# Modelli loglineari

# Introduzione

- I dati da analizzare sono organizzati sotto forma di tabelle di frequenze: le unità statistiche sono classificate sulla base delle modalità di una o più variabili generalmente di tipo **categoriale**
- Le frequenze (assolute) osservate nelle classi (celle) di una tabella di contingenza rappresentano le risposte.
- Ad esempio, consideriamo il caso bidimensionale con una tabella di frequenza di dimensione $I \times J$.
- I dati si presentano nella forma

$$\mathbf{y} = (y_{11}, \ldots, y_{IJ}),$$

dove $y_{ij}, i = 1, \ldots, I$ e $j = 1, \ldots, J$, è il numero di unità statistiche classificate nella $(i, j)$-esima classe

# Two-way contingency tables

- Una tabella di contingenza $I \times J$ riporta le frequenze osservate secondo le combinazioni di due variabili categoriali $(H, X)$

| | X | | | |
|---|---|---|---|---|
| H | 1 | 2 | ... | J |
| 1 | $y_{11}$ | $y_{12}$ | ... | $y_{1J}$ |
| 2 | $y_{21}$ | $y_{22}$ | ... | $y_{2J}$ |
| ... | ... | ... | ... | |
| I | $y_{I1}$ | $y_{I2}$ | ... | $y_{IJ}$ |

- Distribuzione congiunta $(H, X)$, distribuzioni condizionate $(H|X)$, distribuzioni marginali $H, X$

# Ipotesi di indipendenza

- Due variabili sono statisticamente indipendenti se la distribuzione condizionata dell'una non cambia entro i livelli dell'altra
- Nell'ipotesi di indipendenza, le frequenze congiunte sono determinate dalle distribuzioni marginali

$$y_{ij} = \frac{y_{i\cdot} y_{\cdot j}}{n}$$

- Per verificare l'ipotesi di indipendenza
  1. Test esatto di Fisher in tabelle $2 \times 2$ con distribuzione nulla ipergeometrica
  2. In tabelle a due vie, quando almeno una variabile ha più di due livelli, si può utilizzare il test $X^2$ di Pearson o il TRV $G^2$ con distribuzione asintotica nulla $\chi^2_g, g = (I-1)(J-1)$ o anche una versione multivariata del test esatto di Fisher

# Tabelle di dimensione maggiore

- L'ipotesi di indipendenza costituisce solo un caso particolare delle possibili situazioni e problemi che si possono incontrare nell'analisi delle tabelle di frequenze.
- Nella pratica, nelle tabelle di dimensione maggiore di due, è interessante studiare le possibili relazioni tra le variabili rilevate.
- Ad esempio, date tre variabili $X, H, Z$, è possibile valutare se sono tutte e tre indipendenti, se c'è dipendenza solo a coppie, o se dipendono sia a coppie, che tutte e tre insieme.
- In molti studi, l'interesse consiste nello studiare le relazioni tra $H$ e $X$ mantenendo fisso il livello della terza variabile $Z$; in questi casi $Z$ è una variabile di controllo

# Three-way contingency tables

- Tabella di contingenza $I \times J \times K$

| | | X | | | |
|---|---|---|---|---|---|
| Z | H | 1 | 2 | ... | J |
| 1 | 1 | $y_{11(1)}$ | $y_{12(1)}$ | $\cdots$ | $y_{1J(1)}$ |
| | 2 | $y_{21(1)}$ | $y_{22(1)}$ | $\cdots$ | $y_{2J(1)}$ |
| | ... | ... | ... | ... | |
| | I | $y_{I1(1)}$ | $y_{I2(1)}$ | $\cdots$ | $y_{IJ(1)}$ |
| ... | 1 | $y_{11...)}$ | $y_{12(...)}$ | $\cdots$ | $y_{1J(...)}$ |
| | 2 | $y_{21(...)}$ | $y_{22(...)}$ | $\cdots$ | $y_{2J(...)}$ |
| | ... | ... | ... | ... | |
| | I | $y_{I1(...)}$ | $y_{I2(...}$ | $\cdots$ | $y_{IJ(...)}$ |
| K | 1 | $y_{11(K)}$ | $y_{12(K)}$ | $\cdots$ | $y_{1J(K)}$ |
| | 2 | $y_{21(K)}$ | $y_{22(K)}$ | $\cdots$ | $y_{2J(K)}$ |
| | ... | ... | ... | ... | |
| | I | $y_{I1(K)}$ | $y_{I2(K)}$ | $\cdots$ | $y_{IJ(K)}$ |

