



UNIVERSITÀ · STATISTICA

Formulario generale di statistica

Statistica descrittiva e bivariata · probabilità · epidemiologia e test
diagnostici · variabili aleatorie · inferenza · intervalli di confidenza · test di
ipotesi · regressione semplice e multipla · ANOVA · modello logit



infomaat.it

Centro ripetizioni · since 2011



INDICE

- | | |
|-------------------------------------|----------------------------------|
| 1. Statistica descrittiva | 9. Inferenza e stimatori |
| 2. Rappresentazioni grafiche | 10. Intervalli di confidenza |
| 3. Statistica bivariata | 11. Test di ipotesi |
| 4. Indipendenza stocastica | 12. Regressione lineare semplice |
| 5. Dipendenza in media | 13. Regressione lineare multipla |
| 6. Probabilità | 14. ANOVA |
| 7. Epidemiologia e test diagnostici | 15. Modello logit |
| 8. Variabili aleatorie | |

01 Statistica descrittiva

FREQUENZE

- Frequenza assoluta n_i ; frequenza relativa $f_i = n_i/n$, con $\sum_i f_i = 1$.
- Frequenze cumulate: assolute $N_i = \sum_{j \leq i} n_j$, relative $F_i = \sum_{j \leq i} f_j$.
- Per caratteri in classi di ampiezza a_i : densità di frequenza $d_i = n_i/a_i$.

MODA

Modalità a cui è associata la massima frequenza (assoluta, relativa o massima densità se in classi).

- Calcolabile per **tutti** i tipi di carattere.
- È una modalità, non una frequenza; può non esistere o non essere unica.

MEDIANA

Modalità che occupa la posizione centrale nella distribuzione ordinata.

- n dispari: posizione $p = (n + 1)/2$, $Me = x_{(p)}$.
- n pari: $p = n/2$ e $p = n/2 + 1$,
 $Me = (x_{(n/2)} + x_{(n/2+1)})/2$.
- Sulla tabella: prima $N_i \geq p$ (frequenze cumulate).

MEDIANA PER CARATTERI IN CLASSI

$$Me = h_{i-1} + (p - N_{i-1}) \frac{a_i}{n_i} \quad \text{oppure} \quad Me = h_{i-1} + (0.5 - F_{i-1}) \frac{a_i}{f_i}$$

h_{i-1} = estremo inferiore della classe mediana; N_{i-1} , F_{i-1} = cumulate della classe precedente; p = posizione mediana.

QUARTILI E PERCENTILI

- Q_1 : prima modalità con cumulata relativa $F_i \geq 0.25$; $Q_2 = Me$; $F_i \geq 0.5$; Q_3 : $F_i \geq 0.75$.
- La mediana è il percentile di ordine **0.5**. Mediana e percentili: solo caratteri quantitativi o qualitativi ordinati.



MEDIE POTENZIATE

$$\mu^{(r)} = \left(\frac{1}{n} \sum_{i=1}^k (x_i)^r n_i \right)^{1/r} = \left(\sum_{i=1}^k (x_i)^r f_i \right)^{1/r}$$

$r = 1$ media aritmetica $\mu = \frac{1}{n} \sum x_i n_i = \sum x_i f_i$

$r = 2$ media quadratica $\mu^{(2)} = \left(\frac{1}{n} \sum x_i^2 n_i \right)^{1/2}$

$r = 0$ media geometrica $\ln \mu^{(0)} = \frac{1}{n} \sum \ln(x_i) n_i$, poi $\mu^{(0)} = e^{\ln \mu^{(0)}}$

$r = -1$ media armonica $\mu^{(-1)} = \frac{n}{\sum_i \frac{1}{x_i} n_i}$ ($x \neq 0$)

Funzione monotona crescente in r : $x_{\min} \leq \mu^{(-1)} \leq \mu^{(0)} \leq \mu \leq \mu^{(2)} \leq x_{\max}$.

MEDIA ARITMETICA – PROPRIETÀ

- I – **scarti nulli**: $\sum_i (x_i - \mu) n_i = 0$, ovvero $M(X - \mu) = 0$.
- II – **proprietà di minimo**: $\sum_i (x_i - \alpha)^2 n_i$ è minima per $\alpha = \mu$: $M[(X - \mu)^2] = \min$.
- III – **associativa**: la media generale è la media ponderata delle medie parziali: $\mu = \sum_{i=1}^h \mu_i \frac{n_i}{n}$.
- Operatore lineare**: $M(bX + a) = bM(X) + a$. La media è compresa fra $x_{\min}$ e $x_{\max}$ ed è influenzata dai valori estremi.

APPLICABILITÀ DEGLI INDICI DI POSIZIONE

CARATTERE	MODA	MEDIANA	MEDIA	PERCENTILI
Qualitativo nominale	sì	no	no	no
Qualitativo ordinale	sì	sì	no	sì
Quantitativo	sì	sì	sì	sì

INDICI DI VARIABILITÀ GLOBALE

- Campo di variazione: $x_{\max} - x_{\min}$; differenza interquartile: $Q_3 - Q_1$.
- Gli indici di variabilità sono sempre positivi, nulli se la variabile è degenere, invarianti per traslazione ($X + c$).

INDICE DI ETEROGENEITÀ DI GINI

$$E = \sum_{i=1}^k f_i(1 - f_i) = 1 - \sum_{i=1}^k f_i^2 \quad E_N = \frac{E}{E_{\max}} = \frac{1 - \sum f_i^2}{1 - 1/k}$$

$0 \leq E_N \leq 1$: $E_N = 0$ minima mutabilità (tutta la frequenza su una modalità), $E_N = 1$ massima (frequenze uguali).

