissato grado di P all'evento che il valore stimato del parametro si discosti dal suo valore vero per meno di una certa quantità, e in modo equivalente:

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

dove L_i = limite inferiore, L_s = limite superiore, $1-\alpha$ = livello di fiducia ↑ intervallo di fiducia al livello $1-\alpha$

- può essere unilaterale (inferiore / superiore)
- Bilaterale (simmetrico)

Stima per intervalli della media

Caso: varianza popolazione nota

Sia $\{X_1 \dots X_n\}$ campione estratto da una popolazione avente media μ incognita e varianza σ^2 nota

Stimatore più efficiente per la media:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Si è visto che $E[\bar{X}_n] = \mu$ $Var[\bar{X}_n] = \sigma^2/n$, inoltre $\bar{X}_n$ segue una distribuzione normale con parametri μ e σ^2/n dove la v. c:

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$ segue la distribuzione normale standardizzata con parametri $E[Z]=0$, $\sigma_Z^2 = 1$

Possiamo scrivere che

$$P[-z_{\alpha/2} < Z \leq z_{1-\alpha/2}] = 1-\alpha$$

da cui dopo opportuni calcoli si ottiene

$$P\left[\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu \leq \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 33 di 58

Supponiamo $\{X_{11} \dots X_{1n_1}\}$ campione casuale di dimensione n_1 estratto da una popolazione distribuita normalmente con media μ_1 e varianza σ_1^2 e che $\{X_{21}, X_{22} \dots X_{2n_2}\}$ sia anche esso (casuale)

Il migliore stimatore x la differenza delle medie $\mu_1 - \mu_2$ è:

$$\bar{X}_1 - \bar{X}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} \qquad \text{STIMATORE LINEARE}$$

Per le varianze:

a) caso con varianze NOTE

$$\sigma_\Delta^2 = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} \mapsto Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_\Delta}$$

$$P\left[-z_{1-\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_\Delta} < z_{1-\alpha/2}\right] \text{ si ottiene}$$

$$(L_1, L_S) = \left[(\bar{X}_1 - \bar{X}_2) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

b) caso con varianze NON NOTE ma ritenibili =

OGNUNO dei 2 campioni può essere usato per ottenere una STIMA per la varianza comune σ^2

I 2 stimatori corretti per σ_1^2 e σ_2^2 sono:

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 \quad \text{per STIMARE } \sigma_1^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2 \quad \text{per STIMARE } \sigma_2^2$$

Siccome S_1^2 e S_2^2 sono entrambi stimatori corretti, possiamo usarli entrambi per ottenere la stima per σ^2 data da:

$$S_{pool}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

$$\uparrow$$

$$\delta_{pool}^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Ne consegue che la V.C

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \mapsto n_1 + n_2 - 2 \text{ GDL}$$

segue la distribuzione t di STUDENT con $n_1 + n_2 - 2$ GDL

quindi si può affermare che

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 35 di 58

STIMA PER INTERVALLI del RAPPORTO DI 2 VARIANZE (CONFRONTO DI 2 VARIANZE)

$$F = \frac{S_1^2}{\sigma_1^2} \Big/ \frac{S_2^2}{\sigma_2^2}$$

↑ (n_1-1, n_2-1) GDL

DI SOLITO $f_{(n-1),(n-2),\alpha/2}^{m\ n}$ ↦ NON lo trovo sulle tavole ↳ DELLA FISHER

$$f_{n,m,\alpha} \rightsquigarrow = \frac{1}{f_{m,n,1-\alpha/2}}$$

$$P\left[\frac{S_1^2}{S_2^2} f_{n_1-1,n_2-1,\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{S_2^2}{S_1^2} f_{n_1-1,n_2-1,1-\alpha/2}\right] = 1-\alpha$$

↳ $(L_1 L_2) = \left(\frac{S_1^2}{S_2^2} f_{n_2-1,n_2-1,\alpha/2}, \frac{S_2^2}{S_1^2} f_{n_1-1,n_2-1,1-\alpha/2}\right)$

