DATI CATEGORICI

ex)
- TIPO DI SANGUE (8 LIVELLI) $\longrightarrow \{O^+, O^-, A^+, A^-, B^+, B^-, AB^+, AB^-\}$ NO ORDINE
- STATO DI APPARTENENZA (~200 LIVELLI) $\longrightarrow \{\text{ITALIA, GIAPPONE} \dots\}$ NO ORDINE
- RISPOSTA A UN FARMACO (3 LIVELLI) $\longrightarrow \{+, 0, -\}$ SÌ ORDINE (Ha senso dire che '+' > '−')

MA

LA DISTANZA FRA I LIVELLI NON È BEN DEFINITA

VARIABILE BINARIA

Y → 1 (successo)
Y → 0 (insuccesso)

$$f_Y(y) = \begin{cases} p & \text{se } y = 1 \\ 1-p & \text{se } y = 0 \end{cases}$$

$Y_i \overset{iid}{\sim} \text{Bernoulli}(\pi)$

$$N = \sum_{i=1}^{n} Y_i \sim \text{Binomiale}(n, \pi) \qquad f_N(n) = \binom{n}{N} p^N (1-p)^{n-N}$$

n = n° esperimenti
N = n° successi

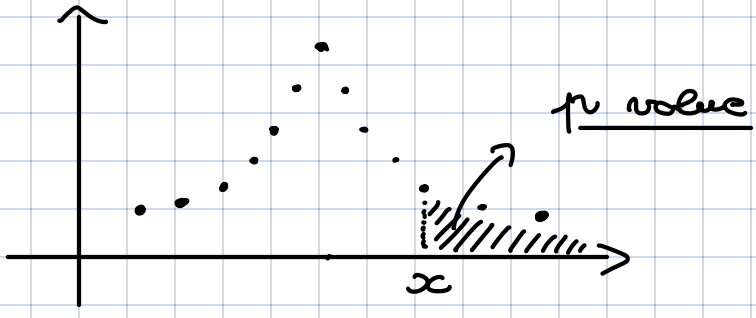
TEST D'IPOTESI

$H_0 : \pi = \pi^0$
$H_A : \pi > \pi^0$

al livello α, unilaterale

- esperimento $\longrightarrow$ x successi
  - RIFIUTA H_0 SE $\sum_{x=m+1}^{N} f_N(x) < \alpha$
  - ACCETTO SE $\sum_{x=m+1}^{N} f_N(x) \geq \alpha$

OSS: Accettare significa non avere abbastanza evidenza per rifiutare



INTERVALLI DI CONFIDENZA (di FIDUCIA)

→ CLOPPER PEARSON ①
→ WALD ②
→ WILSON ③

α : livello di fiducia

① L^{CP} : $\min \{ \pi \mid \mathbb{P}(Bin(n,p) \leq x) > \frac{\alpha}{2} \}$

U^{CP} : $\max \{ \pi \mid \mathbb{P}(Bin(n,p) \geq x) > \frac{\alpha}{2} \}$

OSS: È un intervallo "esatto" perché utilizza la densità binomiale ma ha copertura $> 1-\alpha$ (conservativo)

② $$\boxed{\sqrt{n}\left(\frac{N}{n} - \pi\right) \xrightarrow[TLC]{n \to \infty} \mathcal{N}(0, \pi(1-\pi))}$$

$\frac{N}{n} = \hat{\pi}$

$$\mathbb{P}\left(\underbrace{\left| \frac{\sqrt{n}(\hat{\pi} - \pi)}{\sqrt{\pi(1-\pi)}} \right|}_{(*)} < z_{\frac{\alpha}{2}} \right) = 1-\alpha$$

$$\mathbb{P}\left(-z_{\frac{\alpha}{2}} \leq \underbrace{\frac{\sqrt{n}(\hat{\pi}-\pi)}{\sqrt{\pi(1-\pi)}}}_{\text{SOSTITUIRE } \pi \text{ CON } \hat{\pi} \;!!} \leq z_{\frac{\alpha}{2}} \right) = 1-\alpha$$

$$\mathbb{P}\Bigg(\underbrace{\hat{\pi} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}_{L^{WALD}} \leq \pi \leq \underbrace{\hat{\pi} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}_{U^{WALD}} \Bigg) = 1-\alpha$$

