UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA TRIENNALE IN STATISTICA, POPOLAZIONE E SOCIETÀ

TESI DI LAUREA

ANALISI DI MODELLI STATISTICI PER LA PREVISIONE DEL NUMERO DI RETI SEGNATE DA UNA SQUADRA DI CALCIO

ANALYSIS OF STATISTICAL MODELS FOR FORECASTING NUMBER OF GOALS SCORED BY A FOOTBALL TEAM

RELATORE: CH.MO PROF. STUART GEORGE COLES

LAUREANDO: TADDEO MAURO

MATRICOLA 553620

ANNO ACCADEMICO 2008-09

Indice

Introduzione

Per l'uomo primitivo l'attività fisica consisteva nell'allenare le proprie capacità fisiche in vista della lotta contro l'ambiente circostante o contro i propri simili: il premio di questa sfida era la sopravvivenza.
Con lo sviluppo l'uomo riuscì a vincere la sfida con la natura; nonostante questo, egli continuò a sviluppare e testare le proprie doti fisiche tramite esercitazioni, inizialmente individuali e successivamente collettive. Il passaggio alle esercitazioni collettive e la nascita dello spirito agonistico tra simili diede origine allo sport. L'esercizio più diffuso inizialmente fu la corsa; ma l'originalità dell'uomo e il ricordo della primitiva lotta per la sopravvivenza portarono ad estendere la varietà delle discipline. Nacquero così i salti e i lanci, il nuoto, la canoa, la lotta, il pugilato...
A seconda delle epoche storiche e delle civiltà cambiarono le discipline di riferimento; tuttavia per giungere ad un nozione moderna di sport bisogna fare riferimento all'antica Grecia e alla nascita dei Giochi Olimpici (776 a.C.), che tuttora rimangono l'evento sportivo più importante al mondo. In questa epoca si diffusero anche i primi sport di squadra.
Anche se i Giochi nel 393 furono soppressi, lo spirito agonistico rimase vivo durante l'epoca romana caratterizzandosi per una esasperazione della violenza.
Messa al bando dal Romanticismo e dall'Illuminismo, nell'Ottocento la cultura sportiva ebbe un forte rilancio grazie alla cultura anglosassone.
Le scommesse si diffusero assieme all'evento sportivo. Accanto al piacere di assistere ad una gara gli uomini iniziarono a volere provar l'ebbrezza del rischio e l'esperienza del gioco d'azzardo. L'evoluzione degli eventi sportivi è stata accompagnata da una evoluzione del sistema di scommesse: da un sistema basato sull'oralità si è passati ad uno regolato da norme, col passare del tempo, sempre più precise e rigide. La fase conclusiva di questo processo fu la sua entrata nel contesto della legalità facendo sì che da passatempo riservato a pochi esso diventasse un fenomeno di costume regolato dallo Stato.
Gli scommettitori in epoca moderna hanno mostrato interesse verso moltissimi eventi, di natura diversa, non solo sportivi, come le corse dei cavalli e i risultati

elettorali. Tuttavia l'evento preferito su cui gli scommettitori hanno preferito puntare rimane la partita di calcio. Qualsiasi appassionato di calcio ha almeno una volta scommesso sul risultato di un incontro allo scopo di aggiungere un pizzico di divertimento al semplice piacere di guardare le squadre in campo. La storia del fenomeno delle scommesse cambia di nazione in nazione: Paese per antonomasia della scommessa è l'Inghilterra. Per quanto riguarda in Italia negli anni '90 è emerso il Totocalcio, la famosa schedina, che è diventata nel tempo, per il numero sempre maggiore di appassionati, un vero e proprio fenomeno di costume nazionale.

Ogni gioco d'azzardo ha però la necessità che il sistema funzioni: il banco deve vincere. Per questa ragione sono stati compiuti dei tentativi per prevedere con la massima precisione possibile il risultato di una partita. La determinazione delle varie probabilità da un lato permette che il sistema stia in piedi, dall'altro determina le quote per ciascun risultato di ogni partita.

Queste previsioni vengono effettuate sulla base di metodi statistici. Questi ultimi tengono conto esclusivamente delle partite giocate dalle due squadre in campo dando peso maggiore a quelle più recenti.

La tesi utilizza diverse distribuzioni di probabilità per prevedere i risultati delle partite di calcio. Lo scopo è quello di valutare quale di queste note distribuzioni di probabilità riesca meglio a descrivere i gol segnati da una squadra di calcio.

Un metodo statistico complesso ha lo scopo di studiare congiuntamente i gol segnati dalla squadra di casa e i gol segnati da quella in trasferta. La mia tesi si concentra sugli aspetti preliminari, analizzando marginalmente i gol segnati da una squadra senza considerare né il team avversario né la dipendenza tra il fenomeno dei gol segnati dalla squadra di casa e il fenomeno dei gol segnati dalla squadra in trasferta. Lo studio della dipendenza congiunto viene analizzato nell'ultima parte tramite l'utilizzo di una variabile normale bivariata. Nei restanti casi, il fenomeno di interesse è il numero di gol segnati da una singola squadra.

Per le stime dei parametri si è utilizzato un campione composto da circa 11.500 partite di trenta diverse nazioni del mondo. Nell'utilizzo di questi dati il fattore campo è stato l'unica discriminante. In altri termini, si è analizzato in modo disgiunto la distribuzione di probabilità per la squadra di casa e per la squadra in

trasferta, senza considerare le reali capacità delle squadre in campo. Per alcune analisi sono state selezionate partite giocate in una determinata nazione.