SCOSTAMENTI MEDI DA UN CENTRO C

$$D_r(c) = \left[\sum_{i=1}^k |x_i - c|^r f_i \right]^{1/r}$$

$r = 1, c = \text{Me}$: scostamento medio assoluto dalla mediana. $r = 2, c = \mu$: scarto quadratico medio σ .



VARIANZA

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \overline{x^2} - \mu^2$$

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

- σ^2 varianza della popolazione; s^2 varianza campionaria (corretta, divisore $n-1$). Scarto quadratico medio: $\sigma = \sqrt{\sigma^2}$.
- Proprietà: $\text{Var}(a) = 0$; $\text{Var}(X+a) = \text{Var}(X)$; $\text{Var}(bX) = b^2 \text{Var}(X)$; in sintesi $\text{Var}(bX+a) = b^2 \text{Var}(X)$.

VARIANZA NORMALIZZATA

$$\sigma_N^2 = \frac{\sigma^2}{\sigma_{\max}^2} \in [0, 1]$$

In $[a, b]$: $\sigma_{\max}^2 = (b - \mu)(\mu - a)$. Per n dati non negativi di media μ : $\sigma_{\max}^2 = (n-1)\mu^2$.

COEFFICIENTE DI VARIAZIONE

$$CV = \frac{\sigma}{|\mu|} \cdot 100\% \quad CV = \frac{s}{|\bar{x}|} \cdot 100\%$$

Indice di variabilità relativa: permette confronti fra fenomeni con unità di misura o medie diverse.

TEOREMA DI SCOMPOSIZIONE DELLA VARIANZA

$$\sigma^2 = \sigma_W^2 + \sigma_B^2 \quad \sigma_W^2 = \frac{1}{n} \sum_{l=1}^h \sigma_l^2 n_l \quad \sigma_B^2 = \frac{1}{n} \sum_{l=1}^h (\mu_l - \mu)^2 n_l = \mu_l^2 - \mu^2$$

Dati suddivisi in h gruppi: la varianza totale è la somma della varianza **within** (media delle varianze dei gruppi) e della varianza **between** (varianza delle medie dei gruppi). Media totale: $\mu = \sum_l \mu_l n_l / n$.

INDICI DI FORMA

- Momenti centrali di ordine s : $\bar{\mu}_s = M[(X - \mu_X)^s] = \sum_i (x_i - \mu_X)^s f_i$ ($\bar{\mu}_1 = 0$; $\bar{\mu}_2 = \sigma^2$).
- Momenti dall'origine: $\mu_s = M(X^s) = \sum_i x_i^s f_i$ (μ_1 = media aritmetica).

$$\gamma_1 = \frac{M[(X - \mu)^3]}{\sigma^3} = \frac{\bar{\mu}_3}{\sigma^3}$$

- $\gamma_1 > 0$ asimmetria positiva (obliqua a destra, $\text{Mo} < \text{Me} < \mu$); $\gamma_1 < 0$ asimmetria negativa ($\mu < \text{Me} < \text{Mo}$).
- Distribuzione simmetrica: $\text{Mo} = \text{Me} = \mu$ e momenti centrali di ordine dispari nulli.

ATTENZIONE

$\gamma_1 = 0$ non garantisce la simmetria: l'annullarsi dell'indice è solo un sintomo, non vale il viceversa.

02 Rappresentazioni grafiche

QUALE GRAFICO PER QUALE CARATTERE

CARATTERE	GRAFICO	NOTE
Qualitativo sconnesso	Diagramma a torta o a rettangoli separati	torta preferibile con poche modalità
Qualitativo ordinato	Diagramma a rettangoli separati	rispetta l'ordine delle modalità
Quantitativo discreto	Diagramma a bastoncini	ordinate = frequenze
Quantitativo continuo in classi	Istogramma	con classi di ampiezza diversa usare le densità $d_i = n_i/a_i$

Nell'istogramma con densità l'**area** del rettangolo rappresenta la frequenza: base a_i , altezza d_i , area n_i .

BOX-PLOT

- Scatola da Q_1 a Q_3 con la mediana all'interno: contiene il 50% delle osservazioni.
- Baffi fino al valore più piccolo e più grande non anomali.
- Valori anomali (outlier): oltre $1.5(Q_3 - Q_1)$ da Q_1 o da Q_3 .

GRAFICO DELLE FREQUENZE CUMULATE

- Rappresenta l'andamento di F_i (o N_i) in funzione delle modalità.
- Per caratteri discreti si presenta come un grafico a gradini, non decrescente, da 0 a 1.

03 Statistica bivariata

TABELLA A DOPPIA ENTRATA

- n_{ij} = frequenza congiunta ($X = x_i, Y = y_j$); n = numerosità totale.
- Marginali: di riga $n_{i.} = \sum_{j=1}^k n_{ij}$ (distribuzione di X); di colonna $n_{.j} = \sum_{i=1}^h n_{ij}$ (distribuzione di Y).
- Relative: congiunta $f_{ij} = n_{ij}/n$; marginali $f_{i.} = n_{i.}/n, f_{.j} = n_{.j}/n$.
- Obiettivo dell'analisi bivariata: stabilire se X e Y sono indipendenti o legate. **Dipendenza $\neq$ causalità.**

DISTRIBUZIONI CONDIZIONATE

- Fissata la colonna j : variabile condizionata $X|y_j$ (k variabili condizionate); fissata la riga i : $Y|x_i$ (h variabili).
- Frequenze condizionate relative: $\frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$ (di $X|y_j$) e $\frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$ (di $Y|x_i$).