- K tabelle di contingenza parziali $I \times J$ che descrivono la relazione tra $H$ e $X$, in corrispondenza di ogni livello fissato della variabile $Z$
- Una tabella di contingenza $I \times J$ marginale che descrive la relazione tra $H$ e $X$, ignorando la variabile $Z$
- La tabella marginale non contiene informazioni su $Z$

# Associazioni parziali

- In una tabella di contingenza $I \times J \times K$ possiamo studiare associazioni marginali e condizionali
- Le associazioni condizionali, valutate in ciascuna tabella parziale, possono essere diverse dalle associazioni marginali
- Il risultato che una associazione marginale può avere segno opposto rispetto alle associazioni condizionali è noto come *Paradosso di Simpson*

# Associazioni parziali

- Associazioni marginali e condizionali possono essere valutate mediante OR
- Per ogni livello della variabile $Z$ possiamo stimare gli OR condizionali, che misurano l'associazione condizionale tra le variabili $H$ e $X$
- Consideriamo una tabella $2 \times 2 \times K$

$$OR_{XH(k)} = \frac{y_{11(k)}y_{22(k)}}{y_{12(k)}y_{21(k)}}, k = 1, 2$$

- L'associazione marginale tra tra $H$ e $X$ si misura attraverso l'OR marginale

$$OR_{XH(\cdot)} = \frac{y_{11(\cdot)}y_{22(\cdot)}}{y_{12(\cdot)}y_{21(\cdot)}}$$

# Pena di morte

Classificazione di 674 imputati in processi per omicidio in Florida tra il 1976 e il 1987

| Razza della vittima (Z) | Razza dell'imputato (X) | Pena di morte (Y) Si | No | % Si |
|---|---|---|---|---|
| Bianca | Bianca | 53 | 414 | 11.3 |
| | Nera | 11 | 37 | 22.9 |
| Nera | Bianca | 0 | 16 | 0.0 |
| | Nera | 4 | 139 | 2.8 |
| Marginale | Bianca | 53 | 430 | 11.0 |
| | Nera | 15 | 176 | 7.9 |

Fonte: An introduction to categorical data analysis, A. Agresti, Wiley, 1996

# Pena di morte

- Se teniamo conto della razza della vittima, la percentuale di pene di morte risulta maggiore per imputati di razza Nera rispetto agli imputati di razza Bianca, sia quando la vittima è di razza Bianca che Nera

$$OR_{XY(Z=Bianca)} = \frac{53 \times 37}{414 \times 11} = 0.43$$

$$OR_{XY(Z=Nera)} = \frac{0 \times 139}{16 \times 4} = 0$$

- Quando la razza della vittima è Bianca, la quota di pene di morte per gli imputati bianchi è meno della metà della quota di pene di morte per gli imputati neri
- Quando la razza della vittima è Nera, non ci sono condanne a morte per i bianchi

# Pena di morte

- Se ignoriamo la razza della vittima, l'associazione tra *pena di morte* e *razza dell'imputato* nella tabella marginale cambia segno rispetto a quanto accade nelle tabelle condizionate

$$OR_{XY} = \frac{53 \times 176}{430 \times 15} = 1.45$$

- Il risultato è dovuto all'associazione tra la variabile di controllo *razza della vittima* con le altre due
- In particolare, l'associazione marginale tra razza dell'imputato e razza della vittima è

$$OR_{XZ} = \frac{467 \times 143}{48 \times 16} = 87$$

  La quota di imputati di razza bianca quando la vittima è bianca è 87 volte la stessa quota quando la vittima è nera
- Bianchi tendono ad uccidere bianchi ed uccidere bianchi conduce con probabilità maggiore alla pena di morte

# Indipendenza marginale e condizionale

- Se $Y$ e $X$ sono indipendenti in ciascuna tabella parziale allora sono condizionalmente indipendenti
- Nel caso di una tabella $2 \times 2 \times K$, l'indipendenza condizionale si ha quando

$$OR_{XY(k)} = 1, \forall k, k = 1, 2, \ldots, K$$

- Indipendenza marginale $\iff$ $OR_{XY(\cdot)} = 1$
- L'indipendenza condizionale non implica l'indipendenza marginale

# Associazione omogenea

- Si parla di associazione omogenea in una tabella $2 \times 2 \times K$ quando gli OR non cambiano al cambiare dei livelli della variabile di controllo $Z$, i.e.

$$OR_{XY(k)} = \lambda, \forall k, k = 1, 2, \ldots, K$$

- L'associazione omogenea tra $Y$ e $X$ in tabelle $I \times J \times K$ significa che ogni OR condizionale ottenuto combinando due livelli della variaile $Y$ e due livelli della variabile $X$ non cambia per ciascun livello della variabile $Z$
- Quando gli OR condizionali sono uguali entro i livelli di $Z$, la stessa proprietà vale per le altre associazioni, i.e. gli OR condizionali tra due livelli di $X$ e $Z$ saranno uguali per ciascun livello di $Y$

