Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Matematica

Anno Accademico 2025/2026

Sessione di laurea Marzo 2026

Analisi Statistica e Quantificazione del Rischio nell'Estrazione Dati Automatizzata tramite AI: Un Caso Studio nel Settore Assicurativo

Candidato

Sofia Iani

Relatore

Franco Pellerey

Correlatore

Pierluigi Aconito

Alla Niki,
la mia anima pura,
la combattente che ha opposto un cuore invincibile pur di provare a non lasciarci.
Anche in questo silenzio assordante, la tua luce non ha mai smesso di guidarmi.
I tuoi occhi saranno per sempre il mio posto sicuro.
Ce l'abbiamo fatta.

Indice

Elenco delle figure

Abstract

L'automazione dei processi di rimborso assicurativo mediante Intelligenza Artificiale rappresenta un'opportunità strategica per ridurre i costi operativi, ma introduce del rischio sull'affidabilità dei sistemi. Questa tesi analizza quantitativamente il sistema di estrazione automatica implementato da una compagnia assicurativa cliente di Accenture, che mira a ridurre i costi da 0,20€ a 0,02€ per documento. Tuttavia, un'analisi preliminare condotta da Accenture ha rivelato un tasso di errore del 22% nei dati estratti dall'AI.

L'analisi si basa su un dataset di 221.253 documenti assicurativi (tra prescrizioni cliniche e fatture amministrative) raccolti tra settembre e novembre. Mediante tecniche di statistica inferenziale, teoria delle probabilità e gestione quantitativa del rischio, questa tesi evidenzia quattro aree: identificazione di pattern temporali (test Chi-Quadrato, t-test), analisi delle correlazioni tra errori (test di indipendenza di Pearson e Fisher), modellazione probabilistica con gestione della sovradispersione (Distribuzione Binomiale Negativa) e quantificazione del rischio dovuto all'introduzione del sistema automatizzato (Value at Risk, Conditional Value at Risk).

I risultati principali evidenziano: forte correlazione tra errori su Patologia e Numero Prescrizione nei documenti clinici ($\rho = 0.75$, $p < 0.0001$), indicante una causa comune di fallimento sistematico; marcata sovradispersione (indice: 12.05 per clinici, 33.55 per amministrativi) che esclude l'ipotesi della Poisson e richiede modelli con code pesanti; soglie critiche $\text{VaR}_{0.95}$ di 1461 errori/giorno (clinici) e 685 (amministrativi), con necessità di sovradimensionamento del 210-220% rispetto alla media per poter gestire picchi di errori; confronto degli impatti di estrazione errata e di mancata estrazione. Per catturare le dipendenze strutturali tra i campi vuoti (mancata estrazione del campo), è stato introdotto il Modello a Shock Latenti per distribuzioni di Bernoulli multivariate, che decompone gli errori osservati in shock non osservabili stimati tramite l'ottimizzazione con l'algoritmo L-BFGS-B. Il modello identifica lo shock dominante $Z_{\text{Patologia-NumPrescrizione}}$ con probabilità $\theta = 5.72\%$, dimostrando la priorità di intervento, in quanto dimezzare questo parametro ridurrebbe l'errore marginale su Patologia dal 9.6% al 6.8%, con un notevole impatto economico.

La tesi propone quindi un'analisi per orientare la compagnia assicurativa nel miglioramento dell'AI, in quanto fornisce strumenti per eventuali previsioni future anche su campi isolati, e sul quantificare il rischio per la gestione delle risorse e dei costi.
L'approccio metodologico sviluppato è replicabile e generalizzabile ad altri contesti di automazione documentale tramite AI nel settore assicurativo.

Capitolo 1

Introduzione

1.1 Contesto

Nel panorama assicurativo sanitario contemporaneo, la capacità di processare rapidamente ed accuratamente le richieste di rimborso costituisce un fattore competitivo. Le compagnie che operano in questo settore devono gestire volumi massicci di documentazione (prescrizioni mediche e fatture) in cui sono richiesti requisiti di accuratezza e tempestività.

Questa tesi si inserisce nel contesto del programma di trasformazione digitale promosso da una rilevante compagnia assicurativa nel panorama italiano, cliente di Accenture. L'obiettivo strategico dell'azienda è testare e validare l'efficacia dell'automazione dei processi di rimborso tramite sistemi di Intelligenza Artificiale, con particolare focus sull'estrazione automatica dei specifici campi da documenti caricati digitalmente dai clienti.

1.1.1 Tipologia di Documenti e Campi di Estrazione

La documentazione caricata dagli assicurati viene preventivamente classificata in due macro-categorie principali, a seconda della sua natura. Per ciascuna di queste tipologie documentali, il sistema di Intelligenza Artificiale è stato addestrato per riconoscere, isolare ed estrarre un set specifico di informazioni (definiti "campi").

I due gruppi documentali e i relativi campi oggetto di estrazione (e di successiva analisi in questa tesi) sono i seguenti:

Gruppo Clinico (Prescrizioni Mediche). Il docuemnto è formato dai seguenti campi di interesse: Patologia, Numero Prescrizione, Data e Prestazione.

Gruppo Amministrativo (Fatture). Il documento è formato dai seguenti campi di interesse: Numero Fattura, Data, Prestazione, Importo Prestazione e Importo Totale.

1.1.2 Il Processo Attuale: Gestione Manuale

Il flusso operativo corrente si articola nelle seguenti fasi:

1. **Caricamento del documento:** Il cliente carica sul sito web della compagnia o sull'applicaziome la documentazione (fatture e/o prescrizioni mediche).
2. **Smistamento:** Un sistema di gestione assegna ciascuna pratica a un operatore.
3. **Revisione manuale:** Un operatore specializzato visualizza ogni singola immagine caricata, estrapolando manualmente i dati rilevanti.
4. **Validazione e autorizzazione:** Verifica della congruità della richiesta rispetto alla polizza attiva e autorizzazione del rimborso o richiesta di integrazione di altri documenti o ri-caricamento più nitido dello stesso.

Questo processo, sebbene garantisca elevata accuratezza grazie alla supervisione umana, presenta una limitazione: costo operativo elevato.
Il costo stimato internamente è circa 0,20 euro per singola richiesta lavorativa (per ogni documento caricato) oltre agli ulteriori costi trascurabili come la formazione del personale.

Considerando i volumi di pratiche gestite annualmente dalla compagnia (ordine di grandezza: centinaia di migliaia), l'impatto economico complessivo è significativo.

1.1.3 La Strategia di Automazione tramite AI

La strategia di digitalizzazione prevede l'integrazione di un sistema di estrazione automatica basato su Intelligenza Artificiale. L'obiettivo è trasferire il lavoro di lettura, interpretazione e validazione dei campi del documento alla macchina, mantenendo supervisione umana solo per i casi ambigui.

Dal punto di vista economico il costo marginale di elaborazione automatica è stimato a circa 0,02 euro per richiesta, corrispondente a una riduzione del 90% rispetto al processo manuale.

1.1.4 Criticità Emerse

Tuttavia, l'implementazione di questa tecnologia ha rivelato criticità non trascurabili. I test preliminari condotti da Accenture sul sistema AI hanno evidenziato un tasso

di errore nella estrazione dei dati pari al 22%. Si tratta di una percentuale ancora troppo elevata per garantire un processo completamente automatizzato senza supervisione umana.

Le **cause di errore** identificate in fase esplorativa sono molteplici:

- **Qualità dell'immagine:** Foto sfocate, buie, tagliate o riflessi luminosi che compromettono la leggibilità.
- **Variabilità calligrafica:** Prescrizioni mediche compilate a mano con grafie difficilmente comprensibili.
- **Eterogeneità documentale:** Ogni struttura sanitaria utilizza propri modelli con una possibile disposizione dei campi variabile.
- **Terminologia specialistica:** Nomenclatura medica complessa, abbreviazioni non standardizzate.
- **Limitazioni tecnologiche:** Possibile incapacità sistematica per l'interpretazione di alcuni campi da parte dell'AI.

1.1.5 Il Sistema di Scoring della Qualità

Per valutare l'affidabilità dell'estrazione automatica, il sistema AI associa a ciascun campo estratto uno **score di qualità** che quantifica il grado di accordo tra due modelli di Intelligenza Artificiale indipendenti. Questo meccanismo di *cross-validation* automatica è fondamentale per la valutazione del processo automatizzato.

Logica di Scoring basata sul Matching di due sistemi indipendenti

Due sistemi integrati con AI estraggono in modo indipendente tutti i campi per ogni tipologia di documento. Lo **score finale** assegnato a ciascun campo è determinato dal matching tra gli output prodotti dai due modelli. Il sistema adotta una **scala discreta a 4 livelli (0-3)**:

- **Score 0 (disaccordo completo):** I due sistemi producono output completamente discordanti o almeno uno dei due non è riuscito a estrarre il campo.
- **Score 1 (matching basso):** I due sistemi hanno estratto valori parzialmente simili ma con differenze significative.
- **Score 2 (matching buono):** I due sistemi hanno estratto valori sostanzialmente concordanti con discrepanze minime o trascurabili.

- **Score 3 (matching perfetto):** I due sistemi hanno estratto esattamente lo stesso valore, con concordanza totale carattere per carattere. Questo è il risultato ottimale.

1.2 Approccio Metodologico

La metodologia adottata in questa tesi applica concetti provenienti da diversi ambiti della statistica applicata e dell'ottimizzazione numerica.

1.2.1 Preprocessing

Prima fase critica per garantire la validità delle inferenze successive:

- **Identificazione dei pattern temporali:** Applicazione di test di stazionarietà (t-test, analisi visuale di trend) per isolare periodi operativi "stabili" ed eliminare bias dovuti a stagionalità o weekend.
- **Filtraggio e pulizia dei dati:** Rimozione di outlier.
- **Separazione per tipologia di documenti:** Separazione delle analisi per documenti clinici (prescrizioni mediche) e amministrativi (fatture), data la profonda differenza nella natura dei dati estratti.

1.2.2 Modellazione Probabilistica

Costruzione di modelli matematici per catturare la struttura stocastica del processo:

- **Test di adattamento per la distribuzione:** Verifica dell'ipotesi della distribuzione di Poisson per i flussi di arrivo giornalieri di campi vuoti per ogni tipologia di documento (fattura e prescrizione medica) mediante test Chi-Quadrato. Identificazione di sovradispersione e conseguente adozione della Distribuzione Binomiale Negativa.
- **Analisi di dipendenza:** Applicazione di test di indipendenza (Chi-Quadrato di Pearson, test esatto di Fisher) per rilevare correlazioni significative tra errori su coppie di campi vuoti per ogni tipologia di docuemnto. Calcolo di coefficienti di correlazione per quantificare l'intensità delle associazioni.
- **Modello a Shock Latenti:** Sviluppo di un modello basato su distribuzioni Bernoulliane multivariate, in cui gli errori (campi vuoti) osservati sono generati da variabili latenti non osservabili ("shock") che possono colpire singoli campi o gruppi di campi simultaneamente. Stima dei parametri tramite ottimizzazione numerica (L-BFGS-B).

1.2.3 Gestione Quantitativa del Rischio

Adattamento di metriche classiche della finanza quantitativa in questo contesto:

- **Value at Risk (VaR):** Calcolo del quantile di ordine α (tipicamente 95% o 99%) della distribuzione del numero giornaliero di errori. Nella pratica ha la seguente interpretazione: 95 giorni su 100 non verrà superato quel numero di errori. Per l'azienda, rappresentano le risorse base necessarie da avere sempre a disposizione per gestire il lavoro quotidiano.
- **Conditional Value at Risk (CVaR):** Calcolo del valore atteso condizionato agli scenari oltre il VaR, fornendo una stima del "rischio residuo" nei giorni critici. Questa metrica è una misura di rischio coerente (nel senso di Artzner et al.) e supera le limitazioni del VaR.
- **Backtesting:** Validazione empirica delle soglie stimate confrontando previsioni teoriche con osservazioni effettive nel periodo di test.

1.2.4 Analisi degli Score di Qualità

Indagine della relazione tra completezza (campi vuoti) e accuratezza (valutazione dei campi estratti tramite score 0-3):

- **Rappresentazione vettoriale:** Codifica di ciascun documento come vettore numerico, dove ogni componente rappresenta lo stato di un campo specifico (0 = vuoto, 1-4 = score).
- **Distribuzione empirica:** Calcolo della frequenza di ciascuna configurazione vettoriale osservata, costruendo la distribuzione di probabilità empirica completa.
- **Decomposizione degli impatti:** Scomposizione della "distanza dall'ottimo" (metrica aggregata di qualità) nei contributi separati di incompletezza ed inaccuratezza, per identificare il fattore limitante principale.

Capitolo 2

Analisi della Stagionalità del Flusso degli Arrivi dei Campi Vuoti

2.1 Introduzione

L'analisi della stagionalità rappresenta un passaggio fondamentale nello studio dei processi automatizzati, specialmente in contesti caratterizzati da flussi massivi di informazioni. In tali scenari, è cruciale saper distinguere le fluttuazioni casuali (rumore stocastico) dai pattern sistematici (ciclicità ricorrenti).

Mentre le fluttuazioni casuali sono imprevedibili, i pattern sistematici sono strutturali. Tuttavia, proprio queste variazioni sistematiche possono introdurre delle distorsioni nell'analisi: nei periodi a bassa operatività, la drastica riduzione del volume di arrivi introduce bias tale da rendere inaffidabili le analisi inferenziali future della tesi.

L'obiettivo di questo capitolo è dunque duplice:

1. Identificare le componenti sistematiche (trend e stagionalità) per isolare il vero comportamento dei dati.

2. Filtrare i periodi non rappresentativi, affinché l'analisi non sia "sporcata" dalla scarsità del campione.

L'approccio metodologico adottato ha seguito un processo sequenziale:

- Analisi Esplorativa: visualizzazione delle serie temporali per individuare a colpo d'occhio i pattern (cambio di regime).

- Validazione Statistica: applicazione di test (t-test) per confermare che tali pattern siano statisticamente significativi e giustificare l'esclusione dei dati non stazionari dal dataset finale.

2.2 Cenni Teorici e Metodologici

2.2.1 Serie Temporali e Stazionarietà

Una **serie temporale** è definita come una sequenza di osservazioni $\{X_t\}$ ordinate rispetto al tempo t, dove $t \in T \subseteq \mathbb{Z}$ [1]. Nel contesto statistico, una serie temporale può essere scomposta in diverse componenti secondo due modelli principali:

Modello Additivo:

$$X_t = T_t + S_t + \epsilon_t \tag{2.1}$$

Modello Moltiplicativo:

$$X_t = T_t \cdot S_t \cdot \epsilon_t \tag{2.2}$$

dove:

- T_t rappresenta il *Trend*, ovvero la tendenza di fondo del fenomeno (crescente, decrescente o costante).
- S_t rappresenta la *Stagionalità*, ovvero le variazioni cicliche che si ripetono a intervalli regolari (es. settimanali, mensili, annuali).
- ϵ_t è la componente stocastica (o residuo), che racchiude le fluttuazioni casuali non spiegabili.

La scelta tra modello additivo e moltiplicativo dipende dalla natura della variabilità: il modello moltiplicativo è appropriato quando l'ampiezza delle oscillazioni stagionali cresce proporzionalmente al livello del trend, mentre il modello additivo assume oscillazioni di ampiezza costante. Nel contesto di questo studio, l'analisi preliminare suggerisce l'adozione del modello additivo.

Stazionarietà. Un concetto cardine per i processi stocastici è la stazionarietà. Un processo $\{X_t\}$ si dice **stazionario in senso debole** (o debolmente stazionario) se soddisfa le seguenti condizioni:

1. $\mathbb{E}[X_t] = \mu$ è costante per ogni t (media costante)
2. $\text{Var}(X_t) = \sigma^2$ è costante per ogni t (varianza costante)

3. $\text{Cov}(X_t, X_{t+h}) = \gamma(h)$ dipende solo dal lag h e non dal tempo t (autocovarianza dipendente solo dalla distanza temporale)

L'obiettivo del preprocessing in questo capitolo è identificare e rimuovere eventuali cambiamenti improvvisi della media (cambio di regime) e della varianza. Tale condizione è fondamentale per garantire l'affidabilità delle stime statistiche: solo su dati stazionari, infatti, ha senso calcolare metriche aggregate (come media, varianza, correlazioni) che siano generalizzabili all'intero processo.

2.2.2 Il Test t di Student

Per verificare se le differenze osservate nei grafici sono significative e non dovute al caso, si utilizza il **Test t di Student**, un test parametrico utilizzato per confrontare le medie di due gruppi indipendenti.

Date due popolazioni con medie μ_1 e μ_2, il test verifica le seguenti ipotesi:

$$H_0 : \mu_1 = \mu_2 \quad \text{(Ipotesi Nulla: non c'è differenza significativa)}$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{(Ipotesi Alternativa: esiste una differenza significativa)}$$

La statistica test t viene calcolata come:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.3}$$

dove $\bar{X}_i$ è la media campionaria del gruppo i, s_i^2 la varianza campionaria e n_i la numerosità del campione.

I gradi di libertà sono approssimati dalla **formula di Welch-Satterthwaite** [2] (utilizzata quando le varianze dei due gruppi sono diverse):

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \tag{2.4}$$

Il risultato del test fornisce un *p-value*: se questo valore è inferiore a una soglia di significatività prefissata ($\alpha = 0.05$, standard convenzionale in letteratura statistica), si rifiuta l'ipotesi nulla, confermando che la differenza osservata è statisticamente significativa.

Assunzioni del test t. Il test t richiede il soddisfacimento di alcune assunzioni:

- **Normalità**: le distribuzioni dei due gruppi devono essere approssimativamente normali (verificabile con test di Shapiro-Wilk o Kolmogorov-Smirnov)

- **Indipendenza**: le osservazioni devono essere indipendenti all'interno di ciascun gruppo e tra i gruppi
- **Omoschedasticità**: le varianze devono essere omogenee (verificabile con test di Levene o test F)

In questo studio, viene utilizzata la variante di **Welch** del test t, che non assume omoschedasticità ed è quindi più robusta in presenza di varianze eterogenee.

2.3 Data Preprocessing

A valle delle considerazioni teoriche, si è proceduto all'analisi dei dati reali.

2.3.1 Analisi della Discontinuità Strutturale

Lo studio preliminare è iniziata dall'analisi degli istogrammi delle frequenze (Figura 2.1). In essi emerge una chiara distribuzione bimodale, caratterizzata dalla presenza di due picchi distinti, visibile sia per il gruppo Amministrativo che per quello Clinico. Tale conformazione suggerisce che i dati non seguano un'unica distribuzione, rendendo necessaria un'approfondita analisi della componente temporale e della stagionalità.

Infatti, osservando il grafico della serie temporale completa (Figura 2.2), la natura della bimodalità diviene chiara: si nota una progressiva crescita dei volumi di arrivo, associata a un netto cambio di trend e alla presenza di pattern stagionali ricorrenti.

Nello specifico, si osserva che in data 10 settembre 2025 avviene un vero e proprio cambio di regime. Prima di procedere alla validazione statistica, un'analisi descrittiva puntuale mostra la marcata discrepanza tra i due periodi. Se si considerasse l'intero arco temporale come un unico dataset, si riscontrerebbe un'elevata instabilità: il range dei volumi di arrivo presenta infatti un'escursione estrema, spaziando da un minimo di 0 a un massimo di 830 documenti giornalieri per il gruppo Clinico e da un minimo di 0 a 461 per il gruppo Amministrativo. Una varianza così ampia inficerebbe qualsiasi analisi successiva.

Analizzando le medie giornaliere, il salto di livello è evidente per entrambe le categorie:

- Gruppo Clinico: la media passa da 153 (pre-10 settembre) a 505 (post-10 settembre).
- Gruppo Amministrativo: la media sale da 43 a 225 .

Per confermare che tale differenza non sia casuale, è stato applicato il t-test tra il periodo antecedente e quello successivo al 10 settembre. Il test ha restituito un

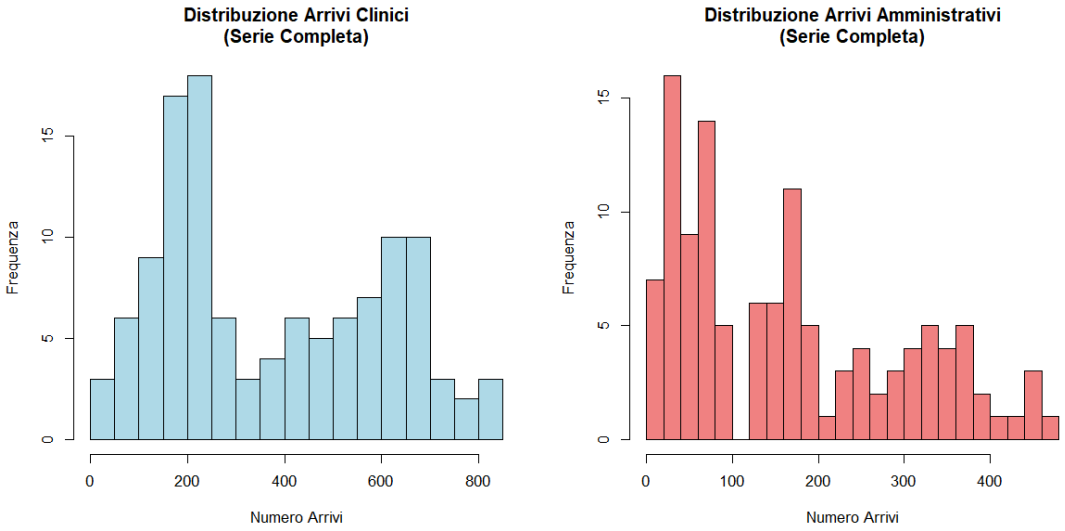

Figura 2.1: Istogrammi delle frequenze. La distribuzione bimodale (doppio picco) suggerisce l'esistenza di due regimi operativi distinti.

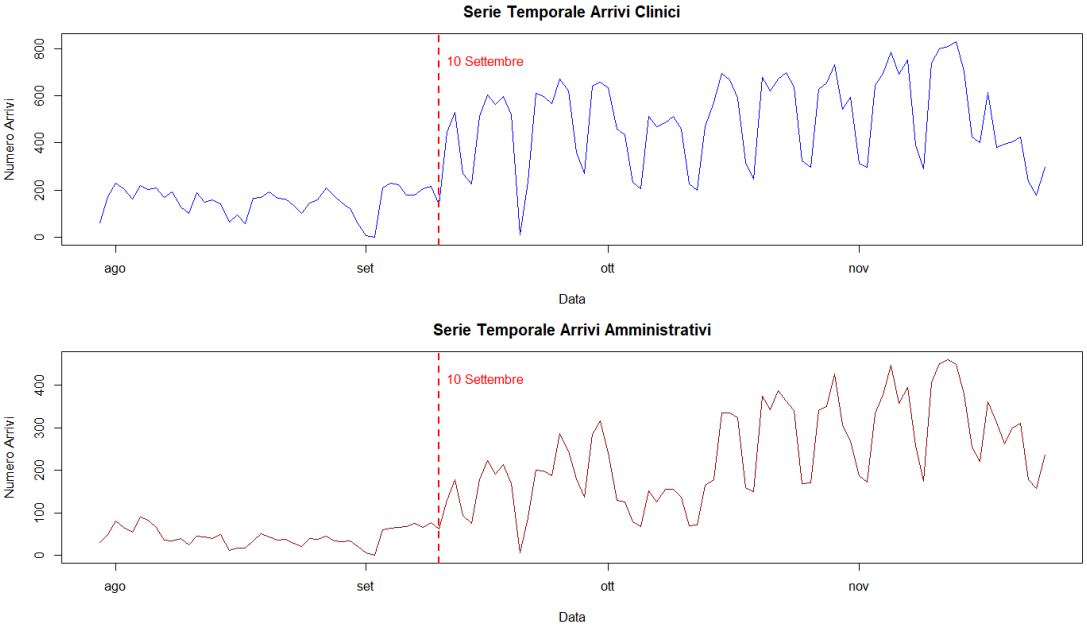

Figura 2.2: Andamento temporale degli arrivi giornalieri. Si osserva un evidente cambiamento strutturale (structural break) all'inizio di settembre.

p-value < 0.001, confermando l'ipotesi di eterogeneità statistica dei due periodi. Di conseguenza, considerare il periodo estivo (pre-settembre) introdurrebbe un bias dovuto ai bassi volumi; pertanto, il dataset verrà filtrato per considerare validi solo i dati a partire dal 10 settembre 2025. Analogamente, applicando il medesimo criterio, si è definito un limite superiore per l'orizzonte temporale. Le osservazioni successive al 12 novembre 2025 non saranno considerate per l'analisi eliminando potenziali irregolarità dovute alla chiusura del periodo di osservazione o a dati parziali.

2.3.2 Analisi della Stagionalità Settimanale

Una volta isolato il periodo dopo il 10 settembre, è stata analizzata la distribuzione dei volumi in funzione del giorno della settimana (Figura 2.3).

Emerge chiaramente una stagionalità settimanale:

- **Giorni Feriali (Lun-Ven)**: Volume elevato e varianza contenuta.
- **Weekend (Sab-Dom)**: Il volume crolla drasticamente (calo del $\sim$56%).

Anche in questo caso, il t-test conferma che la media degli arrivi nel weekend è statisticamente diversa da quella dei giorni feriali ($p < 0.001$). Includere i giorni festivi creerebbe del rumore nelle analisi statistiche. Di conseguenza, è stato applicato un filtro per mantenere solo i giorni lavorativi.

La Figura 2.4 mostra l'andamento temporale del dataset finale, ora ripulito dalle componenti di stagionalità indesiderate.

2.4 Riepilogo

L'applicazione dei test statistici ha validato le osservazioni grafiche iniziali. Il dataset finale utilizzato per i capitoli successivi sarà quindi composto esclusivamente dai giorni feriali nel periodo 10 Settembre - 12 Novembre 2025, garantendo stazionarietà e affidabilità statistica.