La tesi è composta da cinque parti: la prima parte si concentra sul modello di Poisson, ossia quello tradizionalmente usato dalle agenzie di scommesse per prevedere i risultati delle partite. Con lo scopo di una migliore analisi è stato inserito un esempio che mostra come, usando il modello di Poisson, sia possibile tener conto delle reali capacità delle squadre in campo. L'esempio evidenzia inoltre come sia fondamentale, per la previsione del risultato della partita tramite un modello statistico, considerare il fattore campo.

La seconda parte della tesi utilizza la variabile Gaussiana. Dopo avere effettuato alcune trasformazioni, si è potuta sfruttare la variabile statistica più famosa al mondo per descrivere questo fenomeno.

La terza parte si concentra sulla distribuzione Gamma. In seguito alla sua trasformazione in variabile discreta, si è potuto sperimentare la distribuzione Gamma per la descrizione del fenomeno analizzato globalmente nella tesi.

La parte successiva si sofferma sul modello di Weibull. Anche in questo caso si rende necessaria una discretivizzazione affinché la variabile possa descrivere il numero di gol fatti da una squadra in un incontro di calcio.

Giunti a questo punto, dopo aver analizzato le quattro distribuzioni separatamente, vi è un confronto tra esse.

L'introduzione della variabile normale bivariata conclude la tesi. Essa ha lo scopo di studiare la dipendenza tra i fenomeni "gol segnati dalla squadra di casa" e "gol segnati dalla squadra in trasferta", che nel resto della tesi venivano considerati come fenomeni indipendenti.

Il modello Poisson

Introduzione

Il modello utilizzato tradizionalmente per descrivere i gol segnati da una squadra di calcio è quello che si basa sulla distribuzione di Poisson.

La distribuzione di Poisson esprime la probabilità che si verifichi un numero finito di eventi in un periodo di tempo fissato a priori con l'ipotesi che essi accadano indipendentemente l'uno dall'altro.

La mia ricerca prende spunto dall' articolo "Modelling Association Football Scores and Inefficiencies in the Football Betting Market" pubblicato nel 1997 dal professor Mark Dixon e dal professor Stuart Coles. L'articolo, oltre ad introdurre il modello di Poisson, spiega come tener conto delle diverse capacità delle squadre in campo e della dipendenza tra i gol segnati dall' una o dall'altra squadra.

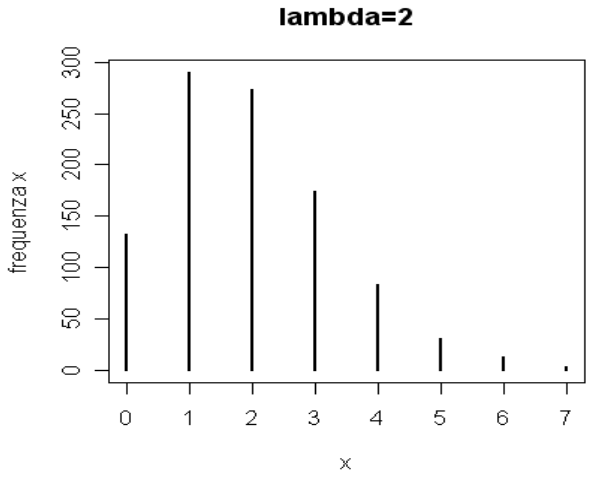
Il parametro di riferimento è λ. Le proprietà di questa variabile casuale sono che la sua media, mediana e varianza sono pari a λ.

Questo aspetto costituisce da un lato un vantaggio dal punto di vista della comodità nell'effettuare le stime del parametro, dall'altro una limitazione poiché la varianza della distribuzione potrebbe essere più o meno ampia rispetto alla media della stessa distribuzione.

Un ulteriore vantaggio di questo modello è rappresentato dal metodo di stima. Il parametro infatti si può stimare con metodo esatto ed ha valore pari alla media aritmetica. Questo costituisce un indubbio risparmio dal punto di vista della complessità dei calcoli e nella velocità con la quale si possono ottenere le stime.

La distribuzione di probabilità per una variabile $X \sim Poisson(\lambda)$ è rappresentato nel grafico 1.

Grafico 1: distribuzione di frequenza della variabile X con lambda fissato a priori



$$P(X=0)=\frac{\lambda^0 \cdot e^{-\lambda}}{0!}$$

$$P(X=1)=\frac{\lambda^1 \cdot e^{-\lambda}}{1!}$$

.

.

.

In generale:

$$P(X=k)=\frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Vediamo, tramite i grafici 2, 3 e 4, che cosa accade alla distribuzione di frequenza al variare del parametro λ. Per ogni esempio è stato creato un vettore di 1000 determinazioni della variabile avente distribuzione di Poisson.