# Associazione omogenea

- L'associazione omogenea è una proprietà simmetrica, che può applicarsi a ciascuna coppia di variabili, entro i livelli della terza
- L'indipendenza condizionale è un caso particolare di associazione omogenea
- Quando non c'è associazione omogenea, gli OR condizionali cambiano al variare dei livelli di $Z$

# Modelli loglineari

- Utili per *modellare* le frequenze in tabelle di contingenza
- Il modello specifica in che modo le frequenze in una cella dipendono dalle modalità ad essa corrispondenti
- La natura della specificazione è connessa alla struttura di dipendenza tra le variabili
- I modelli loglineari descrivono le associazioni e le interazioni tra variabili categoriali

# Costruiamo il modello

- Consideriamo una tabella di frequenze $I \times J$
- L'ipotesi di indipendenza tra le variabili $H$ e $X$ implica il seguente modello additivo

$$\log y_{ij} \quad = \quad \lambda + \lambda_i^H + \lambda_j^X$$

- Il parametro $\lambda$ rappresenta l'effetto costante, e gli $(\lambda_i^H, \lambda_j^X)$ sono gli effetti principali delle due variabili, $H$ e $X$ rispettivamente
- Vincoli sui totali $\Rightarrow$ vincoli sui parametri

$$\sum_{i=1}^{I} \lambda_i^H = 0 = \sum_{j=1}^{J} \lambda_j^X$$

- Il modello è individuato da un totale $1 + (I-1) + (J-1)$ parametri liberi.

# Test d'indipendenza

- Nella pratica l'indipendenza si incontra in poche situazioni
- Per saggiare l'ipotesi di indipendenza abbiamo bisogno di un modello piùgenerale di riferimento
- Inseriamo un *termine d'interazione* al modello di indipendenza

$$\log y_{ij} \quad = \quad \lambda + \lambda_i^H + \lambda_j^X + \lambda_{ij}^{HX}$$

con il vincolo $\sum_{i=1}^{I} \lambda_{ij}^{HX} = \sum_{j=1}^{J} \lambda_{ij}^{HX} = 0$.

- Il modello ha $(IJ - 1)$ parametri in quanto le probabilità congiunte devono unicamente rispettare il vincolo di sommare ad uno
- La differenza tra il numero di parametro sotto l'ipotesi alternativa e sotto l'ipotesi nulla di indipendenza è $(I - 1)(J - 1)$

# Schemi di campionamento

- Per poter verificare l'ipotesi di indipendenza e più in generale per poter attivare le procedure di inferenza proviamo a scrivere la funzione di verosimiglianza associata alle frequenza osservate nelle celle (che, ricordiamo, rappresentano le risposte)
- Le frequenze osservate rappresentano dei **conteggi**
- Qual è il meccanismo probabilistico generatore dei dati?
- I dati riassunti in una tabella di frequenza possono essere generati da diversi schemi di campionamento.
- La numerosità complessiva $n$ può essere prefissata, oppure può ritenersi realizzazione di un processo stocastico

# Osservazione diretta del fenomeno

- Osserviamo il fenomeno per un dato periodo e classifichiamo gli eventi
- Ogni frequenza $y_{ij}$ è la realizzazione di una v.c. di Poisson con valore atteso $\mu_{ij}$, indipendente dalle v.c. associate alle altre frequenza
- La funzione di Verosmiglianza

$$L(\mu) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!}$$

  con $\{\mu_{ij}\}$ parametri incogniti da stimare
- Sotto ipotesi di indipendenza la stima dei $\{\mu_{ij}\}$ corrisponde alla stima di $1 + (I - 1) + (J - 1)$ parametri

# Osservazione per un numero fissato di eventi

- Si decide preliminarmente di raccogliere i dati relativi a $n$ unità
- Il modello statistico appropriato per i dati è la distribuzione Multinomiale $M_d(n, \pi)$, con $d = I \times J$, $\pi = (\pi_{11}, \ldots, \pi_{IJ})$, $0 < \pi_{ij} < 1$ e $\sum_i \sum_j \pi_{ij} = 1$
- La funzione di Verosimiglianza

$$L(\pi) = \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}$$

con $\pi_{ij} = \frac{\mu_{ij}}{n}$

# Una marginale prefissata

- Non si vincola solo il numero totale $n$ di osservazioni, ma si impongono valori prefissati a tutta una riga o una colonna di frequenze marginali.
- Il modello statistico sarà allora il prodotto delle funzioni di probabilità multinomiali relative a ciascuna riga (colonna)

$$L(\pi) = \prod_{j=1}^{J} \left( \prod_{i=1}^{I} \frac{y_{\cdot j}!}{y_{ij}!} \pi_{i|j}^{y_{ij}} \right)$$

con $\pi_{i|j} = \mu_{ij}/\mu_{\cdot j}$. La distribuzione dei dati è data dal prodotto di J funzioni di probabilità di tipo multinomiale.

