Il modello di Poisson

Statistica Aziendale–prof.ssa Maria Grazia Pittau
grazia.pittau@uniroma1.it

Dipartimento di Scienze Statistiche - Sapienza Università di Roma

a.a. 2024–2025

Introduzione al modello di Poisson I

- Il **modello di Poisson** viene usato per modellare una **variabile risposta di conteggio**; ovvero numero di volte in cui si verifica un certo evento:
  - numero di visite al supermercato in una settimana,
  - numero di volte in cui si fa ricorso ad un certo servizio,
  - numero di assicurazioni stipulate nell'ultimo anno,
  - numero di brevetti registrati da un'impresa in un anno,
  - numero di fasi del processo produttivo esternalizzate,
  - numero di pezzi difettosi prodotti da un certo macchinario,
  - numero di assunzioni di personale.
- Quindi la variabile risposta y_i può essere uguale a 0, 1, 2, ...
- I dati di conteggio sono caratterizzati da:
  1. I dati di conteggio sono sempre e solo **positivi**;
  2. La relativa distribuzione di frequenza è discreta $(y = 0, 1, 2, \cdots)$ e caratterizzata da una forte **asimmetria**;
  3. E' sempre presente un'elevata percentuale di zeri;

Introduzione al modello di Poisson II

- L'**obiettivo principale** del modello di Poisson è quello di spiegare il **valore atteso** di y_i in funzione di un insieme di predittori.
- La prima domanda a cui cercheremo di dare risposta è: **Perchè le stime OLS non sono più valide?**
  1. I dati di conteggio sono per loro natura **eteroschedastici**. Si distribuiscono secondo una distribuzione di Poisson che ha media e varianza uguali ma non costanti: $\mathrm{E}\left(y_i|x_i\right) = \theta_i = \mathrm{Var}_i$ con $\mathrm{Var}_i \neq Var$.
  2. I residui non sono normali, o asimmetrici dal momento chge la distribuzione dell'*outcome* risulta fortemente asimmetrica.
  3. Un modello di regressione lineare non ci assicura che i valori stimati siano positivi.

Introduzione al modello di Poisson III

- Pertanto non si considera più una relazione lineare tra il predittore lineare $X_i^{'}\beta$ e la variabile risposta y, ma piuttosto è il logaritmo dell'*outcome* ad espresso come funzione lineare dei predittori:

$$\log\left(\mathrm{E}\left(y_i|x_i\right)\right) = \log(\theta_i) = X_i^{'}\beta$$

  . Abbiamo quindi un **Modello log-lineare**.

- Un'ipotesi distributiva ragionevole per le variabili di conteggio è la **distribuzione di Poisson**, che è definita come segue:

$$\mathrm{prob}\left(Y = y\right) = \frac{\theta^y \mathrm{exp}^{-\theta}}{y!}, \quad y = 0, 1, 2 \cdots$$

- La variabile risposta y_i "numero di eventi osservati" è positiva ed è anche interpretabile come il risultato di y_i successi su n prove.

Introduzione al modello di Poisson IV

- Se questi eventi sono "rari" la distribuzione di Poisson è il limite della distribuzione Binomiale quando il numero di prove tende all'infinito e p a zero:

$$\lim_{n\to\infty} \mathrm{Bin}\left(y, p\right) = \lim_{n\to\infty} \left(\begin{array}{c} n \\ y \end{array} \right) p^y \left(1-p\right)^{n-k} = \frac{\theta^y \mathrm{e}^{-\theta}}{y!}$$