$$I.C.^{WALD} = [L^{WALD}, U^{WALD}]$$

③ $$(*)^2 = \left(\frac{\sqrt{n}(\hat{\pi}-\pi)}{\sqrt{\pi(1-\pi)}} \right)^2 = z^2_{\frac{\alpha}{2}}$$

$$(\hat{\pi}-\pi)^2 = \frac{\pi(1-\pi) z^2_{\frac{\alpha}{2}}}{n}$$

$$\hat{\pi}^2 - 2\pi\hat{\pi} + \pi^2 = \pi \frac{z^2_{\alpha/2}}{n} - \pi^2 \frac{z^2_{\alpha/2}}{n}$$

$$\pi^2\left(1 + \frac{z^2_{\alpha/2}}{n}\right) - \pi\left(2\hat{\pi} + \frac{z^2_{\alpha/2}}{n}\right)\pi + \hat{\pi}^2 = 0$$

EQ II GRADO IN π

$$\text{I.C.}^{\text{WILSON}} : \frac{1}{1+\frac{z^2_{\alpha/2}}{n}} \left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} \pm \frac{z_{\alpha/2}}{2n} \sqrt{4n(1-\hat{\pi})\hat{\pi} + z^2_{\alpha/2}} \right)$$

VARIABILE CATEGORICA MULTILIVELLO

Variabile aleatoria C che può assumere D livelli

- CODIFICA STANDARD
- CODIFICA ONE-HOT

$$C = i \quad i = 1 \dots D$$

$$\vec{B} = (\overbrace{0}^{x_1}, \overbrace{0}^{x_2}, \dots, \overbrace{1}^{x_i}, 0 \dots 0)$$

i-esima posizione

OSS: Sono sufficienti $d = D-1$ bit per codificare $\vec{B}$

I parametri che governano la distribuzione di $\vec{B}$ sono

$$\vec{\pi} = (\pi_1, \dots, \pi_d)$$

$$\pi_i = \mathbb{P}(\vec{B} = (0, \dots, \underset{\substack{\uparrow \\ i\text{-esima} \\ \text{posizione}}}{1}0, \dots 0)) \quad \forall i = 1 \dots d$$

Risultati che ci serviranno dopo

- $\mathbb{E}[X_k] = \pi_k$
- $\text{Cov}(X_k, X_\ell) = \mathbb{E}[X_k X_\ell] - \mathbb{E}[X_k]\,\mathbb{E}[X_\ell]$

$$= \begin{cases} -\pi_k \pi_\ell & k \neq \ell \\ \pi_k - \pi_k^2 & k = \ell \end{cases}$$

$$\pi_k - \pi_k^2 = \pi_k(1-\pi_k)$$

2 LIVELLI

$Y \sim \text{Bernoulli}(\pi)$

D LIVELLI

$\vec{B} \sim \text{Multinomiale}(\vec{\pi})$

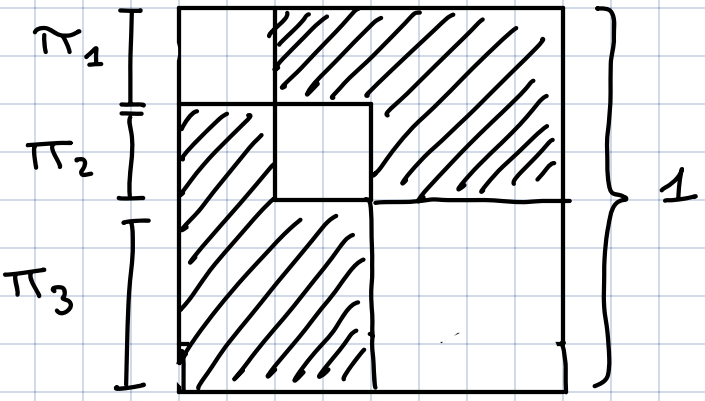
con $\vec{\pi} = (\pi_1 \dots \pi_d)$

OSS: $\sum_{i=1}^{d} \pi_i \neq 1$

$\pi_D = 1 - \sum_{i=1}^{d'} \pi_i$

INDICI DI VARIABILITÀ PER MULTINOULLI

- $E = \sum_{i=1}^{D} \pi_i \log(\pi_i)$ — ENTROPIA DI SHANNON
- $G = 1 - \sum_{i=1}^{D'} \pi_i^2$ — DIVERSITÀ DI GINI