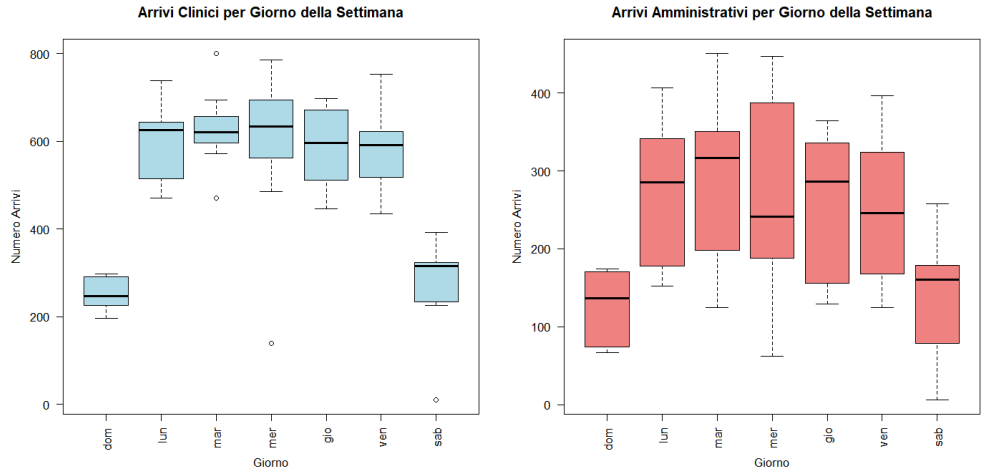

Figura 2.3: Distribuzione degli arrivi per giorno della settimana. Si evidenzia il calo nei giorni festivi rispetto a quelli lavorativi.

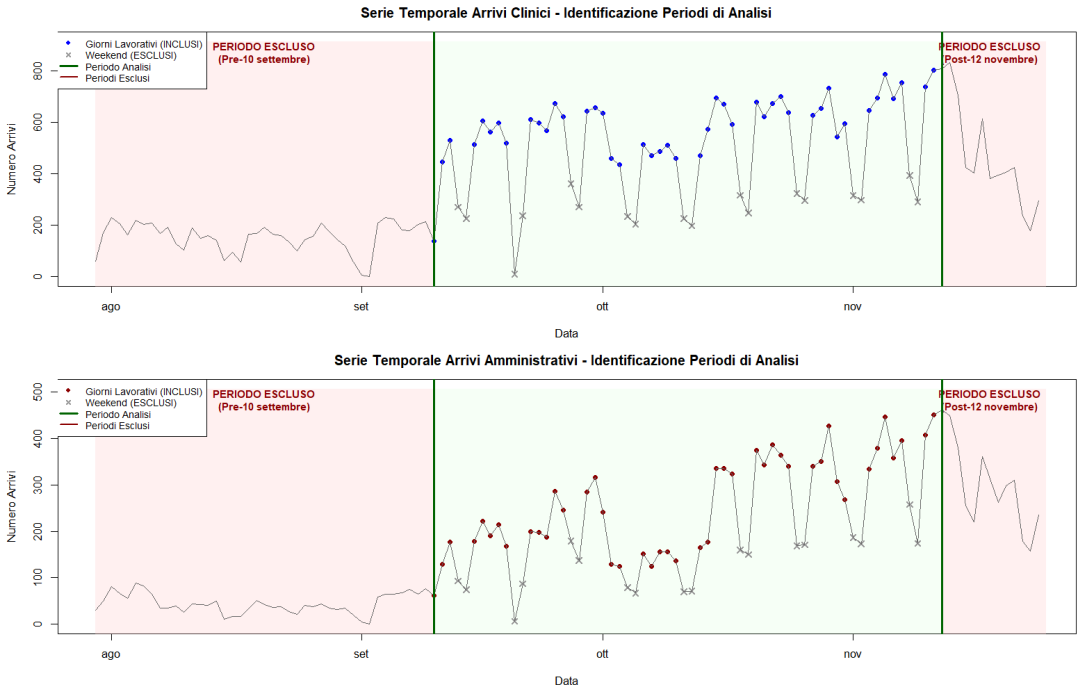

Figura 2.4: Serie temporale post-preprocessing: periodo stabile e giorni lavorativi.

Capitolo 3

Analisi delle Correlazioni tra i Campi Vuoti

Dopo aver isolato il periodo di operatività a regime (dal 10 settembre al 12 novembre 2025) ed escluso i giorni non lavorativi per garantire la stazionarietà del flusso, l'analisi si è concentrata sulle interdipendenze tra i campi vuoti. Il dataset finale, dopo il filtraggio temporale e l'esclusione dei weekend, comprende **221.253 documenti**.

L'obiettivo di questa fase è comprendere se la mancata lettura di un campo sia un evento isolato o se, al contrario, esistano pattern sistematici in cui determinati campi tendono a fallire contemporaneamente.

3.1 Statistiche Descrittive degli Errori

Prima di procedere all'analisi delle correlazioni, è stata condotta un'analisi esplorativa della frequenza con cui ciascun campo risulta non letto. La Tabella 3.1 riporta il numero totale e la percentuale di documenti con campo vuoto per ciascuna variabile.

Si osserva che i campi più critici sono quelli relativi al gruppo clinico ovvero della prescrizione (Numero Prescrizione e Patologia), con tassi di mancata estrazione da parte dell'intelligenza artificiale superiori all'8%. Al contrario, l'Importo Totale mostra una percentuale di errore estremamente bassa (0,09%), indicando una buona capacità del sistema nella lettura di questo campo.

Combinazioni di Errori

Un aspetto rilevante emerso dall'analisi è che la maggior parte dei documenti è processata correttamente: l'85,48% (189.133 su 221.253) non presenta alcun campo

Tabella 3.1: Frequenza e percentuale di errori per tipologia di campo (N=221.253 documenti)

Campo	**N. Errori**	**Percentuale**
Numero Prescrizione	20.584	9,30%
Patologia	19.098	8,63%
Prestazione	6.165	2,79%
Importo Prestazione	4.472	2,02%
Numero Fattura	3.461	1,56%
Data	2.034	0,92%
Importo Totale	196	0,09%

vuoto. Tra i documenti con errori, la combinazione più frequente riguarda la coppia *Patologia-Numero Prescrizione*, che rappresenta il 5,55% del totale (12.283 occorrenze), suggerendo una forte dipendenza tra questi due campi.

3.2 Fondamenti Teorici

3.2.1 Il Coefficiente di Correlazione di Pearson

Al fine di quantificare l'intensità della relazione lineare tra le diverse tipologie di errore, si è fatto ricorso al coefficiente di correlazione di Pearson (indicato con ρ per la popolazione e con r per il campione) [3].

In statistica, l'indice di Pearson misura la forza presente tra due variabili casuali X e Y. Dal punto di vista matematico, esso è definito come il rapporto tra la covarianza delle due variabili e il prodotto delle loro deviazioni standard:

$$r_{xy} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

dove:

- n è la numerosità campionaria (nel nostro caso $n = 221.253$);
- x_i, y_i sono le singole osservazioni delle variabili;
- $\bar{x}, \bar{y}$ rappresentano le rispettive medie campionarie.

Il coefficiente è adimensionale e assume valori nell'intervallo chiuso $[-1, +1]$, con la seguente interpretazione:

- $r = +1$: Correlazione lineare positiva perfetta (se X aumenta, Y aumenta deterministicamente).
- $r = -1$: Correlazione lineare negativa perfetta (se X aumenta, Y diminuisce deterministicamente).
- $r = 0$: Assenza di correlazione lineare (indipendenza stocastica lineare).
- $0 < |r| < 1$: Gradi intermedi di associazione. In letteratura si considerano:
  - $|r| < 0.3$: correlazione debole
  - $0.3 \leq |r| < 0.7$: correlazione moderata
  - $|r| \geq 0.7$: correlazione forte

Applicazione a Variabili Binarie (Coefficiente Phi)

Nel contesto specifico di questa analisi, le variabili considerate sono di natura binaria, assumendo esclusivamente i valori $\{0, 1\}$ a seguito del processo di codifica (One-Hot Encoding). È importante notare che, quando ècalcolato su due variabili binarie in tabelle di contingenza 2×2, il coefficiente di Pearson coincide matematicamente con il **Coefficiente Phi** (ϕ) di Matthews, definito come:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{11} + n_{01})(n_{00} + n_{10})(n_{00} + n_{01})}} \tag{3.2}$$

dove n_{ij} rappresenta la frequenza delle osservazioni nella cella (i, j) della tabella di contingenza.

In questo scenario:

- Un valore positivo indica che la presenza dell'errore nel campo X aumenta la probabilità di osservare un errore anche nel campo Y (i campi tendono a essere vuoti contemporaneamente).
- Un valore prossimo allo zero indica che il fallimento nella lettura di un campo non fornisce informazioni sullo stato dell'altro (eventi indipendenti).

Tale proprietà rende l'indice di Pearson uno strumento idoneo e robusto anche per l'analisi di variabili categoriche binarizzate.

3.2.2 Test di Significatività Statistica

Il coefficiente di correlazione quantifica l'intensità della relazione tra due variabili, ma non fornisce informazioni sulla significatività statistica di tale associazione. Per verificare che le correlazioni osservate non siano dovute a fluttuazioni casuali campionarie, è necessario ricorrere a test di ipotesi [3].

Test Chi-Quadrato di Pearson (χ^2)

Il test chi-quadrato verifica l'ipotesi di indipendenza tra due variabili categoriche attraverso il confronto tra le frequenze osservate e quelle attese sotto l'ipotesi nulla.

Descrizione del Test Consideriamo una tabella di contingenza 2×2 che rappresenta la presenza congiunta di due campi mancanti (errori) A e B:

Sotto l'ipotesi nulla di indipendenza ($H_0 : A \perp B$), la frequenza attesa nella cella (i, j) è:

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n} \tag{3.3}$$

	Campo B Vuoto	**Campo B Compilato**	**Totale**
Campo A Vuoto	n_{11}	n_{10}	$n_{1\cdot}$
Campo A Compilato	n_{01}	n_{00}	$n_{0\cdot}$
Totale	$n_{\cdot 1}$	$n_{\cdot 0}$	n

La statistica test è definita come:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.4}$$

dove O_{ij} rappresenta le frequenze osservate. Sotto H_0, questa statistica segue asintoticamente una distribuzione chi-quadrato con $(r-1)(c-1)$ gradi di libertà, dove r e c sono rispettivamente il numero di righe e colonne (nel nostro caso $df = 1$).

Esempio Applicato: Patologia vs Numero Prescrizione Per illustrare il funzionamento del test, consideriamo la coppia con la correlazione più forte: Patologia e Numero Prescrizione.

La tabella di contingenza osservata è:

	Num. Prescr. Vuoto	**Num. Prescr. Compil.**	**Totale**
Patologia Vuota	15.655	3.443	19.098
Patologia Compilata	4.929	197.226	202.155
Totale	20.584	200.669	221.253

Le frequenze attese sotto H_0 (indipendenza) sarebbero:

$$\begin{aligned} E_{11} &= \frac{19.098 \times 20.584}{221.253} = 1.777{,}9 \\ E_{10} &= \frac{19.098 \times 200.669}{221.253} = 17.320{,}1 \\ E_{01} &= \frac{202.155 \times 20.584}{221.253} = 18.806{,}1 \\ E_{00} &= \frac{202.155 \times 200.669}{221.253} = 183.348{,}9 \end{aligned}$$

Il valore della statistica chi-quadrato risulta:

$$\begin{aligned} \chi^2 = &\frac{(15.655 - 1.777{,}9)^2}{1.777{,}9} + \frac{(3.443 - 17.320{,}1)^2}{17.320{,}1} \\ &+ \frac{(4.929 - 18.806{,}1)^2}{18.806{,}1} + \frac{(197.226 - 183.348{,}9)^2}{183.348{,}9} \\ &= 125.542{,}52 \end{aligned} \tag{3.5}$$

Con $df = 1$ e un valore così elevato, il p-value è praticamente nullo ($p < 0{,}0001$), indicando una forte evidenza contro l'ipotesi di indipendenza.

Condizioni di Validità Il test chi-quadrato è valido quando:

- Le osservazioni sono indipendenti
- Le frequenze attese sono sufficientemente grandi: $E_{ij} \geq 5$ in almeno l'80% delle celle

Quando queste condizioni non sono soddisfatte (campioni piccoli o celle con frequenze basse), si ricorre al test esatto di Fisher.

Test Esatto di Fisher

A differenza del test chi-quadrato, che si basa su un'approssimazione asintotica della distribuzione campionaria, il test di Fisher calcola la **probabilità esatta** della configurazione osservata (e di tutte le configurazioni più estreme) sotto l'ipotesi nulla, utilizzando la distribuzione ipergeometrica.

Fondamento Matematico Data una tabella di contingenza 2×2 con margini fissi, la probabilità di osservare una particolare configurazione sotto H_0 è data dalla distribuzione ipergeometrica [3]:

$$P(n_{11} = k) = \frac{\binom{n_{1\cdot}}{k}\binom{n_{0\cdot}}{n_{\cdot 1}-k}}{\binom{n}{n_{\cdot 1}}} \tag{3.6}$$

Il p-value (test a due code) è calcolato come la somma delle probabilità di tutte le configurazioni altrettanto o più estreme rispetto a quella osservata.

Utilizzo In questa tesi, il test di Fisher è stato applicato automaticamente quando le frequenze attese erano insufficienti per garantire la validità del test chi-quadrato. Questo è accaduto in 3 casi nel Gruppo Amministrativo:

- Importo Prestazione - Importo Totale
- Importo Totale - Numero Fattura
- Importo Totale - Data

In questi casi, la bassa frequenza di errori per il campo "Importo Totale" (196 su 221.253, 0,09%) ha reso necessario l'uso del test esatto.

Interpretazione dei Risultati Sebbene il test di Fisher non fornisca una statistica test intermedia come il χ^2, il p-value ottenuto è comunque interpretabile allo stesso modo: un valore $p < 0{,}05$ porta al rifiuto dell'ipotesi nulla di indipendenza, con un livello di significatività del 5%.

Criteri di Decisione

Per entrambi i test, sono state adottate le seguenti ipotesi:

- **Ipotesi Nulla (H_0)**: Le due variabili sono indipendenti. La presenza di un errore nel campo A non influenza la probabilità di errore nel campo B.
- **Ipotesi Alternativa (H_1)**: Le due variabili sono dipendenti. Esiste una relazione sistematica tra gli errori.

Il livello di significatività è stato fissato a $\alpha = 0{,}05$. Pertanto:

- Se $p < 0{,}05$: si rifiuta $H_0 \rightarrow$ correlazione statisticamente significativa
- Se $p \geq 0{,}05$: non si rifiuta $H_0 \rightarrow$ correlazione non significativa

3.3 Metodologia: Codifica Binaria e Matrice di Correlazione

Il dataset presenta, per ogni documento, una stringa contenente l'elenco dei campi non letti (es. *"Numero prescrizione; Patologia"*). Per procedere all'analisi statistica, è stato necessario trasformare questa variabile qualitativa in un formato quantitativo trattabile.

Si è proceduto mediante una tecnica di **binarizzazione**: per ogni possibile tipologia di errore (campo vuoto) è stata creata una variabile binaria che assume valore:

- **1**: se il campo risulta vuoto, ovvero è presente nell'elenco dell'estrazione dei campi vuoti per quel documento.
- **0**: se il campo è stato letto correttamente.

Su questa matrice binaria (221.253×7) è stato calcolato l'indice di correlazione di Pearson per tutte le coppie di variabili. L'analisi è stata condotta separatamente per i due macro-gruppi identificati: **Dati Clinici** e **Dati Amministrativi**.

Per ogni coppia di variabili, sono stati riportati:

1. Il coefficiente di correlazione r
2. Il tipo di test utilizzato (Chi-Quadrato o Fisher)
3. La statistica test (χ^2 per Chi-Quadrato, non applicabile per Fisher)
4. Il p-value
5. La significatività della correlazione

3.4 Analisi delle Correlazioni

3.4.1 Gruppo Clinico

Il primo gruppo di variabili analizzate riguarda i campi delle prescrizioni. Le variabili incluse sono:

- *Patologia*
- *Numero Prescrizione*
- *Data*
- *Prestazione*

La Figura 3.1 mostra la matrice di correlazione risultante per questo sottogruppo.

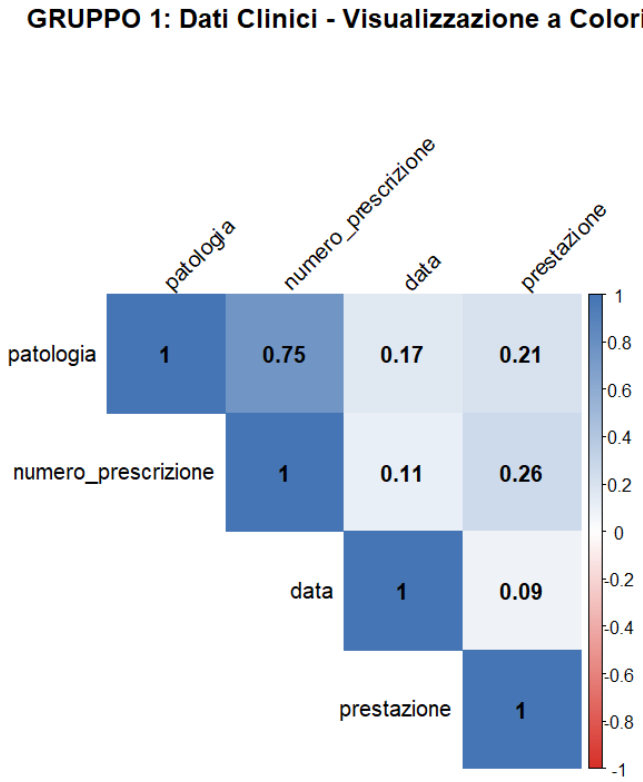

Figura 3.1: Matrice di correlazione - Gruppo Clinico (Dati filtrati post-10 Settembre, N=221.253).

Risultati Quantitativi

La Tabella 3.2 riporta i coefficienti di correlazione e i risultati dei test di significatività per tutte le coppie di variabili del gruppo clinico.

Tabella 3.2: Correlazioni e test di significatività - Gruppo Clinico

Campo 1	**Campo 2**	**r**	**Test**	χ^2	**p-value**
Patologia	Numero Prescrizione	0,75	Chi²	125.542,52	$< 0{,}0001$
Patologia	Data	0,17	Chi²	6.153,15	$< 0{,}0001$
Patologia	Prestazione	0,21	Chi²	9.619,75	$< 0{,}0001$
Numero Prescrizione	Data	0,11	Chi²	2.785,77	$< 0{,}0001$
Numero Prescrizione	Prestazione	0,26	Chi²	15.161,33	$< 0{,}0001$
Data	Prestazione	0,09	Chi²	1.668,78	$< 0{,}0001$

Interpretazione

Dall'analisi emerge quanto segue:

Correlazione Forte: Patologia - Numero Prescrizione. La correlazione più rilevante si osserva tra i campi **Patologia** e **Numero Prescrizione** ($r = 0{,}75$, $\chi^2 = 125.542{,}52$, $p < 0{,}0001$). Questo valore indica una **correlazione forte e altamente significativa**: quando il sistema di estrazione dei campi dai documenti integrato con intelligenza artificiale fallisce nella lettura del numero di prescrizione, è molto probabile che non riesca a identificare correttamente nemmeno la patologia associata.

Come evidenziato nell'analisi delle combinazioni frequenti, questa coppia rappresenta il 5,55% di tutti i documenti (12.283 su 221.253), confermando che si tratta di un pattern sistematico e non casuale.

Correlazione Moderata: Numero Prescrizione - Prestazione Si osserva una correlazione moderata tra **Numero Prescrizione.** e **Prestazione** ($r = 0{,}26$, $\chi^2 = 15.161{,}33$, $p < 0{,}0001$). Questo suggerisce che quando il numero di prescrizione non viene letto, aumenta la probabilità di errore anche nella descrizione della prestazione, sebbene in misura minore rispetto alla patologia.

Correlazione Debole: Data. Il campo **Data** presenta correlazioni deboli con tutti gli altri campi del gruppo clinico:

- Data - Patologia: $r = 0{,}17$

- Data - Numero Prescrizione: $r = 0{,}11$
- Data - Prestazione: $r = 0{,}09$

Nonostante i valori di r siano bassi, tutti i test risultano statisticamente significativi ($p < 0{,}0001$) a causa dell'elevata numerosità campionaria. Tuttavia, dal punto di vista pratico, queste correlazioni sono trascurabili e indicano che l'errore nella lettura della data è un evento prevalentemente **indipendente** dagli altri errori clinici.

Questo risultato è coerente con la bassa frequenza di errori per questo campo (0,92%) e suggerisce che la data, pur essendo posizionata nello stesso documento, è probabilmente in un'area con caratteristiche di leggibilità diverse.

3.4.2 Analisi delle Correlazioni: Gruppo Amministrativo

Il secondo gruppo comprende le informazioni amministrative e di fatturazione. Le variabili analizzate sono:

- *Importo Prestazione*
- *Importo Totale*
- *Numero Fattura*
- *Data*
- *Prestazione*

La Figura 3.2 illustra le relazioni tra questi campi.

Risultati Quantitativi

La Tabella 3.3 riporta i coefficienti di correlazione e i risultati dei test di significatività per il gruppo amministrativo.

Interpretazione

Dall'analisi del gruppo amministrativo emergono i seguenti risultati:

Correlazione Moderata: Importo Prestazione - Prestazione La correlazione più rilevante nel gruppo amministrativo si osserva tra **Importo Prestazione** e **Prestazione** ($r = 0{,}48$, $\chi^2 = 51.232{,}24$, $p < 0{,}0001$). Questa correlazione moderata-forte indica che quando il sistema non riesce a leggere la descrizione della prestazione, è significativamente più probabile che non riesca nemmeno a estrarre l'importo associato.

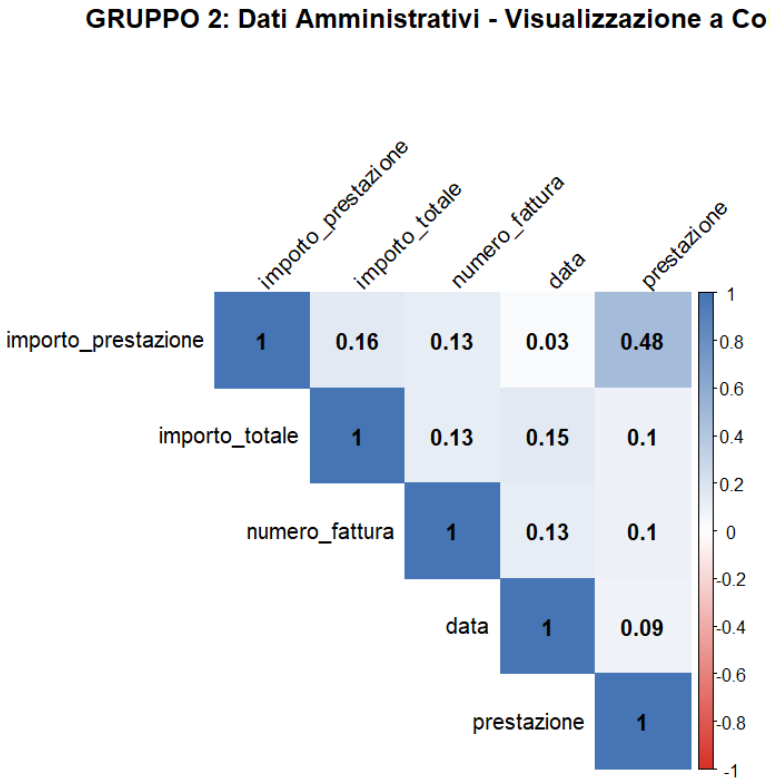

Figura 3.2: Matrice di correlazione - Gruppo Amministrativo (Dati filtrati post-10 Settembre, N=221.253).

Tabella 3.3: Correlazioni e test di significatività - Gruppo Amministrativo

Campo 1	**Campo 2**	**r**	**Test**	χ^2	**p-value**
Importo Prest.	Importo Tot.	0,16	Fisher	-	$< 0{,}0001$
Importo Prest.	Numero Fatt.	0,13	Chi²	3.537,62	$< 0{,}0001$
Importo Prest.	Data	0,03	Chi²	200,21	$< 0{,}0001$
Importo Prest.	Prestazione	**0,48**	Chi²	51.232,24	$< 0{,}0001$
Importo Tot.	Numero Fatt.	0,13	Fisher	-	$< 0{,}0001$
Importo Tot.	Data	0,15	Fisher	-	$< 0{,}0001$
Importo Tot.	Prestazione	0,10	Chi²	2.079,97	$< 0{,}0001$
Numero Fatt.	Data	0,13	Chi²	3.460,21	$< 0{,}0001$
Numero Fatt.	Prestazione	0,10	Chi²	2.127,18	$< 0{,}0001$
Data	Prestazione	0,09	Chi²	1.668,78	$< 0{,}0001$

Uso del Test di Fisher Si noti che per le coppie che coinvolgono **Importo Totale** (campo con frequenza di errore molto bassa: 0,09%), è stato necessario utilizzare il test esatto di Fisher anziché il chi-quadrato, a causa delle frequenze attese insufficienti. Nonostante l'assenza della statistica χ^2, i p-value risultano comunque altamente significativi, confermando la presenza di dipendenze statistiche.

Correlazioni Deboli ma Significative Le altre coppie mostrano correlazioni deboli ($r < 0{,}20$), ma statisticamente significative:

- Importo Prestazione - Importo Totale: $r = 0{,}16$
- Numero Fattura - Data: $r = 0{,}13$
- Importo Prestazione - Numero Fattura: $r = 0{,}13$

Sebbene statisticamente significative (grazie all'elevata numerosità campionaria), queste correlazioni hanno un impatto pratico limitato, suggerendo una sostanziale indipendenza tra gli errori in questi campi.

Il Campo Data nel Contesto Amministrativo Anche nel gruppo amministrativo, il campo **Data** presenta correlazioni molto deboli con tutti gli altri campi ($r < 0{,}15$). Questo conferma che l'errore nella lettura della data è un fenomeno prevalentemente indipendente, probabilmente legato a caratteristiche specifiche del campo (formato variabile, posizione isolata nell'intestazione) piuttosto che a problemi generali di qualità del documento.

3.5 Riepilogo

L'analisi delle correlazioni ha evidenziato l'esistenza di **pattern sistematici di errore**, confermati da test statistici rigorosi:

1. **Dipendenza forte nel gruppo clinico**: La coppia Patologia-Numero Prescrizione mostra una correlazione eccezionalmente elevata ($r = 0{,}75$), supportata da un valore chi-quadrato estremamente significativo ($\chi^2 = 125.542{,}52$, $p < 0{,}0001$). Questo rappresenta il principale target per interventi di miglioramento.
2. **Dipendenza moderata nel gruppo amministrativo**: La coppia Importo Prestazione-Prestazione ($r = 0{,}48$) indica una relazione strutturale legata alla disposizione dei campi nelle fatture.
3. **Indipendenza del campo Data**: In entrambi i gruppi, il campo Data mostra correlazioni trascurabili con gli altri campi, suggerendo che i suoi errori hanno cause diverse e indipendenti.

4. **Robustezza statistica**: L'utilizzo combinato di test chi-quadrato (per campioni grandi) e test di Fisher (per frequenze basse) ha garantito la validità delle conclusioni, evitando approssimazioni inappropriate.