con $\theta = n \cdot p$ e $\mathrm{E}(Y) = \mathrm{Var}(Y)$

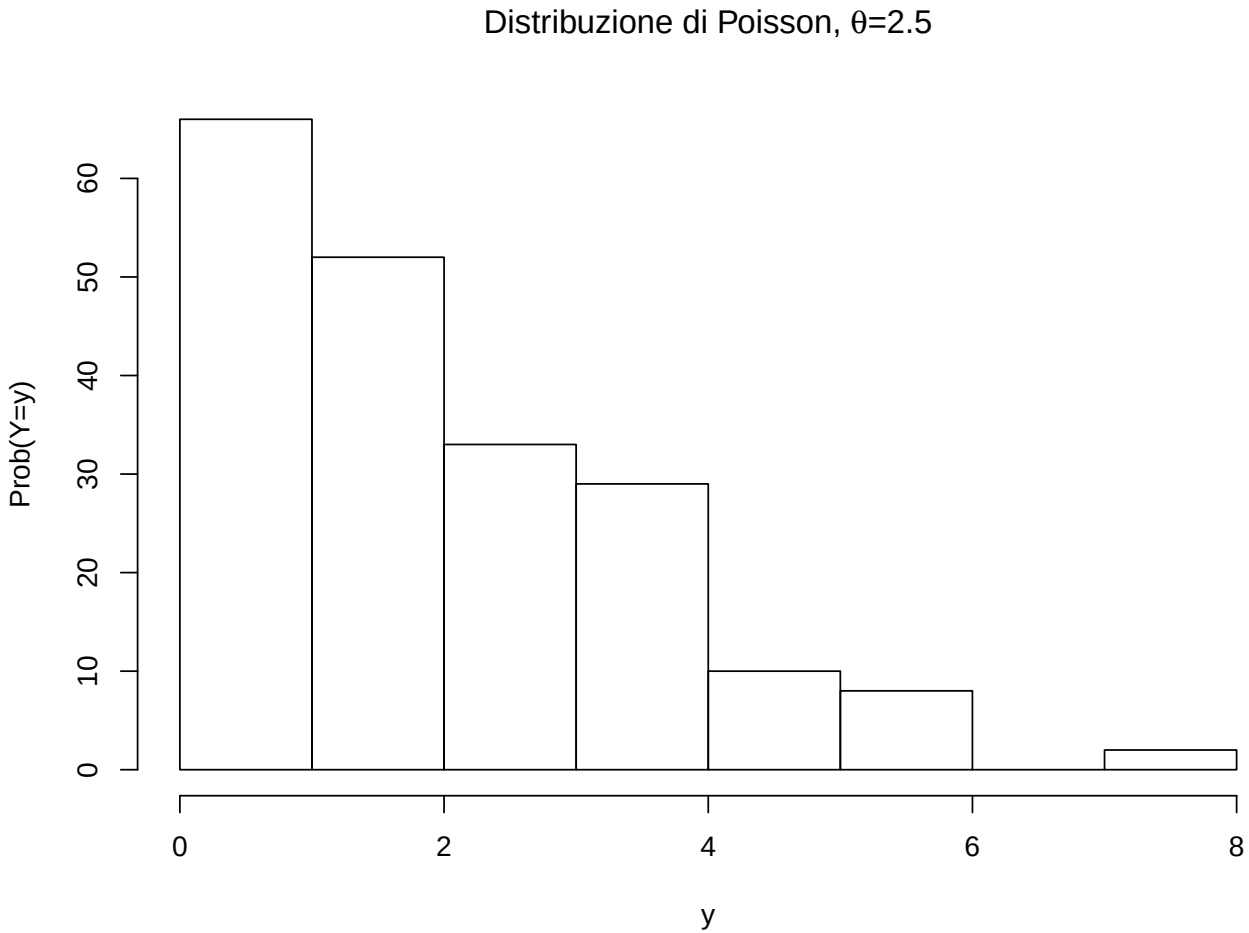

Figure: Distribuzione di Poisson per $\theta = 2.5$

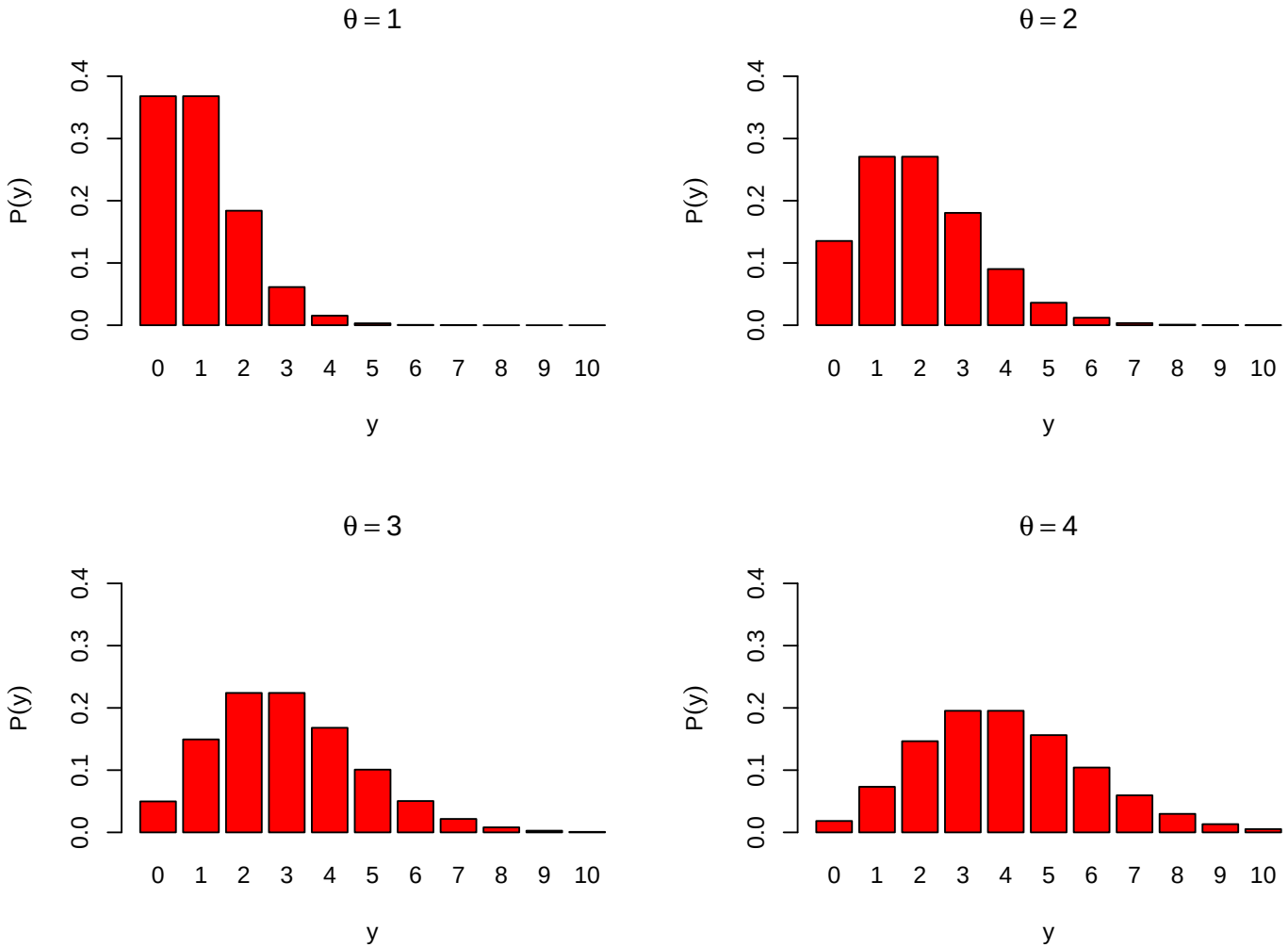

Figure: Distribuzione di Poisson al variare della media θ

Caratteristiche della distribuzione Poisson

- θ è il valore atteso della variabile casuale Y: via via che θ aumenta, la massa della distribuzione si sposta verso destra;
- via via che θ aumenta, diminuisce la probabilità del conteggio 0;
- via via che θ aumenta la distribuzione di Poisson si approssima ad una distribuzione di tipo Normale (si dimostra che l'approssimazione inizia a diventare ragionevole per $\theta = 10.5$).