- Esistono definizioni simili degli schemi di campionamento per tabelle con piùdi due dimensioni.

# Verosimiglianze equivalenti

- Gli schemi di campionamento sono connessi
- Assumiamo di osservare il fenomeno in esame per un dato periodo e che le frequenze osservate $y_{ij}$ in ciascuna cella siano realizzazioni di v.c. di Poisson indipendenti con media $\mu_{ij}$
- La somma $n = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$ è realizzazione di una v.c. di Poisson con parametro $\mu = \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}$
- La distribuzione di $\mathbf{y}$ condizionata a $n$ è di tipo Multinomiale $M_d(n, \pi)$, con $\pi_{ij} = \mu_{ij}/\mu$, $i = 1, \ldots, I$ e $j = 1, \ldots, J$

$$
\begin{aligned}
Pr(\mathbf{Y}{=}\mathbf{y}|n) &= \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} / y_{ij}!}{e^{(\sum_i \sum_j \mu_{ij})} \left( \sum_i \sum_j \mu_{ij} \right)^n / n!} \\
&= \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \left( \frac{\mu_{ij}}{\mu} \right)^{y_{ij}}
\end{aligned}
$$

# Verosimiglianze equivalenti

- Consideriamo la log-verosimiglianza associata al modello Poisson

$$\ell_P(\mu) = \sum_i \sum_j (y_{ij} \log \mu_{ij} - \mu_{ij}) + c$$

- Consideriamo la verosimiglianza associata ad una distribuzione Multinomiale

$$\begin{aligned} \ell_M(\mu) &= \sum_i \sum_j (y_{ij} \log \pi_{ij}) + c \\ &= \sum_i \sum_j (y_{ij} \log E(y_{ij}|n)) + c \\ &= \sum_i \sum_j (y_{ij} \log \mu_{ij} - y_{ij}) + c \end{aligned}$$

- Le due log-verosimiglianze sono equivalenti, sotto la condizione che il modello includa l'intercetta, in modo tale che $\sum_i \sum_j \hat{\mu}_{ij} = \sum_i \sum_j y_{ij}$.

# Verosimiglianze equivalenti

- Guardando i dati è impossibile individuare il modello giusto: si deve sapere il modo in cui sono stati raccolti i dati. Infatti, sappiamo che a seconda del modo in cui i dati sono stati raccolti, abbiamo forme diverse della verosimiglianza.
- **Risultato fondamentale.** L'inferenza basata sulla verosimiglianza è essenzialmente la stessa in ciascun caso di raccolta dei dati. Ciò consente di stimare modelli con distribuzione dei dati di tipo multinomiale, usando la verosimiglianza di tipo Poisson.
- Anche per il caso di una marginale fissata, è possibile sviluppare considerazioni analoghe alle precedenti e constatare l'equivalenza delle verosimiglianze.

# Modello loglineare per tabelle a due vie

- Consideriamo il caso di una tabella di frequenze $I \times J$
- A seconda che il totale delle frequenze $n$ sia casuale (Poisson) o fissato (Multinomiale), le frequenze attese nelle classi sono $E(Y_{ij}) = \mu\pi_{ij}$ e $E(Y_{ij}|n) = n\pi_{ij}$, con $\pi_{ij} = \mu_{ij}/\mu$
- L'ipotesi di indipendenza implica che

$$E(Y_{ij}) = \frac{\mu_{i\cdot}\mu_{\cdot j}}{\mu} = \mu\pi_{i\cdot}\pi_{j\cdot}$$

nel primo caso

$$E(Y_{ij}|n) = \frac{\mu_{i\cdot}\mu_{\cdot j}}{n} = n\pi_{i\cdot}\pi_{j\cdot}$$

nel secondo caso

# Modello loglineare per tabelle a due vie

- Una forma additiva nei parametri si ottiene mediante la trasformazione logaritmica

$$\log E(Y_{ij}) = \eta_{ij} = \log \mu + \log \pi_{i\cdot} + \log \pi_{j\cdot} = \lambda + \lambda_i^H + \lambda_j^X$$

$$\log E(Y_{ij}|n) = \eta_{ij} = \log n + \log \pi_{i\cdot} + \log \pi_{j\cdot} = \lambda + \lambda_i^H + \lambda_j^X$$