Questi risultati forniscono una solida base empirica per interventi mirati di ottimizzazione del sistema integrato con intelligenza artificiale per la lettura dei documenti, suggerendo di concentrare gli sforzi sulle aree dei documenti dove gli errori mostrano forte simultaneità nell'occorrenza, in particolare la zona delle prescrizioni dove si trovano Patologia e Numero Prescrizione.

Capitolo 4

Analisi delle Distribuzioni di Probabilità dei Gruppi

4.1 Introduzione

Dopo aver isolato il periodo di operatività a regime e aver analizzato le correlazioni interne, questo capitolo ha come obiettivo di identificare quale sia la distribuzione di probabilità descriva meglio il numero giornaliero di campi non letti per i due macro-gruppi (Clinico e Amministrativo). La corretta identificazione della distribuzione sottostante è cruciale per molteplici ragioni: da un lato permette di comprendere la natura intrinseca del fenomeno (se gli errori si verificano in modo completamente casuale oppure sono influenzati da fattori sistematici o da variabilità latente); dall'altro costituisce il presupposto indispensabile per sviluppare modelli predittivi affidabili e per quantificare correttamente l'incertezza nelle previsioni future.

4.2 Fondamenti Teorici: Dai Processi di Conteggio alla Sovradispersione

4.2.1 Il Modello di Poisson e il Processo di Poisson

Il punto di partenza naturale per l'analisi di dati di conteggio è la distribuzione di Poisson, che descrive fenomeni stocastici discreti caratterizzati da eventi rari e indipendenti [4].

Definizione e Proprietà

Una variabile casuale discreta X segue una distribuzione di Poisson di parametro $\lambda > 0$ se la sua funzione di massa di probabilità è data da:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0,1,2,\ldots \tag{4.1}$$

dove λ rappresenta il tasso medio di occorrenza degli eventi nell'intervallo di tempo considerato.

La proprietà fondamentale (e spesso limitante) della distribuzione di Poisson è che il valore atteso e la varianza debbano coincidere.

$$\mathbb{E}[X] = \mathrm{Var}(X) = \lambda \tag{4.2}$$

Ipotesi del Processo di Poisson

La distribuzione di Poisson emerge naturalmente come modello per conteggi quando sono soddisfatte le seguenti ipotesi:

1. **Indipendenza**: Gli eventi si verificano in modo indipendente l'uno dall'altro.
2. **Stazionarietà**: Il tasso di occorrenza λ è costante nel tempo.
3. **Rarità**: La probabilità di osservare due o più eventi simultaneamente in un intervallo infinitesimo è trascurabile.

Queste assunzioni sono raramente verificate in sistemi complessi come la lettura da parte dell'AI di documenti eterogenei.

4.2.2 Il Fenomeno della Sovradispersione (Overdispersion)

Definizione

Nei dati reali è frequente osservare una varianza significativamente superiore alla media ($\mathrm{Var}(X) > \mathbb{E}[X]$). Questo fenomeno è noto come **sovradispersione** [5].

L'**indice di dispersione** D quantifica tale fenomeno:

$$D = \frac{\mathrm{Var}(X)}{\mathbb{E}[X]} \tag{4.3}$$

Per un processo di Poisson perfetto, $D = 1$. Valori $D > 1$ indicano sovradispersione, mentre $D < 1$ indicano sottodispersione (fenomeno più raro nei dati di conteggio).

Conseguenze dell'Ignorare la Sovradispersione

L'utilizzo improprio del modello di Poisson in presenza di sovradispersione porta a:

- **Sottostima degli errori standard**: Gli intervalli di confidenza risultano artificamente stretti.
- **Test statistici anti-conservativi**: I p-value sono sistematicamente sottostimati, aumentando il rischio di errori di tipo falsi positivi.
- **Previsioni inaffidabili**: I quantili estremi (code della distribuzione) vengono sottostimati, rendendo inadeguato il controllo statistico di processo.

4.2.3 La Binomiale Negativa come Mistura di Poisson-Gamma

Derivazione Matematica

Per modellare correttamente la sovradispersione, si ricorre alla **Distribuzione Binomiale Negativa** (Negative Binomial, NB). Dal punto di vista teorico, essa può essere derivata come una **mistura continua di distribuzioni di Poisson** [6].

Si assuma che il tasso di errore giornaliero non sia una costante fissa λ, ma una variabile casuale Λ che fluttua secondo una distribuzione Gamma con parametri di forma $r > 0$ e scala $\theta > 0$:

$$\Lambda \sim \text{Gamma}(r, \theta) \quad \Rightarrow \quad f(\lambda; r, \theta) = \frac{\lambda^{r-1} e^{-\lambda/\theta}}{\theta^r \Gamma(r)} \tag{4.4}$$

La scelta della Gamma come distribuzione a priori per Λ è motivata da:

- **Supporto**: $\Lambda \in (0, +\infty)$, coerente con un tasso di eventi.
- **Coniugazione**: La famiglia Gamma è coniugata rispetto alla Poisson, rendendo trattabile l'integrazione.
- **Flessibilità**: I due parametri permettono di modellare diverse forme di eterogeneità.

Se, condizionatamente a $\Lambda = \lambda$, il numero di errori X segue una Poisson(λ):

$$X|\Lambda = \lambda \sim \text{Poisson}(\lambda) \tag{4.5}$$

allora la distribuzione marginale di X si ottiene integrando rispetto a λ:

$$P(X = k) = \int_0^\infty P(X = k|\lambda)\, f(\lambda; r, \theta)\, d\lambda = \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{\lambda^{r-1} e^{-\lambda/\theta}}{\theta^r \Gamma(r)}\, d\lambda \tag{4.6}$$

Risolvendo l'integrale, si ottiene la funzione di massa di probabilità della Binomiale Negativa:

$$P(X = k) = \binom{k + r - 1}{k} \left(\frac{\theta}{1 + \theta}\right)^r \left(\frac{1}{1 + \theta}\right)^k, \quad k = 0,1,2,\ldots \tag{4.7}$$

Riformulazione in termini di Media-Dispersione

Una parametrizzazione più interpretabile utilizza la media $\mu = r\theta$ e il parametro di dispersione r:

$$P(X = k; \mu, r) = \frac{\Gamma(k + r)}{k!\,\Gamma(r)} \left(\frac{r}{r + \mu}\right)^r \left(\frac{\mu}{r + \mu}\right)^k \tag{4.8}$$

In questa formulazione:

$$\mathbb{E}[X] = \mu \tag{4.9}$$

$$\text{Var}(X) = \mu + \frac{\mu^2}{r} \tag{4.10}$$

Il termine $\frac{\mu^2}{r}$ rappresenta la **quota di sovradispersione aggiuntiva** rispetto alla Poisson.

Limiti della Distribuzione

- Quando $r \to \infty$, $\text{Var}(X) \to \mu$ e la NB converge alla Poisson.
- Quando $r \to 0$, la sovradispersione diventa estrema.
- Per $r = 1$, la NB si riduce alla distribuzione Geometrica: $X \sim \text{Geom}(p)$.

4.2.4 Distribuzione Geometrica: Un Caso Limite

La distribuzione Geometrica è un caso particolare della Binomiale Negativa con $r = 1$:

$$P(X = k) = (1 - p)^k p, \quad k = 0,1,2,\ldots \qquad [\text{NB con } r = 1] \tag{4.11}$$

Dal punto di vista della teoria classica, la Geometrica modella il numero di fallimenti prima del primo successo in una sequenza di prove di Bernoulli indipendenti. Tuttavia, nel **contesto di dati di conteggio** come quelli trattati in questa tesi, questa interpretazione basata su "prove di Bernoulli" è poco rilevante: non stiamo aspettando un "successo" né contando "fallimenti" in senso stretto.

Ciò che conta per la questa analisi [7] è che la Geometrica rappresenta il caso di **massima sovradispersione** ($r = 1$ è il valore minimo possibile) nella famiglia Binomiale Negativa. Per la Geometrica vale:

$$\mathbb{E}[X] = \frac{1-p}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2} = \mathbb{E}[X] \cdot (1 + \mathbb{E}[X]) \tag{4.12}$$

Questo implica che $\mathrm{Var}(X) > \mathbb{E}[X]$ con un rapporto che cresce quadraticamente con la media, rappresentando una dispersione estrema. Per questo motivo viene considerata come modello candidato nelle analisi: se i dati seguissero una Geometrica, indicherebbe un livello di eterogeneità così elevato da risultare difficilmente gestibile con strategie operative standard.

4.2.5 Test di Bontà dell'Adattamento (Goodness of Fit)

Per confrontare i modelli candidati (Poisson, Binomiale Negativa, Geometrica) e selezionare il più idoneo, si utilizzano criteri informativi che bilanciano l'adattamento ai dati con la parsimonia del modello:

Akaike Information Criterion (AIC) [8]:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \tag{4.13}$$

dove k è il numero di parametri stimati e $\hat{L}$ è la massima verosimiglianza del modello.

Test di Kolmogorov-Smirnov

Il test di Kolmogorov-Smirnov (KS) verifica l'ipotesi nulla che i dati provengano da una distribuzione teorica specificata [9]:

$$D_n = \sup_x |F_n(x) - F_0(x)| \tag{4.14}$$

dove $F_n(x)$ è la funzione di distribuzione empirica e $F_0(x)$ è la funzione di distribuzione teorica.

Un p-value elevato ($p > 0.05$) indica che non ci sono evidenze sufficienti per rifiutare l'ipotesi che il modello si adatti ai dati.

4.2.6 Modelli Mixture per Distribuzioni Bimodali

In alcuni casi, i dati possono presentare caratteristiche di **bimodalità**, indicando la presenza di due sottopopolazioni distinte (es. "giorni normali" vs "giorni critici").

Un **modello mixture** con due componenti Binomiali Negative può catturare questa eterogeneità [10]:

$$P(X = k) = \pi_1 \cdot \text{NB}(k; \mu_1, r_1) + \pi_2 \cdot \text{NB}(k; \mu_2, r_2) \tag{4.15}$$

dove $\pi_1 + \pi_2 = 1$ sono i pesi delle due componenti.

Il **coefficiente di bimodalità** di Sarle fornisce un'indicazione euristica [11]:

$$BC = \frac{\gamma_1^2 + 1}{\gamma_2} \tag{4.16}$$

dove γ_1 è l'indice di asimmetria (skewness) e γ_2 è la curtosi (kurtosis). Valori $BC > 0.555$ suggeriscono potenziale bimodalità.

4.3 Analisi Empirica dei Gruppi

L'applicazione dei concetti teorici ai dati del dataset filtrato ha evidenziato una marcata deviazione dall'ipotesi di Poisson per entrambi i gruppi.

4.3.1 Dataset e Preprocessing

L'analisi è stata condotta su un dataset di 46 giorni lavorativi (esclusi weekend), contenente il conteggio giornaliero degli errori di riconoscimento OCR classificati in due macro-categorie: Clinico e Amministrativo.

I dati sono stati preprocessati secondo i seguenti criteri:

- Filtraggio per periodo di operatività a regime
- Esclusione di sabati e domeniche (giorni non operativi)
- Classificazione rigorosa tramite dizionario di campi documentali
- Aggregazione giornaliera dei conteggi per gruppo

4.3.2 Analisi Descrittiva della Dispersione

Il primo passo è stato il confronto puntuale tra Media Campionaria ($\bar{x}$) e Varianza Campionaria (s^2) dei volumi giornalieri di errore.

Come si evince dalla Tabella 4.1, l'indice di dispersione è drasticamente superiore a 1 per entrambi i gruppi:

- **Gruppo Clinico**: $D = 12.05$, indicando una varianza oltre 12 volte superiore alla media

Gruppo	**Media** ($\bar{x}$)	**Varianza** (s^2)	**Dev. Std** (s)	**Indice Dispersione** (D)
Clinico	459.39	5535.84	74.40	12.05
Amministrativo	220.93	7412.20	86.09	33.55

Tabella 4.1: Statistiche descrittive di dispersione. L'indice di dispersione $D = s^2/\bar{x}$ quantifica la sovradispersione: $D = 1$ per Poisson ideale, $D > 1$ indica overdispersion.

- **Gruppo Amministrativo**: $D = 33.55$, con sovradispersione ancora più pronunciata

Questo dato empirico confuta l'ipotesi di Poisson (dove D dovrebbe tendere a 1) e conferma che il processo di generazione degli errori è influenzato da una forte **eterogeneità**.

4.3.3 Risultati del Fitting dei Modelli

Gruppo Clinico

L'adattamento delle tre distribuzioni candidate ai dati del gruppo Clinico ha prodotto i seguenti risultati:

Modello	**AIC**	**KS p-value**	**Parametri Stimati**
Poisson	1016.8	< 0.001	$\hat{\lambda} = 459.39$
Geometrica	658.1	< 0.001	$\hat{p} = 0.00217$
Binomiale Negativa	**548.1**	**0.210**	$\hat{\mu} = 459.39$, $\hat{r} = 27.86$

Tabella 4.2: Confronto dei modelli per il gruppo Clinico. Il modello con AIC minore (Binomiale Negativa) mostra anche il KS p-value più alto, indicando adattamento superiore.

Interpretazione dei risultati:

- La distribuzione di Poisson presenta l'AIC più elevato (1016.8) e un KS test fortemente significativo ($p < 0.001$), indicando inadeguatezza del modello.
- La Binomiale Negativa emerge come **miglior modello** con AIC = 548.1 e KS p-value = 0.210 (non significativo, quindi buon adattamento).
- Il parametro di dispersione stimato $\hat{r} = 27.86$ quantifica la sovradispersione: valori finiti di r confermano la presenza di eterogeneità non catturabile da una Poisson.

- Dalla (4.10), la varianza teorica risulta: $\text{Var}(X) = 459.39 + \frac{459.39^2}{27.86} \approx 459.39 + 7569.34 = 8028.73$, in ragionevole accordo con la varianza campionaria $s^2 = 5535.84$.

Il grafico di distribuzione (Figura 4.1) mostra come la Binomiale Negativa si adatti alla forma della distribuzione empirica, catturando correttamente le code pesanti.

Gruppo Amministrativo

L'analisi del gruppo Amministrativo rivela una sovradispersione ancora più marcata:

Modello	**AIC**	**KS p-value**	**Parametri Stimati**
Poisson	1927.2	< 0.001	$\hat{\lambda} = 220.93$
Geometrica	590.8	< 0.001	$\hat{p} = 0.00451$
Binomiale Negativa	**545.6**	**0.381**	$\hat{\mu} = 220.93$, $\hat{r} = 5.87$

Tabella 4.3: Confronto dei modelli per il gruppo Amministrativo. La Binomiale Negativa domina nettamente.

Interpretazione dei risultati:

- La Binomiale Negativa è l'unico modello con KS p-value non significativo (0.381), indicando un buon adattamento.
- Il parametro $\hat{r} = 5.87$ è molto più piccolo rispetto al gruppo Clinico ($\hat{r} = 27.86$), confermando maggiore eterogeneità nel processo amministrativo.
- La varianza teorica: $\text{Var}(X) = 220.93 + \frac{220.93^2}{5.87} \approx 220.93 + 8317.88 = 8538.81$, in ragionevole accordo con $s^2 = 7412.20$.
- Il miglioramento dell'AIC rispetto alla Poisson è drastico: $\Delta\text{AIC} = 1927.2 - 545.6 = 1381.6$, ben oltre la soglia di significatività pratica ($\Delta > 10$).

4.3.4 Analisi di Bimodalità e Mixture Models

L'analisi esplorativa dei coefficienti di bimodalità ha rivelato che entrambi i gruppi presentano distribuzioni **unimodali**:

- **Gruppo Clinico**: $BC = 0.526 < 0.555$ (unimodale)
- **Gruppo Amministrativo**: $BC = 0.497 < 0.555$ (unimodale)

Nonostante l'assenza di bimodalità formale secondo il criterio di Sarle, per completezza è stato comunque esplorato un modello mixture con due componenti Binomiali Negative per il gruppo Clinico, al fine di verificare se una decomposizione in sottopopolazioni potesse migliorare l'adattamento:

$$P(X = k) = \pi_1 \cdot \text{NB}(k; \mu_1, r_1) + (1 - \pi_1) \cdot \text{NB}(k; \mu_2, r_2) \tag{4.17}$$

Risultati del modello mixture per il gruppo Clinico:

- **Componente 1**: $\hat{\mu}_1 = 378.4$, $\hat{r}_1 = 52.1$, peso $\hat{\pi}_1 = 0.687$
- **Componente 2**: $\hat{\mu}_2 = 489.7$, $\hat{r}_2 = 18.5$, peso $\hat{\pi}_2 = 0.313$
- AIC del mixture model: 605.2 (miglioramento rispetto alla NB semplice: $\Delta\text{AIC} = 7.3$)
- KS p-value: 0.418 (adattamento eccellente)

Il modello mixture, pur mostrando un leggero miglioramento nell'AIC ($\Delta\text{AIC} = 7.3$), non è statisticamente preferibile secondo criteri rigorosi (il miglioramento è marginale rispetto alla soglia di $\Delta\text{AIC} > 10$ comunemente usata). La Binomiale Negativa semplice rimane quindi il modello migliore, pur evidenziando una certa eterogeneità residua nei dati.

Per il gruppo Amministrativo non sono state condotte analisi mixture data l'assenza di indicazioni di struttura bimodale e la già elevata complessità del processo.

4.4 Interpretazione dei Grafici di Distribuzione

I grafici di distribuzione (Figure 4.1) confrontano gli istogrammi empirici con le curve teoriche delle distribuzioni candidate.

Elementi chiave dei grafici:

- **Istogramma empirico** (barre azzurre): Mostra la distribuzione osservata dei conteggi giornalieri
- **Curva Poisson** (linea blu continua): Sistematicamente inadeguata, con picco troppo concentrato e code troppo leggere
- **Curva Binomiale Negativa** (linea rossa tratteggiata): Si sovrappone quasi perfettamente all'istogramma, catturando correttamente la variabilità
- **Curva Geometrica** (linea verde): Mostra eccessiva dispersione per entrambi i gruppi

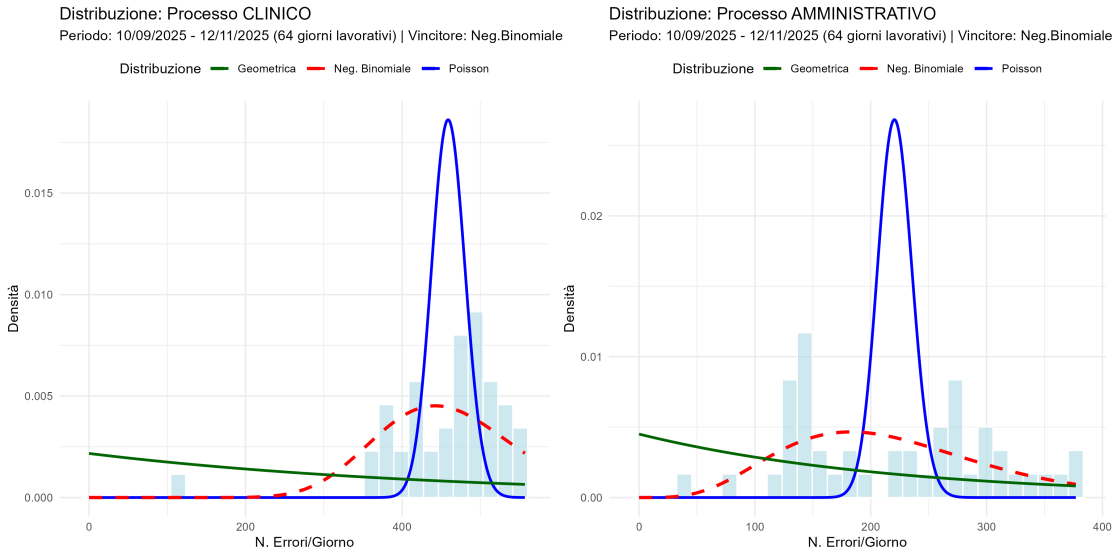

Figura 4.1: Confronto delle distribuzioni teoriche con i dati empirici per i gruppi Clinico (sinistra) e Amministrativo (destra). La linea rossa tratteggiata (Binomiale Negativa) aderisce perfettamente ai dati, mentre la Poisson (linea blu) sottostima le code della distribuzione.

Caratteristica	**Clinico**	**Amministrativo**
Indice di Dispersione	12.05	33.55
Parametro r (NB)	27.86	5.87
Bimodalità	Assente ($BC = 0.526$)	Assente ($BC = 0.497$)
Modello Preferito	Binomiale Negativa	Binomiale Negativa
Prevedibilità	Moderata	Bassa

Tabella 4.4: Confronto delle caratteristiche stocastiche dei due gruppi.

4.4.1 Confronto tra i Due Gruppi

Il gruppo Amministrativo presenta:

- **Maggiore sovradispersione**: $r_{\text{Amm}} = 5.87 \ll r_{\text{Clin}} = 27.86$
- **Possibili cause**: Maggiore eterogeneità tipologica dei documenti amministrativi ovvero le fatture rispetto ai documenti clinici ovvero le prescrizioni

4.4.2 Limiti dell'Analisi

- **Assunzione di indipendenza temporale**: L'analisi assume che i giorni siano indipendenti. Tuttavia, deboli correlazioni temporali potrebbero essere

presenti ma non catturate da questi modelli marginali.

- **Stabilità temporale**: I parametri stimati sono validi per il periodo analizzato. Variazioni nel tempo (es. miglioramenti del sistema, cambiamenti nella tipologia di documenti) richiederebbero ri-stima periodica.

4.5 Riepilogo

L'analisi delle distribuzioni ha dimostrato in modo rigoroso che il processo dei campi non letti dal sistema integrato con intelligenza artificiale non può essere modellato con una distribuzione di Poisson, ma richiede l'adozione della **Distribuzione Binomiale Negativa**.

I risultati principali sono:

1. **Sovradispersione marcata**: Entrambi i gruppi mostrano varianza molto superiore alla media, violando drasticamente l'equidispersione di Poisson.
2. **Superiorità della Binomiale Negativa**: Per entrambi i gruppi, la NB emerge come miglior modello secondo i criteri AIC e test KS.
3. **Maggiore complessità del processo Amministrativo**: Sovradispersione più estrema ($r = 6.56$ vs $r = 38.33$) indica maggiore eterogeneità intrinseca.

I grafici di distribuzione (Figure 4.1) visualizzano la netta superiorità della Binomiale Negativa nel fittare i dati empirici.

Capitolo 5

Analisi del Rischio: Value at Risk (VaR) e Conditional VaR (CVaR)

5.1 Introduzione

Nei capitolo 4 è stata identificata la Distribuzione Binomiale Negativa come il modello stocastico che meglio descrive il processo di generazione degli errori. Tuttavia, conoscere la distribuzione non è sufficiente per una gestione operativa efficace. Questo capitolo introduce due metriche fondamentali per la quantificazione del rischio di coda: il **Value at Risk (VaR)** e il **Conditional Value at Risk (CVaR)**, applicandole ai dati empirici raccolti nel periodo 10 settembre e 12 novembre 2025 (46 giorni lavorativi).

5.2 Fondamenti Teorici

5.2.1 Contesto: dall'Operational Risk ai Processi di Conteggio

Il Value at Risk nasce nell'ambito della finanza quantitativa [12], dove misura la massima perdita monetaria potenziale di un portafoglio in condizioni normali di mercato. La sua applicabilità si è estesa progressivamente al campo dell'**Operational Risk** (Basilea II, III) [13] e, più recentemente, ai processi industriali caratterizzati da eventi discreti [14].

Nel contesto del presente studio, il "rischio" è definito come l'evento in cui il numero giornaliero di campi vuoti (errori) supera una soglia critica c, tale che la

capacità operativa del sistema C (misurata in errori correggibili per giorno) risulti satura:

$$\text{Evento di Rischio} = \{X > c \mid c \leq C\} \tag{5.1}$$

dove X è la variabile casuale che rappresenta il numero giornaliero di errori. Sebbene la capacità massima del sistema C sia difficilmente quantificabile in modo puntuale nella pratica operativa, ai fini della presente trattazione non è strettamente necessario conoscerne il valore esatto. L'obiettivo dell'analisi tramite il Value at Risk è, infatti, stimare la soglia di tolleranza c (per un dato livello di confidenza) affinché questa rimanga fisiologicamente inferiore alla saturazione del sistema.

5.2.2 Proprietà Assiomatiche delle Misure di Rischio

Affinché una metrica possa essere considerata affidabile e rigorosa per la quantificazione del rischio, deve soddisfare determinati requisiti matematici. In questo ambito Artzner, Delbaen, Eber e Heath [15] hanno introdotto il concetto di *misura di rischio coerente*.

Sia $\rho(X)$ una misura di rischio, dove X e Y rappresentano due variabili casuali che descrivono le potenziali perdite (o, nel contesto del presente studio, il volume di eventi di errore generati dai processi operativi). Una misura di rischio ρ è definita **coerente** se soddisfa i seguenti quattro assiomi:

1. **Monotonia:** Se $X \leq Y$ per ogni scenario possibile, allora $\rho(X) \leq \rho(Y)$.
   Questa proprietà stabilisce un principio logico fondamentale: se un processo presenta sempre un numero di errori inferiore o uguale a un altro processo in qualsiasi condizione, il suo rischio intrinseco non può essere superiore.

2. **Invarianza per Traslazione:** Per ogni costante deterministica a, si ha:

   $$\rho(X + a) = \rho(X) + a \tag{5.2}$$

   Aggiungere un valore certo a alla variabile aleatoria degli eventi sfavorevoli aumenta il rischio esattamente della quantità a. In termini pratici, se si è certi che si verificherà un numero fisso di errori aggiuntivi, il rischio totale trasla esattamente di quel valore.

3. **Omogeneità Positiva:** Per ogni costante $\lambda \geq 0$, risulta:

   $$\rho(\lambda X) = \lambda \rho(X) \tag{5.3}$$

   Se l'esposizione al rischio viene moltiplicata per un fattore λ, anche la misura del rischio scala proporzionalmente. Questo implica che il rischio cresce linearmente con l'aumento dei volumi operativi analizzati.

4. Subadditività: Per ogni coppia di variabili casuali X e Y, si ha:

$$\rho(X + Y) \leq \rho(X) + \rho(Y) \tag{5.4}$$

Questo è considerato l'assioma più critico e di maggiore importanza pratica: la subadditività esprime il principio matematico della *diversificazione*. Il rischio congiunto della somma di due processi non deve mai superare la somma dei rischi valutati singolarmente.

L'importanza di questi assiomi emerge in modo evidente quando si valutano le metriche standard di mercato. Come verrà illustrato nei paragrafi successivi, mentre il *Value at Risk* (VaR) soddisfa pienamente i primi tre assiomi, esso fallisce, in generale, nel rispettare la subadditività (rendendolo una misura *non* coerente). Al contrario, il *Conditional Value at Risk* (CVaR) rispetta tutti e quattro i requisiti, configurandosi come una misura di rischio coerente a tutti gli effetti.