Caratteristiche del modello di Poisson I

- Il modello di Poisson viene usato per modellare la variabilità dei dati di conteggio quando è possibile interpretarli come il risultato di k successi su n prove, ovvero come "eventi rari".
- Nel modello di Poisson ogni unità i fa riferimento ad uno **scenario** (che in genere è un luogo o un intervallo temporale) in cui gli eventi y_i vengono osservati.
- La **variabilità degli eventi osservati** in un dato luogo o in un dato tempo **può essere spiegata da un insieme di predittori lineari**.
- Quindi la variabile risposta viene modellata come se si distribuisse secondo una distribuzione di Poisson di parametro $\theta = n \cdot p$ con p pari alla probabilità del singolo successo.
- Quindi:

$$y_i \sim \text{Poisson}(\theta_i)$$

con θ_i pari al valore atteso e alla varianza.

Caratteristiche del modello di Poisson II

- Essendo $\theta_i = \mathrm{E}\left(y_i | x_i\right)$ sembra naturale modellare θ_i in funzione dei predittori lineari.
- Quindi il parametro della distribuzione di Poisson $\theta = n \cdot p$ viene modellato un funzione di un insieme di predittori lineari $\theta_i = X_i^{'}\beta, \;\; i = 1, 2, \cdots, n$
- Tuttavia θ_i è sempre positivo e quindi per assicurare che le previsioni siano sempre positive modelliamo il logaritmo di θ_i piuttosto che il parametro stesso:

$$\log(\theta_i) = X_i^{'}\beta, \;\; i = 1, 2, \cdots, n$$

- Si ha quindi che $y_i \sim \mathrm{Poisson}(\theta_i)$ ma $\log(\theta_i) = X_i^{'}\beta$ e quindi

$$y_i \sim \mathrm{Poisson}\left(\exp\left(X_i^{'}\beta\right)\right)$$

Introduzione del parametro di esposizione I

- La variabile di conteggio può essere interpretata in termini di un valore di riferimento, che viene definito parametro di *exposure* ovvero di esposizione.
- Per esempio nel caso di numero di incidenti questo parametro può essere rappresentato dal numero di veicoli che passano per l'incrocio che stiamo studiando in un dato intervallo di tempo.
- Se si sta studiando il numero di hard disk rotti in un dipartimento universitario, l'*exposure* (esposizione) *exps* all'evento "t" potrebbe rappresentare il numero di ore in cui i diversi computer sono rimasti in funzione. Quindi invece di modellare il numero assoluto di hard disk rotti modelliamo il tasso di fallimento "Y/t".
- Da un punto di vista formale significa introdurre nel modello un nuovo parametro, che traduce la variabile risposta in un tasso. Quindi la variabile risposta non sarà più θ ma $\frac{\theta}{exps}$

Introduzione del parametro di esposizione II

- Il log(exposure), u_i, viene definito come *offset* e viene introdotta nel modello con coefficiente pari ad 1.
- Il modello diventa:

$$\log(\theta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + 1 \cdot \log(exps)$$

ovvero

$$\log(\theta) - \log(exps) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k =$$

$$\log\left(\frac{\theta}{exps}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \mathbf{X}'\boldsymbol{\beta}$$

$$y_i \sim \text{Poisson}(\theta_i, u_i)$$

La sovradispersione nel modello di Poisson I

- Una caratteristica del modello di Poisson è che media e varianza coincidono.
- Quello che in realtà succede è che la varianza osservata è superiore alla media e, di conseguenza, il modello di Poisson non è adeguato ma **sovradisperso**.
- Il rapporto tra la varianza e la media è un indicatore di quanto il modello sia sovra o sotto disperso. Quindi se questo rapporto risulta maggiore di 1 allora il modello sarà **sovradisperso**.
- Dal momento che il modello di Poisson non prevede nessun parametro indipendente per la varianza (la normale al contrario presenta due parametri distinti: *location* (μ) e *scale* (σ), risulta indispensabile valutare il grado di sovradispersione.
- Nella distribuzione di Poisson abbiamo solo un parametro e, in presenza di sovradispersione, $\mathrm{Var}(y_i) > \mathrm{E}(y_i)$, ovvero $\mathrm{Var}(y_i) = \phi\theta_i$.