Grafico 2: distribuzione di frequenza della variabile X con lambda fissato a priori

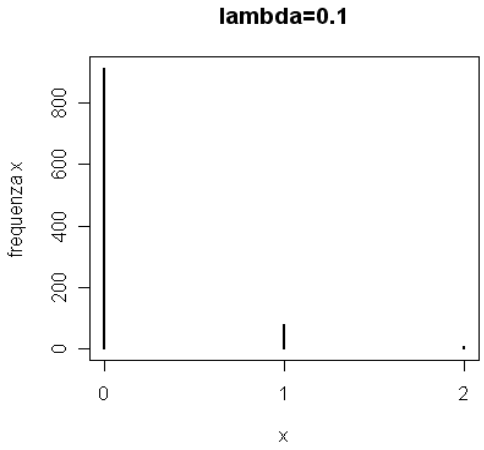


Grafico 3: distribuzione di frequenza della variabile X con lambda fissato a priori

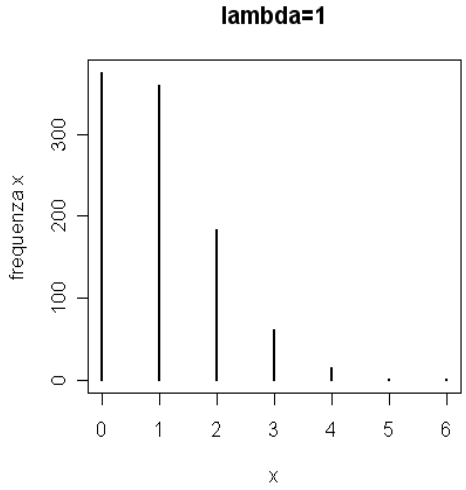
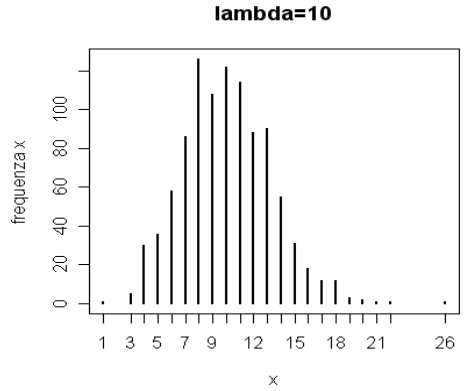


Grafico 4: distribuzione di frequenza della variabile X con lambda fissato a priori



Se λ è pari a 0.1, la quasi totalità dei mille numeri casuali generati tenderà ad assumere valori pari a 0. La probabilità che X assuma un valore differente da 0 e 1 è pari a 4.6%.

Nel secondo caso, pur rimanendo una forte presenza di 0 e di 1, la variabile assume altri valori.

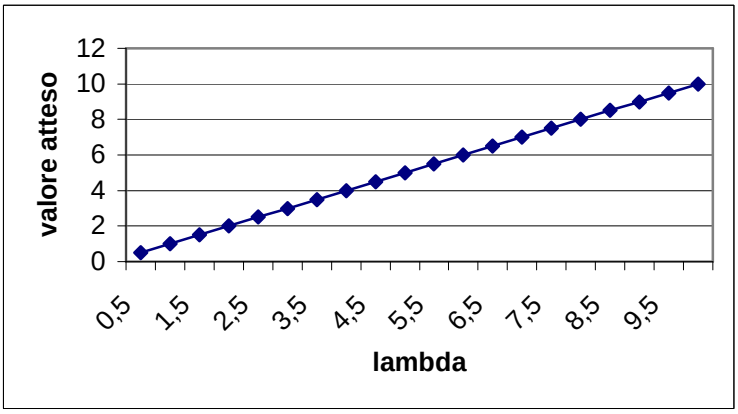
Nell'ultimo caso, cioè quando λ è pari a 10, accade che la distribuzione si sposti verso destra, che la mediana sia 10 e che la probabilità che X sia minore o uguale a 1 sia quasi nulla.

Valore atteso

Per una variabile casuale con distribuzione di Poisson, il valore atteso è pari al parametro di riferimento λ.

Chiaramente esiste una proporzionalità diretta tra il valore atteso e λ.

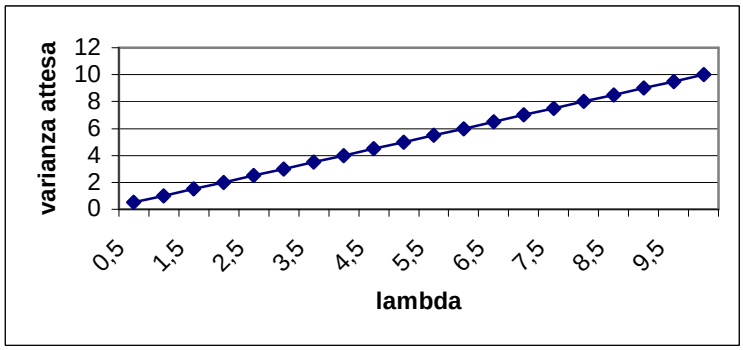
Grafico 5: andamento del valore atteso di una variabile con distribuzione di Poisson al variare di λ



Varianza attesa

Per una variabile casuale con distribuzione di Poisson, la varianza attesa è pari al parametro di riferimento λ.

Grafico 6: andamento della varianza atteso di una variabile con distribuzione di Poisson al variare di λ



Osserviamo quindi che il rapporto tra varianza attesa e valore atteso per questo modello è 1. Come precedentemente affermato, questa è una proprietà caratteristica del modello di Poisson che costituisce una forte limitazione, dovendo presumere la dipendenza tra media e varianza della distribuzione.

Funzione di verosimiglianza

La probabilità che la mia variabile assuma valore x è:

$$P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}.$$

Pertanto la mia funzione di verosimiglianza è:

$$P(X = x_1; x_2; ...; x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!} = \frac{e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} = L(\lambda)$$

Adesso per ricavare la stima di massima verosimiglianza di λ è necessario che io trovi il valore che annulla la derivata prima della funzione. Prima però è conveniente, tramite la trasformazione logaritmica, trasformare la funzione di verosimiglianza in funzione di log-verosimiglianza. Essendo una trasformazione monotona, il valore che annulla la derivata prima della funzione di log-verosimiglianza è lo stesso che annulla la funzione di verosimiglianza.

$$l(\lambda) = \log\left[L(\lambda)\right] = -n\lambda + \sum_{i=1}^{n} x_i \cdot \log(\lambda) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

La derivata prima della log-verosimiglianza rispetto a λ è:

$$\frac{\partial l(\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

Sto cercando la stima di massima verosimiglianza per λ ($\hat{\lambda}$), cioè il valore che annulla la derivata prima.

$$\frac{\partial l(\hat{\lambda})}{\partial \lambda} = 0 \text{ cioè } -n + \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}} = 0$$

E' immediato concludere che la stima di massima verosimiglianza per λ è:

$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ cioè la media aritmetica dei valori x assunti dalla variabile X.