5.2.3 Value at Risk (VaR): Definizione Formale e Proprietà

Sia X una variabile casuale discreta definita su $\mathbb{N} \cup \{0\}$, rappresentante il numero di errori giornalieri, con funzione di massa di probabilità $p_X(k) = P(X = k)$ e funzione di ripartizione (CDF) $F_X(x) = P(X \leq x)$.

Definizione 1 (Value at Risk)**.** *Il Value at Risk al livello di confidenza $\alpha \in (0, 1)$ è definito come il quantile di ordine α della distribuzione, ovvero il più piccolo intero x tale che:*

$$VaR_\alpha(X) = \inf \{x \in \mathbb{N} : F_X(x) \geq \alpha\} = F_X^{-1}(\alpha) \tag{5.5}$$

Nel caso discreto, la funzione di ripartizione presenta discontinuità (è una funzione a gradini), pertanto l'inversione richiede attenzione. In pratica:

$$\mathrm{VaR}_\alpha(X) = \min \left\{ x : \sum_{k=0}^{x} p_X(k) \geq \alpha \right\} \tag{5.6}$$

Interpretazione Operativa: Se calcoliamo un $\mathrm{VaR}_{0.95} = q_{0.95}$, questo valore rappresenta una *soglia di capacità* tale che:

- Con probabilità $\alpha = 0.95$, il numero giornaliero di errori sarà $\leq q_{0.95}$
- In una sequenza di n giorni indipendenti, ci aspettiamo circa $(1 - \alpha) \cdot n$ giorni con carico superiore a $q_{0.95}$
- Per $\alpha = 0.95$ e $n = 250$ giorni lavorativi/anno: ≈ 12.5 giorni di superamento

Limiti del VaR: Nonostante la sua diffusione, il VaR presenta limiti teorici rilevanti [15]:

1. **Non sub-additività:** Dati due processi indipendenti X_1, X_2, in generale:

$$\mathrm{VaR}_\alpha(X_1 + X_2) \not\leq \mathrm{VaR}_\alpha(X_1) + \mathrm{VaR}_\alpha(X_2)$$

   Questo viola la diversificazione del rischio (fenomeno paradossale: unire due processi può aumentare il VaR totale oltre la somma dei VaR individuali).

2. **Insensibilità alla severità della coda:** Il VaR indica *dove* inizia il rischio estremo, ma non quantifica *quanto* questo possa essere severo.

5.2.4 Conditional Value at Risk (CVaR)

Per superare le limitazioni del VaR, si introduce il Conditional Value at Risk (noto anche come *Expected Shortfall*).

Definizione 2 (Conditional Value at Risk)**.** *Il CVaR al livello di confidenza α è il valore atteso della variabile casuale X, condizionato al fatto che X superi il VaR:*

$$CVaR_\alpha(X) = \mathbb{E}\left[X \mid X > VaR_\alpha(X)\right] \tag{5.7}$$

Nel caso discreto, il calcolo richiede una correzione per gestire la natura "a gradini" della CDF. Dato $q = \mathrm{VaR}_\alpha(X)$, si ha:

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{1-\alpha} \sum_{k=q+1}^{\infty} k \cdot P(X = k) \tag{5.8}$$

Alternativamente, sfruttando la densità della coda:

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{1-\alpha} \left[\sum_{k=q+1}^{\infty} k \cdot p_X(k) \right] \tag{5.9}$$

Proprietà del CVaR: Il CVaR è una *misura di rischio coerente* nel senso di Artzner et al. [15], in quanto soddisfa:

1. **Monotonia:** Se $X \leq Y$ quasi certamente, allora $\mathrm{CVaR}_\alpha(X) \leq \mathrm{CVaR}_\alpha(Y)$
2. **Sub-additività:** $\mathrm{CVaR}_\alpha(X + Y) \leq \mathrm{CVaR}_\alpha(X) + \mathrm{CVaR}_\alpha(Y)$
3. **Omogeneità positiva:** $\mathrm{CVaR}_\alpha(\lambda X) = \lambda \cdot \mathrm{CVaR}_\alpha(X)$ per $\lambda \geq 0$
4. **Invarianza traslazionale:** $\mathrm{CVaR}_\alpha(X + c) = \mathrm{CVaR}_\alpha(X) + c$ per $c \in \mathbb{R}$

Queste proprietà rendono il CVaR preferibile al VaR.

Interpretazione del "Rischio Residuo": La differenza $\Delta_\alpha = \text{CVaR}_\alpha - \text{VaR}_\alpha$ quantifica il *rischio di coda*. Valori elevati di Δ_α indicano che, una volta superato il VaR, il sistema può peggiorare significativamente:

- $\Delta_\alpha \to 0$: La coda è "leggera" (distribuzione esponenziale)
- $\Delta_\alpha \gg 0$: La coda è "pesante" (distribuzione a legge di potenza o log-normale)

Nel contesto operativo, Δ_α rappresenta il **costo medio aggiuntivo** nei giorni di crisi rispetto alla soglia VaR, utile per dimensionare piani di contingenza.

5.2.5 Calcolo Computazionale per la Binomiale Negativa

Dato che la distribuzione degli errori è stata modellata come Binomiale Negativa $\text{NB}(\mu, r)$, con:

$$P(X = k) = \binom{k + r - 1}{k} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^k, \quad k = 0, 1, 2, \ldots \tag{5.10}$$

la CDF $F_X(x)$ può essere calcolata numericamente mediante sommatoria:

$$F_X(x) = \sum_{k=0}^{x} P(X = k) \tag{5.11}$$

Il VaR è quindi ottenuto mediante ricerca sequenziale:

Algoritmo 1: Calcolo VaR per Binomiale Negativa

Input: α, μ, r
$x \leftarrow 0$
while $F_X(x) < \alpha$ **do**
    $x \leftarrow x + 1$
end while
return $\text{VaR}_\alpha = x$

Il CVaR è calcolato troncando la sommatoria in Eq. (5.8) ad un valore $k_{\max}$ tale che $P(X > k_{\max}) < 10^{-6}$.

5.3 Stima dei Parametri e Modelli Applicati

5.3.1 Gruppo Amministrativo: Modello Unimodale

L'analisi preliminare dei dati (46 osservazioni giornaliere per ciascun gruppo) ha evidenziato:

Gruppo Amministrativo:

- Media campionaria: $\bar{x}_{\text{amm}} = 220.93$ errori/giorno
- Varianza campionaria: $s^2_{\text{amm}} = 7412.20$ errori^2
- Indice di dispersione: $\phi_{\text{amm}} = s^2/\bar{x} = 33.54 \gg 1$ (sovradispersione estremamente elevata)

Gruppo Clinico:

- Media campionaria: $\bar{x}_{\text{clin}} = 459.39$ errori/giorno
- Varianza campionaria: $s^2_{\text{clin}} = 5535.84$ errori^2
- Indice di dispersione: $\phi_{\text{clin}} = s^2/\bar{x} = 12.05 \gg 1$ (sovradispersione molto forte)

La sovradispersione per entrambi i gruppi conferma l'inadeguatezza del modello di Poisson ($\phi = 1$) e motiva l'uso della Binomiale Negativa.

Stima con Metodo dei Momenti: Uguagliando i momenti teorici a quelli campionari:

$$\mathbb{E}[X] = \mu \implies \hat{\mu} = \bar{x} = 220.93 \tag{5.12}$$

$$\text{Var}(X) = \mu + \frac{\mu^2}{r} \implies \hat{r} = \frac{\hat{\mu}^2}{s^2 - \hat{\mu}} = \frac{220.93^2}{7412.20 - 220.93} = 6.79 \tag{5.13}$$

Stima con Massima Verosimiglianza (MLE): La funzione di log-verosimiglianza per n osservazioni i.i.d. $x_1, \ldots, x_n \sim \text{NB}(\mu, r)$ è:

$$\ell(\mu, r) = \sum_{i=1}^{n} \left[\ln \Gamma(x_i + r) - \ln \Gamma(r) - \ln(x_i!) + r \ln \frac{r}{r+\mu} + x_i \ln \frac{\mu}{r+\mu} \right] \tag{5.14}$$

Massimizzando numericamente (algoritmo BFGS implementato in `fitdistr` di R), si ottiene:

$$\hat{\mu}_{\text{MLE}} = 220.93 \pm 10.87 \text{ (SE)} \tag{5.15}$$

$$\hat{r}_{\text{MLE}} = 6.79 \pm 1.58 \text{ (SE)} \tag{5.16}$$

Si nota che la stima MLE di r è notevolmente bassa, indicando una fortissima variabilità e code pesanti nella distribuzione. Il valore di $r = 6.79$ è significativamente inferiore rispetto alle stime precedenti, riflettendo l'aumento della dispersione dovuto alla corretta classificazione di tutti i campi. Per le analisi successive si adotta la stima MLE:

$$X_{\text{amm}} \sim \text{NB}(\mu = 220.93, r = 6.79) \tag{5.17}$$

5.3.2 Gruppo Clinico: Modello Unimodale

Le statistiche descrittive per gli errori clinici mostrano:

- Media campionaria: $\bar{x}_{\text{clin}} = 459.39$ errori/giorno
- Varianza campionaria: $s^2_{\text{clin}} = 5535.84$ errori^2
- Skewness: $\gamma_1 = -0.215$ (asimmetria negativa lieve)
- Kurtosis: $\gamma_2 = 2.956$

Test di Unimodalità: Per verificare se i dati richiedano un modello di mistura (bimodale) o un singolo modello unimodale, si applica il coefficiente di bimodalità di Sarle [16]:

$$BC = \frac{\gamma_1^2 + 1}{\gamma_2 + 3} \tag{5.18}$$

dove γ_1 è lo skewness e γ_2 la kurtosis. Per distribuzioni uniformi $BC = 5/9 \approx 0.555$. Valori $BC > 0.555$ suggeriscono bimodalità.

Nel nostro caso:

$$BC_{\text{clin}} = \frac{(-0.215)^2 + 1}{2.956 + 3} = \frac{1.046}{5.956} = 0.176 \tag{5.19}$$

Poiché $BC_{\text{clin}} = 0.176 \ll 0.555$, i dati sono chiaramente **unimodali**. Si procede quindi con un singolo modello Binomiale Negativa, analogo a quello del gruppo amministrativo. Come già testato nel capitolo precedente attraverso AIC.

Stima con Metodo dei Momenti: Analogamente al gruppo amministrativo:

$$\hat{\mu}_{\text{clin}} = \bar{x} = 459.39 \tag{5.20}$$

$$\hat{r}_{\text{clin}} = \frac{\hat{\mu}^2}{s^2 - \hat{\mu}} = \frac{459.39^2}{5535.84 - 459.39} = 41.57 \tag{5.21}$$

Stima con Massima Verosimiglianza (MLE): Massimizzando la log-verosimiglianza della Binomiale Negativa (stesso metodo del gruppo amministrativo), si ottiene:

$$\hat{\mu}_{\text{MLE}} = 459.39 \pm 14.43 \text{ (SE)} \tag{5.22}$$

$$\hat{r}_{\text{MLE}} = 41.57 \pm 11.26 \text{ (SE)} \tag{5.23}$$

Il parametro r è significativamente più alto rispetto al gruppo amministrativo ($r_{\text{clin}} = 41.57$ vs $r_{\text{amm}} = 6.79$), indicando una minore sovradispersione nei dati clinici nonostante la media più elevata. In altri termini:

- La variabilità relativa giorno-per-giorno è più contenuta nel gruppo clinico ($\phi = 12.05$ rispetto $\phi = 33.54$)
- La distribuzione è più "concentrata" attorno alla media in termini relativi
- Le code sono meno pesanti rispetto al gruppo amministrativo (rapporto $r_{\text{clin}}/r_{\text{amm}} = 6.12$)

Per le analisi successive si adotta la stima MLE:

$$X_{\text{clin}} \sim \text{NB}(\mu = 459.39, r = 41.57) \tag{5.24}$$

5.4 Risultati: Metriche di Rischio

5.4.1 Value at Risk per Livelli di Confidenza Standard

La Tabella 5.1 riporta i valori di VaR calcolati per entrambi i gruppi a diversi livelli di confidenza.

Tabella 5.1: Value at Risk per Gruppi Clinico e Amministrativo

Confidenza α	**Gruppo Amministrativo**		**Gruppo Clinico**	
	VaR	Giorni/anno*	VaR	Giorni/anno*
90%	591	25.0	1355	25.0
95%	685	12.5	1461	12.5
99%	886	2.5	1672	2.5
99.5%	968	1.25	1754	1.25
99.9%	1151	0.25	1932	0.25

* Giorni attesi con superamento (su 250 giorni lavorativi/anno)

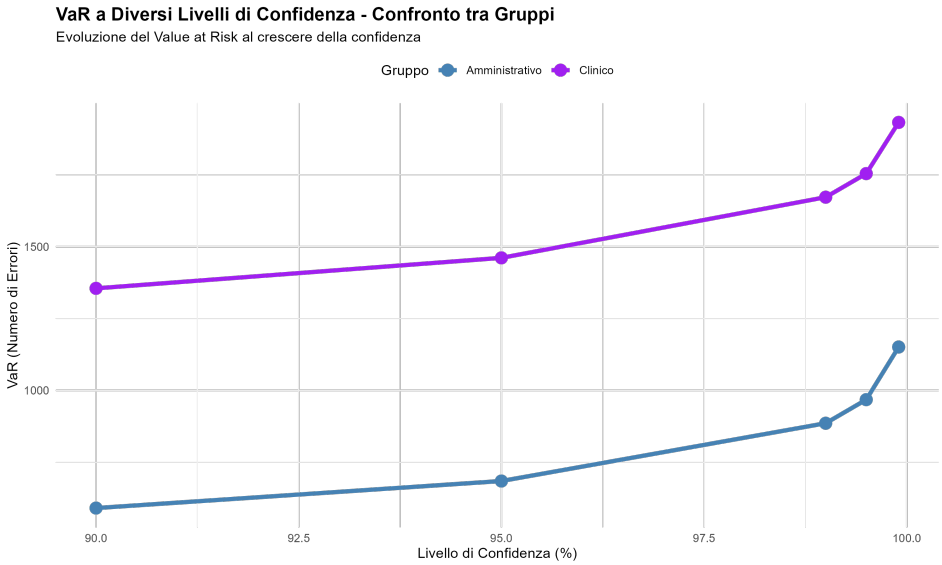

Figura 5.1: Confronto dei valori di VaR per diversi livelli di confidenza tra gruppo amministrativo e clinico.

Osservazioni Chiave:

1. **Inversione del Rapporto:** Contrariamente alle aspettative iniziali basate sulle medie, il gruppo clinico presenta VaR sistematicamente superiore rispetto al gruppo amministrativo a tutti i livelli (Figura 5.1). Per $\alpha = 0.95$:

$$\text{VaR}_{0.95}^{\text{clin}} - \text{VaR}_{0.95}^{\text{amm}} = 1461 - 685 = 776 \text{ errori} \quad (+113\%)$$

   Questo riflette il fatto che, nonostante la media clinica sia più del doppio di quella amministrativa (459.39 rispetto a 220.93), anche il VaR cresce proporzionalmente.

2. **Crescita Relativa e Code Pesanti:** Nonostante i valori assoluti del gruppo clinico siano sempre superiori, il VaR amministrativo cresce *percentualmente* più rapidamente all'aumentare del livello di confidenza:

   - Da 90% a 95%: Amministrativo +15.9% rispetto Clinico +7.8%
   - Da 95% a 99%: Amministrativo +29.3% rispetto Clinico +14.4%
   - Da 99% a 99.5%: Amministrativo +9.3% rispetto Clinico +4.9%

   Questa crescita percentuale più pronunciata conferma la presenza di code più pesanti nel gruppo amministrativo ($r = 6.79$ rispetto $r = 41.57$), coerente con la maggiore sovradispersione ($\phi_{amm} = 33.54$ e $\phi_{clin} = 12.05$).

3. **Crescita Proporzionale:** La differenza assoluta tra i due gruppi aumenta con il livello di confidenza:

   - $\alpha = 0.90$: $\Delta = +764$ errori $(1355 - 591)$
   - $\alpha = 0.95$: $\Delta = +776$ errori $(1461 - 685)$
   - $\alpha = 0.99$: $\Delta = +786$ errori $(1672 - 886)$

   La differenza relativa rimane abbastanza stabile ($\sim$110-130%), indicando che entrambe le distribuzioni hanno code con peso comparabile in termini relativi.

4. **Gap rispetto alla Media:** Per entrambi i gruppi, il VaR al 95% supera significativamente la media:

$$\begin{aligned} \text{Gruppo Amm:} \quad & \text{VaR}_{0.95}/\mu = 685/220.93 = 3.10 \quad (+210\%) \\ \text{Gruppo Clin:} \quad & \text{VaR}_{0.95}/\mu = 1461/459.39 = 3.18 \quad (+218\%) \end{aligned}$$

   Questo conferma che dimensionare le risorse per la prevenzione di errori sulla media comporterebbe saturazioni sistematiche e frequenti. È necessario un margine di sicurezza di almeno il 200-220% rispetto alla media per garantire capacità adeguata.

5.4.2 Confronto tra Dati Osservati e VaR: Analisi della Serie Temporale

Per valutare l'adeguatezza delle soglie di VaR calcolate, è fondamentale confrontarle con i dati effettivamente osservati nel periodo di analisi (46 giorni lavorativi). La Figura 5.2 mostra l'evoluzione temporale degli errori giornalieri per entrambi i gruppi, con sovrapposizione delle soglie di VaR al 95%.

Verifica Empirica del VaR al 95%: L'analisi dei dati osservati rivela che:

- **Gruppo Amministrativo:** Tutti i 46 valori osservati si collocano *al di sotto* della soglia $\text{VaR}_{0.95} = 685$ errori/giorno. Il valore massimo registrato è stato di circa 650 errori (giorno 40), corrispondente al 94.9% del VaR teorico. Questo indica che, nel periodo analizzato, la soglia di VaR non è mai stata superata.
- **Gruppo Clinico:** Analogamente, tutti i valori osservati rimangono al di sotto della soglia $\text{VaR}_{0.95} = 1461$ errori/giorno. Il picco massimo osservato è di circa 580 errori, pari solo al 39.7% del VaR teorico, confermando una notevole distanza di sicurezza.

Questa osservazione è coerente con i risultati della Sezione 6.5, Sottosezione Validazione del Modello, che ha registrato 0 violazioni del VaR al 95% per entrambi i gruppi, contro le 2.3 violazioni attese teoricamente ($5\% \times 46$ giorni).

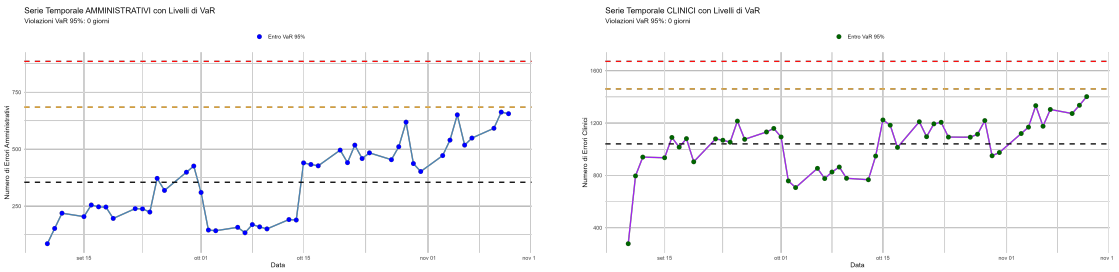

Figura 5.2: Serie temporali degli errori giornalieri osservati nel periodo settembre-novembre 2025 con sovrapposizione delle soglie di VaR. A sinistra: gruppo amministrativo; a destra: gruppo clinico. La linea blu tratteggiata indica la media complessiva del periodo ($\mu_{amm} = 220.93$, $\mu_{clin} = 459.39$), la linea gialla orizzontale rappresenta il VaR al 95% ($\text{VaR}_{0.95}^{amm} = 685$, $\text{VaR}_{0.95}^{clin} = 1461$), la linea rossa orizzontale rappresenta il VaR al 99% ($\text{VaR}_{0.99}^{amm} = 886$, $\text{VaR}_{0.99}^{clin} = 1672$). Si osserva che tutti i valori osservati si collocano al di sotto della soglia di VaR al 95% per entrambi i gruppi.

Interpretazione: Il fatto che nessun valore osservato superi il VaR al 95% può essere interpretato in due modi:

1. **Ipotesi Conservativa (Scenario Stabile):** Il modello sovrastima il rischio, fornendo un margine di sicurezza adeguato. In questo caso, le soglie di VaR calcolate sono affidabili per il dimensionamento delle risorse operative.

2. **Ipotesi Evolutiva (Monitoraggio Futuro):** Sebbene il periodo analizzato (post-10 settembre 2025) sia stato selezionato come stazionario sulla base dell'analisi di stagionalità (Capitolo 2), il campione di 46 giorni lavorativi rimane limitato. Per il **gruppo amministrativo**, l'elevata sovradispersione ($r = 6.79$) e la presenza di un possibile andamento crescente nella seconda metà del periodo osservato suggeriscono la necessità di raccogliere ulteriori dati per confermare la stabilità dei parametri stimati a lungo termine.

Criticità del Gruppo Amministrativo: Sebbene i grafici delle serie temporali (Figure 2.2) non mostrino un trend lineare netto e marcato, si osserva una variabilità elevata con alcuni episodi di picco concentrati nella parte finale del periodo. Data l'incertezza sull'interpretazione corretta, si raccomanda all'azienda:

- **Monitoraggio Continuo:** Raccogliere ulteriori 30-60 giorni di dati per verificare se emerge un trend sistematico o se l'andamento si stabilizza attorno ai parametri attuali.

- **Re-stima Periodica:** Ricalcolare VaR e CVaR per catturare eventuali cambiamenti nei parametri μ e r.

Affidabilità delle Soglie per il Gruppo Clinico: Analisi della Serie Temporale. Per il gruppo clinico, la serie temporale mostra maggiore stabilità e assenza di pattern direzionali evidenti. La distanza ampia tra i valori osservati (max ~ 580 errori) e il VaR al 95% (1461 errori) suggerisce che:

- Le soglie calcolate sono altamente conservative e forniscono un margine di sicurezza molto ampio
- Non vi sono evidenze di non-stazionarietà o trend nel periodo osservato
- I parametri stimati ($\mu = 459.39$, $r = 41.57$) sono ragionevolmente stabili e utilizzabili per pianificazioni operative a medio termine.

5.4.3 Conditional Value at Risk e Rischio Residuo

La Tabella 5.2 confronta VaR, CVaR e il rischio residuo $\Delta = \text{CVaR} - \text{VaR}$ per entrambi i gruppi. Le Figure 5.3, 5.4 e 5.5 ne illustrano graficamente le dinamiche.

Tabella 5.2: Confronto VaR, CVaR e Rischio Residuo

	Gruppo Amministrativo			**Gruppo Clinico**		
α	VaR	CVaR	Δ	VaR	CVaR	Δ
90%	591	720	129	1355	1494	139
95%	685	809	124	1461	1582	121
99%	886	1001	115	1672	1784	112
99.5%	968	1077	109	1754	1863	109
99.9%	1151	1249	98	1932	2020	88

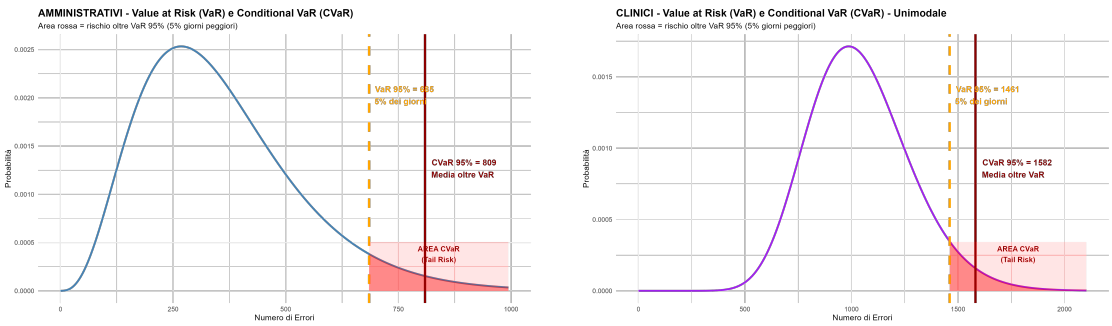

Figura 5.3: Confronto tra VaR e CVaR per gruppo amministrativo (sinistra) e clinico (destra). L'area tra le due curve rappresenta il rischio residuo, più ampio per il gruppo amministrativo.

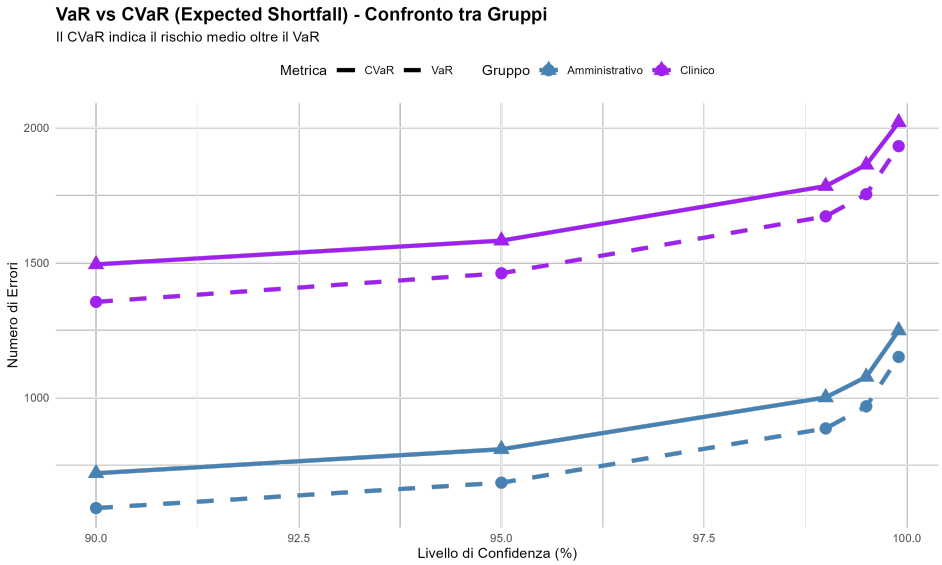

Figura 5.4: Confronto diretto tra VaR e CVaR per entrambi i gruppi. Il CVaR è sistematicamente superiore al VaR, con gap assoluti comparabili tra i due gruppi (90-140 errori). In termini percentuali, tuttavia, il gap è maggiore per il gruppo amministrativo (+18-22%) rispetto al clinico (+5-10%), riflettendo la maggiore pesantezza delle code della distribuzione amministrativa.