- $\lambda_i^H$ e $\lambda_j^X$ sono l'effetto riga e l'effetto colonna, rispettivamente
- $\lambda_i^H$ e $\lambda_j^X$ sono detti *effetti principali*
- L'adozione del legame logaritmico giustifica il nome di questa classe di modelli

# Modello loglineare per tabelle a due vie

- Per saggiare l'ipotesi di indipendenza espressa dal modello loglineare con i soli effetti principali bisogna dichiarare un'ipotesi più generale di riferimento
- Nel caso di tabelle di contingenza a due dimensioni, il modello con $(IJ - 1)$ parametri liberi corrisponde al modello saturo
- Il modello massimale può essere scritto come

$$\log \mu_{ij} = \lambda + \lambda_i^H + \lambda_j^X + \lambda_{ij}^{HX}$$

- $\lambda_{ij}^{HX}$ sono termini di associazione che riflettono le deviazioni dall'indipendenza.

# Modello loglineare per tabelle a due vie

- $\lambda_i^H$ e $\lambda_j^X$ sono i coefficienti di variabili dummy per $(I-1)$ e $(J-1)$ modalità, rispettivamente
- $\lambda_{ij}^{HX}$ è il coefficiente del prodotto di variabili dummy
- Il modello log-lineare che specifica la relazione di indipendenza tra le due variabili ha $1+(I-1)+(J-1)$ parametri
- Il modello saturo ha $(I-1)(J-1)$ parametri in più rappresentati dai coefficienti $\lambda_{ij}^{HX}$
- Verificare l'ipotesi di indipendenza equivale a verificare l'ipotesi di nullità di questi parametri

# Interpretazione dei parametri

- Consideriamo una tabella $I \times 2$, la variabile X ha $I$ livelli, la variabile $H$ è dicotomica, con $\pi_{i1} = Prob(H = 1|X = x_i)$
- Il modello loglineare che specifica indipendenza è il seguente

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^H, i = 1, 2, \ldots, I, j = 1, 2$$

$$\begin{aligned} logit(\pi) &= \log \frac{\mu_{i1}}{\mu_{i2}} = \log \mu_{i1} - \log \mu_{i2} = \\ &= (\lambda + \lambda_i^X + \lambda_1^H) - (\lambda + \lambda_i^X + \lambda_2^H) \\ &= \lambda_1^H - \lambda_2^H = const. \end{aligned}$$

- Il logit non dipende dai livelli di $X$ ma assume lo stesso valore in ciascuna riga
- I modelli loglineari sono davvero utili quando la variabile $Y$ non è dicotomica e quando si vogliono studiare le associazioni tra più di due variabili

# Interpretazione dei parametri

- Il modello saturo è invece caratterizzato dai parametri di associazione $\lambda_{ij}^{HX}$ che sono direttamente connessi ai log OR
- Per semplicità consideriamo una tabella di contingenza $2 \times 2$

$$\begin{aligned}
\log(OR) &= \log \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\
&= (\lambda + \lambda_1^X + \lambda_1^H + \lambda_{11}^{XH}) + (\lambda + \lambda_2^X + \lambda_2^H + \lambda_{22}^{XH}) \\
&- (\lambda + \lambda_1^X + \lambda_2^H + \lambda_{12}^{XH}) - (\lambda + \lambda_2^X + \lambda_1^H + \lambda_{21}^{XH}) \\
&= \lambda_{11}^{XH} + \lambda_{22}^{XH} - \lambda_{12}^{XH} - \lambda_{21}^{XH}
\end{aligned}$$

- I parametri $\lambda_{ij}^{XH}$ determinano gli OR
- Quando sono nulli allora $OR = 1$ e $Y \perp X$

# Modelli non saturi

- In pratica, l'obiettivo è quello di descrivere i dati mediante modelli non saturi, che conducano a più semplici interpretazioni
- Nelle tabelle a più vie, modelli non saturi possono includere termini d'interazione
- I parametri che corrispondono alle frequenze marginali fissate devono essere sempre inclusi nel modello
- Ad esempio, se disponiamo di una tabella $I \times J \times K$ per i tre fattori $X, H, Z$ sapendo che le osservazioni campionarie sono state ottenute separatamente per ogni combinazione $H \times Z$, allora $y_{i\cdot(k)}$ sono fissati e dovremmo includere l'interazione $HZ$ nel modello.
- **Struttura gerarchica**: il modello include tutti i termini di ordine inferiore riguardanti variabili coinvolte in termini di ordine superiore.