Le stime

A questo punto diviene necessario stimare il valore dei parametri per ciascuna nazione.

La stima, come spiegato in precedenza, è ottenuta tramite il calcolo della media aritmetica dei gol segnati.

È stato inoltre calcolato l'intervallo di confidenza al 95% per ciascun valore stimato.

Si sottolinea che la stima del parametro in ciascun campionato non tiene conto delle reali capacità delle squadre in campo. L'unico fattore tenuto in considerazione riguarda se la squadra gioca in casa oppure fuori-casa. Inoltre, i fenomeni "gol segnati dalla squadra in casa" e "gol segnati dalla squadra fuori casa" sono considerati indipendenti, l'analisi è effettuata marginalmente.

I dati, come detto nella parte introduttiva, riguardano partite di calcio di trenta nazioni diverse. Come appare dalla tabella 1, esistono forti differenze nelle numerosità delle partite nelle varie nazioni.

Tabella 1: stima del parametro Lambda e del rispettivo intervallo di confidenza per la squadra 1 e numerosità campionaria in ciascuna nazione

Nazione	Limite inferiore	Stima Lambda	Limite superiore	numerosità
Argentina	1,349	1,475	1,601	337
Australia	0,7600	1,429	2,097	14
Austria	1,562	1,699	1,837	379
Belgio	1,506	1,622	1,739	490
Brasile	1,622	1,705	1,787	904
Cile	1,414	1,580	1,747	212
Croazia	1,108	1,846	2,584	26
Finlandia	1,395	1,588	1,780	383
Francia	1,264	1,343	1,422	835
Germania	1,501	1,588	1,676	976
Giappone	1,290	1,468	1,645	186
Grecia	1,306	1,512	1,718	170
Inghilterra	1,417	1,483	1,549	1388
Irlanda	1,230	1,374	1,519	300
Islanda	1,341	1,563	1,785	129
Italia	1,388	1,505	1,621	410
Messico	1,376	1,579	1,781	152
Norvegia	1,754	1,951	2,149	412
Olanda	1,728	1,839	1,949	644
Polonia	1,305	1,506	1,706	178
Portogallo	1,238	1,360	1,482	378
Rep.Ceka	1,144	1,359	1,574	103
Romania	1,252	1,391	1,530	253
Russia	1,386	1,513	1,639	399
Scozia	1,476	1,682	1,889	173
Spagna	1,293	1,389	1,485	630
Svezia	1,380	1,498	1,615	476
Svizzera	1,514	1,726	1,939	168
Turchia	1,345	1,557	1,769	149
Usa	1,036	1,380	1,724	50
generale	1,525	1,548	1,572	11304

I valori assunti dalla stima del parametro sono compresi tra 1.3 e 2. L'ampiezza degli intervalli di confidenza dipende in modo importante dal numero di partite a disposizione per il calcolo. L'intervallo risulta così particolarmente ampio per il campionato australiano (14 partite a disposizione) e croato (26 partite).

Grafico 7: stima del parametro Lambda e del rispettivo intervallo di confidenza per la squadra 1 in ciascuna nazione

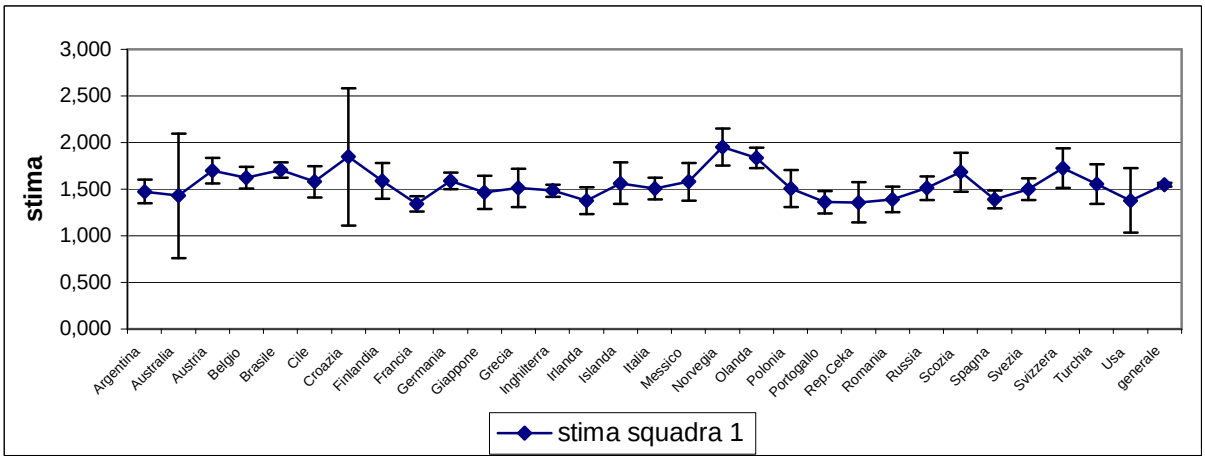
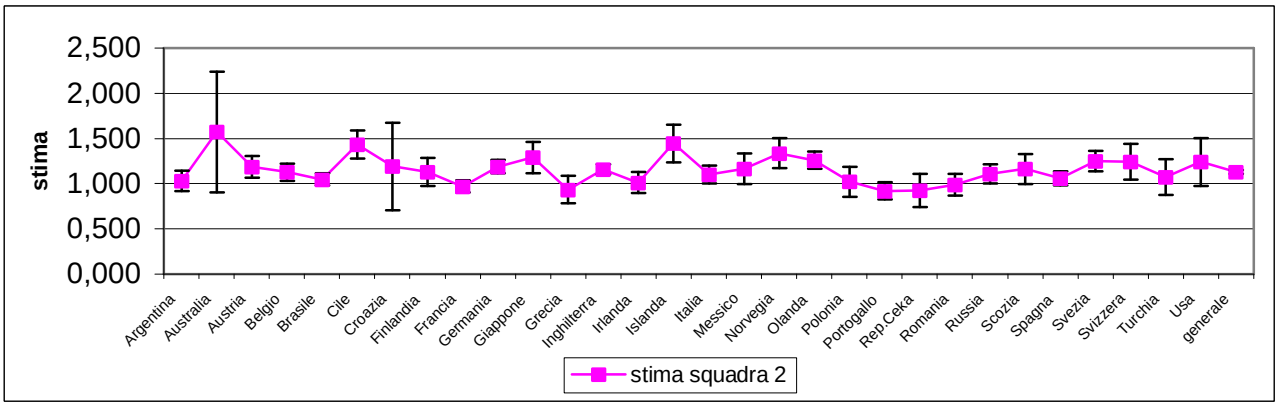


