Formulario statistica descrittiva

MEDIA

(solo per caratteri QUANTITATIVI!)

FR ASSOLUTA

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$$

FR. RELATIVA

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

MEDIA ARITMETICA PONDERATA

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

MEDIA GEOMETRICA

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$ → produttoria $= x_1 \cdot x_2 \cdot x_3 \dots$

FR ASSOLUTA

$$\bar{x}_g = \sqrt[n]{(x_1)^{n_i} \cdot \dots \cdot (x_k)^{n_k}}$$

FR RELATIVA

$$\bar{x}_g = \sqrt[n]{(x_1)^{g_i} \cdot \dots \cdot (x_k)^{g_k}}$$

MODA → modalità che presenta la MASSIMA FREQUENZA

MEDIANA → SOLO PER CARATTERI ORDINABILI!

↓ corrisponde alla modalità per cui $F(x) = 0{,}5$

VARIANZA

$$\sigma^2 = E\left[(x-\bar{x})^2\right]$$

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

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

VARIANZA DI UNA TRASFORMAZIONE LINEARE → $Y = \alpha X + \beta$

$$\sigma^2_Y = \alpha^2 \sigma^2_X \rightarrow V(Y) = a^2 V(X)$$

$$\boxed{V(aX) = a^2 V(X)}$$

$$\boxed{V(aX+bY) = a^2V(X) + b^2V(Y) + 2ab\,\sigma_{XY}}$$ → se X e Y sono indipendenti $V(X \pm Y) = V(X) + V(Y)$

DEVIAZIONE STANDARD o scarto quadratico medio

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

COEFFICIENTE DI VARIAZIONE

$$CV = \frac{\sigma}{|\bar{x}|} \cdot 100$$

STANDARDIZZAZIONE

⇒ rendo qualsiasi variabile con:
- MEDIA NULLA
- VARIANZA UNITARIA

$$z = \frac{x_i - \bar{x}}{\sigma}$$

CAMPO DI VARIAZIONE

$R = X(n) - X(1)$ → differenza tra il valore max assunto dalla variabile e il valore min

DIFFERENZA INTERQUARTILICA

$W = Q_3 - Q_1$ → include il 50% delle osservazioni

MEDIA CONDIZIONATA

$$\bar{y}_{x=x_i} = \frac{1}{n_{i.}} \sum_{j=1}^{k} y_j \cdot n_{ij} \longleftrightarrow \bar{y} = \frac{1}{n} \sum_{i=1}^{H} \bar{y}_{x=x_i} \cdot n_{i.}$$

media totale della popolaz.

VARIANZA CONDIZIONATA

$$\sigma^2_{Y|X=x_i} = \frac{1}{n_{i.}} \sum_{i=1}^{H} \left(y_j - \bar{y}_{x=x_i}\right)^2 \cdot n_{ij}$$

BARICENTRO DI UNA DISTRIBUZIONE DOPPIA

punto che ha coordinate $(\bar{x}; \bar{y})$

Indipendenza statistica

REGOLA DI FATTORIZZAZIONE

$$n_{ij}' = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$ → frequenza teorica che troverei nella tabella X/Y nel caso di indipendenza

Dipendenza e interdipendenza perfetta

X \ Y	y_1	y_2	y_3
x_1	0	0	•
x_2	0	•	0
x_3	•	0	0
x_4	•	0	0

$H \neq R$

X \ Y	y_1	y_2
x_1	•	0
x_2	0	•

$H = R$

→ relazione BIDIREZIONALE

CONTINGENZE

$$c_{ij} = n_{ij} - n_{ij}'$$ → frequenza teorica in caso di indipendenza

INDICE DI ASSOCIAZIONE CHI-QUADRO DI PEARSON

$$X^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{c_{ij}^2}{n_{ij}'}$$ ⟶ in caso di indipendenza $X^2 = 0$