# Modello loglineare per tabelle a tre vie

- Le frequenze attese in ciascuna cella le indichiamo con $\mu_{ijk}$
- I termini singoli li indichiamo con $\lambda_i^X$, $\lambda_j^H$, $\lambda_k^Z$ e corrispondono alle distribuzioni marginali
- I termini d'interazione li indichiamo con $\lambda_{ij}^{XH}$, $\lambda_{ik}^{XZ}$, $\lambda_{jk}^{HZ}$ e descrivono le associazioni parziali tra le variabili
- Consideriamo il modello, che indichiamo come $(XZ, HZ)$

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^H + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{HZ}$$

- Dato che include il termine $XZ$, il modello specifica associazione tra $X$ e $Z$, controllando per $H$, ed allo stesso modo permette associazione tra $H$ e $Z$, controllando per $X$
- Il modello non contiene il termine $XH$, per cui specifica indipendenza condizionale tra $X$ e $H$, controllando per $Z$
- Modello di **indipendenza condizionale** tra $X$ e $H$

# Modelli loglineari per tabelle a tre vie

- Il modello che contiene i soli termini corrispondenti ai fattori principali, $(X, H, Z)$ è detto modello di mutua indipendenza: le variabili sono indipendenti a coppie
- Il modello in base al quale tutte le coppie di variabili sono condizionalmente dipendenti, che indichiamo con $(XZ, XH, HZ)$ è

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^H + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XH} + \lambda_{jk}^{HZ}$$

  Si parla di modello di **associazione omogenea**
- Questo modello specifica una situazione di associazione omogenea, i.e. gli OR condizionali, calcolati combinando due livelli di due variabili, sono uguali per ogni livello della terza variabile

# Modelli loglineari per tabelle a tre vie

- Il modello saturo, $(XHZ)$, contiene i 3 termini corrispondenti ai fattori principali, i 3 termini di interazione a coppie ed il termine d'interazione $XHZ$

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^H + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XH} + \lambda_{jk}^{HZ} + \lambda_{ijk}^{XHZ}$$

- Prevede che gli OR condizionali tra le coppie di livelli di due variabili cambino al variare dei livelli della terza variabile
- Il modello saturo non ha utilità pratica
- L'obiettivo è quello di descrivere le associazioni tra le variabili con modelli non saturi che contengano termini d'interazione (almeno nel caso di tabelle con più di due variabili)
- Rispetto del principio gerarchico

# Interpretazione dei parametri

- Consideriamo una tabella $2 \times 2 \times k$, le variabili X e H sono dicotomiche, Z, invece, assume $K$ modalità
- Assumiamo il modello loglineare che specifica associazione omogenea $(XH, HZ, XZ)$
- L'OR condizionale che misura l'associazione tra X e H in ciascuna delle $K$ tabelle parziali è $\theta_{XH(k)}$
- Si verifcica che

$$\begin{aligned} log\theta_{XH(k)} &= \log \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \\ &= \lambda_{11}^{XH} + \lambda_{22}^{XH} - \lambda_{12}^{XH} - \lambda_{21}^{XH} \end{aligned}$$

- L'OR non dipende dai livelli di $Z$ ma assume lo stesso valore in ciascuna tabella parziale

# Consumo di Alcohol, Sigarette e Marijuana

Studenti dell'ultimo anno di *High School* a Dayton, Ohio in 1992

| Alcohol use (A) | Cigarette use (C) | Marijuana (M) Si | No |
|---|---|---|---|
| SI | Si | 911 | 538 |
| | No | 44 | 456 |
| No | Si | 3 | 43 |
| | No | 2 | 279 |

Fonte: An introduction to categorical data analysis, A. Agresti, Wiley, 1996

# Consumo di Alcohol, Sigarette e Marijuana

- Confrontiamo alcuni possibili modelli
- Per ciascun modello calcoliamo le frequenza attese e gli OR attesi (con rispettivi standard error)
- Riportiamo gli OR attesi

| | Conditional Ass. | | | Marginal Ass. | | |
|---|---|---|---|---|---|---|
| Model | AC | AM | CM | AC | AM | CM |
| (A,C,M) | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| (AC,M) | 17.7 | 1.0 | 1.0 | 17.7 | 1.0 | 1.0 |
| (AM,CM) | 1.0 | 61.9 | 25.1 | 2.7 | 61.9 | 25.1 |
| (AM,AC,CM) | 7.8 | 19.8 | 17.3 | 17.7 | 61.9 | 25.1 |

# Inferenza

- I modelli sono stimati mediante metodo della massima verosimiglianza
- Solo in alcuni modelli è possibile procedere a delle stime *dirette*
- Ad esempio nel modello (XZ,HZ) di indipendenza condizionale

$$\hat{\mu}_{ijk} = \frac{y_{i\cdot k} y_{\cdot jk}}{y_{\cdot\cdot k}}$$

- Verifica della bontà del modello stimato, mediante $G^2$ o $X^2$
- Nell'esempio degli studenti di Dayton, il modello stimato di associazione omogenea fornisce un buon adattamento ai dati