STIMA PER INTERVALLI delle proporzioni (INTERVALLI ASINTOTICI)

$$Z = \frac{P-p_0}{\sqrt{\frac{P(1-P)}{n}}}$$

↳ $\left(P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}},\ P + z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\right)$

NOTA: Campioni provenienti dalla stessa popolazione se: le σ^2 possono ritenersi uguali ovvero se la STATISTICA $\frac{\sigma_2^2}{\sigma_1} = 1$ e se l'INTERVALLO delle μ CONTIENE lo 0 (ovvero $\mu_2 - \mu_1 = 0$)

Cap 9 INFERENZA STATISTICA (TEST D'IPOTESI)

BASI LOGICHE dei TEST

Si suppone UN NUOVO TRATTAMENTO che RENDE RESISTENTE a TRAZIONE 80 N (IN TEORIA)

⇝ Si PRELEVANO 49 pezzi (campione)
se CARICO MEDIO ROTTURA > 78,5 ACCETTO l'AFFERMAZIONE
se " " " < " RIFIUTO "

Si presentano 4 situazioni possibili:

1) CARICO ROTTURA MEDIO > 78,5 ↦ ACCETTO NUOVO TRATTAM.
2) CARICO ROTTURA MEDIO < 78,5, ↦ RIFIUTO " "
   ↳ FACCIO IN REALTÀ IL TRATTAMENTO FUNZIONA
   ERRORE DI 1ª SPECIE
3) CARICO ROTTURA MEDIO > 78,5 ↦ ACCETTO " "
   ↳ NON IN REALTÀ NON FUNZIONA
   ERRORE DI 2ª SPECIE
4) CARICO ROTTURA MEDIO < 78,5 ↦ RIFIUTO " "

Si suppone anche che la DISTRIBUZIONE sia una normale con $\sigma^2 = 21,4\ N^2$ e non cambia x effetto del nuovo trattamento

[L]posso valutare P ERRORE I e II SPECIE!

Infatti date le assunzioni ⇝ $\mu = 80$ $\sigma_{\bar{x}} = \sqrt{\frac{21,4}{49}} = 0,66$

Il rischio errore I SPECIE rappresentato dall'area sottesa dalla curva NORMALE a SX DI 78,5 N (equivale ad avere MEDIE dei campioni DI $n=49$ < DI 78,5 N: (opportuni calcoli e ottengo P = 1%)

Ora: DISTRIBUZIONE $\mu = 79,5$ $\sigma_{\bar{x}} = 0,66$

P errore II specie è quello di avere μ dei campioni DI $n=49$ > DI 78,5 N ↦ nella curva normale area sottesa a DX del valore 78,5

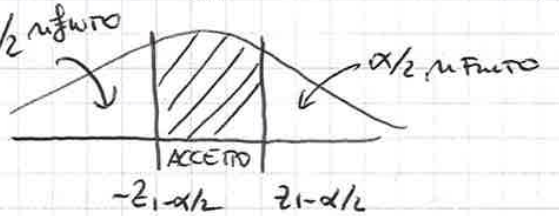
© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 37 di 58

Il problema è quello di testare l'Hp nulla $H_0: \mu = \mu_0$ contro l'Hp alternativa $H_A: \mu_0 \neq \mu_0$ (test bilaterale)
l'Hp nulla H_0 è da rifiutare se il valore ossevato di Z è > di $z_{1-\alpha/2}$ o minore di $-z_{1-\alpha/2}$

$z_{1-\alpha/2}$ quantile tale x cui l'area sottesa dalla curva standard dist. alla sua DX è pari a $\alpha/2$

NON
REGIONE DI RIFIUTO:

$$I_{NL_S} = \left(\mu_0 - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \; ; \; \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