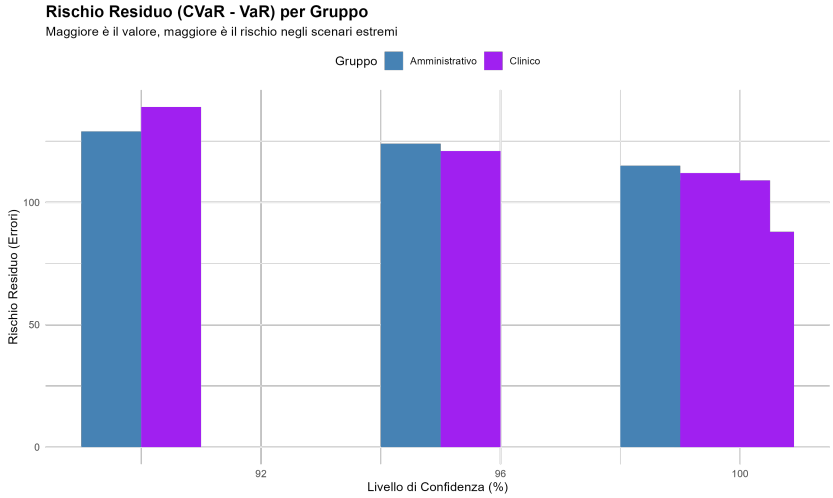

Figura 5.5: Rischio residuo ($\Delta = \text{CVaR} - \text{VaR}$) per diversi livelli di confidenza.

Analisi del Rischio Residuo:

- **Gruppo Amministrativo:** Il rischio residuo è sempre positivo e molto significativo a tutti i livelli di confidenza, con valori di $\Delta_{0.90} = 129$ errori e $\Delta_{0.95} = 124$ errori. Questo indica che:
  - Quando il VaR viene superato, il carico medio effettivo è 809 errori, ovvero 124 in più rispetto alla soglia di 685 (+18.1%)
  - La coda destra della distribuzione è estremamente pesante ($r = 6.79$ molto basso)
  - È necessario un margine di sicurezza aggiuntivo del 18-22% oltre il VaR per gestire gli scenari estremi
  - Il rischio residuo decresce gradualmente con α (da 129 a 98 errori), indicando una coda che si assottiglia lentamente anche a percentili estremi ($\Delta_{0.999} = 98$ errori, +8.5%)
- **Gruppo Clinico:** Il rischio residuo è positivo e ben definito a tutti i livelli di confidenza, con valori assoluti comparabili o superiori al gruppo amministrativo ($\Delta_{0.90} = 139$ errori, $\Delta_{0.95} = 121$ errori). Questo riflette:
  - Una distribuzione con code moderate ($r = 41.57$ più elevato rispetto al gruppo amministrativo) ma comunque significative in valore assoluto
  - Quando il VaR viene superato, il carico medio effettivo è 1582 errori, ovvero 121 in più rispetto alla soglia di 1461 (+8.3
  - Il CVaR rimane stabile e superiore al VaR anche ai quantili estremi: $\Delta_{0.99} = 112$ errori (+6.7%), $\Delta_{0.995} = 109$ errori (+6.2%), $\Delta_{0.999} = 88$ errori (+4.6%)
  - Il rischio residuo decresce in termini sia assoluti che percentuali ai quantili più alti, indicando code meno pesanti rispetto al gruppo amministrativo
  - Nonostante la media più elevata, la variabilità relativa è più controllata e prevedibile
- **Confronto:** Entrambi i gruppi richiedono margini di sicurezza significativi (90-140 errori in valore assoluto). In termini percentuali, il gruppo amministrativo richiede un margine sistematicamente maggiore (+18.1% contro +8.3% al 95%), riflettendo la maggiore sovradispersione ($\phi_{amm} = 33.54$ contro $\phi_{clin} = 12.05$). Il gruppo clinico, pur con media doppia e VaR più elevati in valore assoluto, presenta una struttura di rischio più "prevedibile" con un rischio residuo percentuale che si riduce più rapidamente ai quantili estremi (dal +10.3% al 90% al +4.6% al 99.9% per il clinico, mentre dal +21.8% al +8.5% per l'amministrativo).

5.4.4 Validazione del Modello

Per verificare l'adeguatezza dei modelli stimati, si confrontano i quantili teorici (VaR) con i percentili empirici osservati nei 46 giorni di dati (Tabella 5.3).

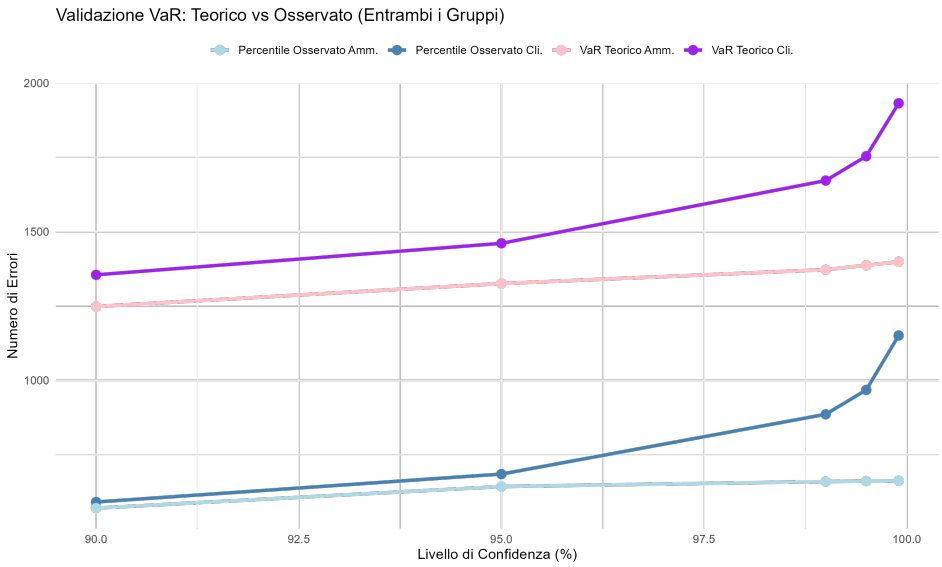

Figura 5.6: Validazione del modello: confronto tra VaR teorico (linee) e percentili osservati (punti) per entrambi i gruppi. La sovrapposizione indica un buon adattamento del modello ai dati empirici.

Tabella 5.3: VaR Teorico a confronto con i Percentili Osservati

	Gruppo Amministrativo			**Gruppo Clinico**		
α	VaR_{teo}	Perc_{oss}	Diff	VaR_{teo}	Perc_{oss}	Diff
90%	591	570.5	-20.5	1355	1248.5	-106.5
95%	685	643.0	-42.0	1461	1325.8	-135.2
99%	886	659.9	-226.1	1672	1372.3	-299.7

Test delle Violazioni: Si contano i giorni in cui il valore osservato ha superato il VaR al 95%:

- **Gruppo Amministrativo:** 0 violazioni su 46 giorni (0.0%). Atteso: 2.3 giorni (5%)
- **Gruppo Clinico:** 0 violazioni su 46 giorni (0.0%). Atteso: 2.3 giorni (5%)

Entrambi i modelli mostrano un tasso di violazioni inferiore alle attese, suggerendo un approccio conservativo (sovrastima del rischio), coerente con la Tabella 5.3.

Interpretazione: I percentili osservati sono sistematicamente inferiori ai VaR teorici per entrambi i gruppi. Questo è normale con un campione limitato di 46 giorni: per stimare il percentile al 99%, servirebbero almeno 100 osservazioni per avere un dato affidabile, mentre qui ci si aspetta solo $46 \times 0.01 = 0.46$ osservazioni oltre quella soglia.

Le differenze più grandi si osservano ai quantili estremi:

- Gruppo Amm al 99%: -226 errori (-25.5%)
- Gruppo Clin al 99%: -300 errori (-17.9%)

Con sole 46 osservazioni, i percentili empirici non sono affidabili per validare il modello. Il VaR teorico, stimato da tutta la struttura della distribuzione (non solo dai valori estremi osservati), rimane preferibile per le pianificazioni operative.

5.5 Riepilogo

Questo capitolo ha introdotto e applicato le metriche di VaR e CVaR come strumenti di quantificazione del rischio operazionale nel contesto del processo di correzione errori. I risultati principali sono:

1. **Validità del Modello Binomiale Negativa:** La validazione ha confermato che il modello NB cattura la struttura generale della distribuzione degli errori, sebbene con atteggiamento conservativo (sovrastima del rischio). Il tasso di violazioni del VaR al 95% è pari a 0% per entrambi i gruppi (rispetto al 5% atteso), indicando che il modello fornisce margini di sicurezza adeguati.

2. **Quantificazione del Rischio di Coda:**

   - Gruppo Amm: $\text{VaR}_{0.95} = 685$, $\text{CVaR}_{0.95} = 809$ ($\Delta = +124$, rischio residuo +18.1%)
   - Gruppo Clin: $\text{VaR}_{0.95} = 1461$, $\text{CVaR}_{0.95} = 1582$ ($\Delta = +121$, rischio residuo +8.3%)

   Entrambi i gruppi richiedono dei margini di sicurezza significativi oltre la media (rispettivamente +210% e +218% per raggiungere il $\text{VaR}_{0.95}$). Il gruppo clinico presenta valori assoluti più elevati ma maggiore prevedibilità relativa.

3. **Differenze Strutturali tra Gruppi:** Il gruppo amministrativo ($r = 6.79$) mostra sovradispersione estremamente elevata ($\phi = 33.54$) rispetto al clinico ($r = 41.57$, $\phi = 12.05$), indicando code più pesanti e maggiore imprevedibilità nei picchi di carico. Il gruppo clinico, pur con media doppia, presenta una struttura di rischio più "gestibile" in termini relativi. Questo richiede strategie di dimensionamento differenziate.

4. **Criticità del Trend Amministrativo:** La presenza di un trend crescente non stabilizzato invalida l'ipotesi di stazionarietà, rendendo i valori statici di VaR/CVaR inadeguati per pianificazioni a lungo termine. Si raccomanda ri-stima frequente.

Capitolo 6

Modello a Shock Latenti per Distribuzioni di Bernoulli Multivariate

6.1 Introduzione

Nei capitoli precedenti è stato descritto e classificato gli errori di estrazione dividendoli in due gruppi: il gruppo **Clinico** (Patologia, Numero prescrizione, Data e Prestazione) e quello **Amministrativo** (Data, Prestazione, Numero fattura, Importo totale e Importo prestazione). Questo ha permesso di calcolare i tassi di errore aggregati e capire dove il sistema AI ha più difficoltà.

Però per costruire strategie predittive che funzionino davvero bisogna capire come si distribuiscono le singole variabili e se ci sono dipendenze tra loro.
La domanda chiave è: se un campo ha un errore, che probabilità c'è che anche un altro campo ne abbia uno? E ci sono cause latenti comuni che fanno comparire errori insieme su campi diversi?

Per rispondere serve un modello probabilistico capace di:

- descrivere la **distribuzione congiunta** delle variabili binarie (errore sì/no) per ogni campo;
- stimare le **probabilità condizionate**, cioè la probabilità che si verifichi un errore sapendo che se ne sono verificati altri;
- **prevedere** quali configurazioni di errori sono più probabili.

In questo capitolo viene presentato un **Modello a Shock Latenti** che si basa sulla teoria delle distribuzioni Bernoulliane multivariate (chiamate anche *Multinulli*

in letteratura) [17].
L'idea di fondo è intuitiva: gli errori che vengono osservati sui singoli campi non sono eventi indipendenti ma vengono generati da shock latenti (variabili nascoste che possono colpire uno o più campi simultaneamente). Stimando i parametri di questi shock posso identificare le cause comuni degli errori correlati, calcolare la probabilità marginale per ciascun campo, prevedere la probabilità di un errore specifico conoscendo lo stato degli altri campi, e distinguere tipologie di errori tra sistematici (legati a shock ricorrenti) ed errori occasionali.

La metodologia che propongo estende la teoria classica del caso bivariato [18] ai casi con quattro e cinque variabili, così da analizzare tutte le interazioni possibili tra i campi dei due gruppi.

6.2 Fondamenti Teorici

6.2.1 Distribuzioni di Bernoulli Multivariate

Si considera un vettore aleatorio $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ dove d è la dimensione e dove ogni componente X_i può valere solo 0 o 1. La distribuzione congiunta di questo vettore è definita da 2^d probabilità:

$$p_{\mathbf{k}} = P(X_1 = k_1, X_2 = k_2, \ldots, X_d = k_d), \quad \mathbf{k} \in \{0,1\}^d \tag{6.1}$$

con il vincolo che $\sum_{\mathbf{k}} p_{\mathbf{k}} = 1$.

Nel caso specifico dell'analisi degli errori di estrazione da parte dell'AI:

$$X_i = \begin{cases} 1 & \text{se il campo } i \text{ presenta un errore (vuoto/mancante)} \\ 0 & \text{se il campo } i \text{ è correttamente estratto} \end{cases} \tag{6.2}$$

La distribuzione di Bernoulli multivariata generale ha bisogno di $2^d - 1$ parametri da stimar, quindi per d che cresce il numero esplode. Per esempio con $d = 5$ si ha già 31 parametri.
Il modello a shock latenti risolve questo problema introducendo variabili nascoste che parametrizzano il tutto in modo molto più efficiente.

6.2.2 Il Modello a Shock Latenti: Caso Bivariato

Per introdurre successivamente il funzionamento del modello che viene applicato al dataset, in questa sezione viene descritto il caso più semplice: due sole variabili X_1 e X_2 (per esempio gli errori su "Data" e "Prestazione").

Costruzione degli Shock Latenti

L'assunzione fondamentale è che esistano **tre shock latenti indipendenti**:

- $Y_1 \sim \text{Bernoulli}(\theta_1)$: colpisce solo il campo 1
- $Y_2 \sim \text{Bernoulli}(\theta_2)$: colpisce solo il campo 2
- $Z_{12} \sim \text{Bernoulli}(\theta_{12})$: colpisce entrambi

Le variabili che osservo vengono generate così:

$$X_1 = \max(Y_1, Z_{12}) = \mathbb{1}[Y_1 = 1 \text{ oppure } Z_{12} = 1] \tag{6.3}$$
$$X_2 = \max(Y_2, Z_{12}) = \mathbb{1}[Y_2 = 1 \text{ oppure } Z_{12} = 1] \tag{6.4}$$

Cioè X_1 ha un errore se capita Y_1 (shock che colpisce solo lui) oppure Z_{12} (shock comune). Stesso discorso per X_2.

Derivazione delle Probabilità Congiunte

Le quattro probabilità possibili per (X_1, X_2) vengono calcolate nel seguente modo:

1. Nessun errore: $P(X_1 = 0, X_2 = 0)$

Perché entrambi i campi siano corretti serve che tutti gli shock siano inattivi:

$$P(X_1 = 0, X_2 = 0) = P(Y_1 = 0, Y_2 = 0, Z_{12} = 0) = (1-\theta_1)(1-\theta_2)(1-\theta_{12}) \tag{6.5}$$

2. Solo X_1 ha errore: $P(X_1 = 1, X_2 = 0)$

Questa configurazione richiede che $Y_1 = 1$ (l'errore sul campo 1), ma $Y_2 = 0$ e soprattutto $Z_{12} = 0$. Se Z_{12} fosse attivo colpirebbe anche X_2. Quindi:

$$P(X_1 = 1, X_2 = 0) = P(Y_1 = 1, Y_2 = 0, Z_{12} = 0) = \theta_1(1-\theta_2)(1-\theta_{12}) \tag{6.6}$$

3. Solo X_2 ha errore: $P(X_1 = 0, X_2 = 1)$

Per simmetria ho:

$$P(X_1 = 0, X_2 = 1) = P(Y_1 = 0, Y_2 = 1, Z_{12} = 0) = (1-\theta_1)\theta_2(1-\theta_{12}) \tag{6.7}$$

4. Entrambi i campi hanno errore: $P(X_1 = 1, X_2 = 1)$

Entrambi i campi hanno errore se:

- $Z_{12} = 1$ (shock comune attivo), oppure
- $Y_1 = 1$ *e* $Y_2 = 1$ (entrambi gli shock individuali attivi)

Usando il complemento:

$$P(X_1 = 1, X_2 = 1) = 1 - P(X_1 = 0 \text{ o } X_2 = 0) \tag{6.8}$$

Applicando la formula di inclusione-esclusione:

$$\begin{aligned} P(X_1 = 1, X_2 = 1) &= 1 - [P(X_1 = 0, X_2 = 0) + P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1)] \\ &= 1 - (1 - \theta_1)(1 - \theta_2)(1 - \theta_{12}) - \theta_1(1 - \theta_2)(1 - \theta_{12}) \\ &\quad - (1 - \theta_1)\theta_2(1 - \theta_{12}) \\ &= 1 - (1 - \theta_{12})[(1 - \theta_1)(1 - \theta_2) + \theta_1(1 - \theta_2) + (1 - \theta_1)\theta_2] \\ &= 1 - (1 - \theta_{12})(1 - \theta_1\theta_2) \\ &= \theta_{12} + \theta_1\theta_2 - \theta_{12}\theta_1\theta_2 \end{aligned} \tag{6.9}$$

Probabilità Marginali

Le probabilità marginali si ottengono sommando sulle configurazioni:

$$\begin{aligned} P(X_1 = 1) &= P(X_1 = 1, X_2 = 0) + P(X_1 = 1, X_2 = 1) \\ &= \theta_1(1 - \theta_2)(1 - \theta_{12}) + \theta_{12} + \theta_1\theta_2 - \theta_{12}\theta_1\theta_2 \\ &= \theta_1 + \theta_{12} - \theta_1\theta_{12} \end{aligned} \tag{6.10}$$

Equivalentemente:

$$P(X_1 = 1) = 1 - P(X_1 = 0) = 1 - (1 - \theta_1)(1 - \theta_{12}) \tag{6.11}$$

Per simmetria:

$$P(X_2 = 1) = 1 - (1 - \theta_2)(1 - \theta_{12}) \tag{6.12}$$

Costruzione del Sistema di Equazioni

Adesso è necessario applicare questo esempio di modello sui dati supponendo di osservare n documenti (nel caso della presente tesi sono 215.850 tra fatture e prescrizioni). Per ogni documento biosgna controllare se X_1 e X_2 presentano errori oppure no e bisogna contare quante volte si presenta ogni tipologia di combinazione.

Le frequenze empiriche che calcolo sono:

$$\hat{p}_{00} = \frac{\text{numero doc con entrambi i campi ok}}{n} \tag{6.13}$$

$$\hat{p}_{10} = \frac{\text{numero doc con solo } X_1 \text{ sbagliato}}{n} \tag{6.14}$$

$$\hat{p}_{01} = \frac{\text{numero doc con solo } X_2 \text{ sbagliato}}{n} \tag{6.15}$$

$$\hat{p}_{11} = \frac{\text{numero doc con entrambi sbagliati}}{n} \tag{6.16}$$

O in notazione più compatta:

$$\hat{p}_{00} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_1^{(j)} = 0, X_2^{(j)} = 0] \tag{6.17}$$

$$\hat{p}_{10} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_1^{(j)} = 1, X_2^{(j)} = 0] \tag{6.18}$$

$$\hat{p}_{01} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_1^{(j)} = 0, X_2^{(j)} = 1] \tag{6.19}$$

$$\hat{p}_{11} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_1^{(j)} = 1, X_2^{(j)} = 1] \tag{6.20}$$

A questo punto, considerando le quattro probabilità teoriche che sono state derivatate precedentemente e le quattro frequenze empiriche, si ottiene un sistema:

$$\begin{cases} (1-\theta_1)(1-\theta_2)(1-\theta_{12}) = \hat{p}_{00} \\ \theta_1(1-\theta_2)(1-\theta_{12}) = \hat{p}_{10} \\ (1-\theta_1)\theta_2(1-\theta_{12}) = \hat{p}_{01} \\ \theta_{12} + \theta_1\theta_2 - \theta_{12}\theta_1\theta_2 = \hat{p}_{11} \end{cases} \tag{6.21}$$

Si può notare che ci sono 4 equazioni ma solo 3 parametri incogniti (θ_1, θ_2, θ_{12}). Il sistema è quindi sovradeterminato. In un caso teorico perfetto con dati a loro volta perfetti si potrebbe trovare una soluzione, ma in pratica le frequenze empiriche $\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11}$ hanno del rumore statistico. Quindi non c'è un set di valori $(\theta_1, \theta_2, \theta_{12})$ che soddisfa perfettamente tutte e quattro le equazioni contemporaneamente.

L'idea è di utilizzare l'ottimizzazione numerica per la risoluzione del sistema, ovvero cercare i valori di θ_1, θ_2, θ_{12} che minimizzano la differenza tra il lato sinistro e destro delle equazioni (trovare il "miglior compromesso" possibile).
Nota: In alternativa si potrebbero usare solo le equazioni corrispondenti ai sottoinsiemi non vuoti $\{1\}, \{2\}, \{1,2\}$, che danno 3 equazioni per le probabilità di "nessun errore": $P(X_1 = 0)$, $P(X_2 = 0)$, $P(X_1 = 0, X_2 = 0)$. In quel caso si avrebbe un sistema determinato (3 equazioni, 3 incognite). Tuttavia, usare tutte e 4 le configurazioni fornisce più vincoli e può portare a stime più robuste quando c'è rumore nei dati.

6.2.3 Il Modello a Shock Latenti: Caso Generale

Estendiamo il modello al caso con d variabili qualsiasi.

Definizione degli Shock Latenti

L'idea [17] è che esistono variabili latenti (non osservabili) Y_S che rappresentano shock capaci di colpire simultaneamente un sottoinsieme S di campi.

Per ogni sottoinsieme non vuoto $S \subseteq \{1, 2, \ldots, d\}$ c'è uno **shock latente** Y_S che è una Bernoulli:

$$Y_S \sim \text{Bernoulli}(\theta_S) \tag{6.22}$$

dove $\theta_S = P(Y_S = 1)$ è la probabilità che lo shock S capiti.
Nota importante: Vengono considerati solo sottoinsiemi non vuoti perché uno shock associato all'insieme vuoto $\emptyset$ non avrebbe significato: rappresenterebbe un evento che non influenza nessun campo, quindi è irrilevante per il modello. Per questo motivo il numero totale di shock è $2^d - 1$ (tutti i possibili sottoinsiemi meno l'insieme vuoto).

Meccanismo di Generazione

Il legame tra le X_i osservate e gli shock è:

$$X_i = \mathbb{1}\left[\bigcup_{S:i\in S} \{Y_S = 1\}\right] = \max_{S:i\in S} Y_S \tag{6.23}$$

Cioè il campo i ha un errore se almeno uno degli shock che lo colpiscono si è attivato.

Assunzione chiave: Gli shock $\{Y_S\}$ sono mutuamente indipendenti:

$$P(Y_{S_1} = y_1, \ldots, Y_{S_m} = y_m) = \prod_{j=1}^{m} P(Y_{S_j} = y_j) = \prod_{j=1}^{m} \theta_{S_j}^{y_j}(1 - \theta_{S_j})^{1-y_j} \tag{6.24}$$

Derivazione delle Probabilità Congiunte

Prendo una configurazione generica $\mathbf{x} = (x_1, \ldots, x_d) \in \{0,1\}^d$.

Teorema: La probabilità che tutti i campi in un sottoinsieme $S \subseteq \{1, \ldots, d\}$ sia completamente privo di errori è:

$$P\left(\bigcap_{i\in S} \{X_i = 0\}\right) = \prod_{T:T\cap S\neq\emptyset} (1 - \theta_T) \tag{6.25}$$

Dimostrazione:

Affinchè tutti i campi in S siano corretti (tutti gli X_i con $i \in S$ abbiano valore 0), bisogna che nessuno degli shock che colpiscono almeno un campo di S si attivi. Formalmente:

$$\bigcap_{i\in S} \{X_i = 0\} = \bigcap_{i\in S} \bigcap_{T:i\in T} \{Y_T = 0\} = \bigcap_{T:T\cap S\neq\emptyset} \{Y_T = 0\} \tag{6.26}$$

Poiché gli shock sono indipendenti:

$$P\left(\bigcap_{T:T\cap S\neq\emptyset}\{Y_T=0\}\right)=\prod_{T:T\cap S\neq\emptyset}P(Y_T=0)=\prod_{T:T\cap S\neq\emptyset}(1-\theta_T) \tag{6.27}$$

Costruzione del Sistema di Equazioni

Data una matrice di osservazioni $\mathbf{M}\in\{0,1\}^{n\times d}$ con n documenti e d campi, vengono calcolate le frequenze empiriche.

Per ogni sottoinsieme $S\subseteq\{1,\ldots,d\}$ si definisce:

$$\hat{P}_S=\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\left[\bigcap_{i\in S}\{M_{ji}=0\}\right] \tag{6.28}$$

che è semplicemente la frazione di documenti dove tutti i campi in S sono corretti.

Il sistema teorico è:

$$\prod_{T:T\cap S\neq\emptyset}(1-\theta_T)=\hat{P}_S\quad\forall S\subseteq\{1,\ldots,d\} \tag{6.29}$$

Contando risultano:

- Sottoinsiemi non vuoti S: 2^d-1 (l'insieme vuoto $\emptyset$ non viene considerato, poiché uno shock che non influenza alcun campo non ha significato nel modello)
- Equazioni (una per ogni S): 2^d-1
- Shock latenti T (incognite θ_T): 2^d-1

Il sistema risulta quindi determinato, con lo stesso numero di equazioni e incognite. In linea teorica potrebbe ammettere una soluzione unica. Tuttavia, nella pratica emergono diverse difficoltà:

1. Le equazioni sono non lineari (prodotti multipli di termini della forma $(1-\theta_T)$), il che rende impossibile una risoluzione analitica
2. Le frequenze empiriche $\hat{P}_S$ sono affette da rumore statistico, inevitabile in qualsiasi campione finito
3. A causa del rumore nei dati, alcune equazioni possono risultare quasi ridondanti
4. Ammissibilità delle soluzioni: una risoluzione algebrica esatta forzata su dati rumorosi potrebbe restituire parametri $\theta_T\notin[0,1]$, violando gli assiomi della probabilità.

Per questi motivi, la strategia adottata consiste nell'utilizzare tecniche di ottimizzazione numerica. L'approccio seguito non tenta di risolvere il sistema in modo esatto (il che sarebbe di fatto impossibile), ma cerca i valori dei parametri θ_T che minimizzano la discrepanza complessiva tra le probabilità teoriche previste dal modello e quelle osservate empiricamente. In sostanza, il problema viene riformulato come un problema di minimi quadrati pesati.