# Inferenza

- Il test di associazione omogenea si traduce nel confronto fra modello saturo (ACM) e modello (AM,AC,CM)
- Test di associazione parziale consiste nel confrontare il modello (AM,AC,CM) con i modelli annidati al suo interno
- Ad esempio, supponiamo di voler verificare l'ipotesi di non associazione parziale tra uso di alcohol e sigarette: $\lambda^{AC} = 0$
- Confrontiamo i modelli nidificati (AM,CM) e (AM, AC,CM) mediante TRV

$$W = Dev(AM, CM) - Dev(AM, AC, CM)$$

- $Dev(AM, CM)$ sono associati 2 gdl, Dev(AM,AC, CM) 1 gdl
- $W^{oss} = 187.8 - 0.4 = 187.4$
- $pval = Prob\left(\chi^2_1 > 187.4\right) < .001$

# Legame con i modelli logit

- I modelli loglineari non distinguono tra variabile risposta e variabili esplicative
- I modelli logit descrivono in che misura una risposta binaria dipende da un insieme di esplicative
- Tuttavia i modelli sono connessi
- Per un modello loglineare si possono costruire i logit per una variabile per aiutare l'interpretazione del modello
- Un modello logit con esplicative categoriali ha un equivalente modello loglineare

# Legame con i modelli logit

- Come esempio, consideriamo un modello di associazione omogenea (XH,HZ, XZ)
- Assumiamo che $H$ sia binaria e consideriamola come variabile risposta
- Sia $\pi = Prob(H = 1|X = i, Z = k)$
- Si verifica che

$$logit(\pi) = \log \frac{\mu_{i1k}}{\mu_{i2k}} = (\lambda_1^H - \lambda_2^H) + (\lambda_{i1}^{XH} - \lambda_{i2}^{XH}) + (\lambda_{1k}^{HZ} - \lambda_{2k}^{HZ})$$

- Il logit ha la forma additiva

$$logit(\pi) = \alpha + \beta_i^X + \beta_k^Z$$

# Sovradispersione

- Nell'analisi di dati di conteggio, la variabilità osservata può eccedere quella prevista da un modello di Poisson, per il quale $Var(Y) = E(Y) = \mu$. Si parla di sovradispersione
- Le SMV sono consistenti ma gli errori standard sottostimati
- L'eterogeneità è una causa comune: in corrispondenza di valori fissati delle variabili esplicative, i valori attesi condizionati variano per effetto di variabili non osservate
- La popolazione dalla quale è estratto il campione non è omogenea ma caratterizzata dalla presenza di *sotto-popolazioni* (**cluster**), per cui la risposta può esprimersi come la somma di un numero casuale di contributi casuali i.i.d.

$$Y = Z_1 + Z_2 + \ldots, Z_N, \;\; Z_i \; i.i.d., \;\; N \sim Pois(\mu_N), N \perp Z_i, \forall i$$

# Sovradispersione

- Dato che

$$\begin{aligned} E(Y|N=n) &= nE(Z) \\ Var(Y|N=n) &= nVar(Z) \end{aligned}$$

- Otteniamo che

$$\begin{aligned} E(Y) &= E_N[E(Y|N)] = E(N)E(Z) = \mu_N E(Z) \\ Var(Y) &= E_N[Var(Y|N)] + Var_N[E(Y|N)] \\ &= E(N)Var(Z) + Var(N)E(Z)^2 \\ &= \mu_N[Var(Z) + E(Z)^2] = \mu_N E(Z^2) \end{aligned}$$

- Il processo di Poisson viene osservato lungo un periodo di tempo di natura casuale e non fisso
- Più raramente, esistono fenomeni di sottodispersione

# Modello mistura Poisson-Gamma

- **Problema**: come modellare la risposta tenendo conto della possibile presenza di sovradispersione nei dati?
- Consideriamo un modello mistura in base al quale, condizionatamente al valore osservato delle variabili esplicative, il valore atteso della risposta cambia per effetto di variabili non osservate
  1. $Y \sim Pois(\lambda)$
  2. $\lambda \sim Gamma(\mu, \theta)$, con $\theta$ parametro di forma, per cui

$$E(\lambda) = \mu \ , Var(\lambda) = \frac{\mu^2}{\theta}$$

- Marginalmente, questa mistura conduce alla distribuzione *Binomiale Negativa*

$$p_Y(Y = y; \mu, \theta) = \frac{\Gamma(y + \theta)}{y!\Gamma(\theta)} \frac{\theta^\theta \mu^y}{(\mu + \theta)^{y+\theta}}, \ y = 0, 1, 2, \ldots,$$