$$X^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{n_{ij}}{n_{ij}'} - n$$

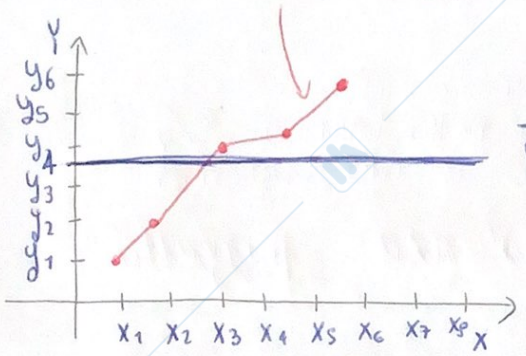
$$X^2 = n\left(\sum_{i=1}^{H} \sum_{j=1}^{K} \frac{n_{ij}}{n_{i\cdot} \, n_{\cdot j}} - 1 \right)$$ → sottraggo 1 una volta sola

INDICE NORMALIZZATO

$$\tilde{X}^2 = \frac{X^2}{n \cdot \min\{(H-1), (K-1)\}}$$ ⟶ il suo campo di variazione è $[0, 1]$: 0 = indipend. statistica, 1 = dipendenza o interdipend. perfetta

Indipendenza in media

SPEZZATA DI REGRESSIONE



$\bar{Y}$ → nel caso di indipendenza la spezzata assume questo valore

SCOMPOSIZIONE DELLA VARIANZA

$$\sigma^2 Y = \sigma^2 \text{media}(Y|x) + \text{media}(\sigma^2 Y|x)$$

VARIANZA SPIEGATA

$$\sigma^2 \text{media}(Y/x) = \frac{1}{n} \sum_{i=1}^{H} (\bar{y}_{x=x_i} - \bar{y})^2 n_{i\cdot}$$

→ varianze delle medie di Y condizionate ad X

VARIANZA RESIDUA

$$\text{media}(\sigma^2 Y|x) = \frac{1}{n} \sum_{i=1}^{H} \sigma^2 Y|_{x=x_i} \cdot n_{i\cdot}$$

→ media delle varianze condizionate della Y date le X

RAPPORTO DI CORRELAZIONE

$$\eta^2_{Y|x} = \frac{\sigma^2 \text{media}(Y|x)}{\sigma^2 Y}$$

$$\eta^2_{Y|x} = 1 - \frac{\text{media}(\sigma^2 Y|x)}{\sigma^2 Y}$$

→ il suo campo di variazione è:

$[0, 1]$

0: indipendenza statistica

1: dipendenza statistica

Incorrelazione

COVARIANZA ⟶ misura ⟶ la forza di legame lineare che ∃ tra le variabili

misura ⟶ concordanza, discordanza

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

$$\sigma_{XY} = E[(X-\bar{X})(Y-\bar{Y})]$$

$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x}$$

⟶ il suo campo di variazione oscilla tra $[-\sigma_X\sigma_Y ; +\sigma_X\sigma_Y]$ (prodotto degli scarti quadratici medi)

$\sigma_{XY} = E[XY] - E[X]\cdot E[Y]$ $\quad \sigma(a+bX)(cY) = bc\,\sigma(XY)$

TRASFORMAZIONE LINEARE DELLA COVARIANZA

se $X \to aX+b$, $Y \to cY+d$ $\Rightarrow$ $\sigma_{X'Y'} = a\cdot c\cdot\sigma_{XY}$

COEFFICIENTE DI CORRELAZIONE LINEARE DI BRAVAIS E PEARSON → indica l'addensamento dei dati (clustering) nello scatter plot

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y}$$

⟶ il suo campo di variazione oscilla tra $[-1 ; 1]$

- -1: perfetto legame lineare con caratteri DISCORDI
- 1: perfetto legame lineare con caratteri CONCORDI

se X e Y sono correlate allora non è detto che X sia la causa di Y o viceversa ↓ potrebbe esistere un terzo fattore Z che è la causa di entrambe

$\rho_{XY} = 0$ ⇓ è possibile che ci sia indipendenza statistica ma non certo, in quanto questo indice non è abbastanza forte da poterlo determinare

Probabilità

$0 \leq P(\cdot) \leq 1$ (evento nullo ✓, evento certo $P(\Omega)$)

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

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

DEFINIZIONE DI PROBABILITÀ

$$P(E) = \frac{\text{n. di casi favorevoli}}{\text{n. di casi possibili}}$$ → questi devono essere EQUIPROBABILI!!!