$Y_i \sim \text{Bernoulli}(\pi)$ $\qquad N = \sum_{i=1}^{n} Y_i \sim \text{Bin}(n, \pi)$

$\vec{B}_i \sim \text{Multinoulli}(\vec{\pi})$ $\qquad \vec{N} = \sum_{i=1}^{n} \vec{B}_i \sim \text{Multin}(n, \vec{\pi})$

$$f_N(n_1, \dots, n_d) = \frac{n!}{n_1! n_2! \dots n_D!} \prod_{i=1}^{D} \pi_i^{n_i}$$

RISULTATO ASINTOTICO (TCL MULTIVARIATO)

$$\sqrt{n}\left(\frac{N_1}{n} - \pi_1, \dots, \frac{N_d}{n} - \pi_d\right) \xrightarrow{n \to \infty} N_d(\vec{0}, \Sigma)$$

con $\Sigma = [\sigma_{ij}] = \begin{cases} \pi_i(1-\pi_i) & i = j \\ -\pi_i \pi_j & i \neq j \end{cases}$

TEST PER $\vec{\pi}$

$H_0 : \vec{\pi} = \vec{\pi}^\circ$
$H_A : \vec{\pi} \neq \vec{\pi}^\circ$

① Caso binomiale

RIFIUTO H_0 SE $\left| \frac{N - m\pi^\circ}{\sqrt{m\pi^\circ(1-\pi^\circ)}} \right| > z_{\frac{\alpha}{2}}$

O EQUIVALENTEMENTE $\left(\frac{N - m\pi^\circ}{\sqrt{m\pi^\circ(1-\pi^\circ)}} \right)^2 > \underbrace{z^2_{\alpha/2}}_{\equiv \chi^2_\alpha(1)}$

$$\frac{N^2 - 2Nm\pi^\circ + m^2\pi^{\circ 2}}{m\pi^\circ(1-\pi^\circ)} = \dots = \frac{(N - \pi^\circ m)^2}{\pi^\circ m} + \frac{(m - N - m(1-\pi^\circ))^2}{m(1-\pi^\circ)}$$

$$= \sum_{i=1}^{2} \frac{(N_i - m\pi^\circ_i)^2}{m\pi^\circ_i} = X^2$$

CHI-QUADRO DI PEARSON

dove π°_1 = probabilità successo sotto H°
π°_2 = probabilità insuccesso sotto H°

TEST CHI-QUADRO

$X^2 \overset{?}{>} \chi^2_\alpha$ RIFIUTO H_0

$X^2 \leq \chi^2_\alpha$ ACCETTO H_0

② Caso multinomiale

TEO

$$\sum_{i=1}^{D} \frac{(N_i - m\pi_i^0)^2}{m\pi_i^0} \longrightarrow \chi^2(d)$$

Dim) ① $\sum_{i=1}^{D} \frac{(N_i - m\pi_i^0)^2}{m\pi_i^0} = \left(\frac{N_1}{m} - \pi_1, \dots\right) \Sigma^{-1} \left(\frac{N_1}{m} - \pi_1, \dots\right)^T$

× CASA

HINT: $\Sigma^{-1} = [\tau_{ij}] = \begin{cases} \frac{1}{\pi_i} + \frac{1}{\pi_D} & \text{se } i=j \\ \frac{1}{\pi_D} & \text{se } i \neq j \end{cases}$

② $\left(\frac{N_1}{m} - \pi_1, \dots\right) \Sigma^{-1} \left(\frac{N_1}{m} - \pi_1, \dots\right)^T \longrightarrow \chi^2(d)$