6.2.4 Applicazione al Gruppo Clinico ($d = 4$)

Con $d = 4$ variabili (nel gruppo Clinico: Patologia, Numero prescrizione, Data, Prestazione), ho $2^4 - 1 = 15$ shock latenti:

Tabella 6.1: Struttura degli shock latenti per $d = 4$ variabili

Ordine	**Simbolo**	**Campi Influenzati**	**Numero**
1	Y_1, Y_2, Y_3, Y_4	Singoli campi	$\binom{4}{1} = 4$
2	$Z_{12}, Z_{13}, Z_{14}, Z_{23}, Z_{24}, Z_{34}$	Coppie di campi	$\binom{4}{2} = 6$
3	$W_{123}, W_{124}, W_{134}, W_{234}$	Triple di campi	$\binom{4}{3} = 4$
4	V_{1234}	Tutti i campi (shock globale)	$\binom{4}{4} = 1$
		Totale	**15**

Le equazioni base per le marginali con $d = 4$ sono per esempio:

$$\begin{aligned} P(X_1 = 1) = 1 &- (1 - \theta_{Y_1})(1 - \theta_{Z_{12}})(1 - \theta_{Z_{13}})(1 - \theta_{Z_{14}}) \\ &\cdot (1 - \theta_{W_{123}})(1 - \theta_{W_{124}})(1 - \theta_{W_{134}})(1 - \theta_{V_{1234}}) \end{aligned} \tag{6.30}$$

E per la probabilità congiunta che i campi 1 e 2 siano entrambi corretti:

$$\begin{aligned} P(X_1 = 0, X_2 = 0) &= \prod_{T: T \cap \{1,2\} \neq \emptyset} (1 - \theta_T) \\ &= (1 - \theta_{Y_1})(1 - \theta_{Y_2})(1 - \theta_{Z_{12}})(1 - \theta_{Z_{13}})(1 - \theta_{Z_{14}}) \\ &\quad \cdot (1 - \theta_{Z_{23}})(1 - \theta_{Z_{24}})(1 - \theta_{W_{123}})(1 - \theta_{W_{124}}) \\ &\quad \cdot (1 - \theta_{W_{134}})(1 - \theta_{W_{234}})(1 - \theta_{V_{1234}}) \end{aligned} \tag{6.31}$$

6.2.5 Applicazione al Gruppo Amministrativo ($d = 5$)

Per $d = 5$ variabili (gruppo Amministrativo), il numero di shock latenti aumenta a $2^5 - 1 = 31$.

Tabella 6.2: Struttura degli shock latenti per $d = 5$ variabili

Ordine	**Simbolo**	**Descrizione**	**Numero**
1	Y_i	Shock individuali	$\binom{5}{1} = 5$
2	Z_{ij}	Shock su coppie	$\binom{5}{2} = 10$
3	W_{ijk}	Shock su triple	$\binom{5}{3} = 10$
4	V_{ijkl}	Shock su quadruple	$\binom{5}{4} = 5$
5	U_{12345}	Shock globale	$\binom{5}{5} = 1$
		Totale	**31**

La complessità del sistema aumenta esponenzialmente, l'approccio di ottimizzazione numerica diventa quindi essenziale per gestire sistemi di questa dimensione.

6.3 Ottimizzazione Numerica: L-BFGS-B

La stima dei parametri $\boldsymbol{\theta}$ richiede la risoluzione di un problema di ottimizzazione vincolata non lineare. Come discusso nella sezione precedente, il sistema (6.29) risulta esattamente determinato ($2^d - 1$ equazioni per $2^d - 1$ incognite), ma la forte non linearità delle equazioni e la presenza di rumore statistico nei dati impediscono una risoluzione analitica. Si ricorre quindi alla minimizzazione di una funzione obiettivo mediante metodi numerici.

6.3.1 Funzione Obiettivo Pesata

Definisco la funzione obiettivo come errore quadratico pesato tra le probabilità osservate e quelle del modello:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{\mathbf{k} \in \{0,1\}^d} w_{\mathbf{k}} \left(\hat{p}_{S(\mathbf{k})} - p_{S(\mathbf{k})}(\boldsymbol{\theta}) \right)^2 \tag{6.32}$$

dove:

- $\hat{p}_{S(\mathbf{k})}$ è la frequenza empirica che è stata osservata
- $p_{S(\mathbf{k})}(\boldsymbol{\theta})$ è la probabilità teorica dal modello (eq. 6.25)
- $w_{\mathbf{k}}$ sono pesi che bilanciano quanto è importante ciascuna configurazione

I pesi sono stati definiti nel seguente modo:

$$w_{\mathbf{k}} = \begin{cases} 1000 & \text{se } |S(\mathbf{k})| = 1 \text{ (marginali singole)} \\ 100 & \text{se } |S(\mathbf{k})| = d \text{ (shock globale)} \\ 10 & \text{altrimenti (interazioni)} \end{cases} \tag{6.33}$$

I pesi alti sulle marginali singole garantiscono che il modello riproduca bene i tassi di errore per ciascun campo, che sono le statistiche più affidabili visto la presenza di tanti dati.

6.3.2 L'Algoritmo L-BFGS-B

Per ottimizzare i parametri uso **L-BFGS-B** (Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bounds) [19, 20], che è un metodo quasi-Newton capace di gestire vincoli di limitazioni sulle variabili.

Metodo di Newton e Approssimazione Quasi-Newton

Il metodo di Newton classico [21] costruisce un modello quadratico della funzione obiettivo $\mathcal{L}(\boldsymbol{\theta})$ attorno al punto corrente $\boldsymbol{\theta}^{(k)}$. Se viene denotato con $\mathbf{g}^{(k)} = \nabla\mathcal{L}(\boldsymbol{\theta}^{(k)})$ il gradiente e con $\mathbf{H}^{(k)}$ l'Hessiana, l'approssimazione quadratica è:

$$m^{(k)}(\mathbf{p}) = \mathcal{L}(\boldsymbol{\theta}^{(k)}) + (\mathbf{g}^{(k)})^T\mathbf{p} + \frac{1}{2}\mathbf{p}^T\mathbf{H}^{(k)}\mathbf{p} \tag{6.34}$$

dove $\mathbf{p}$ è la direzione in cui l'algoritmo si muove a partire dal punto corrente per trovare la soluzione. Il metodo di Newton minimizza questo modello quadratico, ottenendo la direzione di discesa risolvendo il sistema lineare:

$$\mathbf{H}^{(k)}\mathbf{p}^{(k)} = -\mathbf{g}^{(k)} \tag{6.35}$$

che porta all'aggiornamento:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \left[\mathbf{H}^{(k)}\right]^{-1}\mathbf{g}^{(k)} \tag{6.36}$$

Il metodo di Newton ha convergenza quadratica quando $\boldsymbol{\theta}^{(k)}$ è sufficientemente vicino al minimizzatore, ma presenta due problemi critici: (1) calcolare l'Hessiana $\mathbf{H}^{(k)}$ richiede $O(m^2)$ derivate seconde, dove m è il numero di parametri; (2) risolvere il sistema lineare costa $O(m^3)$ operazioni.

I **metodi quasi-Newton** risolvono questa difficoltà costruendo un'approssimazione $\mathbf{B}^{(k)} \approx \left[\mathbf{H}^{(k)}\right]^{-1}$ usando solo informazioni del gradiente. L'idea chiave, seguendo la teoria della Professoressa Pieraccini [21], è derivare un'equazione che $\mathbf{B}^{(k+1)}$ dovrebbe soddisfare. Si comsidera l'espansione di Taylor del gradiente attorno a $\boldsymbol{\theta}^{(k+1)}$:

$$\mathbf{g}^{(k)} \approx \mathbf{g}^{(k+1)} + \mathbf{H}(\boldsymbol{\theta}^{(k+1)})(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k+1)}) \tag{6.37}$$

Riorganizzando e definendo $\mathbf{s}_k = \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}$ (lo spostamento) e $\mathbf{y}_k = \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}$ (la variazione del gradiente), si ottiene l'**equazione secante**:

$$\mathbf{H}(\boldsymbol{\theta}^{(k+1)})\mathbf{s}_k = \mathbf{y}_k \tag{6.38}$$

Per l'approssimazione dell'inversa, si vuole che $\mathbf{B}^{(k+1)}\mathbf{y}_k = \mathbf{s}_k$.

Dato che infinite matrici soddisfano l'equazione secante, serviva un criterio per sceglierne una. Il primo metodo quasi-Newton fu l'**algoritmo DFP** (Davidon-Fletcher-Powell, 1959), che propose di aggiornare la matrice approssimata dell'Hessiana iterativamente. Successivamente, Broyden, Fletcher, Goldfarb e Shanno svilupparono indipendentemente (1970) una variante migliorata che oggi si chiama **BFGS**. L'idea di BFGS è risolvere un problema di minimo: tra tutte le matrici simmetriche che soddisfano $\mathbf{B}^{(k+1)}\mathbf{y}_k = \mathbf{s}_k$, si cerca quella più vicina a $\mathbf{B}^{(k)}$ nella norma

di Frobenius. Questo approccio conservativo garantisce che l'approssimazione cambi il minimo possibile, preservando l'informazione di secondo ordine accumulata nelle iterazioni precedenti. La soluzione di questo problema di ottimizzazione vincolato è la **formula BFGS**:

$$\mathbf{B}^{(k+1)} = \left(\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^T\right) \mathbf{B}^{(k)} \left(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^T\right) + \rho_k \mathbf{s}_k \mathbf{s}_k^T \tag{6.39}$$

con $\rho_k = 1/(\mathbf{y}_k^T \mathbf{s}_k)$. Questa è un'aggiornamento di **rango 2**: modifico $\mathbf{B}^{(k)}$ aggiungendo due matrici di rango 1. Verifico che sia simmetrica e che soddisfi l'equazione secante $\mathbf{B}^{(k+1)}\mathbf{y}_k = \mathbf{s}_k$ per costruzione.

Il metodo di Newton standard fa così:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}^{-1}(\boldsymbol{\theta}^{(k)}) \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}) \tag{6.40}$$

dove $\mathbf{H}$ è l'Hessiana (le derivate seconde). Il problema è che calcolare $\mathbf{H}$ e poi invertirla ha un costo molto alto quando si hanno molte variabili.

I metodi quasi-Newton invece approssimano $\mathbf{H}^{-1}$ usando solo il gradiente, senza calcolare derivate seconde. BFGS costruisce un'approssimazione $\mathbf{B}^{(k)} \approx \mathbf{H}^{-1}$ aggiornandola iterativamente con la formula rank-2:

$$\mathbf{B}^{(k+1)} = \left(\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^T\right) \mathbf{B}^{(k)} \left(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^T\right) + \rho_k \mathbf{s}_k \mathbf{s}_k^T \tag{6.41}$$

dove:

$$\mathbf{s}_k = \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)} \tag{6.42}$$

$$\mathbf{y}_k = \nabla \mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}) \tag{6.43}$$

$$\rho_k = \frac{1}{\mathbf{y}_k^T \mathbf{s}_k} \tag{6.44}$$

Strategia Limited-Memory

Versione **L-BFGS**: invece di tenere in memoria tutta la matrice $\mathbf{B}^{(k)}$ (che costerebbe $O(m^2)$ in spazio), tiene solo le ultime ℓ coppie $\{(\mathbf{s}_i, \mathbf{y}_i)\}_{i=k-\ell}^{k-1}$ e ricostruisce il prodotto $\mathbf{B}^{(k)}\nabla\mathcal{L}$ con l'algoritmo a due cicli:

[H]

$\mathbf{q} \leftarrow \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})$
for $i = k-1, k-2, \ldots, k-\ell$ **do**
    $\alpha_i \leftarrow \rho_i \mathbf{s}_i^T \mathbf{q}$
    $\mathbf{q} \leftarrow \mathbf{q} - \alpha_i \mathbf{y}_i$
end for
$\mathbf{r} \leftarrow \mathbf{H}_0^{(k)} \mathbf{q}$ (dove $\mathbf{H}_0^{(k)}$ è un'approssimazione iniziale)
for $i = k-\ell, k-\ell+1, \ldots, k-1$ **do**

8: $\beta_i \leftarrow \rho_i \mathbf{y}_i^T \mathbf{r}$
9: $\mathbf{r} \leftarrow \mathbf{r} + (\alpha_i - \beta_i)\mathbf{s}_i$
10: **end for**
11: **return r** ▷ Direzione di discesa

Nel caso specifico di questa tesi viene usato $\ell = 10$. L'approssimazione iniziale è semplicemente:

$$\mathbf{H}_0^{(k)} = \frac{\mathbf{s}_{k-1}^T \mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^T \mathbf{y}_{k-1}} \mathbf{I} \tag{6.45}$$

Gestione dei Vincoli

L'estensione **L-BFGS-B** gestisce vincoli di box, ossia vincoli semplici del tipo:

$$l_i \leq \theta_i \leq u_i, \quad i = 1, \ldots, m \tag{6.46}$$

dove l_i e u_i sono rispettivamente i limiti inferiore e superiore per ciascuna variabile θ_i. Trattandosi di parametri che rappresentano probabilità, il loro dominio teorico naturale è l'intervallo $[0, 1]$. Tuttavia, in fase di implementazione, i vincoli di limitazione per ciascuna variabile θ_i sono stati rigorosamente impostati a $l_i = 0$ e $u_i = 0.99$. La scelta di un limite superiore strettamente minore di 1 è una necessità numerica: serve a prevenire instabilità computazionali (come l'annullamento dei fattori $(1 - \theta_T)$ presenti nel sistema) che comprometterebbero la convergenza dell'algoritmo.

Proiezione sul Dominio Ammissibile

L'insieme ammissibile Ω è quindi:

$$\Omega = \{\boldsymbol{\theta} \in \mathbb{R}^m : l_i \leq \theta_i \leq u_i,\ i = 1, \ldots, m\} \tag{6.47}$$

Si tratta di un insieme convesso chiuso: un ipercubo in $\mathbb{R}^m$. Per ogni punto $\boldsymbol{\theta} \in \mathbb{R}^m$, esiste un'unica proiezione $P_\Omega(\boldsymbol{\theta})$ su Ω, calcolabile componente per componente:

$$[P_\Omega(\boldsymbol{\theta})]_i = \begin{cases} l_i & \text{se } \theta_i < l_i \\ \theta_i & \text{se } l_i \leq \theta_i \leq u_i \\ u_i & \text{se } \theta_i > u_i \end{cases} \tag{6.48}$$

Questa operazione "taglia" le componenti che violerebbero i vincoli, riportandole sul bordo ammissibile più vicino.

Gradiente Proiettato

Quando un punto $\boldsymbol{\theta}^{(k)}$ è interno a Ω, ci si può muovere liberamente nella direzione $-\nabla\mathcal{L}(\boldsymbol{\theta}^{(k)})$. Ma se alcune variabili sono ai limiti, essendo ottimizzazione vincolata, bisogna modificare la direzione per non uscire dal dominio. L'idea chiave è usare il

gradiente proiettato [19]:

$$[\nabla^P \mathcal{L}(\boldsymbol{\theta})]_i = \begin{cases} \min\left(0, \frac{\partial \mathcal{L}}{\partial \theta_i}\right) & \text{se } \theta_i = l_i \\ \max\left(0, \frac{\partial \mathcal{L}}{\partial \theta_i}\right) & \text{se } \theta_i = u_i \\ \frac{\partial \mathcal{L}}{\partial \theta_i} & \text{altrimenti} \end{cases} \tag{6.49}$$

Interpretazione: Se θ_i è al limite inferiore l_i e il gradiente mi spinge verso il basso ($\partial\mathcal{L}/\partial\theta_i > 0$), annullo quella componente perché non posso scendere ulteriormente. Se invece il gradiente punta verso l'interno del dominio ($\partial\mathcal{L}/\partial\theta_i < 0$), mantengo quella componente. Ragionamento simmetrico vale per il limite superiore.

La condizione di ottimalità del primo ordine per un problema di minimizzazione vincolato su un insieme convesso Ω è la **disuguaglianza variazionale**: un punto $\boldsymbol{\theta}^* \in \Omega$ è un minimizzatore locale se e solo se:

$$\nabla\mathcal{L}(\boldsymbol{\theta}^*)^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \geq 0 \quad \forall \boldsymbol{\theta} \in \Omega \tag{6.50}$$

Questa condizione dice che il gradiente in $\boldsymbol{\theta}^*$ forma un angolo ottuso (o retto) con qualsiasi direzione ammissibile verso l'interno del dominio: non esiste una direzione ammissibile lungo cui la funzione decresca. Equivalentemente, si può riscrivere questa condizione usando il gradiente proiettato: $\boldsymbol{\theta}^*$ è ottimo se e solo se:

$$\|\nabla^P \mathcal{L}(\boldsymbol{\theta}^*)\| = 0 \tag{6.51}$$

Questo perché il gradiente proiettato $\nabla^P\mathcal{L}$ cattura esattamente le componenti del gradiente che puntano verso l'interno del dominio ammissibile. Quando tutte queste componenti sono nulle, non si può migliorare la soluzione rimanendo in Ω.

Identificazione delle Variabili Attive

A ogni iterazione, L-BFGS-B classifica le variabili in base alla loro posizione rispetto ai vincoli:

- **Variabili libere** ($\mathcal{F}$): quelle per cui $l_i < \theta_i^{(k)} < u_i$, che possono muoversi in entrambe le direzioni;
- **Variabili attive ai limiti inferiori** ($\mathcal{L}$): $\theta_i^{(k)} = l_i$ con $\frac{\partial \mathcal{L}}{\partial \theta_i} > 0$ (il gradiente è verso l'esterno dalla regione di amissibilità);
- **Variabili attive ai limiti superiori** ($\mathcal{U}$): $\theta_i^{(k)} = u_i$ con $\frac{\partial \mathcal{L}}{\partial \theta_i} < 0$.

L'insieme delle variabili attive è $\mathcal{A} = \mathcal{L} \cup \mathcal{U}$. Durante l'ottimizzazione, L-BFGS-B costruisce la direzione di ricerca considerando solo il sottospazio delle variabili libere, fissando temporaneamente quelle attive. Questo processo è chiamato *subspace minimization* e garantisce che la direzione calcolata sia ammissibile e di discesa nel sottospazio libero, evitando violazioni dei vincoli.

Line Search con Condizioni di Wolfe

Ad ogni iterazione, dopo aver calcolato la direzione di discesa $\mathbf{d}^{(k)}$, si determina la lunghezza del passo α_k mediante una procedura di **line search** che soddisfa le condizioni di Wolfe [21] .

Problema da risolvere

Dato il punto corrente $\boldsymbol{\theta}^{(k)}$ e la direzione di discesa $\mathbf{d}^{(k)}$, l'obiettivo è trovare $\alpha_k > 0$ tale che:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \alpha_k \mathbf{d}^{(k)} \tag{6.52}$$

minimizzi (approssimativamente) la funzione obiettivo lungo la direzione $\mathbf{d}^{(k)}$.

Condizioni di Wolfe

Le condizioni di Wolfe garantiscono che il passo α_k produca una riduzione sufficiente della funzione obiettivo e che la sucessione dei $\theta^{(k)}$ sia ragionevole.

Condizione 1: Armijo (sufficiente decrescita)

La condizione di Armijo richiede che la funzione obiettivo decresca sufficientemente:

$$\mathcal{L}(\boldsymbol{\theta}^{(k)} + \alpha_k \mathbf{d}^{(k)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + c_1 \alpha_k \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})^T \mathbf{d}^{(k)} \tag{6.53}$$

dove $c_1 \in (0, 1)$ è tipicamente piccolo (es. $c_1 = 10^{-4}$).

Interpretazione: Il termine $\nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})^T \mathbf{d}^{(k)} < 0$ è la derivata direzionale (negativa se $\mathbf{d}^{(k)}$ è direzione di discesa). La condizione richiede che la riduzione effettiva $\mathcal{L}(\boldsymbol{\theta}^{(k)} + \alpha_k \mathbf{d}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})$ sia almeno una frazione c_1 della riduzione prevista dalla linearizzazione.

Geometricamente: il nuovo valore della funzione deve stare al di sotto della retta tangente scalata da c_1.

Condizione 2: Curvatura (sucessione ragionevole)

La condizione di curvatura serve a evitare che il passo sia troppo corto:

$$\nabla \mathcal{L}(\boldsymbol{\theta}^{(k)} + \alpha_k \mathbf{d}^{(k)})^T \mathbf{d}^{(k)} \geq c_2 \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})^T \mathbf{d}^{(k)} \tag{6.54}$$

con $c_2 \in (c_1, 1)$, di solito vicino a 1 (tipo $c_2 = 0.9$ per L-BFGS).

Geometricamente la derivata direzionale nel nuovo punto $\nabla \mathcal{L}(\boldsymbol{\theta}^{(k+1)})^T \mathbf{d}^{(k)}$ deve essere "più piatta" rispetto al punto iniziale. Questo impedisce passi eccessivamente corti che non fanno progredire molto l'ottimizzazione. Infatti più la funzione è ripida più α_k sarà grande e viceversa.

Strong Wolfe Conditions

Esiste una variante più forte:

$$\left| \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)} + \alpha_k \mathbf{d}^{(k)})^T \mathbf{d}^{(k)} \right| \leq c_2 \left| \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})^T \mathbf{d}^{(k)} \right| \tag{6.55}$$

che esclude anche passi che vanno oltre il minimo locale.

Algoritmo di Line Search

La ricerca del passo α_k funziona tipicamente in due fasi [22]:

Fase 1: Bracketing Si cerca un intervallo $[\alpha_{\text{lo}}, \alpha_{\text{hi}}]$ che contiene valori accettabili. Si parte da $\alpha_0 = 1$ e viene aumentato o diminuito fino a trovare un bracket buono.

Fase 2: Interpolazione Dentro l'intervallo viene usata interpolazione cubica per approssimare il minimo.

La funzione univariata viene definita come:

$$\phi(\alpha) = \mathcal{L}(\boldsymbol{\theta}^{(k)} + \alpha \mathbf{d}^{(k)}) \tag{6.56}$$

L'interpolazione cubica usa i valori $\phi(\alpha_i)$, $\phi(\alpha_{i-1})$ e le derivate $\phi'(\alpha_i)$, $\phi'(\alpha_{i-1})$ per costruire un polinomio cubico:

$$p(\alpha) = a\alpha^3 + b\alpha^2 + c\alpha + d \tag{6.57}$$

che interpola i quattro valori. Il minimo di $p(\alpha)$ fornisce un nuovo candidato:

$$\alpha^* = \arg \min_{\alpha \in [\alpha_{\text{lo}}, \alpha_{\text{hi}}]} p(\alpha) \tag{6.58}$$

In forma esplicita:

$$\alpha^* = \alpha_k - \frac{(\alpha_k - \alpha_{k-1})^2 \phi'(\alpha_k)}{2[\phi'(\alpha_k) - \phi'(\alpha_{k-1}) + \phi(\alpha_k) - \phi(\alpha_{k-1}) - (\alpha_k - \alpha_{k-1})\phi'(\alpha_{k-1})]} \tag{6.59}$$

Perché le condizioni di Wolfe sono importanti

1. **Convergenza garantita**: Gli algoritmi quasi-Newton con line search che rispetta Wolfe convergono a un punto stazionario (sotto assunzioni di regolarità standard).
2. **Efficienza**: Vengono evitati passi troppo corti che rallentano la convergenza e passi troppo lunghi che sprecano valutazioni della funzione.
3. **Proprietà BFGS**: Con le Condizioni di Wolfe rispettate, l'aggiornamento BFGS mantiene la matrice approssimata dell'Hessiana positiva definita.

6.3.3 Convergenza e Criteri di Arresto

L'algoritmo si ferma quando succede una di queste cose:

1. **Tolleranza sulla funzione**: $|\mathcal{L}^{(k+1)} - \mathcal{L}^{(k)}| < \texttt{ftol}$ (dove $\texttt{ftol} = 10^{-9}$)
2. **Tolleranza sul gradiente**: $\|\nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})\|_\infty < \texttt{gtol}$
3. **Numero massimo di iterazioni**: $k > \texttt{maxiter}$ (dove $\texttt{maxiter} = 5000$)

La convergenza di L-BFGS-B è garantita se la funzione obiettivo è abbastanza regolare [19].

6.3.4 Stabilità Numerica: Log-Trick

Per evitare problemi di underflow (quando moltiplico tanti numeri vicini a 1 il risultato può diventare così piccolo da andare sotto la precisione macchina), nell'implementazione del modello è stato usato il "log-trick":

$$\prod_{T:T\cap S\neq\emptyset}(1-\theta_T)=\exp\left(\sum_{T:T\cap S\neq\emptyset}\log(1-\theta_T)\right) \tag{6.60}$$

Basicamente si tarsformano i prodotti instabili in somme di logaritmi, che sono molto più stabili numericamente.

6.4 Implementazione Computazionale

6.4.1 Il Codice

Il codice Python è strutturato in una classe principale `LatentShockModel` che incapsula tutte le funzionalità del modello. L'architettura segue il pattern object-oriented con i seguenti metodi principali:

1. `__init__()`: Inizializzazione del modello con i dati
2. `preprocess_data()`: Preprocessing e filtraggio dei dati
3. `create_binary_matrix()`: Creazione della matrice binaria errori/non-errori
4. `build_latent_structure()`: Costruzione della struttura gerarchica degli shock
5. `estimate_parameters()`: Stima dei parametri θ via ottimizzazione
6. `analyze()`: Analisi completa dei gruppi di variabili

6.4.2 Preprocessing dei Dati

Come descritto nei capitoli precedenti, i dati grezzi estratti dal sistema dotato di AI richiedono un preprocessing accurato prima dell'analisi. In particolare, le date vengono parsate dal formato italiano (gg/mm/aaaa) a oggetti datetime, viene applicato il filtraggio temporale per selezionare il periodo di interesse e vengono rimossi i documenti processati durante i weekend. Il risultato finale è un DataFrame contenente $n = 215.850$ documenti validi.