PROBABILITÀ CONDIZIONATA DI A DATO B

$$P(A|B) = \frac{\text{n. di casi favorevoli ad } (A \cap B)}{\text{n. di casi favorevoli a } B} = \frac{P(A \cap B)}{P(B)}$$

PRINCIPIO DELLE PROBABILITÀ COMPOSTE

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

INDIPENDENZA TRA EVENTI

nel caso di distribuzione di probabilità congiunta ⇓ SI HA INDIPENDENZA SE LE PROBABILITÀ CONDIZIONALI SONO UGUALI!

$P(B|A) = P(B) \qquad P(A|B) = P(A)$

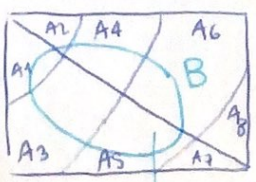
$P(A \cap B) = P(A) \cdot P(B)$

TEOREMA DI BAYES → utile per problemi a più casi

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

Probabilità a priori → verosimiglianza

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



$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$ ← teorema delle probabilità totali

Variabili casuali discrete

FUNZIONE DI RIPARTIZIONE

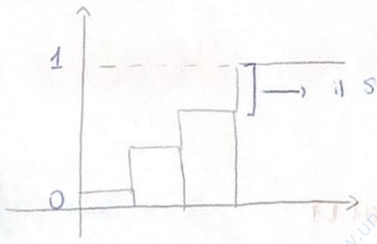
↳ si basa sulle probabilità cumulate

$$F(x) = P(X \leq x) = \sum_{w \leq x} P(X = w) \qquad x \in \mathbb{R}$$

dominio $\mathbb{R}$ — codominio $[0,1]$

↓

funzione a gradini



→ il singolo gradino corrisponde all'ultima probabilità che viene aggiunta

FUNZIONE DI PROBABILITÀ

↓

associa ad ogni possibile valore di x la probabilità che questo si verifichi → $P(X = x_i)$

VALORE ATTESO → misura dove i valori tendono a concentrarsi

$$E(x) = \sum_i x_i \cdot P(x_i)$$

VARIANZA

$$V(x) = \sum_i [x_i - E(x)]^2 \cdot P(x_i)$$

$$V(x) = E[(x - E(x))^2]$$

$$V(x) = \underbrace{E(x^2)}_{\text{momento secondo}} - [E(x)]^2$$

$$\Rightarrow E(x^2) = \sum_i x_i^2 \cdot P(x_i)$$

FUNZIONE LINEARE DI UNA VARIABILE CASUALE

↓

$$Y = a + bX$$

$$E(Y) = a + b(E(x))$$

$$V(Y) = b^2 (V(x))$$

$$SD(Y) = \sqrt{b^2 V(x)}$$

DISTRIBUZIONE UNIFORME DISCRETA

$X \sim Ud(a, S)$ → il numero dei possibili valori

valore più piccolo assumibile

funzione di probabilità

$P(X=x) = \frac{1}{S}$ → l'intervallo in cui questa funzione è definita è $[a, a+S-1]$