$$X^2 = \sum_{i=1}^{D} \frac{(N_i - m\pi_i^0)^2}{m\pi_i^0}$$

TEST CHI-QUADRO

$X^2 > \chi^2_\alpha(d)$ RIFIUTO H_0

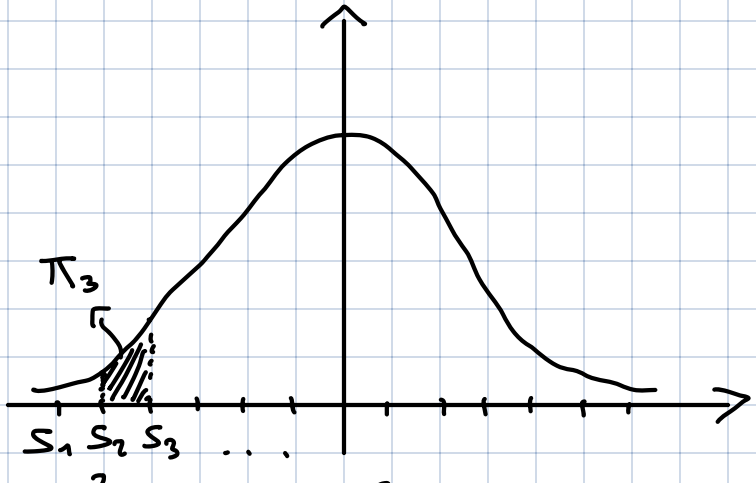
$X^2 \leq \chi^2_\alpha(d)$ ACCETTO H_0

TEST DI NORMALITÀ (μ, σ^2 NOTI)

$y_i \overset{iid}{\sim} ? \; N(0,1)$

$$\pi_i = \frac{1}{\sqrt{2\pi}} \int_{s_{i-1}}^{s_i} e^{-y^2/2} \, dy$$

$i = 1 \dots D$



TEST: $X^2 = \sum_{i=1}^{D} \frac{(N_i - m\pi_i)^2}{m\pi_i} \sim \chi^2(d)$

TEO BIS Se le probabilità $\vec{\pi}$ dipendono da un vettore di r parametri $\hat{\theta}$

$$X^2(\hat{\theta}) = \sum_{i=1}^{D} \frac{(N_i - m\pi_i(\hat{\theta}))^2}{m\pi_i(\hat{\theta})} \longrightarrow \chi^2(d-r)$$

TEST DI NORMALITÀ

$$y_i \overset{iid}{\sim} ? \; N(\mu, \sigma^2) \qquad \hat{\theta} = (\bar{Y}, S^2)$$

$\bar{Y}$: media campionaria; S^2: varianza campionaria

$$\pi_i(\hat{\theta}) = \frac{1}{\sqrt{2\pi S^2}} \int_{S_{i-1}}^{S_i} \exp\left(\frac{1}{S^2}(y - \bar{Y})^2\right) dy$$

sono nel caso del TEO BIS

TEST:

$$X^2(\hat{\theta}) = \sum_{i=1}^{D} \frac{(N_i - m\pi_i(\hat{\theta}))^2}{m\pi_i(\hat{\theta})}$$

RIFIUTO SE $X^2(\hat{\theta}) > \chi^2_\alpha(d-2)$

ACCETTO H_0 SE $X^2(\hat{\theta}) \leq \chi^2_\alpha(d-2)$

DUE VARIABILI CATEGORICHE

A \ C	1	2	...	J	
1	N_{11}	N_{12}	...	N_{1J}	$n_{1+} = \sum_{J=1}^{J} n_{1J}$
2	N_{21}				
⋮	⋮				
I	N_{I1}				
	N_{+1}	N_{+2}			n

TEST DI OMOGENEITÀ

(mi chiedo se le probabilità del fattore J sono omogenee lungo fattore i)

$$H_0 : \pi_{J|1} = \pi_{J|2} = \dots = \hat{\pi}_{+J} \quad \forall J = 1 \dots J$$

Sono parametri stimati a partire dai dati !!! (Sono nel caso del TEC BIS)

$$\hat{\pi}_{+J} = \frac{N_{+J}}{n} \quad \forall J = 1 \dots J$$

n° param. $= I(J-1)$

n° vincoli $= J - 1$

$$X^2 = \sum_{i=1}^{I} \sum_{J=1}^{J} \frac{(N_{iJ} - n_{i+}\hat{\pi}_{+J})^2}{n_{i+}\hat{\pi}_{+J}}$$

RIFIUTATO SE $X^2 > \chi^2_{\alpha}((I-1)(J-1))$

↘ ACCETTATO SE $X^2 \leq \chi^2_{\alpha}((I-1)(J-1))$