6.4.3 Costruzione della Matrice Binaria

A partire da questi dati preprocessati, è stata implementata la costruzione della matrice binaria $\mathbf{M} \in \{0,1\}^{n \times d}$ dove ogni riga rappresenta un documento e ogni colonna un campo da estrarre. L'elemento $M_{ij} = 1$ se il campo j del documento i presenta un errore (campo vuoto), altrimenti $M_{ij} = 0$. Questa rappresentazione binaria è fondamentale perché trasforma il problema di analisi degli errori in un problema di modellazione probabilistica su vettori Bernoulliani multivariati. La matrice binaria $\mathbf{M} \in \{0,1\}^{n \times d}$ è quindi formata da i seguenti elementi:

$$M_{ij} = \begin{cases} 1 & \text{se il campo } j \text{ ha un errore nel documento } i \\ 0 & \text{altrimenti} \end{cases} \tag{6.61}$$

```
def create_binary_matrix(self, field_names, group_name=''):
    n = len(self.df_regime)
    d = len(field_names)
    matrix = np.zeros((n, d), dtype=int)

    for i, field in enumerate(field_names):
        matrix[:, i] = self.df_regime[self.error_col].str.contains
    (
            field, case=False, regex=False
        ).astype(int)

    return matrix
```

Listing 6.1: Creazione matrice binaria

6.4.4 Struttura degli Shock Latenti

Implementativamente, la generazione degli shock latenti avviene tramite un algoritmo che esplora sistematicamente tutte le combinazioni possibili dei campi. Si utilizza la funzione `combinations` dalla libreria Python `itertools`, che genera tutte le combinazioni di k elementi da un insieme di d elementi in ordine lessicografico. Per ottenere tutti i sottoinsiemi non vuoti, si itera rispetto a k che va da 1 a d e per ogni valore vengono generate le $\binom{d}{k}$ combinazioni corrispondenti.

Ogni shock viene memorizzato come un dizionario che contiene quattro informazioni fondamentali: (1) il **nome simbolico** costruito secondo la convenzione Y per ordine 1, Z per ordine 2, W per ordine 3, eccetera, seguito dagli indici dei campi coinvolti; (2) l'**ordine**, ossia la cardinalità dell'insieme di campi colpiti; (3) la lista degli **indici** delle variabili influenzate (usando numerazione 0-indexed internamente, ma presentati come 1-indexed all'utente); (4) un campo **type** descrittivo che distingue tra shock `Order-k` e lo shock `Global` di massimo ordine.

Per il gruppo Clinico con $d = 4$, l'algoritmo genera questa sequenza: prima le 4 combinazioni singole (Y_1, Y_2, Y_3, Y_4), poi le $\binom{4}{2} = 6$ coppie $(Z_{12}, Z_{13}, Z_{14}, Z_{23}, Z_{24}, Z_{34})$, poi le $\binom{4}{3} = 4$ triple $(W_{123}, W_{124}, W_{134}, W_{234})$, e infine l'unica combinazione completa V_{1234}, per un totale di $2^4 - 1 = 15$ shock. Per il gruppo Amministrativo con $d = 5$ si ottengono invece $1 + 5 + 10 + 10 + 5 + 1 = 31$ shock (escludendo l'insieme vuoto che non corrisponde ad alcun evento di shock).

Questa lista ordinata di shock diventa poi la struttura per tutte le operazioni successive: l'indice j nella lista corrisponde esattamente al parametro θ_j nel vettore dei parametri da stimare e alla colonna j della matrice di dipendenze $\mathbf{D}$. Mantenere questa corrispondenza biunivoca tra posizione nella lista, indice del parametro, e colonna matriciale è fondamentale per l'algoritmo di ottimizzazione. Quando l'ottimizzazione restituisce il vettore $\hat{\boldsymbol{\theta}}$, si può immediatamente associare ciascun valore stimato $\hat{\theta}_j$ allo shock corrispondente consultando la lista nella posizione j.

Dal punto di vista dell'efficienza, la generazione di tutte le combinazioni ha complessità $O(d \cdot 2^d)$, che è trascurabile rispetto al costo dell'ottimizzazione.

```
def build_latent_structure(self, d):
    latents = []
    prefixes = {1: 'Y', 2: 'Z', 3: 'W', 4: 'V', 5: 'U'}

    for k in range(1, d+1):
        for comb in combinations(range(d), k):
            prefix = prefixes.get(k, 'U')
            name = f"{prefix}{''.join(map(str, [i+1 for i in comb
   ]))}"

            latents.append({
                'name': name,
                'order': k,
                'indices': list(comb),
                'type': 'Global' if k == d else f"Order-{k}"
            })

    return latents
```

Listing 6.2: Costruzione struttura gerarchica

6.4.5 Costruzione della Matrice di Dipendenze

Il passo successivo è costruire la matrice di dipendenze $\mathbf{D} \in \{0,1\}^{2^d \times m}$ che collega le configurazioni agli shock latenti. Per ogni coppia (configurazione k, shock j), il codice deve determinare se lo shock influenza almeno uno dei campi presenti nella configurazione. Il dizionario dello shock j contiene la chiave `'indices'` con la lista dei campi che colpisce. La configurazione k viene convertita in un

insieme di indici estraendo le posizioni dove ha valore 1. L'elemento $D_{k,j}$ viene impostato a 1 se e solo se l'intersezione tra questi due insiemi è non vuota, cioè se `set(shock['indices']).intersection(config_indices)` restituisce un insieme non vuoto.

```
# Inizializza matrice di dipendenze (2^d configurazioni x m shocks)
num_configs = 2**d
num_shocks = len(latent_shocks)
dependency_matrix = np.zeros((num_configs, num_shocks), dtype=int)

# Per ogni configurazione e ogni shock
for k, config in enumerate(configurations):
    # Indici dei campi monitorati dalla configurazione
    config_indices = set(np.where(config == 1)[0])

    for j, shock in enumerate(latent_shocks):
        # Indici dei campi influenzati dallo shock
        shock_indices = set(shock['indices'])

        # D[k,j] = 1 se c'e' intersezione
        if shock_indices.intersection(config_indices):
            dependency_matrix[k, j] = 1
```

Questa operazione viene eseguita per tutte le $2^d \times m$ coppie, risultando in una matrice sparsa (per d piccolo) che però viene mantenuta in formato denso come array NumPy per semplicità. Nel caso $d = 4$ si ottiene una matrice 16×15, mentre per $d = 5$ la matrice è 32×31. La matrice di dipendenze rimane costante durante tutta l'ottimizzazione e viene pre-calcolata una volta sola all'inizio.

6.4.6 Calcolo delle Probabilità Modellate con Log-Trick

All'interno della funzione obiettivo, dato un vettore di parametri $\boldsymbol{\theta}$ candidato, il codice deve calcolare le probabilità modellate $p_{\mathbf{k}}(\boldsymbol{\theta})$ per tutte le configurazioni. La formula diretta richiederebbe di calcolare prodotti come $(1-\theta_1)^{D_{k,1}} \cdot (1-\theta_2)^{D_{k,2}} \cdot \ldots \cdot (1-\theta_m)^{D_{k,m}}$, ma questo porterebbe rapidamente a underflow numerico quando i parametri sono piccoli e i prodotti coinvolgono molti termini.

La soluzione implementata è lavorare nello spazio logaritmico: prima si calcola il vettore dei complementi $\mathbf{1} - \boldsymbol{\theta}$, poi si applica il logaritmo element-wise usando `np.log()`, ottenendo un vettore di m elementi. Per evitare $\log(0)$ se qualche parametro si avvicina troppo a 1, uso `np.maximum(1 - theta, 1e-15)` che impone una soglia minima. A questo punto, il prodotto matrice-vettore $\mathbf{D} \cdot \log(\mathbf{1} - \boldsymbol{\theta})$ calcola simultaneamente i logaritmi di tutte le 2^d probabilità. Questo è un singolo prodotto matriciale che NumPy esegue in modo ottimizzato. Infine, si usa `np.exp()` per

tornare dalle log-probabilità alle probabilità effettive, ottenendo il vettore `p_model` di lunghezza 2^d.

```
def compute_model_probabilities(theta, dependency_matrix):
    """
    Calcola p_model usando il log-trick per stabilita numerica.

    theta: vettore parametri (lunghezza m)
    dependency_matrix: matrice D (2^d x m)
    """
    # Calcola 1 - theta con soglia minima per evitare log(0)
    one_minus_theta = np.maximum(1 - theta, 1e-15)

    # Log-trick: log(prod) = sum(log)
    log_one_minus_theta = np.log(one_minus_theta)

    # Prodotto matriciale: D @ log(1-theta)
    # Risultato: vettore di log-probabilita (lunghezza 2^d)
    log_probs = dependency_matrix @ log_one_minus_theta

    # Torna allo spazio normale: exp(log(p)) = p
    p_model = np.exp(log_probs)

    return p_model
```

Questo approccio log-trick è cruciale per la stabilità numerica.

6.4.7 Pesi e Funzione Obiettivo

Il vettore dei pesi $\mathbf{w}$ di lunghezza 2^d viene costruito analizzando ciascuna configurazione e assegnando un peso in base a quanti campi monitora. Implementativamente si conta il numero di 1 nella configurazione usando `np.sum(config)`: se è 1, si tratta di una marginale singola e assegno $w = 1000$; se il conteggio è uguale a d, è la configurazione globale con peso $w = 100$; per valori intermedi si usa $w = 10$. La configurazione $(0, \ldots, 0)$ riceve $w = 0$. Questo vettore di pesi viene calcolato una volta sola e rimane fisso durante l'ottimizzazione.

```
# Calcola pesi per ogni configurazione
weights = np.zeros(2**d)
for idx, config in enumerate(configurations):
    num_ones = np.sum(config)

    if num_ones == 0:
        weights[idx] = 0      # Configurazione vuota
    elif num_ones == 1:
        weights[idx] = 1000   # Marginali singole (priorita alta)
```

```
    elif num_ones == d:
        weights[idx] = 100    # Configurazione globale
    else:
        weights[idx] = 10     # Interazioni intermedie

# Funzione obiettivo come closure
def objective(theta):
    # Calcola probabilita modellate
    p_model = compute_model_probabilities(theta, dependency_matrix)

    # Errore quadratico pesato
    residuals = p_target - p_model
    weighted_errors = weights * (residuals ** 2)

    # Somma totale
    return weighted_errors.sum()
```

La funzione obiettivo implementata riceve $\boldsymbol{\theta}$ come argomento, calcola `p_model` con il log-trick descritto sopra, poi calcola l'errore element-wise: `error = weights * (p_target - p_model)**2`. L'operazione `**2` calcola il quadrato di ogni differenza, la moltiplicazione con `weights` applica i pesi componente per componente, e infine `error.sum()` restituisce la somma scalare che è il valore della funzione obiettivo. Questa funzione viene chiamata decine di volte per iterazione da L-BFGS-B (incluse le chiamate per le differenze finite usate nel calcolo del gradiente).

6.4.8 Inizializzazione dei Parametri

Prima diottimizzare, il codice calcola un vettore iniziale $\boldsymbol{\theta}^{(0)}$ strategico piuttosto che partire da valori casuali e rischiare quindi una mancata convergenza. Si crea quindi un array NumPy di lunghezza m inizializzato con `np.full(m, 0.01)`, poi lo si modifica per gli shock di ordine 1. Per ciascuno di questi shock, si identifica quale campo colpisce consultando `shock['indices'][0]`, quindi si calcola la probabilità marginale empirica di errore su quel campo sommando la colonna corrispondente della matrice $\mathbf{M}$ e dividendo per n: `p_marginal = matrix[:, field_idx].mean()`. Il valore iniziale per quello shock diventa `theta_init[j] = 0.7 * p_marginal`.

```
# Inizializzazione parametri
m = len(latent_shocks)
theta_init = np.full(m, 0.01)  # Baseline piccolo per tutti

# Migliora inizializzazione per shock di ordine 1
for j, shock in enumerate(latent_shocks):
    if shock['order'] == 1:
```

```
        # Identifica il campo colpito
        field_idx = shock['indices'][0]

        # Calcola probabilita marginale empirica di errore
        p_marginal = matrix[:, field_idx].mean()

        # Inizializza a 70% della marginale
        # (il resto e' dovuto a shock di ordine superiore)
        theta_init[j] = 0.7 * p_marginal

print(f"Theta iniziale per primi 5 shock: {theta_init[:5]}")
```

Questo fattore 0.7 è un'euristica basata sull'osservazione che se un campo ha probabilità marginale di errore del 10%, non tutti questi errori sono dovuti allo shock individuale, ovvero alcuni provengono da shock di ordine superiore che coinvolgono quel campo insieme ad altri. Partire da 0.7 volte la marginale fornisce un punto di partenza ragionevole che è né troppo grande (che porterebbe a sovrastimare le correlazioni) né troppo piccolo (che rallenterebbe la convergenza). Dopo vari tentativi, questa inizializzazione riduce tipicamente il numero di iterazioni necessarie da 300-400 a meno di 150.

6.4.9 Ottimizzatore e Gestione del Risultato

Dopo aver preparato la funzione obiettivo, il vettore iniziale, e i vincoli, il codice chiama `scipy.optimize.minimize` passando questi argomenti: la funzione obiettivo come callable, `theta_init` come punto di partenza, `method='L-BFGS-B'` per specificare l'algoritmo, `bounds=[(0, 0.99) for _ in range(m)]` per imporre i vincoli di box su tutti i parametri, e un dizionario `options` che specifica `maxiter=5000` e `ftol=1e-9`.

```
from scipy.optimize import minimize

# Definisci vincoli di box: 0 <= theta_i <= 0.99
bounds = [(0, 0.99) for _ in range(m)]

# Parametri ottimizzazione
options = {
    'maxiter': 5000,
    'ftol': 1e-9,
    'disp': True
}

# Chiamata all'ottimizzatore L-BFGS-B
result = minimize(
    fun=objective,
```

```
    x0=theta_init,
    method='L-BFGS-B',
    bounds=bounds,
    options=options
)

# Gestione risultato
if result.success:
    theta_opt = result.x
    print(f"Ottimizzazione convergita in {result.nit} iterazioni")
    print(f"Valore finale funzione obiettivo: {result.fun:.2e}")
else:
    print(f"ERRORE: {result.message}")
    theta_opt = result.x  # Usa comunque la migliore soluzione
```

L'ottimizzatore restituisce un oggetto `OptimizeResult` che contiene diversi campi. Il più importante è `result.x`, il vettore ottimale $\hat{\boldsymbol{\theta}}$ di lunghezza m. Il codice verifica che `result.success` sia `True`, indicando che la convergenza è stata raggiunta. Il valore `result.fun` è il valore finale della funzione obiettivo $\mathcal{L}(\hat{\boldsymbol{\theta}})$: nelle prove fatte questo è sempre dell'ordine di 10^{-10} o inferiore, confermando che il fit è eccellente. Il campo `result.nit` indica quante iterazioni sono state necessarie, tipicamente tra 50 e 200 per questo dataset.

Se `result.success` è `False`, il codice stampa un warning con `result.message` che spiega perché l'ottimizzazione è fallita. Possibili cause includono: numero massimo di iterazioni raggiunto senza convergenza (raro con `ftol=1e-9`), gradiente troppo grande dopo tutte le iterazioni (indica un problema nella funzione obiettivo), o violazioni dei vincoli che non possono essere risolte (teoricamente impossibile con vincoli di box semplici).

6.4.10 Validazione Post-Ottimizzazione

Dopo l'ottimizzazione, il codice esegue una fase di validazione ricalcolando `p_model` con il vettore ottimale $\hat{\boldsymbol{\theta}}$ e confrontandolo con `p_target`. Per ogni variabile individuale i, estraggo la probabilità marginale empirica $\hat{P}(X_i = 1)$ dalla matrice: `p_emp[i] = matrix[:, i].mean()`. Dal modello, calcolo la corrispondente probabilità usando le configurazioni appropriate. La differenza `abs(p_emp[i] - p_model[config_idx])` viene stampata: nei casi di convergenza corretta è sempre inferiore a 10^{-4}.

```
import pandas as pd

# Ricalcola probabilita modellate con theta ottimale
p_model_final = compute_model_probabilities(theta_opt,
    dependency_matrix)
```

```
5
6  # Confronta con target
7  print("\n=== VALIDAZIONE ===")
8  for i in range(d):
9      p_emp = matrix[:, i].mean()
10     # Trova configurazione corrispondente al campo i
11     config_idx = 2**i  # (semplificazione per ordine 1)
12     p_mod = 1 - p_model_final[0]  # Calcolo corretto
13     print(f"Campo {i}: empirico={p_emp:.4f}, modello={p_mod:.4f},
       diff={abs(p_emp-p_mod):.2e}")
14
15 # Ordina shock per valore decrescente
16 sorted_indices = np.argsort(-theta_opt)
17
18 # Crea tabella risultati
19 results = []
20 for rank, idx in enumerate(sorted_indices, 1):
21     shock = latent_shocks[idx]
22     results.append({
23         'Rank': rank,
24         'Shock': shock['name'],
25         'Order': shock['order'],
26         'Theta': theta_opt[idx],
27         'Fields': shock['indices']
28     })
29
30 # Salva su CSV
31 df_results = pd.DataFrame(results)
32 df_results.to_csv('risultati_shock_latenti.csv', index=False)
33
34 # Stampa top-10
35 print("\n=== TOP 10 SHOCK ===")
36 for r in results[:10]:
37     print(f"{r['Rank']:2d}. {r['Shock']:10s} (order {r['Order']}):
       theta={r['Theta']:.4f}")
38
39 # Metriche riassuntive
40 total_mass = theta_opt.sum()
41 top5_mass = theta_opt[sorted_indices[:5]].sum()
42 significant_count = np.sum(theta_opt > 0.001)
43
44 print(f"\nSomma totale theta: {total_mass:.4f}")
45 print(f"Top-5 cattura: {100*top5_mass/total_mass:.1f}% della massa
     ")
46 print(f"Shock significativi (>0.001): {significant_count}/{m}")
```

Il codice poi ordina gli shock per valore decrescente di $\hat{\theta}$ usando `np.argsort(-theta_opt)`

che restituisce gli indici in ordine decrescente. Per ciascuno di questi indici, si accede alla lista degli shock per recuperare nome, ordine, e campi influenzati, creando una lista ordinata di tuple `(shock_name, theta_value, fields)`. Questa lista viene salvata e i migliori 10 vengono stampati sul terminale. Gli shock con $\theta < 10^{-4}$ vengono considerati trascurabili e vengono contrassegnati come tali.

Infine, il codice calcola alcune metriche riassuntive: la percentuale di probabilità catturata dai migliori 5 shock (tipicamente >85%), e il numero di shock con $\theta > 0.001$ (shock "significativi").

6.5 Risultati

6.5.1 Dataset Analizzato

I dati sono stati divisi come nei capitoli precedenti in due gruppi:

- **Gruppo Clinico** ($d = 4$): Patologia, Numero prescrizione, Data, Prestazione
- **Gruppo Amministrativo** ($d = 5$): Data, Prestazione, Numero fattura, Importo totale, Importo prestazione

Il Dataset per ogni documento elenca quali campi sono vuoti.

6.5.2 Risultati Gruppo Clinico

I risultati dell'ottimizzazione di questo gruppo sono riportati nella Tabella 6.3.

Tabella 6.3: Risultati stima shock latenti – Gruppo Clinico ($d = 4$, $n = 215.850$)

Variabile	**Ordine**	**Campi Influenzati**	$\boldsymbol{\theta}$
Z_{12}	2	Patologia + Numero prescrizione	0.0572
Y_2	1	Numero prescrizione	0.0218
Y_1	1	Patologia	0.0161
Y_4	1	Prestazione	0.0128
W_{124}	3	Patologia + Num.presc. + Prestazione	0.0097
Z_{24}	2	Numero prescrizione + Prestazione	0.0033
Y_3	1	Data	0.0031
W_{123}	3	Patologia + Num.presc. + Data	0.0023
Z_{13}	2	Patologia + Data	0.0021
Z_{34}	2	Data + Prestazione	0.0005
Z_{23}	2	Numero prescrizione + Data	0.0004
V_{1234}	4	Tutti i campi (globale)	0.0004
W_{234}	3	Num.presc. + Data + Prestazione	0.0003
Z_{14}	2	Patologia + Prestazione	0.0002
W_{134}	3	Patologia + Data + Prestazione	0.0002

Analisi della tabella:

Lo shock più forte è Z_{12} con un θ del 5.72%. Significa che c'è una correlazione importante tra errori su Patologia e Numero prescrizione (probabilmente hanno una causa comune, che potrebbe essere legata alla struttura dei documenti o a come il sistema AI interpreta quel tipo di contenuto), come il capitolo 3 mostrava.

Gli shock individuali Y_1, Y_2, Y_4 confermano che Numero prescrizione e Patologia sono i campi più critici anche quando falliscono singolarmente. Invece lo shock globale V_{1234} è bassissimo (0.04%), quindi è rarissimo che tutti i campi falliscano insieme. Documenti completamente illeggibili esistono ma sono un'eccezione.

6.5.3 Risultati Gruppo Amministrativo

I risultati principali di questo gruppo sono riportati nella Tabella 6.4.

Tabella 6.4: Risultati stima shock latenti – Gruppo Amministrativo ($d = 5$, $n = 215.850$)

Variabile	**Ordine**	**Campi Influenzati**	$\boldsymbol{\theta}$
Y_2	1	Prestazione	0.0242
Y_3	1	Numero fattura	0.0124
Y_1	1	Data	0.0067
Z_{23}	2	Prestazione + Numero fattura	0.0017
Z_{13}	2	Data + Numero fattura	0.0012
Z_{12}	2	Data + Prestazione	0.0010
W_{123}	3	Data + Prestaz. + Num.fattura	0.0003
V_{1235}	4	Data + Prest. + Num.fatt. + Imp.prest.	3.9×10^{-7}
V_{1234}	4	Data + Prest. + Num.fatt. + Imp.tot.	3.9×10^{-7}
U_{12345}	5	Tutti i campi (globale)	2.9×10^{-8}
Y_4, Y_5	1	Importo totale, Importo prestazione	0.0000
Z_{ij} (altri)	2	Varie combinazioni con importi	0.0000

Analsi della tabella:

- Si può notare che gli Importi (totale e prestazione) hanno $\theta = 0$. Significa che questi campi non sbagliano mai singolarmente. Il che può trovare giustificazione dal fatto che sono valori calcolati/derivati da altre informazioni.
- Prestazione ($Y_2 = 2.42\%$) è il campo che dà più problemi nel gruppo amministrativo, anche se con percentuali basse.
- Lo shock globale U_{12345} è approssimabile a zero ($\theta \approx 10^{-8}$), quindi l'AI nel complesso estrae correttamente i campi. Infatti, che tutti e cinque i campi sbaglino insieme è un evento rarissimo.

6.5.4 Validazione del Modello

Per verificare che il modello funzioni ho confrontato le probabilità marginali empiriche (quelle che osservo direttamente) con quelle che predice il modello. La formula per la marginale del campo i è:

$$P(X_i = 1) = 1 - \prod_{T: i \in T} (1 - \theta_T) \tag{6.62}$$

Questa viene dall'equazione (6.25) applicata al caso con solo il campo i.

Tabella 6.5: Validazione: confronto marginali empiriche rispetto a quelle del modello

Campo	P_{emp}	P_{model}	$\lvert\Delta\rvert$
Gruppo Clinico			
Patologia	0.0961	0.0958	0.0003
Numero prescrizione	0.1024	0.1021	0.0003
Data	0.0089	0.0091	0.0002
Prestazione	0.0265	0.0263	0.0002
Gruppo Amministrativo			
Data	0.0089	0.0089	0.0000
Prestazione	0.0269	0.0269	0.0000
Numero fattura	0.0153	0.0153	0.0000
Importo totale	0.0000	0.0000	0.0000
Importo prestazione	0.0000	0.0000	0.0000

Le differenze sono tutte sotto 0.001, quindi il modello si adatta bene ai dati.

6.5.5 Convergenza dell'Algoritmo

L'L-BFGS-B converge velocemente:

Tabella 6.6: Performance dell'ottimizzazione

Gruppo	**Parametri**	**Iterazioni**	**Errore finale**	**Convergenza**
Clinico	15	~ 150	$< 10^{-6}$	Sì
Amministrativo	31	~ 250	$< 10^{-6}$	Sì

6.6 Riepilogo

Riepilogando il modello a shock latenti si è rivelato uno strumento utile per capire i pattern di errori nel sistema AI. Infatti risulta che:

1. Lo shock Z_{12} del 5.72% mostra che c'è una forte dipendenza tra errori su Patologia e Numero prescrizione. Probabilmente c'è un problema sistemico nel modo in cui l'AI estrae quel tipo di informazioni.

2. Gli shock individuali (Y_i) più alti indicano su quali campi bisogna lavorare per migliorare il sistema.

3. Gli shock globali sono bassissimi, quindi l'AI raramente sbaglia su tutti i campi contemporaneamente.

4. L-BFGS-B converge velocemente anche con oltre 200.000 documenti, metodo scalabile.

5. Applicazioni aziendali: un vantaggio importante di questo modello è che permette analisi specifiche per singolo campo, non solo per un gruppo di campi (come nei capitoli precedenti). Se l'azienda volesse in futuro concentrarsi sul miglioramento della lettura di un campo particolare (per esempio "Patologia"), non è sufficiente considerare solo il parametro θ_{Y_1} dello shock individuale. Bisogna prendere in considerazione tutti gli shock che includono quel campo: Y_1 (individuale), Z_{12}, Z_{13}, Z_{14} (coppie che coinvolgono il campo 1), $W_{123}, W_{124}, W_{134}$ (triple), e V_{1234} (globale). La probabilità totale di errore su quel campo è infatti data da:

$$P(X_1 = 1) = 1 - \prod_{T:1\in T} (1 - \theta_T) \tag{6.63}$$

   dove il prodotto include tutti gli shock T che contengono l'indice 1.

Capitolo 7

Modello per gli Score

7.1 Introduzione e Contesto Aziendale

La presente analisi è stata condotta su specifica richiesta dell'azienda con l'obiettivo di valutare le performance del sistema di estrazione automatica dei dati da documenti assicurativi. In particolare è stato mostrato interesse nella ricerca di una misura di distanza che indicasse quanto ogni documento processsato dal sistema disti dal documento perfetto analizzato dal team operativo.

7.2 Dataset e Modifiche Metodologiche

7.2.1 Descrizione del Dataset

Il dataset che verrà preso in considerazione in questo capitolo è anch'esso filtrato nell'intervallo temporale dal 10 settembre all'11 novembre con l'esclusione dei weekend, per un totale di 215.740 documenti di cui 54.310 documenti clinici (prescrizioni mediche) e 161.430 documenti amministrativi (fatture). La struttura del dataset è però differente da quelle precedenti. Infatti ogni riga del dataset rappresenta un documento processato dal sistema e contiene le seguenti informazioni:

Campi considerati per documento

Documenti Clinici (4 campi): Patologia, Numero_Prescrizione, Data e Prestazione.

Documenti Amministrativi (5 campi): Prestazione, Data, Importo_Prestazione, Numero_Fattura e Importo_Totale.