$E(X) = a + \frac{S-1}{2}$

$V(X) = \frac{S^2-1}{12}$

DISTRIBUZIONE DI BERNOULLI

$X \sim Ber(\pi)$ → probabilità di successo

si riferisce ad osservazioni BINARIE

$x=0$ quando l'evento si VERIFICA

$x=1$ quando l'evento non si verifica

x	0	1
$P(x_i)$	$1-\pi$	π

$P(X=x) = \pi^x (1-\pi)^{1-x}$

$E(X) = \pi$

$V(X) = \pi(1-\pi)$

DISTRIBUZIONE BINOMIALE → Successione di n prove di tipo Bernoulliano

$X \sim Bin(n; \pi)$

$P(X=x) = \underbrace{\binom{n}{x}}_{\text{BINOMIO DI NEWTON}} \pi^x (1-\pi)^{[n-x]}$

numero di prove effettuate

[n-x] → numero di successi ottenuti

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$

$E(X) = n\pi$

$V(X) = n\pi(1-\pi)$

DISTRIBUZIONE DI POISSON → applicazione nel conteggio di un evento casuale in uno specifico Δt

$X \sim Poi(\lambda)$

$$P(X=x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$$

$E(x) = \lambda$
$V(x) = \lambda$
} generalmente EQUIDISPERSO

Variabili casuali continue

$P(X=x) = 0$ → non è possibile calcolare la probabilità di un singolo evento

FUNZIONE DI DENSITÀ

$$P(x) = \int_a^b f(x)\,dx$$

FUNZIONE DI RIPARTIZIONE

$$F(x) = \int_{-\infty}^{x} f(w)\,dw$$

VALORE ATTESO

$$E(x) = \int_a^b x \cdot f(x)\,dx$$

VARIANZA

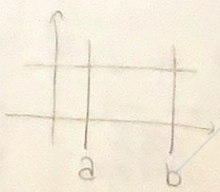
$$V(x) = \int_a^b (x - E(x))^2 f(x)\,dx$$

$$V(x) = E(x^2) - [E(x)]^2 = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - [E(x)]^2$$

DISTRIBUZIONE UNIFORME CONTINUA

$X \sim U(a,b)$ — a: valore più piccolo, b: valore più grande

$$f(x) = \frac{1}{b-a}$$



$$E(x) = \frac{a+b}{2} \qquad V(x) = \frac{(a-b)^2}{12}$$

DISTRIBUZIONE NORMALE → simmetrica rispetto a μ

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

FUNZIONE DI DENSITA'

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

VALORE ATTESO → PARAMETRO DI LOCAZIONE
determina la posizione della campana su $\overrightarrow{x}$

$E(x) = \mu$

VARIANZA → PARAMETRO DI SCALA
più σ^2 aumenta più la curva è schiacciata

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

la sua standardizzazione è

$\frac{x-\mu}{\sigma} = z \sim N(0,1)$ → distribuzione che uso per le tavole

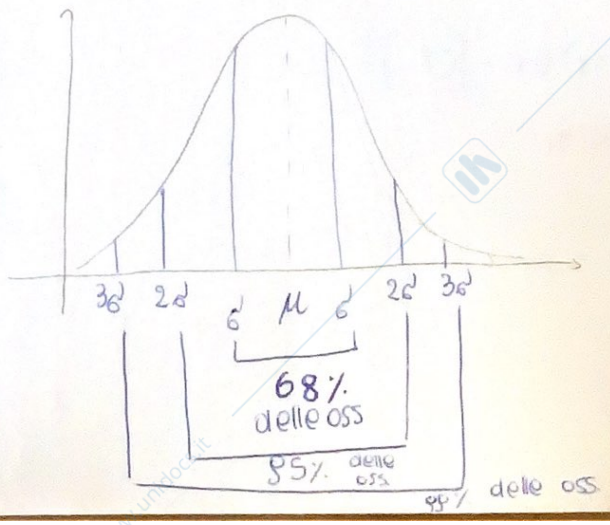
TRASFORMAZIONE LINEARE DI $z \sim N(0,1)$

$Y = a + bX \sim N(a, b^2)$ → quando $X \sim N(0,1)$

SOMMA DI DUE NORMALI

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

LEGGE DEI TRE SIGMA



teorema del limite centrale

I.I.D. o

DATA UNA SUCCESSIONE DI VARIABILI CASUALI TUTTE UGUALI NE CONSIDERO LA MEDIA CAMPIONARIA.