Tabella 2: stima del parametro Lambda e del rispettivo intervallo di confidenza per la squadra 2 in ciascuna nazione

Nazione	Limite inferiore	Stima Lambda	Limite superiore
Argentina	0,9179	1,030	1,141
Australia	0,9029	1,571	2,240
Austria	1,066	1,185	1,303
Belgio	1,030	1,127	1,223
Brasile	0,9795	1,048	1,116
Cile	1,281	1,434	1,587
Croazia	0,7074	1,192	1,677
Finlandia	0,9714	1,128	1,285
Francia	0,9025	0,9701	1,038
Germania	1,116	1,189	1,262
Giappone	1,117	1,290	1,463
Grecia	0,7815	0,9353	1,089
Inghilterra	1,101	1,158	1,215
Irlanda	0,8936	1,011	1,128
Islanda	1,238	1,444	1,651
Italia	1,003	1,102	1,202
Messico	0,9970	1,164	1,332
Norvegia	1,172	1,336	1,501
Olanda	1,165	1,259	1,353
Polonia	0,8561	1,022	1,189
Portogallo	0,8241	0,9206	1,017
Rep.Ceka	0,7394	0,9223	1,105
Romania	0,8691	0,9881	1,107
Russia	1,005	1,111	1,216
Scozia	0,9955	1,162	1,328
Spagna	0,9802	1,060	1,140
Svezia	1,136	1,250	1,364
Svizzera	1,044	1,244	1,444
Turchia	0,8791	1,074	1,269
Usa	0,9734	1,240	1,507
generale	1,108	1,128	1,149

I valori assunti dalla stima di lambda per la squadra 2 sono nella quasi totalità dei casi più bassi rispetto a quelli assunti dalla squadra 1.
Poiché all'aumentare di λ aumentano i gol attesi, è evidente che la squadra di casa abbia maggiori probabilità di segnare più reti della squadra in trasferta grazie al fattore campo.

Grafico 8: stima del parametro Lambda e del rispettivo intervallo di confidenza per la squadra 2 in ciascuna nazione



Si osservi che le nazioni con un intervallo di confidenza ampio sono le stesse che abbiamo trovato per la squadra 1. Questo aspetto è dovuto alla numerosità campionaria della nazione analizzata.

Ora è importante andare ad analizzare le differenze tra la distribuzione empirica dei gol segnati e quella che si ottiene tramite il nostro modello basato sulla distribuzione di Poisson.

Confronto fra la distribuzione empirica e la distribuzione ottenuta tramite modello

Giunti a questo punto, è stata confrontata la distribuzione empirica campionaria con la distribuzione simulata tramite il modello di Poisson. Lo scopo è vedere se la distribuzione osservata nel campione possa essere descritta opportunamente da una distribuzione ottenuta dal modello di Poisson avente come parametro il valore stimato.

La probabilità empirica per un dato evento è ottenuta come rapporto tra il numero di eventi di interesse nel campione e numero totale di partite appartenenti al campione. Quindi, ad esempio, la probabilità che la squadra di casa nel campionato italiano segni 2 gol si ottiene dal rapporto tra il numero di partite del campione giocate in Italia in cui la squadra di casa segna 2 reti e il numero totale di partite del campione giocate in Italia.

La distribuzione simulata tramite modello si ottiene invece con il calcolo della probabilità esatta tramite la formula $P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$ in cui x assume il valore dell'evento di interesse e lambda il valore precedentemente stimato per quel determinato campione.

Nel nostro esempio, squadra 1 italiana che segna 2 gol, x è pari a 2 e lambda vale 1,505.