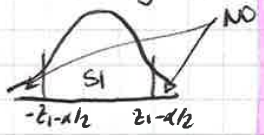
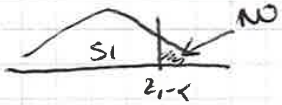
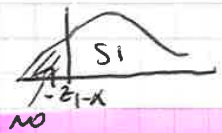


NOTA: oltre al test bilaterale esiste anche l'UNILATERALE superiore e INFERIORE

⇓

rispetto al bilaterale sono vantaggiosi x gli errori di II specie

Tab riassuntiva

H_0	H_A	Regione rifiuto H_0	Schema grafico
$H_0: \mu = \mu_0$ Bilaterale	$H_A: \mu \neq \mu_0$	$z > z_{1-\alpha/2}$ o $z < -z_{1-\alpha/2}$	
$H_0: \mu \geq \mu_0$ normale superiore	$H_A: \mu > \mu_0$	$z > z_{1-\alpha}$	
$H_0: \mu = \mu_0$ normale INFERIORE	$H_A: \mu < \mu_0$	$z < -z_{1-\alpha}$	

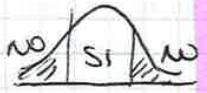
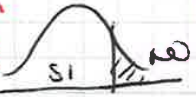
Media, Varianza incognite

Considero campione casuale $\{X_1 \dots X_n\}$ di dimensione n proveniente da una distribuzione normale, volendo sottopone a test l'Hp che proviene da una popolazione con media $\mu = \mu_0$ e σ^2 non nota:

$$T = \frac{\bar{X}_n - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{\bar{X}_n - \mu_0}{S} \sqrt{n}$$

↗ T STUDENT (n-1) GDL

Tabella

H_0	H_A	Rifiuto H_0	Schema grafico
$H_0: \mu = \mu_0$	$H_A: \mu \neq \mu_0$	$T > t_{n-1, 1-\alpha/2}$ o $T < -t_{n-1, 1-\alpha/2}$ (se non c'è $-t_{n-1, \alpha/2}$)	
$H_0: \mu = \mu_0$	$H_A: \mu > \mu_0$	$T > t_{n-1, 1-\alpha}$ (se non c'è $-t_{n-1, \alpha}$)	
$H_0: \mu = \mu_0$	$H_A: \mu < \mu_0$	$T < -t_{n-1, \alpha}$	

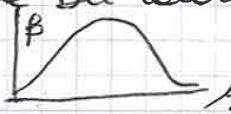
CURVA CARATTERISTICA OPERATIVA E POTENZA DEL TEST

Procedimento per il calcolo del rischio di errore di II specie (ipotesi in toto l'ipotesi nulla)

$$\beta = P\left[\bar{x}_{min} < \bar{X}_n \leq \bar{x}_{max} \,\middle|\, H_A : \mu \right]$$

dato che vera e H_0 è falsa

⇒ importando su di un diagramma cartesiano i valori β (α-β) in funzione dei valori della H_A si ottiene il grafico della CURVA CARATTERISTICA OPERATIVA



TEST D'IP sulle proporzioni (ASINTOTICO)

- Campione BERNOULLIANO vs se campione elevato (P. RITENIAMO approssimare)

H_0: $p = p_0$ con livello di significatività α.

STATISTICA: $Z = \dfrac{\hat{P} - p_0}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}}$ che segue DISTRIB NORMALE STANDARD

se $|z_{calc}| > z_{1-\alpha/2}$ SI RIFIUTA H_0 a favore di H_a

Se SOTTOPONGO IL TEST: $H_0: P_1 - P_2$
(con n_1 e n_2 due campioni INDIPENDENTI DI 2 popolaz)