CONNESSIONE — INDICE CHI-QUADRO

$$E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \quad \chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = n \left\{ \sum_i \sum_j \frac{n_{ij}^2}{n_{i.} \cdot n_{.j}} - 1 \right\}$$

- E_{ij} = frequenze attese (teoriche) nell'ipotesi di indipendenza; $c_{ij} = n_{ij} - E_{ij}$ = contingenze.
- $\chi^2_{\min} = 0$: indipendenza stocastica (osservate = teoriche). $\chi^2_{\max} = n \cdot \min(r-1, c-1)$: dipendenza funzionale.

INDICI NORMALIZZATI

$$\chi^2_N = \frac{\chi^2}{n \min(r-1, c-1)} \in [0, 1] \quad V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} \in [0, 1]$$

V di Cramer: 0 = variabili non associate (indipendenti); 1 = perfettamente dipendenti. Lo studio della connessione è simmetrico: $\chi^2_N(X|Y) = \chi^2_N(Y|X)$.

COVARIANZA

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) = \overline{xy} - \bar{x} \bar{y} \quad S_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{n}{n-1} \left(\frac{\sum x_i y_i}{n} - \bar{x} \bar{y} \right)$$

- $\sigma_{XY} > 0$: direttamente correlate; $\sigma_{XY} < 0$: inversamente correlate; $\sigma_{XY} = 0$: non linearmente correlate.
- Indice simmetrico: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$; limiti: $-\sigma_X \sigma_Y \leq \sigma_{XY} \leq \sigma_X \sigma_Y$; risente dell'unità di misura.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$; con $W = a + bY, Z = c + dX$:
 $\text{Cov}(W, Z) = bd \text{Cov}(X, Y)$.

COEFFICIENTE DI CORRELAZIONE DI PEARSON

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [-1, 1] \quad r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Esprime tipo (diretta/inversa) e intensità della relazione **lineare**. $\rho = \pm 1$: legame lineare perfetto; $\rho = 0$: incorrelazione.

04 Indipendenza stocastica

DEFINIZIONE

X e Y sono **stocasticamente indipendenti** se tutte le distribuzioni condizionate (relative) sono uguali fra loro e uguali alle rispettive marginali:

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} = f_{i.} \quad \forall i, j \quad \Longleftrightarrow \quad f_{ij} = f_{i.} \cdot f_{.j} \quad (\text{teorema di fattorizzazione}) \quad \Longleftrightarrow \quad \chi^2 = 0$$

RELAZIONE FRA LE FORME DI INDIPENDENZA

indipendenza **stocastica** (simmetrica) $\Rightarrow$ indipendenza **in media** (non necessariamente simmetrica) $\Rightarrow$ indipendenza **lineare** / incorrelazione (simmetrica)

Non vale il viceversa in nessuno dei due passaggi.

- Indipendenza stocastica $\Rightarrow \eta_Y^2 = \eta_X^2 = 0$; dipendenza funzionale biunivoca $\Rightarrow \eta_Y^2 = \eta_X^2 = 1$.
- $\rho = 0$ (incorrelazione) **non** implica indipendenza stocastica: esclude solo il legame lineare.

05 Dipendenza in media

FUNZIONE DI REGRESSIONE

$$g(x_i) = M(Y|x_i) \quad \text{minimizza} \quad L[y-g(x)] = M[(Y-g(X))^2] = \frac{1}{n} \sum_i \left[\sum_j (y_j - g(x_i))^2 \frac{n_{ij}}{n_{i.}} \right] n_{i.}$$

Il miglior modello in media è la funzione (spezzata di regressione) che in x_i assume il valore della media condizionata $M(Y|x_i)$.

INDIPENDENZA IN MEDIA

Y è indipendente in media da X se le medie condizionate sono tutte uguali (e uguali alla marginale):

$\mu_Y(x_1) = \dots = \mu_Y(x_h) = \mu_Y$. La relazione **non è simmetrica**: indipendenza di Y da X non implica quella di X da Y .

II TEOREMA DI SCOMPOSIZIONE DELLA VARIANZA

$$\sigma_Y^2 = \underbrace{\bar{\sigma}_Y^2}_{\text{spiegata}} + \underbrace{\sigma_Y^{2*}}_{\text{residua}} \quad \bar{\sigma}_Y^2 = \text{Var}(\mu_Y(x_i)) = M(\mu_Y(x_i)^2) - \mu_Y^2 \quad \sigma_Y^{2*} = M(\sigma_Y^2(x_i))$$

Varianza **spiegata** = varianza delle medie condizionate (fra i gruppi); varianza **residua** = media delle varianze condizionate (nei gruppi).

RAPPORTO DI CORRELAZIONE

$$\eta_Y^2 = \frac{\bar{\sigma}_Y^2}{\sigma_Y^2} = 1 - \frac{\sigma_Y^{2*}}{\sigma_Y^2} \quad 0 \leq \eta_Y^2 \leq 1$$

- $\eta_Y^2 = 0$: indipendenza in media di Y da X (varianza delle medie condizionate nulla).
- $\eta_Y^2 = 1$: dipendenza funzionale di Y da X (varianze condizionate tutte nulle: a ogni x corrisponde una sola y).
- In generale $\eta_Y^2 \neq \eta_X^2$; l'indice è simmetrico nei casi estremi (indipendenza stocastica: entrambi 0; dipendenza funzionale biunivoca: entrambi 1).