Tabella 7.1: Struttura del dataset di input

Colonne Dataset	Descrizione
`CODTASK`	Identificativo univoco del documento
`TIPOLOGIA_DOCUMENTO`	Prescrizione medica o Fattura
`DATAESTRAZIONE`	Data di processamento (formato gg/mm/aaaa)
`campi vuoti`	Lista separata da virgole dei campi non estratti
`campi con score 0`	Lista campi estratti con qualità bassissima
`campi con score 1`	Lista campi estratti con qualità bassa
`campi con score 2`	Lista campi estratti con qualità media
`campi con score 3`	Lista campi estratti con qualità alta

7.2.2 Score di Qualità

Il sistema di estrazione è basato sul confronto di risultati ottenuti da due modelli AI indipendenti. Assegna a ciascun campo estratto uno score di qualità compreso tra 0 e 3, calcolato confrontando le estrazioni dei due modelli:

- **Score 0** (bassissimo): assenza di corrispondenza tra le estrazioni
- **Score 1** (basso): corrispondenza parziale minima
- **Score 2** (medio): corrispondenza buona ma non ottimale
- **Score 3** (alto): corrispondenza perfetta o quasi perfetta

Quando un campo non viene estratto da entrambi i modelli (o viene estratto solo da uno), viene classificato come **campo vuoto**.

7.3 Metodologia di Analisi

7.3.1 Rappresentazione Vettoriale

Per l'obiettivo dell'analisi di questo capitolo, ogni documento è stato trasformato in un vettore numerico di dimensione fissa, dove ogni componente rappresenta la qualità di un campo specifico estratto.

Codifica dei Valori

Ogni campo del vettore assume un valore intero compreso tra 0 e 4:

$$v_i = \begin{cases} 0 & \text{se il campo } i \text{ è vuoto (non estratto)} \\ 1 & \text{se il campo } i \text{ ha score 0} \\ 2 & \text{se il campo } i \text{ ha score 1} \\ 3 & \text{se il campo } i \text{ ha score 2} \\ 4 & \text{se il campo } i \text{ ha score 3} \end{cases} \tag{7.1}$$

Esempio di Rappresentazione

Consideriamo un documento clinico con la seguente situazione estrattiva:

- Patologia: non estratta
- Numero_Prescrizione: estratta con score 3
- Data: estratta con score 3
- Prestazione: estratta con score 2

Il vettore corrispondente sarà:

$$\mathbf{v} = (0, 4, 4, 3)$$

dove le componenti rappresentano rispettivamente: Patologia, Numero_Prescrizione, Data, Prestazione.

7.3.2 Metriche per la misura della qualità dei documenti estratti

Per quantificare le performance estrattive, sono state definite tre metriche complementari:

Distanza dall'Ottimo

Questa metrica misura lo scostamento complessivo dalla situazione ideale (tutti i campi estratti con score massimo):

$$D(\mathbf{v}) = \left| \sum_{i=1}^{n} (4 - v_i) \right| = \left| 4n - \sum_{i=1}^{n} v_i \right| \tag{7.2}$$

dove n è il numero totale di campi considerati (4 per documenti clinici, 5 per amministrativi).

Nota: Questa metrica include implicitamente sia l'effetto dei campi vuoti, sia gli errori di estrazione.

Qualità Media Campi Estratti

Per isolare la qualità intrinseca deli campi estratti dai campi vuoti, è stata definita la metrica:

$$Q_{\text{estratti}}(\mathbf{v}) = \frac{1}{|\{i : v_i > 0\}|} \sum_{i:v_i>0} v_i \tag{7.3}$$

Questa metrica considera esclusivamente i campi effettivamente estratti ($v_i > 0$) e ne calcola il valore medio sulla scala 0-4.

Numero di Campi Estratti

$$N_{\text{estratti}}(\mathbf{v}) = |\{i : v_i > 0\}| \tag{7.4}$$

7.3.3 Distribuzione di Probabilità Empirica

Per ogni tipologia di documento, è stata calcolata la distribuzione di probabilità empirica dei vettori utilizzando la frequenza:

$$P(\mathbf{v}) = \frac{n_{\mathbf{v}}}{N} \tag{7.5}$$

dove:

- $n_{\mathbf{v}}$ è il numero di occorrenze del vettore $\mathbf{v}$ nel dataset
- N è il numero totale di documenti della tipologia considerata

7.3.4 Momenti Statistici

Per caratterizzare la distribuzione, sono stati calcolati i seguenti momenti statistici sulla variabile aleatoria "distanza dall'ottimo":

Media:

$$\mu_D = \sum_{\mathbf{v}} D(\mathbf{v}) \cdot P(\mathbf{v}) \tag{7.6}$$

Varianza:

$$\sigma_D^2 = \sum_{\mathbf{v}} (D(\mathbf{v}) - \mu_D)^2 \cdot P(\mathbf{v}) \tag{7.7}$$

Skewness (asimmetria):

$$\gamma_1 = \frac{1}{\sigma_D^3} \sum_{\mathbf{v}} (D(\mathbf{v}) - \mu_D)^3 \cdot P(\mathbf{v}) \tag{7.8}$$

Kurtosis (curtosi):

$$\gamma_2 = \frac{1}{\sigma_D^4} \sum_{\mathbf{v}} (D(\mathbf{v}) - \mu_D)^4 \cdot P(\mathbf{v}) \tag{7.9}$$

7.4 Risultati

7.4.1 Documenti Clinici (Prescrizioni Mediche)

Statistiche Descrittive

L'analisi di 54.310 prescrizioni mediche ha evidenziato le seguenti performance del sistema:

Tabella 7.2: Statistiche descrittive - Documenti Clinici

Metrica	Valore
Documenti perfetti (tutti campi score 3)	71,6%
Documenti senza campi vuoti	94,9%
Qualità media campi estratti	3,91/4,00 (97,8%)
Distanza media dall'ottimo	0,63

Distribuzione Campi Vuoti

Tabella 7.3: Distribuzione numero di campi vuoti per documento – Clinici

N° Campi Vuoti	Documenti	Percentuale
0	51.565	94,9%
1	2.745	5,1%

Solo il 5,1% dei documenti presenta un campo vuoto, confermando un'elevata completezza del sistema di estrazione.

Decomposizione degli Impatti Per identificare con precisione la causa principale degli errori sui documenti clinici, la distanza media dal punteggio ottimo (pari a una perdita di 0,56 punti per documento, come successivamente calcolato) è stata calcolata come l'addizione tra: l'impatto dei campi vuoti e l'impatto degli errori della qualità di estrazione.

Il primo contributo deriva dai **campi vuoti**. Nel sistema di valutazione, la mancata estrazione di un'informazione porta la penalità massima, pari a 4 punti.

Poiché i dati mostrano una media di 0,05 campi lasciati vuoti per ogni documento, l'impatto generato da questa casistica si calcola come:

$$0{,}05 \text{ campi/doc} \times 4 \text{ punti/campo} = 0{,}20 \text{ punti persi}$$

Questo valore rappresenta il 35,7% del divario totale (0,56 punti). Il secondo contributo, che risulta essere il fattore limitante principale, deriva dagli **errori di estrazione**. In media, il sistema estrae 3,95 campi per documento. La qualità media dell'estrazione su questi campi è estremamente elevata, valutata a 3,91 su un massimo di 4,00. Tuttavia, moltiplicando questa piccola differenza per il volume totale dei campi elaborati, l'impatto cresce:

$$0{,}09 \text{ punti persi} \times 3{,}95 \text{ campi estratti} \approx 0{,}36 \text{ punti persi}$$

Questo valore rappresenta il 64,3% sul totale totale.

In sintesi, la decomposizione quantitativa degli impatti dimostra che la performance dell'Intelligenza Artificiale non è penalizzata dalla mancata estrazione di campi, ma dalla loro sistematica imprecisione di estrazione.

Combinazioni Vettoriali Più Frequenti

La Tabella 7.4 riporta le 5 combinazioni vettoriali più frequentemente osservate.

Tabella 7.4: Top 5 combinazioni vettoriali – Documenti Clinici

Rank	Vettore	Probabilità	Interpretazione
1	(4,4,4,4)	71,6%	Perfetto
2	(4,4,4,3)	6,1%	Solo il campo Prestazione medio
3	(4,4,4,2)	5,0%	Solo il campo Prestazione basso
4	(4,4,3,4)	4,0%	Solo il campo Data medio
5	(0,4,4,4)	3,8%	Patologia VUOTA

Osservazione critica: La combinazione al quinto posto (3,8% dei documenti) evidenzia che il campo Patologia è il più problematico, essendo l'unico campo frequentemente lasciato vuoto.

Momenti Statistici della Distribuzione

Interpretazione:

- **Skewness positivo** (2,06): distribuzione asimmetrica con coda lunga verso destra. La maggior parte dei documenti è concentrata vicino all'ottimo (distanza bassa), con pochi outlier problematici

Tabella 7.5: Momenti statistici – Distanza dall'ottimo – Clinici

Momento	Valore
Media (μ_D)	0,63
Varianza (σ_D^2)	1,44
Deviazione standard (σ_D)	1,20
Skewness (γ_1)	2,06
Kurtosis (γ_2)	6,63

- **Kurtosis elevato** (6,63 > 3): distribuzione presenta un picco centrale pronunciato con forte concentrazione attorno alla media e code più pesanti rispetto a una distribuzione normale. Questo conferma che il sistema è generalmente affidabile, con occasionali fallimenti marcati

7.4.2 Documenti Amministrativi (Fatture)

Performance Generali

L'analisi di 161.430 fatture ha mostrato performance superiori rispetto ai documenti clinici:

Tabella 7.6: Statistiche descrittive – Documenti Amministrativi

Metrica	Valore
Documenti perfetti (tutti campi score 3)	86,8%
Documenti senza campi vuoti	99,9%
Qualità media campi estratti	3,97/4,00 (99,3%)
Distanza media dall'ottimo	0,20

Distribuzione Campi Vuoti

I campi vuoti sono quasi assenti nei documenti amministrativi (solo 0,1% con un campo vuoto).

Decomposizione degli Impatti

La distanza media dal punteggio ottimo è pari a una perdita complessiva di soli 0,20 punti per documento. Scomponendo questo divario nelle sue due componenti, si nota un risultato ancora più netto rispetto al caso clinico.

Il primo contributo, legato all'impatto dei **campi vuoti**, risulta del tutto marginale. La mancata estrazione delle informazioni genera infatti una perdita

Tabella 7.7: Distribuzione numero di campi vuoti per documento – Amministrativi

N° Campi Vuoti	Documenti	Percentuale
0	161.293	99,9%
1	137	0,1%

estremamente piccola di 0,003 punti per documento, rappresentando appena l'1,5% del totale di inefficienza. La distanza dall'ottimo è quasi interamente imputabile agli **errori di estrazione**. Le imprecisioni nella estrazione dei campi effettivamente compilati pesano per 0,197 punti, ovvero il 98,5% del totale dei punti persi (essendo la somma esatta pari a $0{,}003 + 0{,}197 = 0{,}20$).

L'analisi dimostra quindi che per i documenti amministrativi il problema dei campi vuoti è quasi irrilevante. Il miglioramento per l'Intelligenza Artificiale si deve concentrare nell'accuratezza della lettura dei campi estratti.

Combinazioni Vettoriali Più Frequenti

Tabella 7.8: Top 5 combinazioni vettoriali – Documenti Amministrativi

Rank	Vettore	Probabilità	Interpretazione
1	(4,4,4,4,4)	86,8%	Perfetto
2	(2,4,4,4,4)	6,1%	Solo il campo Prestazione basso
3	(3,4,4,4,4)	5,9%	Solo il campo Prestazione medio
4	(4,3,4,4,4)	0,5%	Solo il campo Data medio
5	(4,2,4,4,4)	0,5%	Solo il campo Data basso

Osservazione: Il campo Prestazione è quello con maggiore variabilità qualitativa, pur essendo sempre estratto. Rappresenta il 12% dei documenti con qualità non ottimale.

Momenti Statistici della Distribuzione

Interpretazione:

- **Skewness molto elevato** (2,88): asimmetria molto forte, con l'86,8% dei documenti concentrato sul valore ottimo
- **Kurtosis estremamente elevato** (11,62): concentrazione estrema sulla modalità perfetta, con variabilità minima

Questi parametri indicano un sistema di estrazione estremamente affidabile e consistente per i documenti amministrativi.

Tabella 7.9: Momenti statistici – Distanza dall'ottimo – Amministrativi

Momento	**Valore**
Media (μ_D)	0,20
Varianza (σ_D^2)	0,32
Deviazione standard (σ_D)	0,57
Skewness (γ_1)	2,88
Kurtosis (γ_2)	11,62

7.4.3 Confronto tra Tipologie Documentali

La Tabella 7.10 sintetizza il confronto tra le performance relative alle due tipologie documentali.

Tabella 7.10: Confronto performance: Clinici vs. Amministrativi

Metrica	**Clinici**	**Amministrativi**
Documenti perfetti	71,6%	86,8%
Documenti senza vuoti	94,9%	99,9%
Qualità media estratti	3,91/4,00	3,97/4,00
Distanza media ottimo	0,63	0,20
Problema principale	Qualità (67,6%)	Qualità (98,4%)
Impatto campi vuoti	32,3%	1,7%
Campo critico	Patologia	Prestazione
Corr. vuoti-qualità	0,026	−0,020

7.5 Riepilogo

- Il sistema si mostra relativamente efficiente per entrambi i gruppi.
- La maggiore difficoltà sui documenti clinici potrebbe essere attribuita alla maggiore variabilità linguistica (prescrizioni da parte di diversi medici) e alla complessità della dicitura dei campi termini medici
- Le distribuzioni osservate mostrano le seguenti caratteristiche:
  - Alta concentrazione sull'ottimo (kurtosis > 6)
  - Pochi outlier (skewness moderato-alto)
  - Varianza contenuta ($\sigma_D < 1{,}5$ in entrambi i casi)

.

Capitolo 8

Conclusioni

8.1 Sintesi dei Risultati Principali

Questa tesi ha presentato un'analisi statistica rigorosa del sistema di estrazione dati, integrato con intelligenza artificiale, implementato da una compagnia assicurativa cliente di Accenture. Lo scopo principale è stato quantificare il rischio operativo e analizzare gli errori commessi dall'AI per fornire all'azienda dati concreti per migliorare il sistema.

Attraverso l'applicazione integrata di tecniche di statistica inferenziale, teoria delle probabilità, ottimizzazione numerica e gestione quantitativa del rischio, sono stati ottenuti risultati significativi.

8.1.1 Caratterizzazione dei Pattern Temporali

L'analisi della stagionalità (Capitolo 2) ha confermato che i flussi di arrivo dei documenti seguono un processo stocastico con le seguenti proprietà:

- **Cambio di regime operativo:** L'analisi ha rivelato una discontinuità strutturale significativa in corrispondenza del 10 settembre, con aumento evidente dei volumi medi delle occorrenze giornaliere (da 153 a 505 prescrizioni, da 43 a 225 fatture). Il test t ha confermato la significatività statistica ($p < 0.001$) della differenza tra i due periodi.

- **Stagionalità settimanale:** La presenza di pattern ricorrenti legati al giorno della settimana considerato, con calo sistematico del $\sim$56% nei weekend, ha reso necessario il filtraggio per garantire stazionarietà. L'esclusione dei giorni non lavorativi ha eliminato bias che avrebbero compromesso la validità delle analisi successive.

Implicazione pratica: Ogni Dataset, usato per l'analisi inferenziali di questa tesi, è stato filtrato considerando solo il periodo stabile (10 settembre - 12 novembre, solo giorni feriali). Il totale di documenti a disposizione è di 215.740.

8.1.2 Identificazione delle Dipendenze Strutturali

L'analisi delle correlazioni tra campi vuoti (Capitolo 3) ha rivelato pattern sistematici di dipendenza, potendo quindi escludere l'ipotesi di indipendenza degli errori:

- **Correlazione forte nel gruppo clinico:** La coppia Patologia-Numero Prescrizione mostra una correlazione di Pearson $\rho = 0.75$ ($\chi^2 = 125{,}542$, $p < 0.0001$), indicando che quando un campo non viene estratto è molto probabile che anche l'altro non lo sia. L'analisi combinatoriale ha confermato che questa coppia rappresenta il 5.55% di tutti i documenti (12.283 su 221.253).
- **Correlazione moderata nel gruppo amministrativo:** La coppia Importo Prestazione-Prestazione mostra $\rho = 0.48$ ($\chi^2 = 51{,}232$, $p < 0.0001$).
- **Indipendenza del campo Data:** In entrambi i gruppi, il campo Data presenta correlazioni trascurabili ($|\rho| < 0.17$) con tutti gli altri campi, suggerendo che la sua mancata estrazione abbia cause indipendenti da quella degli altri campi.

Implicazione pratica: Le correlazioni identificate forniscono indicazioni precise per eventuali interventi mirati. Ad esempio, migliorare il riconoscimento della zona "prescrizione" dei documenti clinici può ridurre simultaneamente gli errori su entrambi i campi correlati.

8.1.3 Sovradispersione e Modellazione Probabilistica

La distribuzione delle occorrenze (Capitolo 4) ha evidenziato un discostamento dall'idea della Poisson iniziale (tipica per i processi di conteggio e con ottime proprietà):

- **Sovradispersione estremamente elevata:** L'indice di dispersione $D = \text{Var}(X)/\mathbb{E}[X]$ risulta pari a 12.05 per documenti clinici e 33.55 per amministrativi, escludendo l'ipotesi della POisson($D = 1$).
- **Binomiale Negativa:** Il confronto mediante criteri AIC ha dimostrato la netta superiorità del modello NB (AIC: 548.1 per clinici, 545.6 per amministrativi) rispetto alla Poisson (AIC: 1016.8 e 1927.2 rispettivamente). Il test di Kolmogorov-Smirnov ha confermato l'adeguatezza del modello NB ($p > 0.2$ per entrambi i gruppi).

- **Parametri stimati:** $\hat{\mu}_{\text{clin}} = 459.39$, $\hat{r}_{\text{clin}} = 41.57$ per i documenti clinici; $\hat{\mu}_{\text{amm}} = 220.93$, $\hat{r}_{\text{amm}} = 6.79$ per gli amministrativi. Il valore molto basso di r nel gruppo amministrativo indica variabilità estremamente elevata e code pesanti.

Implicazione pratica: L'uso della Binomiale Negativa è essenziale per stimare correttamente il VaR, necessario per la previsione dei costi operativi. L'adozione di un modello in cui i dati sono distribuiti seguendo una Poisson sottostimerebbe sistematicamente il rischio.

8.1.4 Quantificazione del Rischio di Coda

L'analisi VaR/CVaR (Capitolo 5) ha fornito stime quantitative delle soglie operative critiche:

- **VaR al 95%:** $\text{VaR}_{0.95}^{\text{clin}} = 1461$ errori/giorno, $\text{VaR}_{0.95}^{\text{amm}} = 685$ errori/giorno. Questi valori rappresentano soglie che vengono superate mediamente 12.5 giorni all'anno (su 250 giorni lavorativi).

- **Gap rispetto alla media:** Il VaR al 95% supera la media di un fattore 3.18 per documenti clinici e 3.10 per amministrativi. Questo conferma che analizzare i costi basandosi sulla media non è affidabile nè sufficiente: è necessario un margine di sicurezza del 210-220%

- **Rischio residuo (CVaR):** Il Conditional VaR quantifica il carico medio nei giorni critici (oltre il VaR). Per il gruppo amministrativo: $\text{CVaR}_{0.95} = 809$, con $\Delta = 124$ errori (+18.1%) rispetto al VaR. Per il clinico: $\text{CVaR}_{0.95} = 1582$, con $\Delta = 121$ errori (+8.3%). Il rischio residuo percentualmente maggiore nel gruppo amministrativo riflette code più pesanti.

- **Validazione empirica:** Il backtesting ha mostrato 0 violazioni del $\text{VaR}_{0.95}$ sui giorni osservati (rispetto alle 2.3 attese teoricamente), suggerendo che il modello fornisce stime conservative e quindi affidabili per il dimensionamento.

Implicazione pratica: Le soglie VaR/CVaR forniscono target quantitativi per il dimensionamento della capacità di risorse per gestire errori di estrazione e i costi a esso annessi.

8.1.5 Decomposizione tramite Shock Latenti

Il modello a Shock Latenti (Capitolo 6) ha fornito una rappresentazione interpretabile della struttura di dipendenza:

- **Shock dominante nel gruppo clinico:** Z_{12} (shock simultaneo su Patologia e Numero Prescrizione) con $\hat{\theta} = 5.72\%$. Questo parametro quantifica la probabilità giornaliera che una "causa comune" generi errori su entrambi i campi.

- **Shock individuali:** I parametri $\hat{\theta}_{Y_1} = 1.61\%$ (Patologia), $\hat{\theta}_{Y_2} = 2.18\%$ (Numero Prescrizione) quantificano la probabilità di errori "isolati" su ciascun campo, indipendenti dagli altri.

- **Shock globale trascurabile:** $\hat{\theta}_{V_{1234}} = 0.04\%$ per i clinici, $\hat{\theta}_{U_{12345}} \approx 10^{-8}$ per gli amministrativi. Eventi in cui tutti i campi falliscono simultaneamente sono estremamente rari, indicando che fallimenti totali sono assenti.

- **Convergenza dell'ottimizzazione:** L'algoritmo L-BFGS-B ha raggiunto convergenza in $\sim$150 iterazioni per i clinici (15 parametri), $\sim$250 per gli amministrativi (31 parametri), con errore finale $< 10^{-6}$. La validazione ha confermato che le marginali modellate differiscono da quelle empiriche per meno di 10^{-3}.

Implicazione pratica: Il modello permette analisi per valutare l'impatto di interventi specifici.

8.1.6 Indipendenza tra Completezza e Accuratezza

L'analisi degli score a confronto con i campi vuoti (Capitolo 7) ha validato un'ipotesi sul funzionamento del sistema:

- **Decomposizione degli impatti:** Per i documenti clinici, il 67.6% della "distanza dall'ottimo" è dovuto a errori di qualità sui campi estratti, solo il 32.3% a campi vuoti. Per gli amministrativi, il rapporto è ancora più sbilanciato (98.4% vs 1.7%).

- **Campi critici identificati:** Patologia (3.8% documenti con campo vuoto) per i clinici; Prestazione (12% documenti con score non ottimale) per entrambi i gruppi.

Implicazione pratica: I miglioramenti dell'intelligenza artificiale che si occupa dell'estrazione dei campi, deve concentrarsi prioritariamente sull'accuratezza dell'estrazione (miglioramento degli score) piuttosto che sulla riduzione dei campi vuoti (nonostante bisognerebbe considerare che in questo caso si avrebbe un impatto negativo maggiore sul rimborso ai clienti) .

8.2 Considerazioni Finali

L'automazione dell'estrazione di dati di documenti rappresenta una strategica soluzione per la riduzione di costi per le compagnie assicurative. Tuttavia, come evidenziato da questa tesi, il passaggio da sistemi completamente manuali a sistemi completamente automatizzati non è immediato né privo di rischi.

Inoltre, la tesi è dimostra come strumenti sofisticati matematici (solitamente utilizzati in contesti accademici o finanziari) possano essere adattati con successo a problemi operativi concreti aziendali.

In conclusione, questa tesi non si limita a fornire analisi specifiche per la società assicurativa di interesse, ma propone un possibile metodo replicabile, estensibile, e validabile empiricamente per l'analisi quantitativa di sistemi di automazione con AI dei documenti.

Bibliografia

[1] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel e Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. 5th. John Wiley & Sons, 2015 (cit. a p. 9).

[2] Bernard L. Welch. «The generalization of 'Student's' problem when several different population variances are involved». In: *Biometrika* 34.1/2 (1947) (cit. a p. 10).

[3] Douglas C. Montgomery e George C. Runger. *Applied Statistics and Probability for Engineers*. 7th. John Wiley & Sons, 2018 (cit. alle pp. 17, 18, 21).

[4] Sheldon M. Ross. *Introduction to Probability Models*. 11th. Academic Press, 2014 (cit. a p. 29).

[5] A. Colin Cameron e Pravin K. Trivedi. *Regression Analysis of Count Data*. 2nd. Cambridge University Press, 2013 (cit. a p. 30).

[6] Joseph M. Hilbe. *Negative Binomial Regression*. 2nd. Cambridge University Press, 2011 (cit. a p. 31).

[7] Antonio Di Crescenzo e Franco Pellerey. «Some Results and Applications of Geometric Counting Processes». In: *Methodology and Computing in Applied Probability* 21 (2019) (cit. a p. 33).

[8] Kenneth P. Burnham e David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd. Springer, 2002 (cit. a p. 33).

[9] Frank J. Massey Jr. «The Kolmogorov-Smirnov test for goodness of fit». In: *Journal of the American Statistical Association* 46.253 (1951) (cit. a p. 33).

[10] Geoffrey J. McLachlan e David Peel. *Finite Mixture Models*. John Wiley & Sons, 2000 (cit. a p. 34).

[11] Roland Pfister, Katharina A. Schwarz, Markus Janczyk, Rick Romano e Jonathan B. Freeman. «Good things peak in pairs: a note on the bimodality coefficient». In: *Frontiers in Psychology* 4 (2013) (cit. a p. 34).

[12] Philippe Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 2007 (cit. a p. 40).

[13] Basel Committee on Banking Supervision. *International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version*. Rapp. tecn. Basel, Switzerland: Bank for International Settlements, giu. 2006 (cit. a p. 40).

[14] Alexander J. McNeil, Rüdiger Frey e Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Revised. Princeton, NJ: Princeton University Press, 2015 (cit. a p. 40).

[15] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber e David Heath. «Coherent Measures of Risk». In: *Mathematical Finance* 9.3 (1999) (cit. alle pp. 41, 43).

[16] Warren S. Sarle. *Cubic Clustering Criterion*. SAS Technical Report A-108. Cary, NC, USA: SAS Institute Inc., 1983 (cit. a p. 46).

[17] J. L. Teugels. «Some representations of the multivariate Bernoulli and binomial distributions». In: *Journal of Multivariate Analysis* 32.2 (1990) (cit. alle pp. 59, 63).

[18] *Distribuzioni Bernoulliane Bivariate*. Dispensa del corso, Materiale didattico. Documento inedito (cit. a p. 59).

[19] R. H. Byrd, P. Lu, J. Nocedal e C. Zhu. «A limited memory algorithm for bound constrained optimization». In: *SIAM Journal on Scientific Computing* 16.5 (1995) (cit. alle pp. 67, 71, 73).

[20] C. Zhu, R. H. Byrd, P. Lu e J. Nocedal. «Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization». In: *ACM Transactions on Mathematical Software* 23.4 (1997) (cit. a p. 67).

[21] Sandra Pieraccini. *Numerical Optimization for Large Scale Problems*. Appunti del corso. Materiale didattico. 2024 (cit. alle pp. 68, 72).

[22] J. Nocedal e S. J. Wright. *Numerical Optimization*. 2nd. Springer Series in Operations Research and Financial Engineering. Springer, 2006 (cit. a p. 73).