La sovradispersione nel modello di Poisson II

- La sovradispersione nel modello di Poisson ha le stesse conseguenze che si avrebbero nella regressione se i venisse meno l'ipotesi di omoschedasticità: le stime sono ancora consistenti ma non sono più efficienti. Siamo in presenza di una forte **sottostima** degli errori standard con una conseguente inflazione delle statistiche test $t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$ e quindi una tendenza al rifiuto della ipotesi nulla di indipendenza ($\beta = 0$) più severa del necessario.
- Se $y_i \sim \text{Poisson}(\theta_i, u_i)$ per valutare il grado di sovradispersione (ovvero se il modello è un buon modello) valuto i residui, o meglio i **residui standardizzati**, chiamati anche *residui di Pearson*.
- Dati i valori stimati $\hat{y}_i = \text{E}(y_i|x_i) = \hat{\theta}_i\hat{u}_i$, i residui $\epsilon_i = y_i - \hat{y}_i$ avranno media e varianza $\hat{\theta}_i\hat{u}_i$.

La sovradispersione nel modello di Poisson III

- Quindi i residui standardizzati:

$$r_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)} = \frac{y_i - \hat{u}_i\hat{\theta}_i}{\sqrt{\hat{u}_i\hat{\theta}_i}}$$

  sotto l'ipotesi che il modello di Poisson sia un buon modello dovrebbero distrinuirsi in modo indipendente con media nulla e varianza unitaria.
- Nel caso in cui, invece, siamo in presenza di sovradispersione il loro andamento sarà caratterizzato da un andamento crescente.
- Considero il sistema di ipotesi:
  $H_0 : \theta_i = \mathrm{Var}(y_i)$ (MEDIA=VARIANZA)
  $H_1 : \mathrm{Var}(y_i) = \phi\theta_i$ (VARIANZA>MEDIA)
  Per cui $\phi = \frac{\mathrm{Var}(y_i)}{\theta_i}$ rappresenta il **fattore di distorsione**

La sovradispersione nel modello di Poisson IV

- Per stimare ϕ, si considera la somma dei residui standardizzati che, a meno del parametro di esposizione e sotto l'ipotesi nulla in cui $\phi = 1$, sono pari a:

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(y_i - \theta_i)^2}{\mathrm{Var}(y_i)} = \sum_{i=1}^{n} \frac{(y_i - \theta_i)^2}{\theta_i}$$

- La $\sum_{i=1}^{n} r_i^2$ può essere alternativamente vista come:

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(\frac{(\mathrm{observed}_i - \mathrm{expected}_i)^2}{\mathrm{expected}_i} \right)$$

si distribuisce come una distribuzione χ^2 con una dato numero di gradi di libertà:

$$\sum_{i=1}^{n} r_i^2 \sim \chi^2_{n-k}$$

La sovradispersione nel modello di Poisson V

e, di conseguenza, risulta semplice effettuare un test statistico per accettare o rifuitare l'ipotesi nulla mediante il confronto tra la statistica test $w = \sum_{i=1}^{n} r_i^2$ e il valore del χ^2 teorico per un dato fissato livello di significatività.

- Inoltre, sotto l'ipotesi nulla di uguaglianza tra media e varianza, ovvero di **equidispersione**, $\phi = 1$ e quindi $\sum_{i=1}^{n} r_i^2$ segue una distribuzione Chi quadrato. Ricordando che la media della distribuzione del Chi quadrato risulta pari al valore dei gradi di libertà e, che quindi sotto H_0 $\mathrm{E}(\chi^2) = n - k$, e di conseguenza $\hat{\phi} = \frac{\sum_{i=1}^{n} r_i^2}{(n-k)}$ rappresenta una buona stima della dispersione presente nei dati.
- Il problema della sovradispersione viene risolto **moltiplicando** gli errori standard delle stime per la radice quadrata del fattore di dispersione:

$$\text{standard error}^{\text{new}} = \text{standard error} \times \sqrt{\hat{\phi}}$$

La sovradispersione nel modello di Poisson VI

- Per esempio, supponiamo che il valore della statistica test stimata su un campione di $n = 225$ unità sia $w = \sum_{i=1}^{n} r_i^2 = 2700$, la quale deve essere compatibile con $\mathrm{E}(\chi^2) = n - k = 148$, pari alla differenza tra $n = 225$ e 77 il numero di parametri stimati dal modello. Il fattore stimato di sovradipersione sarà $\frac{2700}{148} = 18.2$ con un associato p-value pari a 1, quindi la probabilità che una variabile casuale estratta da una distribuzione χ^2_{148} assuma un valore pari a 2700, risulta pari a zero.