06 Probabilità

DEFINIZIONI DI BASE

- **Esperimento casuale:** processo la cui singola esecuzione (prova) dà un risultato non prevedibile. **Spazio campionario** Ω : insieme di tutti gli esiti possibili (discreto o continuo).
- **Evento elementare:** un possibile esito; **evento:** qualsiasi sottoinsieme di Ω . Eventi **disgiunti/incompatibili**: $A \cap B = \emptyset$.
- Leggi di De Morgan: $\overline{A \cup B} = \bar{A} \cap \bar{B}$; $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

ASSIOMI E CONSEGUENZE

- Assiomi: $P(A) \geq 0$; $P(\Omega) = 1$; se $A \cap B = \emptyset$ allora $P(A \cup B) = P(A) + P(B)$.
- Conseguenze: $P(\bar{A}) = 1 - P(A)$; $P(\emptyset) = 0$; $0 \leq P(A) \leq 1$.
- **Regola della somma:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- Assegnazione: **classica** $P(A)$ = casi favorevoli / casi possibili (equiprobabili); **frequentista** $P(A) \approx n$. successi / n. prove (per molte prove).

PROBABILITÀ CONDIZIONATA

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0 \qquad P(A \cap B) = P(A|B) P(B) = P(B|A) P(A)$$

In generale $P(A|B) \neq P(B|A)$.

PROBABILITÀ TOTALE E TEOREMA DI BAYES

$$P(A) = \sum_{i=1}^k P(A|E_i) P(E_i) \qquad P(E_i|A) = \frac{P(A|E_i) P(E_i)}{\sum_{i=1}^k P(A|E_i) P(E_i)}$$

con $\{E_i\}$ partizione di Ω : $\bigcup_i E_i = \Omega$, $E_i \cap E_j = \emptyset$ per $i \neq j$. Caso di due eventi:

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \text{ e } P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}.$$

INDIPENDENZA FRA EVENTI

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

- Campionamento casuale semplice **con** reinserimento: prove indipendenti; **senza** reinserimento: prove non indipendenti.

ATTENZIONE

Incompatibilità ($A \cap B = \emptyset$) e indipendenza sono concetti diversi: due eventi incompatibili con probabilità positive **non** sono mai indipendenti; $A \cap B \neq \emptyset$ non implica indipendenza.

07 Epidemiologia e test diagnostici

TABELLA ESPOSIZIONE × MALATTIA

FATTORE DI RISCHIO	D (MALATTIA PRESENTE)	$\bar{D}$ (ASSENTE)	TOT.
R (esposti)	a	b	a+b
$\bar{R}$ (non esposti)	c	d	c+d
Tot.	a+c	b+d	n

RISCHIO RELATIVO (RR)

$$RR = \frac{P_{\text{esposti}}}{P_{\text{non esposti}}} = \frac{P(D|R)}{P(D|\bar{R})} = \frac{a/(a+b)}{c/(c+d)}$$

- Rapporto fra incidenza negli esposti e nei non esposti: misura la forza di associazione fra fattore e malattia.
- $RR > 1$: fattore di rischio; $RR < 1$: fattore di protezione; $RR = 1$: nessuna associazione.

ODDS RATIO (OR)

$$OR = \frac{ODD_{\text{esposti}}}{ODD_{\text{non esposti}}} = \frac{a/b}{c/d} = \frac{a \cdot d}{c \cdot b}$$

- Rapporto fra gli odds di malattia negli esposti e nei non esposti.
- $OR > 1$: fattore di rischio; $OR < 1$: fattore di protezione.

TIPO DI STUDIO E MISURA DA USARE

- Studio prospettico (di coorte)**: gruppi già classificati per fattore di rischio, seguiti nel tempo per verificare l'insorgenza della malattia (noti $a + b$ e $c + d$) → si calcola il **rischio relativo**.
- Studio retrospettivo (caso-controllo)**: individui già classificati per insorgenza della malattia, si risale al fattore di rischio → si calcola l'**odds ratio**.

TEST DIAGNOSTICI

	M (MALATO)	$\bar{M}$ (SANO)	TOT.
T ⁺	VP	FP	VP+FP
T ⁻	FN	VN	FN+VN
Tot.	VP+FN	FP+VN	n

Sensibilità $P(T^+|M) = \frac{VP}{VP+FN} = \frac{P(M|T^+) P(T^+)}{P(M)}$

Specificità $P(T^-|\bar{M}) = \frac{VN}{FP+VN} = \frac{P(\bar{M}|T^-) P(T^-)}{P(\bar{M})}$

Falso positivo $P(T^+|\bar{M}) = 1 - \text{specificità}$

Falso negativo $P(T^-|M) = 1 - \text{sensibilità}$

VALORI PREDITTIVI (BAYES)

$$VPP = P(M|T^+) = \frac{P(T^+|M) P(M)}{P(T^+)} = \frac{VP}{VP+FP}$$

$$VPN = P(\bar{M}|T^-) = \frac{P(T^-|\bar{M}) P(\bar{M})}{P(T^-)} = \frac{VN}{FN+VN}$$

Probabilità di test errato: $P(N) = P(M \cap T^-) + P(\bar{M} \cap T^+)$. I valori predittivi dipendono dalla prevalenza $P(M)$, sensibilità e specificità no.

08 Variabili aleatorie

DEFINIZIONE

Una **variabile casuale** (aleatoria) è una funzione che associa a ogni evento elementare di Ω uno e un solo numero reale (determinazione), a cui risulta associata la probabilità del suo verificarsi. È **discreta** se Ω è discreto, **continua** se Ω è continuo.