SE $n \to \infty$ LA MEDIA CAMPIONARIA, A PRESCINDERE DELLA DISTRIBUZIONE DELLE VARIABILI, QUANDO E' STANDARDIZZATA, SI DISTRIBUISCE COME UNA NORMALE STANDARD $N(0,1)$

SE $n \to \infty \longrightarrow z_n = \dfrac{(\bar{x}_n - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$

z_n: media campionaria standardizzata

- SE $\bar{x}$ NON FOSSE STANDARDIZZATA

se $n \to \infty \quad \bar{x}_n \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

- SE CONSIDERASSIMO LA SOMMA DI n V.C. I.I.D

se $n \to \infty \quad S_n \sim N(n\mu; n\sigma^2)$

CURTOSI

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

- se $\gamma < 3$ la distribuzione e' IPERNORMALE, ovvero ha la coda piu' spessa rispetto alla normale
- $\gamma = 3$ la distribuzione e' NORMALE
- se $\gamma > 3$ la distribuzione e' IPONORMALE ovvero i casi estremi sono ancora piu' estremi

DISTRIBUZIONE CHI-QUADRO → corrisponde al quadrato della distribuzione normale standard

$$f(x) = \frac{1}{2^{\frac{g}{2}}\,\Gamma\left(\frac{g}{2}\right)} \cdot x^{\frac{g}{2}-1} \cdot e^{-\frac{x}{2}}$$

$$\chi^2_g \sim \sum_{i=1}^{g_1} x_i^2 \rightarrow N(0,1)^2$$

↳ somma g volte delle normali standard al quadrato

VALORE ATTESO

$E(x) = g$

VARIANZA

$V(x) = 2g$

DISTRIBUZIONE t DI STUDENT

$$X = \frac{z}{\chi^2_g}$$

DISTRIBUZIONE F di FISHER

$$F_{g_1, g_2} = \frac{\chi^2 g_1}{\chi^2 g_2}$$

4 # Inferenza statistica

FRAZIONE DI CAMPIONAMENTO

$\frac{n}{N}$ → n → dimensione campionaria
N → numerosità della popolazione

DISTRIBUZIONE DELLA MEDIA CAMPIONARIA → $\bar{X}$ NELLE POPOLAZIONI INFINITE

se $X \sim N(\mu; \sigma^2)$ allora $\bar{x} \sim N(\underbrace{\mu}_{E(\bar{x})}; \underbrace{\frac{\sigma^2}{n}}_{V(\bar{x})})$

se $X \sim Ber(\pi)$ allora $\bar{x} \sim \frac{1}{n} Bin(n; \pi)$

$n > 30$

per T.L.C. $\bar{X} \sim N(\mu; \frac{\sigma^2}{n})$

DISTRIBUZIONE DELLA $\bar{X}$ NELLE POPOLAZIONI FINITE

se $X \sim N(\mu; \sigma^2)$ allora $\bar{x}_n \sim N(\mu; \underbrace{\frac{(N-n)}{(n-1)}}_{\text{FATTORE DI CORREZIONE}} \frac{\sigma^2}{n})$

PROPRIETÀ DEGLI STIMATORI

↓
QUANTITÀ TEORICA
- V.C. utilizzata per stimare una determinata caratteristica θ della popolazione

- **CORRETTEZZA** → la media della distr. che ipotizzo è centrata nella media reale

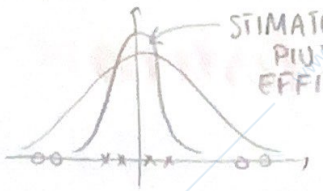
$E(T) = \theta \quad \forall \theta$ possibile

↓

se $E(T) \neq \theta \Rightarrow T$ sarebbe uno stimatore **DISTORTO**

↓
DISTORSIONE
$B(T) = E(T) - \theta$

- EFFICIENZA



(σ² minore)