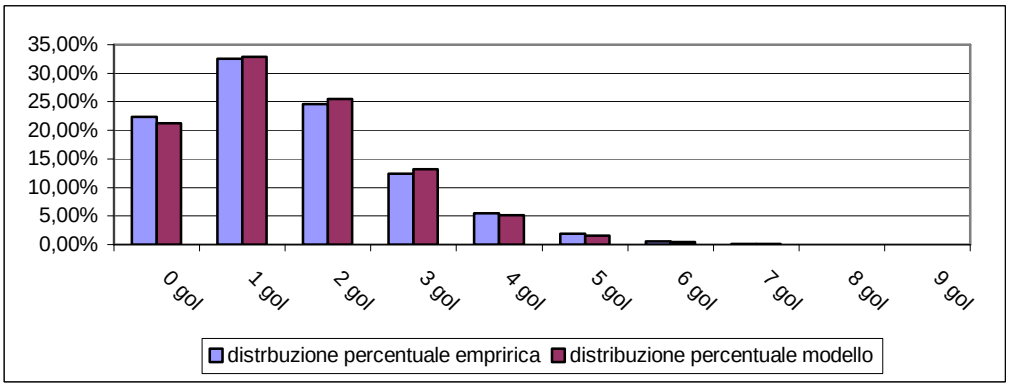
Tabella 3: frequenza percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	Frequenza empirica	Frequenza modello
0 gol	22,40%	21,27%
1 gol	32,54%	32,92%
2 gol	24,60%	25,48%
3 gol	12,39%	13,15%
4 gol	5,43%	5,09%
5 gol	1,90%	1,58%
6 gol	0,55%	0,41%
7 gol	0,11%	0,09%
8 gol	0,05%	0,02%
9 gol	0,02%	0,00%

Il nostro modello riesce a descrivere l'andamento in modo corretto. L'errore per ciascuna classe è inferiore al punto percentuale. Possiamo quindi ipotizzare che il modello così costruito sia in grado di descrivere il fenomeno.

Questo buon adattamento ai dati ci viene confermato dal seguente grafico.

Grafico 9: frequenza percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione simulata tramite modello



Tuttavia un buon adattamento potrebbe dipendere dai dati utilizzati per il confronto tra distribuzione empirica e simulata.
Utilizziamo ora i dati a disposizione relativi alla squadra 2.

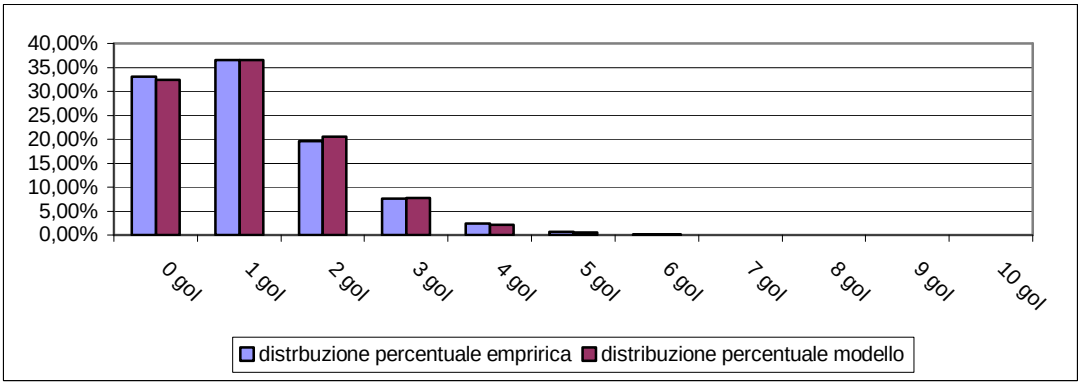
Tabella 4: frequenza percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	Frequenza empirica	Frequenza modello
0 gol	33,08%	32,37%
1 gol	36,49%	36,51%
2 gol	19,54%	20,59%
3 gol	7,62%	7,74%
4 gol	2,39%	2,18%
5 gol	0,61%	0,49%
6 gol	0,17%	0,09%
7 gol	0,05%	0,01%
8 gol	0,02%	0,00%
9 gol	0,02%	0,00%
10 gol	0,01%	0,00%

Anche in questo caso la distribuzione simulata ripropone l'andamento della distribuzione empirica.

Tuttavia è presente una sottostima per l'evento "0 gol" bilanciata da una sovrastima per l'evento "2 gol". Inoltre la coda della nostra distribuzione fa pesare troppo poco gli eventi caratterizzati da un alto numero di reti.

Grafico 10: frequenza percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione simulata tramite modello



Il grafico permette di evidenziare il problema relativo alle classi vicine all'origine.

Proviamo ad analizzare cosa accade al variare del parametro lambda: scegliamo la distribuzione caratterizzata dal valore stimato più alto: i gol segnati dalla squadra di casa nel campionato norvegese.

Tabella 5: frequenza percentuale del numero di gol segnati dalla squadra 1 nel campionato norvegese per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	Frequenza empirica	Frequenza modello
0 gol	17,70%	14,21%
1 gol	26,11%	27,73%
2 gol	23,45%	27,05%
3 gol	18,58%	17,59%
4 gol	7,08%	8,58%
5 gol	4,42%	3,35%
6 gol	2,65%	1,09%
7 gol	0%	0,30%
8 gol	0%	0,07%