VARIABILI CASUALI DISCRETE

- Funzione di probabilità: $p_i = P(X = x_i)$ con $p_i \geq 0$ e $\sum_i p_i = 1$.
 - Funzione di ripartizione: $F(x_0) = P(X \leq x_0) = \sum_{x_i \leq x_0} p_i$; $F(-\infty) = 0$, $F(+\infty) = 1$,
 $P(x' < X \leq x'') = F(x'') - F(x')$.
- $$\mu = E(X) = \sum_i x_i p_i \quad \sigma^2 = \text{Var}(X) = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 p_i = E(X^2) - \mu^2$$

PROPRIETÀ DI VALORE ATTESO E VARIANZA

- $E(a + bX) = a + bE(X)$; $\text{Var}(a + bX) = b^2 \text{Var}(X)$.
- $E(X + Y) = E(X) + E(Y)$ sempre; se $X \perp Y$: $E(XY) = E(X)E(Y)$ e $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
- X e Y indipendenti: conoscere il valore di una non modifica le probabilità dell'altra.

BERNOULLI

$$X \in \{0, 1\} \quad P(X=1) = p$$
$$E(X) = p; \text{Var}(X) = p(1 - p).$$

BINOMIALE

$$X \sim \text{Bin}(n, p) \quad P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Numero di successi in n prove indipendenti;

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad E(X) = np;$$

$$\text{Var}(X) = np(1 - p).$$

IPERGEOMETRICA

$$p_k = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \quad k = 0, 1, \dots, n$$

Estrazioni **senza** reinserimento: N palline totali, M successi, n estrazioni. $E(X) = n \frac{M}{N}$;
 $\text{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}$.

POISSON

$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

Conteggio di eventi rari; $E(X) = \text{Var}(X) = \lambda$.
Approssima la binomiale per n grande e p piccolo
con $\lambda = np$.

VARIABILI CASUALI CONTINUE

Modello $\{x, f(x)\}$ con funzione di densità di probabilità $f(x) \geq 0$. Il calcolo delle probabilità avviene per integrazione (aree sotto la curva):

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du \quad \int_{-\infty}^{+\infty} f(u) du = 1 \quad P(a \leq X \leq b) = \int_a^b f(u) du = F(b) - F(a)$$

$P(X = x) = 0$ per ogni singolo punto, quindi $P(a \leq X \leq b) = P(a < X < b)$. Momenti:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad \text{Var}(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = E(X^2) - \mu^2$$

UNIFORME CONTINUA

$$f(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

$$F(x) = \frac{x-a}{b-a} \text{ su } [a, b]; \quad E(X) = \frac{a+b}{2};$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}.$$

ESPOENZIALE

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$$F(x) = 1 - e^{-\lambda x}; \quad E(X) = \frac{1}{\lambda};$$

$$\text{Var}(X) = \frac{1}{\lambda^2}. \text{ Assenza di memoria:}$$

$$P(X > s + t | X > s) = P(X > t). \text{ Tempi di attesa fra eventi di Poisson.}$$

NORMALE (GAUSSIANA)

$$X \sim N(\mu; \sigma^2) \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Simmetrica con centro in μ : $\mu = E(X) = \text{Me} = \text{Mo}$; parametri μ (media) e σ^2 (varianza).
- **Standardizzazione:** $Z = \frac{X - \mu}{\sigma} \sim N(0; 1)$, quindi $P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma}) = \Phi(z)$ (tavole).
- Simmetria: $\Phi(-z) = 1 - \Phi(z)$; $P(Z > z) = 1 - \Phi(z)$; $P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1)$.
- Ogni combinazione lineare di normali indipendenti è normale; in particolare $\bar{X} \sim N(\mu; \sigma^2/n)$.

DISTRIBUZIONI DERIVATE DALLA NORMALE

χ^2_n somma dei quadrati di n normali standard indipendenti: $\chi^2_n = \sum_{i=1}^n Z_i^2$; $E = n$, $\text{Var} = 2n$; asimmetrica positiva. Usi: varianze, test chi-quadro.

t_n $t_n = \frac{Z}{\sqrt{\chi^2_n/n}}$; simmetrica intorno a 0, code più pesanti della normale; $t_n \rightarrow N(0; 1)$ per n grande. Usi: inferenza sulla media con varianza ignota.

$F_{m,n}$ $F_{m,n} = \frac{\chi^2_m/m}{\chi^2_n/n}$; asimmetrica positiva. Usi: confronto di varianze, ANOVA, test F della regressione.

09 Inferenza e stimatori

IMPOSTAZIONE

- Obiettivo: trarre conclusioni su una **popolazione** a partire dai risultati di un **campione** (stima e verifica di ipotesi).
- Parametri**: funzioni di tutte le osservazioni della popolazione (es. μ, σ^2, π). **Statistiche campionarie**: funzioni dei soli dati campionari (es. $\bar{X}, S^2, \hat{p}$); al variare del campione sono variabili casuali con una propria **distribuzione campionaria**.

MEDIA CAMPIONARIA

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad E(\bar{X}) = \mu \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (\text{errore standard})$$

L'errore standard cresce con la varianza di X nella popolazione e diminuisce al crescere di n . Se $X_1, \dots, X_n$ provengono da una popolazione normale, $\bar{X}$ ha distribuzione normale esatta.