SI COSTRUISCONO I 2 STIMATORI DI MAX VEROSIMIGLIANZA $\hat{P}_1$ e $\hat{P}_2$. LO STIMATORE DIFFERENZA $\hat{P} = \hat{P}_1 - \hat{P}_2$ essendo la DIFF. DI 2 V.C ASINTOTICAMENTE NORMALI ED INDIPENDENTI segue ancora UNA DISTRIBUZIONE NORMALE con MEDIA $P_1 - P_2$ e VARIANZA $\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$

Se H_0 vera ($P_1 = P_2 = P$) si ha:

$$\hat{P} \sim N\left(0, \sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

RIFIUTO H_0 se $|z_{calc}| > z_{1-\alpha/2}$

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 41 di 58

Regressione

OUTPUT

Results for: dati_A

Regression Analysis: Y1 versus X1

Analisi di regressione con 1 solo variabile indipendente → X1

The regression equation is:

$Y1 = 0,235 + 0,932\ X1$

Quello dello stimatore l' H0 di prenderli = 0 ← P-value ≃ 0 per entrambi coefficienti altamente significativi

Predictor	Coef	SE Coef	T	P
Constant	0,23471	0,08288	2,83	0,006
X_1	0,93224	0,02758	33,80	0,000

$H_0: \beta_0 = 0$
$H_1: \beta_1 = 0$

Stima della radice quadrata della varianza residua $\hat{\sigma}$

S = 0,300090   R-Sq = 93,2%   R-Sq(adj) = 93,1%

$R^2 = 93,2$, ottima percentuale di variabilità spiegata dal modello che ho adottato cioè la retta

Analysis of variance:

Source	DF (n-2)	SS	MS	F	P
Regression	1	102,89	102,89	1142,50	0,000
Residual Error	84	7,56	0,09		
Lack of fit	79	7,51	0,10	9,16	0,010
Pure Error	5	0,05	0,01		
Total	85	110,45			

$H_0: \beta_1 = 0$ comp. riguard. al fatto che nel modello abbia introdotto β_0

77 rows with no replicates

P-value → 1%
comincio ad avere dubbi che il mio modello sia adeguato

livello di significatività ← se avessi preso 5% come livello di significatività, rifiuterei l'adeguatezza del modello

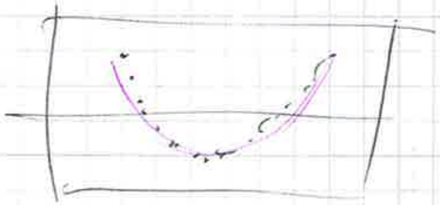
© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 43 di 58

Ma non ci dice qual è il nuovo modello da adottare.

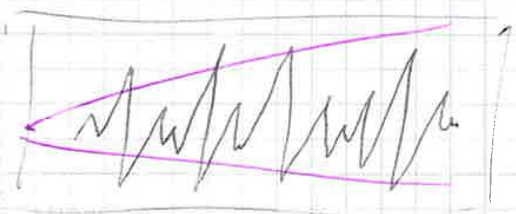
↳ il nuovo modello da adottare è:

↳ devo mettere un termine del 2° ordine e nei grafici si evidenzia un

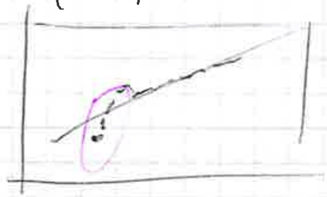
↳ termine del 2° ordine nel leverage fits



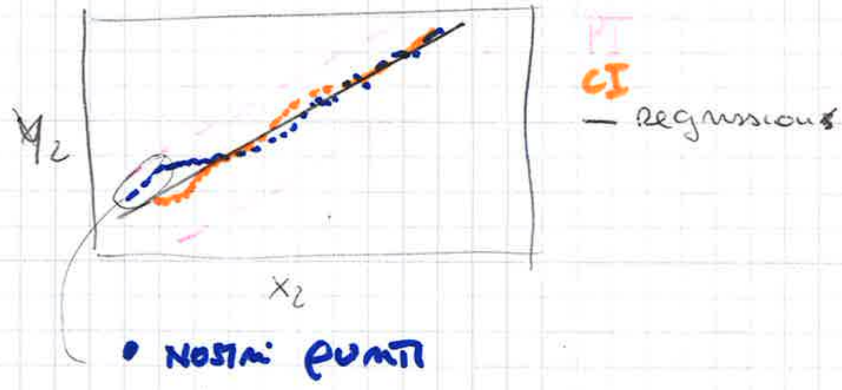
e anche nell' leverage order, si vede che variabilità non costante, aumenta:



e anche l'assunzione di normalità è un po' forzata



Fitted line plot, si evidenzia intervallo di previsione e intervallo di confidenza



Si vede che la retta di regressione non coglie i nostri pt, ci sono fuori da PI le punti, (e il 95% dovrebbe starci dentro!)

(inoltre) ↳ stanno anche quasi fuori dall'intervallo di previsione

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 45 di 58

le sono:

Source	DF	SS	MS	F	P
a	2	843,6	421,81	2,57	0,097
b	3	4551,7	1517,23	9,25	0,000
Interaction	6	4702,5	783,75	4,78	0,002
Error	24	3936,3	164,0		
Total	35	14034,1			

S = 12,81 Rsq = 71,95% R-Sq(adj) = 59,10%

S è la $\sqrt{MS_{error}}$

R^2 è la % di variabilità osservata nelle risposte y spiegata dai 2 fattori e dalla loro interazione

⇓

71.95% $= \frac{SS_a + SS_b + SS_{interaz}}{SS_{tot}}$

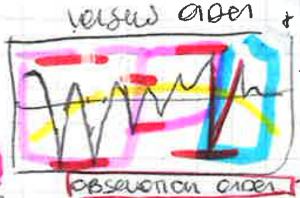
II) al contrario misurerà la miglior stima che abbiamo della varianza comune a tutte le popolazioni

ne ho tante quanti sono i livelli del fattore colonna e quanti del fattore riga

↳ quello che rimane lo addebito all'errore

GRAFICI

1° grafico rappresenta i residui ovvero le realizzazioni degli errori casuali $\hat{E}_i$



⇨ se sono vere le assunzioni fatte, questi residui dovrebbero essere realizzazioni di una v.c. che segue una distribuz normale

↳ ovvero rispetto all'ordine di osservazione

pg op

e poi o meno dobbiamo avere 50% sotto e 50% sopra al pt (minitab li collega con 1 spezzata)

↳ ad occhio vedo che + e - il 50% sta sotto e 50% sta sopra quindi ok, ma la stranezza è che: questa parte (rosa) stanno centrati ma un po deviati verso il basso, qui invece: (rosa) sono centrati ma deviati verso l'alto e per (azzurro) ha addirittura uno molto ≠

↳ OUT LAYER

vanno tenuti sotto controllo
→ possono essere errori di battitura
→ possono essere errori di campionamento

Inoltre si introduce un trend nei residui (prima tutti su e poi tutti giù) ↦ non è una bella cosa perché i residui sono determinaz. delle variabili casuali e proprio perchè sn v.c. non devono seguire alcun trend

Inoltre, anche la variabilità non è costante

↳ questo grafico x gnt detto dei residui non mi piace molto non mi sembrano realizzazioni di una variabile casuale normale con media = 0 e varianza costante!

| GRAFICO INTERAZIONI |

3 spezzate che corrispondono ai 3 livelli del fattore 'a' al variare dei livelli del fattore 'b'

ovvero sono le risposte medie ottenute

× SIMILI
× la spezzata del livello 3 ha tutt'un altro comportamento

• differenza tra le risposte ai vari livelli

rocche tra i le segue comb.

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 47 di 58

ne in realtà basta che non sia concatenato solo con 1.