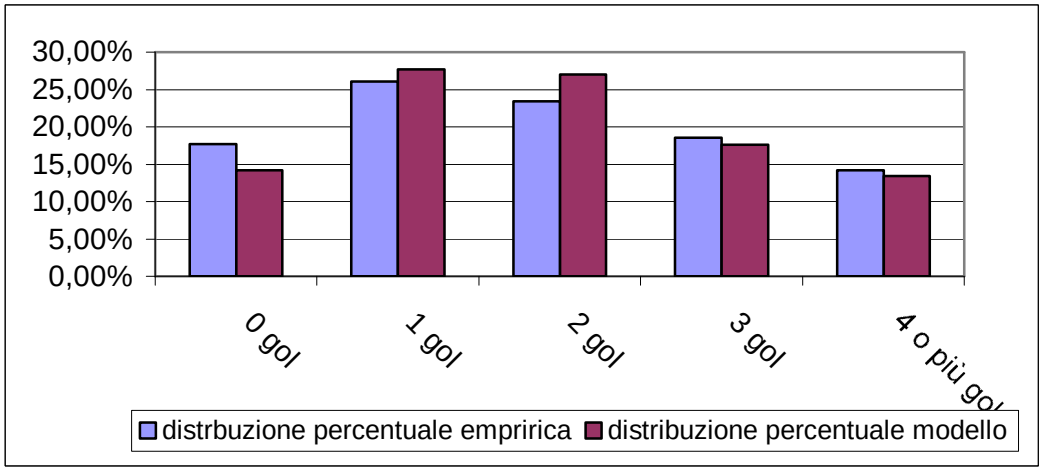
Per gli eventi vicini all'origine osserviamo delle probabilità nella distribuzione tramite modello troppo basse.

Anche in questo caso vi è una distorsione nella coda. Tuttavia nell'esempio precedente è presente un sottodimensionamento delle probabilità per un alto numero di gol mentre in questo esempio è presente un sovradimensionamento. È bene sottolineare che si tratta comunque di un numero limitato di casi e che la somma delle frequenze percentuali relative ad un numero maggiore o uguale a 4 gol è quasi la stessa.

Tabella 6: frequenza percentuale del numero di gol segnati dalla squadra 1 nel campionato norvegese per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	Frequenza empirica	Frequenza modello
0 gol	17,70%	14,21%
1 gol	26,11%	27,73%
2 gol	23,45%	27,05%
3 gol	18,58%	17,59%
4 o più gol	14,16%	13,39%

Grafico 11: frequenza percentuale del numero di gol segnati dalla squadra 1 nel campionato norvegese per la distribuzione empirica e per la distribuzione simulata tramite modello



Dal grafico 11 è possibile quindi apprezzare il problema di questo esempio: sottostima dell' evento "0 gol" e sovrastima dei successivi due eventi.

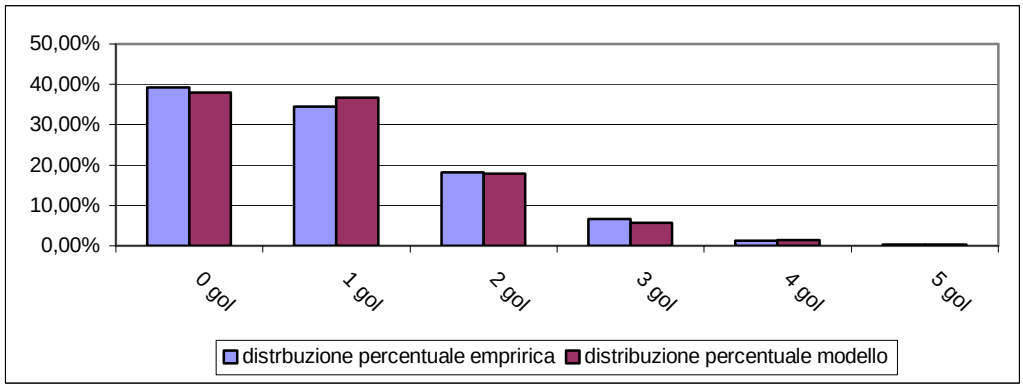
Che cosa accade se invece il parametro lambda assume un valore stimato particolarmente basso. Proviamo adesso ad andare in Francia per vedere che cosa accade alla distribuzione dei gol per la squadra in trasferta.

Tabella 7: frequenza percentuale del numero di gol segnati dalla squadra 2 nel campionato francese per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	Frequenza empirica	Frequenza modello
0 gol	39,16%	37,90%
1 gol	34,49%	36,77%
2 gol	18,20%	17,84%
3 gol	6,71%	5,77%
4 gol	1,20%	1,40%
5 gol	0,24%	0,27%

È subito evidente che vi è uno sfasamento per i primi due eventi: l'evento "0 gol" nel nostro modello pesa troppo poco rispetto alla distribuzione empirica. Viceversa l'evento "1 gol" ha una frequenza percentuale troppo alta rispetto a ciò che accade nella realtà.

Grafico 12: frequenza percentuale del numero di gol segnati dalla squadra 2 nel campionato francese per la distribuzione empirica e per la distribuzione simulata tramite modello



Per concludere, possiamo sottolineare che il modello ottenuto con una distribuzione di Poisson riesce a descrivere in modo opportuno il fenomeno studiato.
Il problema principale riscontrato riguarda la stima della frequenza percentuale per un basso numero di gol. Per questa ragione, infatti, il perfezionamento di questo modello prevede una correzione per le partite caratterizzate da un basso numero di reti (0-0, 1-0, 0-1, 1-1). Questa correzione ha lo scopo di considerare la dipendenza tra i due fenomeni studiati

Simulazione di dati

L'analisi del modello di Poisson prosegue con una simulazione dei dati che ha lo scopo di illustrare come sia possibile, tramite questo modello, considerare le reali capacità delle squadre in campo.
Questa è l'unica parte della tesi in cui si differenziano le squadre non solo in base al fatto che giochino in casa o fuori casa ma anche sulla base delle reali caratteristiche.
Prendendo spunto dall'articolo "Modelling Association Football Scores and Inefficiencies in the Football Betting Market", ho deciso di introdurre nella tesi un esempio di come il modello di Poisson sia in grado di descrivere le differenze dovute alla forza di un team, distinguendo tra capacità offensiva e capacità difensiva.
In questo modo è possibile anche tenere conto del fatto che se una squadra forte in attacco gioca contro una squadra debole in difesa tenderà ad avere una alta probabilità di segnare tante reti. In questo modo è quindi possibile tenere conto della reciproca dipendenza tra le capacità difensive ed offensive della squadre in campo.