ZIP Poisson model I

- Questo modello assume che i dati provengano da una mistura di 2 diversi processi generatori: uno che genera solo zeri e un altro che di tipo Poisson.
- Una distribuzione Bernoulliana viene usata per determinare quale dei due processi ha genererato i dati.
- Quindi ci sono due possibili processi generatori dei dati, e il risultato di una prova di tipo Bernoulliana determina quale tipo di processo è generatore dei dati.
- Per ciascuna osservazione i il processo 1 è generato con probabilità P_i e il processo 2 con probabilità $(1 - P_i)$. Il processo 1 genera solo valori pari a zero, mentre il processo 2 una Poisson o una Binomiale negativa (semplicemente un modello più generale che introduce un parametro di eterogenità individuale) per tener conto della sovradispersione.

ZIP Poisson model II

- Osserviamo quindi che esistono due fonti di zero: gli zero che arrivano dalla distribuzione che ha una massa in zero e gli zero che invece sono propri della distribuzione di Poisson per dati di conteggio.
- Quindi la nostra variabile *outcome* Y viene considerata come proveniente da una **mistura** di due distribuzioni: una che ha massa in zero e una distribuzione di Poisson per dati di conteggio:

$$\text{Prob}(Y = y_i) = \omega f_1 + (1 - \omega) f_2$$

dove

$$f_1 = \begin{cases} 1 \text{ se } y = 0 \\ 0 \text{ altrimenti} \end{cases}$$

e

$$f_2 = \frac{\theta^y \exp^{-\theta}}{y!}$$

ZIP Poisson model III

e quindi:

$$\text{Prob}(Y = y_i) = \begin{cases} \omega 1 + (1-\omega)\,\exp^{-\theta}\frac{\theta^0}{0!} & \text{se } \; y_i = 0 \\ (1-\omega)\,\exp^{-\theta}\frac{\theta^y}{y!} & \text{se } \; y_i \neq 0 \end{cases}$$

ovvero

$$\text{Prob}(Y = y_i) = \begin{cases} \omega + (1-\omega)\,\exp^{-\theta} & y_i = 0 \\ (1-\omega)\,\exp^{-\theta}\frac{\theta^y}{y!} & \text{se } \; y_i \neq 0 \end{cases}$$

- I pesi ω sono probabilità. In particolare $\omega = \text{Prob}\{Y = y_i\}$ provenga dal processo 1 e $1-\omega$ la probabilità che $Y = y_i$ provenga dal processo 2.

ZIP Poisson model IV

- Poiché ω dipende da y_i può essere scritta come:

$$\omega_i = \text{Prob}\{y_i = 0\} = \varphi\left(Z_i'\gamma\right)$$

dove $\varphi(\cdot)$ rappresenta una funzione *link* (in genere Normale o Logistica) che lega un valore di probabilità ad un predittore lineare $Z_i'\gamma$.

- si noti che Z rappresenta l'insieme dei **predittori degli zero**.
- Quindi, in generale, il modello ZIP può essere scritto come:

$$\text{Prob}(Y = y_i|Z_i, X_i) = \begin{cases} \varphi(Z_i'\gamma) + \left(1 - \varphi(Z_i'\gamma)\right) g(0|X_i'\beta) & y_i = 0 \\ \left(1 - \varphi(Z_i'\gamma)\right) g(y_i|X_i'\beta) & y_i \neq 0 \end{cases}$$

dove $g(\cdot)$ è una distribuzione di Poisson.

ZIP Poisson model V

- Un modelllo di questo tipo è noto come **Zero Inflated Poisson model**. Spesso la presenza elevata di zeri può essere la causa di sovradispersione.