ERRORE QUADRATICO MEDIO

$$MSE(T) = E\left[(T-\theta)^2\right]$$

$$MSE(T) = \underbrace{V(T)}_{\text{varianza}} + [B(T)]^2 \quad \text{(distorsione)}$$

$$V(T) = E\left[(T-E(T))^2\right]$$

nella scelta tra due stimatori prediligerò quello che presenta un MSE minore → i suoi valori saranno più prossimi a θ

- CONSISTENZA (∞) → all'aumentare della n cresce la precisione dello stimatore

LEGGE DEI GRANDI NUMERI: uno stimatore è consistente se la probabilità che una sua realizzazione cada in un intervallo che contiene il vero valore nella popolazione tende ad uno al limite e quindi che la numerosità del campione aumenta

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

→ UNO STIMATORE SI DICE CONSISTENTE IN MEDIA QUADRATICA SE $V(T_n)$ E $B(T_n)$ TENDONO A ZERO

- CORRETTEZZA ASINTOTICA (∞)

$$\lim_{n\to\infty} B(T) = \lim_{n\to\infty} E(T_n) - \theta = 0$$

STIMA PUNTUALE DELLA MEDIA DI UNA POPOLAZIONE

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

stimatore naturale della media

MEDIA CAMPIONARIA $\bar{x}$ / $E(\bar{x}) = \mu$

$\bar{x}$ → Stimatore → corretto / consistente

STIMA PUNTUALE DELLA PROPORZIONE IN UNA POPOLAZIONE

$$\hat{\pi} = \bar{x}$$

$$E(\bar{x}) = \pi \qquad V(\bar{x}) = \frac{\pi(1-\pi)}{n}$$

Stimatore → corretto / consistente

STIMA PUNTUALE DELLA VARIANZA DELLA POPOLAZIONE

$\hat{\sigma}^2$ (var. campionaria) ⟶ STIMATORE DISTORTO

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$ → VARIANZA CAMPIONARIA CORRETTA

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

⇓ equivale a

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

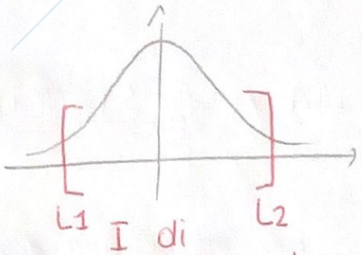
($\frac{1}{n} \sum_{i=1}^{n} x_i^2$: MEDIA QUADRATICA)

} stimatore non corretto di σ^2

→ per passare da uno stimatore non corretto $\hat{\sigma}^2$ allo stimatore corretto S^2

$$S^2 = \hat{\sigma}^2 \cdot \frac{n}{n-1}$$

STIMA PER INTERVALLO



I di congidenza di livello $1-\alpha$] → arbitrario per il parametro θ

↓

la sua realizzazione in corrispondenza del valore osservato è detto

INTERVALLO DI CONFIDENZA STIMATO

↓

NON abbiamo la certezza che il nostro intervallo stimato attraverso i dati del campione contenga θ

$1-\alpha$ %. dei campioni possibili producono un I che contiene il vero valore di θ

α %. dei campioni casuali che posso estrarre dalla popolazione producono un I che contiene θ

I DI CONFIDENZA PER LA MEDIA μ (σ^2 NOTA)

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

LUNGHEZZA DELL'INTERVALLO DI CONFIDENZA

$$\text{lunghezza} = 2 \cdot z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

↓ Dipende da

- n → se n aumenta la lunghezza diminuisce → più precisione
- $1-\alpha$ → se aumenta la lunghezza dell'I aumenta → paga la maggiore confidenza con una perdita di precisione

INTERVALLO DI CONFIDENZA PER LA MEDIA μ (σ^2 IGNOTA)

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \; ; \; \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right]$$

(S: SD campionaria corretta)

t di student con $n-1$ gdl

MARGINE D'ERRORE (semilunghezza dell'I di confidenza): MAX DIFFERENZA ASSOLUTA CHE CI ASPETTIAMO TRA LA NOSTRA STIMA E IL VERO PARAMETRO

INTERVALLO DI CONFIDENZA PER LA PROPORZIONE (campioni elevati)

↳ CONDIZIONE PER L'USO DEL T.L.C.