TEOREMA DEL LIMITE CENTRALE

Per n sufficientemente grande, **qualunque** sia la distribuzione di X :

$$\bar{X} \approx N\left(\mu; \frac{\sigma^2}{n}\right) \quad \hat{p} \approx N\left(\pi; \frac{\pi(1-\pi)}{n}\right)$$

PROPORZIONE CAMPIONARIA

$$\hat{p} = \frac{X}{n} \quad E(\hat{p}) = \pi \quad \text{Var}(\hat{p}) = \frac{\pi(1-\pi)}{n}$$

X = numero di successi nel campione, π = proporzione (non nota) nella popolazione.

VARIANZA CAMPIONARIA

$$S_c^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{n}{n-1} (X^2 - \bar{X}^2) \quad E(S_c^2) = \sigma^2$$

Con divisore n ($\tilde{S}^2$) lo stimatore è distorto: $E(\tilde{S}^2) = \frac{n-1}{n} \sigma^2$.

PROPRIETÀ DEGLI STIMATORI

Correttezza $E(T) = \theta$: lo stimatore è non distorto (es. $\bar{X}$ per μ , S_c^2 per σ^2).

Efficienza fra stimatori corretti è più efficiente quello con varianza minore.

Consistenza al crescere di n lo stimatore converge al parametro (errore quadratico medio $\rightarrow 0$).

10 Intervalli di confidenza

INTERPRETAZIONE

Estraendo ripetutamente campioni, il $100(1 - \alpha)\%$ degli intervalli costruiti contiene il vero valore (sconosciuto) del parametro; il $100\alpha\%$ non lo contiene. Livello di confidenza tipico: $1 - \alpha = 0.95$, $z_{0.975} = 1.96$.

MEDIA, VARIANZA NOTA

$$IC_{\alpha}(\mu) : \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad A = 2 z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}} \quad ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}$ = quantile della normale standard che lascia $\alpha/2$ nella coda destra; A = ampiezza; ME = margine di errore.

MEDIA, VARIANZA NON NOTA

$$IC_{\alpha}(\mu) : \left[\bar{x} - t_{n-1, \alpha/2} \frac{s_c}{\sqrt{n}}; \bar{x} + t_{n-1, \alpha/2} \frac{s_c}{\sqrt{n}} \right] \quad A = 2 t_{n-1, \alpha/2} \sqrt{\frac{S_c^2}{n}}$$

Si sostituisce σ con s_c e la normale con la t di Student a $n - 1$ gradi di libertà.

NUMEROSITÀ CAMPIONARIA

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{ME^2}$$

Numerosità minima per ottenere un margine di errore ME prefissato (arrotondare per eccesso).

PROPORZIONE

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Valido per n grande (TCL). Numerosità:

$$n = z_{\alpha/2}^2 \hat{p}(1 - \hat{p}) / ME^2 \text{ (caso peggiore } \hat{p} = 0.5).$$

DIFFERENZA FRA DUE MEDIE, VARIANZE NOTE

$$IC_{\alpha}(\mu_1 - \mu_2) : \left[(\bar{x}_1 - \bar{x}_2) \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

DIFFERENZA FRA DUE MEDIE, VARIANZE UGUALI NON NOTE

$$S_p^2 = \frac{S_{1c}^2(n_1 - 1) + S_{2c}^2(n_2 - 1)}{n_1 + n_2 - 2} \quad IC_{\alpha}(\mu_1 - \mu_2) : \left[(\bar{x}_1 - \bar{x}_2) \mp t_{n_1+n_2-2, \alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right]$$

S_p^2 = varianza pooled (combinata); gradi di libertà $n_1 + n_2 - 2$.

DATI APPAIATI (CAMPIONI DIPENDENTI)

$$D_i = X_i - Y_i \quad S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1} \quad IC : \left[\bar{d} \mp t_{n-1, \alpha/2} \frac{s_D}{\sqrt{n}} \right] \quad ME = t_{n-1, \alpha/2} \frac{s_D}{\sqrt{n}}$$

Si lavora sulla variabile differenza: l'inferenza si riduce al caso di una sola media.

11 Test di ipotesi

IMPOSTAZIONE

- **Ipotesi nulla** H_0 (nei test di significatività: semplice, un solo valore del parametro) e **ipotesi alternativa** H_1 (composta). Si fissa α e si usa una **statistica test** funzione dei dati campionari.
- **Errore di I tipo:** $\alpha = P(\text{rifiutare } H_0 \mid H_0 \text{ vera})$; **errore di II tipo:** $\beta = P(\text{non rifiutare } H_0 \mid H_0 \text{ falsa})$; **potenza** $= 1 - \beta$.
- L'intervallo dei valori della statistica test si divide in **zona di accettazione** e **zona di rifiuto**, delimitate dal **valore critico**.

TEST SULLA MEDIA

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (\text{Z-test, } \sigma \text{ nota}) \qquad U = \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \quad (\text{T-test, } \sigma \text{ non nota})$$

H_0	H_1	RIFIUTO H_0 SE (Z-TEST)	RIFIUTO H_0 SE (T-TEST)
$\mu = \mu_0$	$\mu \neq \mu_0$	$ U \geq z_{1-\alpha/2}$	$ U \geq t_{n-1, 1-\alpha/2}^{n-1}$
$\mu = \mu_0$ ($\text{o } \mu \leq \mu_0$)	$\mu > \mu_0$	$U \geq z_{1-\alpha}$	$U \geq t_{n-1, 1-\alpha}^{n-1}$
$\mu = \mu_0$ ($\text{o } \mu \geq \mu_0$)	$\mu < \mu_0$	$U \leq -z_{1-\alpha}$	$U \leq -t_{n-1, 1-\alpha}^{n-1}$

TEST SULLA PROPORZIONE

$$U = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0;1)$$

Stesse regioni di rifiuto del **Z**-test (bilaterale: $|U| \geq z_{1-\alpha/2}$; unilaterali: $U \geq z_{1-\alpha}$ oppure $U \leq -z_{1-\alpha}$).