↳ e questo evidenzia l'alta significatività di questo fattore

↳ quando si verifica questa situazione uso i CONTRASTI per indagare meglio cosa capire

↳ o contrasto è un parametro oscuro

$\Gamma = \sum_{i=1}^{k} c_i \mu_i$ — combinazione lineare delle medie tale per cui $\sum_{i=1}^{k} c_i = 0$

GAMMA maiuscola

↓

stimatore $C = \sum c_i \bar{Y}_{i..}$ — stimatore di μ_i

nel mio output x vedere il problema su b

↳ $\frac{\mu_1 + \mu_2 + \mu_3}{3} - \mu_4$

$\frac{1}{3} + \frac{1}{3} + \frac{1}{3} - 1 = 0$ è un contrasto

↳ la distribuzione dello stimatore C è una normale
v.c.
in quanto è una combinaz. di v.c. normali

$C \sim N(\mu_c, \text{var}[c])$

$\mu_c = E[\sum c_i \bar{Y}_{i..}] = \sum_{i=1}^{k} c_i \mu_i = \Gamma$ è uno stimatore corretto

$\text{var}[c] = \text{var}\left[\sum_{i}^{k} c_i \bar{Y}_{i..}\right] = \sum_{i=1}^{k} c_i^2 \frac{\sigma^2}{nb}$

elevo al^2

avendo μ_c e σ posso fare test hp e intervalli di confidenza

NOTA

$Y_{\cdot (i)}$ ← colonna

↑ righe ↓ il punto dove che lo faccio x tutte le righe

valori nelle celle

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 49 di 58

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi. 55 - Torino / Pagina 51 di 58

COME FARE

Analisi della varianza: Anova one-way

N° GDL del LACK OF FIT

$n-2 = n-K+X$

$\hookrightarrow n-2-n-K \Rightarrow K-2$

$$\frac{SS_{LOF}}{K-2} = MS_{LOF} = \hat{\sigma}^2$$

GDL

stima corretta di σ^2 solo sotto l'HP nullo che il modello sia adeguato $\rightsquigarrow$ altrimenti è $\hat{\sigma}^2 + b(x)$ ↗ dovuto al modello inadeguato

$$Y_i - \hat{Y}_i = \{(Y_i - \bar{Y}_i) - E(Y_i - \bar{Y}_i)\} + E(Y_i - \bar{Y}_i)$$

R_i — errore sistematico

SS_{PE}: $MS_{PE} = \hat{\sigma}^2$

errore puro

$\sum_{i=1}^{K} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$

stima corretta di σ^2 sempre (sia modello adeguato che non)

$SS_{GDL} \quad \sum_{i=1}^{K} (n_i - 1) = n - K$

n° tot punti

$SS_{LOF} = SS_{RES} - SS_{PE}$

$SS_{RES} = SS_{PE} + SS_{LOF}$

LACK OF FIT → mancanza di adattamento

↳ errore sistematico di scostamento del modello vero

Ho 2 stime di σ^2 una sempre corretta e una solo se il modello è adeguato

↳ le confronto se sono ≈ modello adeguato, altrimenti NO

H_0: modello adeguato
H_1 = non =

$$\frac{MS_{LOF}}{MS_{PE}} \approx 1 \text{ se modello adeguato} \qquad F_{K-2,\, n-K}$$

i DATI ↦ + R^2-ADJUSTED che per un parametro è ≃ R^2 ma per più parametri è migliore perché tiene conto delle numerosità delle variabili usate

S ↦ non molto utile perché è semplicemente lo ∇ di MS ERROR

CI ↦ intervalli di confidenza per le medie

POOLED ST DEV ↦ media delle St dev dei singoli fattori

TWO WAY ANOVA — PARTE 1

↦ una caratteristica da misurare 2 fattori

STESSA CONDIZIONE:

$H_0: \mu_1 = \mu_2 \dots \mu_K$

$H_a: \mu_i \neq \mu_K$

DF ↦ FATTORI: 1° → (K−1), 2° → (n−1) × ↦ n° di celle

TOT: valore ↦ valore + 1 n° dati

se ≠ vuol dire che ho più dati in ogni cella (replicazioni p)