- $n \cdot \hat{\pi} \geq 5$
- $n(1-\hat{\pi}) \geq 5$

$\hat{\pi}/\bar{x} \sim \frac{1}{n} \text{Bin}(n; \pi)$ ↳ distr. reale

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \; ; \; \bar{x} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]$$

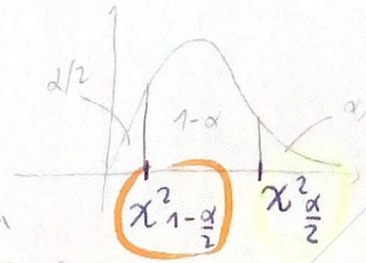
distribuz. della varianza campionaria corretta S^2 ↓ ignota! → trasformiamolo in modo da poterlo studiare più facilmente

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

INTERVALLO DI CONFIDENZA PER LA VARIANZA σ^2

$$\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \; , \; \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\right]$$

- $\chi^2_{\frac{\alpha}{2}}$ → chi-quadro con $n-1$ gdl
- $\chi^2_{1-\frac{\alpha}{2}}$



NON È SIMMETRICA → devo identificare separatamente entrambi i valori

→ se dovessi stimare σ dovrei mettere tutto sotto $\sqrt{\ }$

passi da seguire per verifica ipotesi

Hp. sulla quale si basa la verifica d'ipotesi

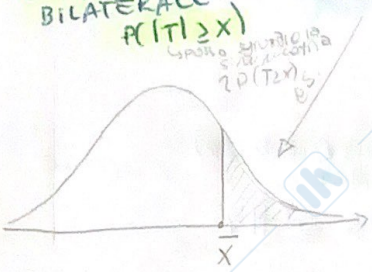
1. che il campione sia estratto casualmente (ogni xi è indipendente dalle altre)
2. che le variabili siano tra loro INDIPENDENTI

SOSPETTO SUL PARAMETRO

1. Formulare il sistema d'ipotesi, H_0 e H_1
2. Determinare la statistica-test più appropriata e determinare la distribuzione campionaria sotto H_0
   - es. se sto facendo un test sulla proporzione o sulla media μ la statistica test sarà basata sulla media campionaria $\bar{X}$
3. Scelgo il livello di significabilità ecc... e la numerosità campionaria → connesso all'errore di I tipo
   - se è importante che questo errore sia minimo significa che α piccolo
4. Determino la regione di rifiuto del test, calcolandone i/il valori/e critici/o
   - a seconda delle alternative
5. Estraggo il campione e svolgo tutte le analisi
6. Calcolo il valore campionario della statistica-test
7. Vedo se questo valore cade nella regione di accettazione o di rifiuto del test
8. Prendo una decisione (statistica) per poterne descrivere le conseguenze nei termini del problema reale che si sta affrontando.

P-VALUE → Prob. di osservare un valore della statistica-test uguale o più estremo del valore ottenuto dal campione sotto l'Hp nulla

se l'Ha è BILATERALE $P(|T| \geq x)$



se $P(T \geq \bar{x})$ è MOLTO GRANDE → non rifiuto l'Ho → $p > 0,1$

se $P(T \geq \bar{x})$ è PICCOLO $(p < \alpha)$ → rifiuto l'Ho

ERRORI DI PRIMO E SECONDO TIPO

(os falsi negativi) RIFIUTARE l'Hp NULLA QUANDO È VERA → + imp!!