CONFRONTO FRA DUE MEDIE

$$U = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (\text{varianze note}) \qquad U = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2} \quad (\text{varianze uguali non note})$$

Sotto $H_0 : \mu_1 = \mu_2$. Per dati appaiati: $U = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t_{n-1}$ con $H_0 : \mu_D = 0$.

P-VALUE

Probabilità, sotto H_0 , di osservare un valore della statistica test uguale o più estremo di quello ottenuto. Si rifiuta H_0 se $p\text{-value} < \alpha$. Con statistica u^* osservata: bilaterale $2(1 - \Phi(|u^*|))$; unilaterale destro $1 - \Phi(u^*)$; sinistro $\Phi(u^*)$ (per il **T**-test si usa F_{n-1} al posto di Φ).

TEST CHI-QUADRO DI INDIPENDENZA

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \quad \text{rifiuto } H_0 \text{ se } \chi^2 \geq \chi_{\alpha; (r-1)(c-1)}^2$$

H_0 : le due variabili sono indipendenti; H_1 : sono associate. Regione di rifiuto sulla coda destra, gradi di libertà $(r-1)(c-1)$.

TEST CHI-QUADRO DI ADATTAMENTO

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n p_i^0)^2}{n p_i^0} \quad \text{gdl} = k - 1 - m$$

H_0 : le frequenze osservate seguono la distribuzione ipotizzata (probabilità p_i^0); m = numero di parametri stimati dai dati.

12 Regressione lineare semplice

RETTE DEI MINIMI QUADRATI

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{XY}}{S_X^2} = r_{XY} \frac{S_Y}{S_X} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Minimizza la somma degli scarti quadratici fra valori osservati e previsti: $\sum_i (y_i - \hat{y}_i)^2$.
- La pendenza è la variazione stimata nel valore medio di Y per una variazione unitaria di X ; segno concorde con r_{XY} .
- La retta passa per il punto $(\bar{x}, \bar{y})$.

DECOMPOSIZIONE DELLA VARIANZA E R^2

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\widetilde{SST}} = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\widetilde{SSE}} + \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\widetilde{SSR}} \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = (r_{XY})^2 \in [0, 1]$$

R^2 misura la quota di variabilità totale di Y spiegata dal modello: $R^2 \rightarrow 1$ ottimo adattamento lineare, $R^2 \rightarrow 0$ la retta non spiega nulla. Inoltre $SST = (n-1) s_Y^2$ e $SSR = (n-1) r_{XY}^2 s_Y^2$.

VARIANZA RESIDUA DEI MODELLI

$y = a$ stima $\hat{a} = \mu_Y$; varianza residua $\sigma_R^2 = M[(Y - \mu_Y)^2] = \sigma_Y^2$

$y = bx$ stima $\hat{b} = \frac{M(XY)}{M(X^2)}$; $\sigma_R^2 = M(Y^2) - \hat{b} M(XY)$

$y = a + bx$ stima $\hat{b} = \frac{\text{Cov}(X, Y)}{\sigma_X^2}$, $\hat{a} = M(Y) - \hat{b} M(X)$; $\sigma_R^2 = \sigma_Y^2 (1 - \rho^2)$; varianza spiegata

$$\sigma_{SP}^2 = \frac{[\text{Cov}(X, Y)]^2}{\text{Var}(X)}$$

Con $E = Y - Y^*$ (residui): $M(E) = 0$ e varianza residua $\sigma_{RE}^2 = V(E) = M(E^2) = M[Y - Y^*]^2$. Il modello migliore è quello con varianza residua minore.

MODELLI NON LINEARI LINEARIZZABILI

MODELLO NON LINEARE	MODELLO LINEARE NELLE VARIABILI TRASFORMATE
$y = a e^{bx}$	$(\ln y) = (\ln a) + b x$
$y = a x^b$	$(\ln y) = (\ln a) + b (\ln x)$
$y = a b^x$	$(\ln y) = (\ln a) + x (\ln b)$
$y = a + b/x$	$y = a + b (1/x)$
$y = 1/(a + bx)$	$(1/y) = a + b x$
$y = x/(a + bx)$	$(1/y) = b + a (1/x)$

Sui modelli trasformati si può calcolare R^2 o ρ^2 ; in generale per i modelli non lineari non vale il teorema di scomposizione della varianza.

INFERENZA SULLA PENDENZA

$$s^2 = \frac{SSE}{n-2} \quad se(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} \quad t = \frac{\hat{\beta}_1 - \beta_1^0}{se(\hat{\beta}_1)} \sim t_{n-2} \quad IC: \hat{\beta}_1 \mp t_{n-2, \alpha/2} se(\hat{\beta}_1)$$

Test di significatività della pendenza: $H_0: \beta_1 = 0$ (nessuna relazione lineare) contro $H_1: \beta_1 \neq 0$.

13 Regressione lineare multipla
MODELLO

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

Ogni $\hat{\beta}_j$ è l'effetto **parziale** di x_j su y : variazione stimata del valore medio di Y per una variazione unitaria di x_j , a **parità delle altre variabili**. Ipotesi classiche: errori a media nulla, omoschedastici, incorrelati, (normali per l'inferenza esatta).

 R^2 E R^2 AGGIUSTATO

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1}$$

R^2 non diminuisce mai aggiungendo regressori; R_{adj}^2 penalizza i regressori inutili ed è l'indice corretto per confrontare modelli con un numero diverso di variabili.