P-value ↦ se qualche fattore ha p-value ≤ 10% è debolmente significativo che vuol dire che le differenze sulle risposte che si ottengono cambiando i livelli del fattore in questione non sono molto significative quanto visto

lo si può confrontare anche con SS che in quel fattore sarà notevolmente < che negli altri

↳ viceversa se il p-value ≤ 5% ↦ altamente significativo

P-value ↦ INTERACTION ⇝ se altamente significativo vado ad analizzare l'interaction plot per verificare in che modo influisce sulle risposte

R^2 ↦ (stessa def) $= \dfrac{SS_{fattore1} + SS_{fattore2} + SS_{interazione}}{SS_{TOT}}$

R^2_{ADJ} ↦ (")

NOTA: $SS_{TOT} - SS_{fat1} - SS_{fat2} - SS_{fat3} = SS_{errori}$

S ↦ $\sqrt{MS_{E}}$

PARTE 2: INTERVALLI DI FIDUCIA

qui verifico al livello di fiducia indicato sopra gli intervalli se essi sono tra loro concorrenti

| Se sono tutti conc | ⇒ $H_0: \mu_1 = \mu_2 \dots$ NON può essere rifiutata a tale livello di fiducia

| se non tutti concor | → uso i contrasti per capire meglio cosa capita (Γ)

~~PARTE B~~ analizza anche come si distaccano (dati vedi pg 255)

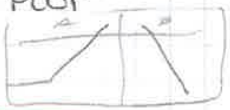
© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 53 di 58

MAIN EFFECTS PLOT: Disegna le medie delle risposte ottenute rispetto ai livelli del 1° e del 2° fattore separatamente

↳ in assenza di interazione è utile x vedere in che modo ottenere la risposta massima

(es: se pt più alto del grafico corrisponde ad un livello 200 del fattore A vuol dire che posso ottenere il max della risposta con quel fattore)

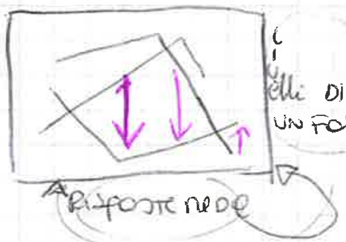
↳ se presente interazione tra i fattori ed essa è significativa, bisogna prima controllare l'INTERACTION PLOT



cfr 263

INTERACTION PLOT

risposte che corrispondono ai valori del fattore a ai valori del fattore b (o viceversa)



↦ se linee // non c'è interazione
se linee non // c'è interaz.

Per vedere come ottenere la risposta max con un certo livello del fattore in asse x basta vedere quale delle linee disegnate sta + in alto.

• differenze tra le risposte ai vari livelli, se ad un certo pt segno cambia, interazione forte

!, **FITTED LINE PLOT**: si evidenzia intervallo di previsione e di confidenza

REGRESSIONE (lineare semplice)

→ ho equazione della retta con 1 variabile dipendente e una indipendente (e+)

→ sotto ho in colonna i coeff. predittori e vengono x entrambi testate $H_0: \beta_0 = 0$ e $H_0: \beta_1 = 0$ $H_A: \beta_0 \neq 0$ $H_A: \beta_1 \neq 0$

→ se p-value altamente significativo rifiuto H_0 (< 0,05)

↦ se coef = $\sqrt{\hat{var}[B_i]}$ ↦ standard error

abbiamo sempre: S, R-Sq, R-sq (adj)

→ posso avere: T = Tcalc (che ci conferma che minitab sta facendo un test di Hp)

nella sezione dell'analisi della varianza: (studio β_1 subordinato alla presenza di β_0)

p-value del LACK OF FIT ↦ se non significativo modello buono (è un indicatore della bontà del modello ↦ non posso rifiutare H_0: modello adeguato)