# Modello Binomiale-Negativa

- Per $\theta$ fissato, è un elemento di una famiglia di dispersione esponenziale, con parametro naturale $\log \frac{\mu}{\mu+\theta}$
- $E(Y) = E_\lambda[E(Y|\lambda)] = \mu$, per cui è preservata l'ipotesi sul valore atteso
- La varianza è una funzione **quadratica** di $\mu$

$$Var(Y) = Var_\lambda[E(Y|\lambda)] + E_\lambda[Var(Y|\lambda)] = \frac{\mu^2}{\theta} + \mu$$

- Al diminuire di $\theta$ aumenta la sovradispersione, viceversa il modello Binomiale-Negativa converge alla distribuzione di Poisson
- Se il parametro di forma $\theta$ dipende dal valore atteso, ad esempio nella forma $\theta = \tau\mu$, allora la varianza della risposta è una funzione **lineare** di $\mu$

$$Var(Y) = \frac{\mu}{\tau} + \mu = \frac{1+\tau}{\tau}\mu = \phi\mu, \ \phi > 1$$

- Il modello NB lineare non rientra nella famiglia dei GLM

# GLM con risposta Binomiale Negativa

- Il modello di regressione è specificato in termini di $\mu = \mu(\beta)$, nella forma $g(\mu) = X\beta$
- Le equazioni di verosimiglianza sono diverse da quelle che si ottengono a partire dal modello Poisson
- Le differenze sono trascurabili per piccoli livelli di sovradispersione
- Per il modello NB quadratico, la stima di $\theta$ si può ottenere mediante il metodo dei momenti, risolvendo in $\theta$ l'equazione di stima

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1 + \hat{\mu}_i/\theta)} = n - p$$

# Eccesso di conteggi nulli

- La frequenza di conteggi nulli eccede quella prevista dal modello di Poisson
- Zero può rappresentare la moda della distribuzione empirica, in disaccordo con la moda prevista dal modello di Poisson che coincide con la parte intera della media
- Dati con eccesso di conteggi nulli (*zero-inflated*) presentano troppe risposte nulle ma anche troppe risposte molto grandi, così che la media sia posizionata lontano da zero
- L'eccesso di conteggi nulli è meno problematico rispetto ad un'ipotesi distributiva di tipo NB, in quanto, in questo caso, la moda può essere zero indipendentemente dal valore della media
- I dati possono esibire una distribuzione bi-modale, con una moda a zero ed una lontana da zero: una frazione della popolazione assume valore nullo, mentre la rimanente parte ha una certa distribuzione
- Un modello NB può non fornire un buon adattamento ai dati in presenza di una distribuzione bi-modale

# Modello ZIP

- $p(y) = (1 - \pi_i)\Delta(0) + \pi_i \; Pois(\mu)$

$$\begin{cases} \log \mu_i = x_i\beta \\ \text{logit}(\pi_i) = w_i\gamma \end{cases}$$

- Gli insiemi di esplicative $x$ e $w$ possono anche coincidere
- I contributi individuali alla verosimiglianza sono:
- $Prob(Y_i = 0) = (1 - \pi_i) + \pi_i \exp(-\mu_i)$
- $Prob(Y_i = y) = \pi_i \exp(-\mu_i)\frac{\mu_i^y}{y!}, \; y = 1, 2, \ldots,$

# Modello ZIP

- L'eccesso di conteggi nulli comporta sovradispersione rispetto al modello Poisson
- Consideriamo la variabile latente

$$Z = \begin{cases} 0, & y = 0 \\ 1, & Y \sim Pois(\mu) \end{cases}$$

$$\begin{aligned} E(Y) &= E_Z E(Y|Z=z) = \pi\mu \\ Var(Y) &= Var_Z E(Y|Z=z) + E_Z Var(Y|Z=z) \\ &= \pi\mu[1+(1-\pi)\mu] \end{aligned}$$

- Osserviamo che $Var(Y) > E(Y)$
- Spesso, la sovradispersione esiste anche limitatamente alla parte positiva del modello. Un modello ZINB è una scelta appropriata per evitare di sottostimare gli errori standard

# Modello Hurdle

- Modello consta di due parti: una regressione logistica (o probit) per modellare la probabilità che la risposta sia nulla o positiva; se la risposta è positiva, si utilizza un modello troncato nella seconda parte del modello
- I contributi alla verosimiglianza sono:
- $Prob(Y_i = 0) = (1 - \pi_i)$
- $Prob(Y_i = y) = \pi_i \frac{p(y;\mu_i)}{1-p(0;\mu_i)}, \ y = 1, 2, \ldots,$

$$\begin{cases} \log \mu_i = x_i \beta \\ \text{logit}(\pi_i) = w_i \gamma \end{cases}$$