TEST F GLOBALE

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k, n-k-1}$$

Verifica la significatività complessiva del modello; si rifiuta H_0 per $F \geq F_{\alpha; k, n-k-1}$ (coda destra).

TEST T SUI SINGOLI COEFFICIENTI

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} \quad H_0 : \beta_j = 0 \quad IC : \hat{\beta}_j \mp t_{n-k-1, \alpha/2} se(\hat{\beta}_j)$$

MULTICOLLINEARITÀ

Regressori fortemente correlati fra loro: stime instabili, errori standard gonfiati, coefficienti con segni controintuitivi (con F globale magari significativo e t singoli no). Diagnostica: $VIF_j = \frac{1}{1 - R_j^2}$, dove R_j^2 è l' R^2 della regressione di x_j sulle altre esplicative; valori oltre 5-10 segnalano un problema.

14 ANOVA

ANOVA A UNA VIA

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k \quad H_1 : \text{almeno una media diversa}$$

Confronto fra le medie di k gruppi. Ipotesi: campioni indipendenti, popolazioni normali, varianze uguali (omoschedasticità). Con n_j osservazioni nel gruppo j , $N = \sum_j n_j$:

$$SSB = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 \quad SSW = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \quad SST = SSB + SSW$$

$$F = \frac{MSB}{MSW} = \frac{SSB/(k-1)}{SSW/(N-k)} \sim F_{k-1, N-k} \quad \text{rifiuto } H_0 \text{ se } F \geq F_{\alpha; k-1, N-k}$$

TABELLA ANOVA (UNA VIA)

FONTE	SOMMA DEI QUADRATI	GDL	MEDIA DEI QUADRATI	F
Fra i gruppi (between)	SSB	$k - 1$	$MSB = SSB/(k-1)$	$F = MSB/MSW$
Entro i gruppi (within)	SSW	$N - k$	$MSW = SSW/(N-k)$	
Totale	SST	$N - 1$		

È la scomposizione della varianza in chiave inferenziale: between = spiegata dal fattore, within = residua.

ANOVA A DUE VIE SENZA REPLICAZIONE

Due fattori: A su r righe, B su c colonne, una osservazione per cella (x_{ij}).

$$SSA = c \sum_{i=1}^r (\bar{x}_{i.} - \bar{x})^2 \quad SSB = r \sum_{j=1}^c (\bar{x}_{.j} - \bar{x})^2 \quad SSE = SST - SSA - SSB$$

$$F_A = \frac{SSA/(r-1)}{SSE/[(r-1)(c-1)]} \sim F_{r-1, (r-1)(c-1)} \quad F_B = \frac{SSB/(c-1)}{SSE/[(r-1)(c-1)]} \sim F_{c-1, (r-1)(c-1)}$$

Si verifica separatamente l'effetto di ciascun fattore (H_0 : medie di riga uguali; H_0 : medie di colonna uguali).

ANOVA A DUE VIE CON REPLICAZIONE (INTERAZIONE)

Con $m > 1$ osservazioni per cella si può stimare anche l'interazione $A \times B$:

$$SST = SSA + SSB + SS_{AB} + SSE \quad F_{AB} = \frac{SS_{AB}/[(r-1)(c-1)]}{SSE/[rc(m-1)]} \sim F_{(r-1)(c-1), rc(m-1)}$$

gdl: fattore A = $r - 1$; fattore B = $c - 1$; interazione = $(r - 1)(c - 1)$; errore = $rc(m - 1)$; totale = $rcm - 1$. Se l'interazione è significativa, gli effetti principali vanno interpretati con cautela.

15 Modello logit

PERCHÉ NON IL MODELLO LINEARE

Con Y binaria (0/1) il modello lineare $p = \beta_0 + \beta_1 x$ può produrre probabilità fuori da $[0, 1]$, con errori non normali ed eteroschedastici. Serve una funzione che comprima $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ in $(0, 1)$.

MODELLO LOGISTICO

$$p = P(Y=1 | \mathbf{x}) = \frac{e^\eta}{1 + e^\eta} = \frac{1}{1 + e^{-\eta}} \quad \text{logit}(p) = \ln \frac{p}{1-p} = \eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

La trasformazione logit rende lineare il legame fra probabilità e regressori; $p \in (0, 1)$ per costruzione, curva a S (sigmoide).

ODDS E ODDS RATIO

$$\text{odds} = \frac{p}{1-p} = e^\eta \quad OR_j = \frac{\text{odds}(x_j + 1)}{\text{odds}(x_j)} = e^{\beta_j}$$

- $\beta_j > 0$ ($OR_j > 1$): l'aumento di x_j aumenta la probabilità dell'evento; $\beta_j < 0$ la riduce; $\beta_j = 0$ nessun effetto.
- e^{β_j} è l'odds ratio per un incremento unitario di x_j , a parità delle altre variabili.

STIMA E VERIFICA

- Stima per **massima verosimiglianza** (non OLS): si massimizza $\ln L(\beta)$ numericamente.
- Test di Wald** sul singolo coefficiente: $z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \approx N(0; 1)$, con $H_0: \beta_j = 0$.
- Confronto fra modelli: test del rapporto di verosimiglianza $LR = -2(\ln L_0 - \ln L_1) \sim \chi_q^2$ (q = vincoli).
- Adattamento: pseudo- R^2 di McFadden = $1 - \frac{\ln L_1}{\ln L_0}$ (non ha l'interpretazione dell' R^2 classico).