↳ confronto con R-Sq

NOTA: Pure error ↦ la SS: posso calcolare
LACK OF FIT ↦ SS: la ottengo come differenza tra Residual error e Pure error

NOTA: se ho il (Tcalc) il p-value è: $2P[T_{n-2} \geq 1{,}08]$ $_{1-\alpha/2}$
↳ si ottiene $\dfrac{b_i - H_0}{\sqrt{\hat{var}[B_i]}}$

NOTA: p-value regression altamente significativo mi dice che devo rifiutare β_1 condizionatamente a β_0

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 55 di 58

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\left[(y_i-\bar{y})-(\hat{y}_i-\bar{y})\right]^2$$

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

- $\sum_{i=1}^{n}(y_i-\bar{y})^2$: misura della variazione totale → DEVIANZA TOT SS_{TC}
- $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$: misura della variazione delle risposte previste mediante il modello assunto rispetto alla media delle risposte sperimentali → SS_{REG}
- $\sum_{i}^{n}(y_i-\hat{y}_i)^2$: deviazione residua SS_{RES} → sarebbe 0 se tutti i valori cadessero sulla retta stimata, non accade

R^2 = frazione della variabilità totale della risposta rispetto al loro valor medio spiegata dal modello di regress.

$$R^2=\frac{\sum_{i}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{SS_{REG}}{SS_{TC}} \quad (*) \quad 0\le R^2\le 1$$

(*) tutti i valori previsti coincidono con la media dei valori osservati

$0\le R^2\le 1$: tutti i valori previsti coincidono con i valori osservati

$\bar{y}$ | $\hat{y}=b_0$

$\hat{y}=b_0+b_1x$

$$R^2_{adj}=1-\frac{MS_{ERR}}{SS_{tot}/n-1}$$

$R^2=r_{xy}^2$

Coincide con il quadrato del coeff. di correlaz. lineare tra risp sperimentale e variabile indip

Residui: $\varepsilon_i=y_i-\hat{y}_i$ differenza tra valori osservati della risposta e valori previsti dal modello

→ $\sum(y_i-\hat{y}_i)=0$ NOTA

$$\varepsilon_i=\{(y_i-\hat{y}_i)-E[y_i-\hat{y}_i]\}+E[y_i-\hat{y}_i]$$

- $\{(y_i-\hat{y}_i)-E[y_i-\hat{y}_i]\}$: variabile casuale in quanto differenza tra v.c. $(y_i-\hat{y}_i)$ e il suo valor atteso
- $E[y_i-\hat{y}_i]$: costante → errore sistematico introdotto dall'utilizzo di un modello non adatto

$$\Downarrow$$

$$\varepsilon_i=R_i+d_i$$

Inoltre $\sum_{i=1}^{n}R_i^2$ segue una distribuzione $\sigma^2\chi^2_{n-2}$

con valor atteso $(n-2)\sigma^2$

Elevando al quadrato tutti i membri si ha

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

Da cui:

$$E\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-2)\right]=\sigma^2+\frac{1}{n-2}\sum_{i=1}^{n}d_i^2$$

↦ LA SOMMA DEI QUADRATI CORRISPONDENTI ALLA VARIABILITÀ NATURALE (STIMA DI σ^2) INDIPENDENTE DELLA REGRESSIONE ↭ PURE ERROR ↦ SS_E STIMA CORRETTA DI σ^2 sempre

$$MS_{PE}=s^2_{PE}=\frac{\sum_{i}^{k}\sum_{j}^{n_i}(y_{ij}-\bar{y}_i)^2}{(n-k)\ GDL}\quad SS_{PE}$$

k = n° pt sperimentali distinti

n = n°tot risposte osservate

Mean square

© Proprietà riservata dell'autore - Digitalizzazione e distribuzione a cura del CENTRO APPUNTI - Corso Luigi Einaudi, 55 - Torino / Pagina 57 di 58