α = prob di commettere l'errore di I tipo

⇓

$1-\alpha$ = COEFF. DI CONFIDENZA DEL TEST

NON RIFIUTARE l'Hp NULLA QUANDO È FALSA (os falsi positivi)

β = prob di commettere l'errore di II tipo

⇓

$1-\beta$ = POTENZA DEL TEST

α e β sono INVERSAMENTE PROPORZIONALI

↓

non possono essere minimizzati entrambi

↓

è molto più imp. minimizzare gli errori di I tipo, anche a discapito di quelli di II tipo

Test per $\mu_1 - \mu_2$ (σ^2

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

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2 \sim N(\mu_2, \sigma_2^2)$

- SOMMA DI NORMALI

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

- DIFFERENZA TRA NORMALI

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

↓ sempre la somma delle σ^2

↓

Se pongo $\mu_D = \mu_1 - \mu_2$

allora

$\begin{cases} H_0: \mu_D = 0 \\ H_1: \mu_D \gtrless 0 \end{cases}$ ⟶ mi riconduco al test per le medie classico

Test $\mu_1 - \mu_2$ (σ^2 ignote ma UGUALI)

↓

$\sigma_1^2 = \sigma_2^2$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{H_0}{\sim} t_{(n_1+n_2-2)}$$

↓

STIMATORE CONGIUNTO (pooled) → corrisponde alla media ponderata di s_1 e s_2 con pesi proporzionali alla loro numerosità.

$$S_p^2 = \frac{(n_1-1)\, s_1^2 + (n_2-1) s_2^2}{n_1 + n_2 - 2}$$

test per la media (σ^2 nota)

$\rightarrow X \sim N(\mu, \sigma^2)$
oppure in tutti i casi in cui $n \geq 120$

Si sotto H_0: $\mu_1 = \mu_2$ $\quad Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1)$

test per la media (σ^2 ignota)

$\rightarrow X \sim N(\mu; \sigma^2)$

sotto H_0: $\mu_1 = \mu_2$ $\quad T = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$

stimatore della varianza
$S^2 = \dfrac{1}{n-1} \sum (x_i - \bar{x})^2$

test per la proporzione

$\rightarrow X \sim Ber(\pi)$

$H_0: \pi = \pi_0$

devo poter applicare il T.L.C.
↓
verifico che $n\pi_0 \geq 5$; $n(1-\pi_0) \geq 5$

$$Z = \frac{\bar{x} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \overset{H_0}{\sim} N(0,1)$$

test per la varianza

$$\chi = (n-1)\frac{S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2_{n-1} \rightarrow$$

la chi-quadro NON è simmetrica!!!

Test per il rapporto tra due varianze

→ x_1 e x_2 devono distribuirsi come $N(\mu; \sigma^2)$

Sotto H_0: $\sigma_1^2 = \sigma_2^2$

varianze corrette

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1)(n_2-1)}$$

→ F di Fisher (no simm.)

↳ DECIDO QUALE METTERE SOPRA O SOTTO IN BASE A QUALE MI PERMETTE DI AVERE IL RAPPORTO > 1 ↳ meglio lavorare a più grandi

Test per la differenza tra due proporzioni

→ devo supporre un campione GRANDE ⇒ T.L.C.

$x_1 \sim Ber(\pi_1)$

$x_2 \sim Ber(\pi_2)$

Sotto H_0: $\pi_1 = \pi_2$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\bar{x}_p(1-\bar{x}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

STIMATORE CONGIUNTO DI π

$$\bar{x}_p = \frac{n_1 \cdot \bar{x}_1^{\hat{\pi}_1} + n_2 \cdot \bar{x}_2^{\hat{\pi}_2}}{n_1 + n_2}$$

↓

bisogna sempre verificare che

$$\begin{cases} n\hat{\pi}_A \geq 5 \\ n(1-\hat{\pi}_A) \geq 5 \end{cases} \quad e \quad \begin{cases} n\hat{\pi}_B \geq 5 \\ n(1-\hat{\pi}_B) \geq 5 \end{cases}$$

Se un test passa dall'essere bilaterale ($\neq$)
all'essere unilaterale ($\lessgtr$)

↓

il valore critico passa da

$z_{\frac{\alpha}{2}} / t_{\frac{\alpha}{2}} \ldots \rightarrow t_\alpha / t_\alpha \rightarrow$ È PIÙ PICCOLO

la regione di rifiuto
AUMENTA