Supponiamo che X sia la variabile che descrive i gol segnati dalla squadra i quando gioca contro la squadra j.
Una distribuzione poissoniana dipende da un unico parametro λ.
Quindi:

$$X_{i,j} \sim \text{Poisson}(\lambda_x)$$

Resta però ora da capire da che cosa dipenda il parametro λ_x.

Anche i non esperti di calcio potrebbero giungere alla conclusione che il numero di gol segnati da una squadra dipende dall' abilità in attacco della squadra in questione e dall'abilità in difesa della squadra avversaria. Pertanto possiamo scomporre λ in prodotto di fattori che descrivano l' abilità offensiva della squadra studiata e l'abilità difensiva della squadra che le si oppone.

Pertanto λ_x dipende dal parametro α_i che descrive la forza in attacco della squadra *i* e dal parametro β_j che descrive la capacità in difesa della squadra *j*.

Per tutte queste premesse, è evidente che una squadra forte in attacco, quindi con α avente valore elevato, che si trova di fronte una squadra scarsa in difesa, con un β avente valore alto, avrà una probabilità alta di fare un elevato numero di reti.

Viceversa una squadra già di suo scarsa, valore di α basso, contro una squadra abile in difesa, valore di β basso, difficilmente riuscirà a segnare più di un gol.

Vediamo che cosa accade simulando un quadrangolare tra squadre con grosse differenze nella abilità offensive e difensive.

In sintesi per ciascun match avremo:

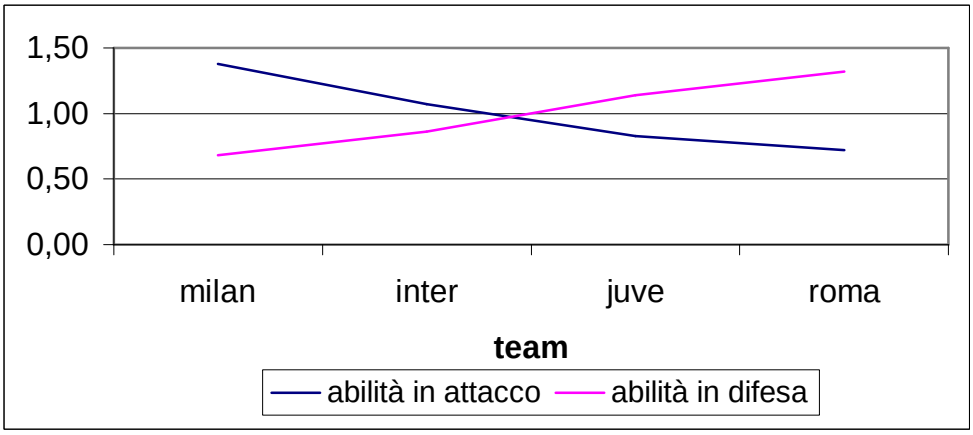
$$\lambda_x = \alpha_i \times \beta_j$$

$$\lambda_y = \alpha_j \times \beta_i$$

Tabella 8: valori assunti dai parametri α e β per ciascuna squadra

team	α	β
milan	1,38	0,68
inter	1,07	0,86
juve	0,83	1,14
roma	0,72	1,32

Grafico 13: abilità in attacco e abilità in difesa al variare della squadra



Notiamo innanzi tutto che la somma dei parametri sia di attacco che di difesa è pari a n (il numero delle squadre considerate). Questo è necessario per evitare la sovra-parametrizzazione del modello.

Inoltre è chiaro che le squadre sono disposte in ordine decrescente sia per il parametro α, cioè abilità offensiva, che β, abilità difensiva. In altre parole, ci aspettiamo che il Milan stravinca questo quadrangolare.

Infatti, se andiamo a far disputare uno di questi incontri virtuali, ad esempio Milan – Roma ci rendiamo conto che:

$X_{milan-roma} \sim$ Poisson(λ_x) con $\lambda_x = \alpha_{milan} \cdot \beta_{roma} = 1.82$

$Y_{milan-roma} \sim$ Poisson (λ_y) con $\lambda_y = \alpha_{roma} \cdot \beta_{milan} = 0.49$

Per questa ragione ci aspettiamo che la partita finisca con una vittoria milanista. Simulando il match, come previsto, otteniamo un 2-0.

Andando a simulare interamente il torneo, otteniamo, come prevedibile, la stessa classifica che avevamo per i nostri parametri.

Si sottolinea che è presente una ipotesi di indipendenza tra il fenomeno "gol team 1" ed il fenomeno "gol team 2". In realtà l'evidenza empirica suggerisce il contrario. Tuttavia in questa analisi preliminare i due fenomeni sono considerati indipendenti.

Tabella 9: gol segnati dalla squadra 1 e gol segnati dalla squadra 2 per ciascuna partita del quadrangolare

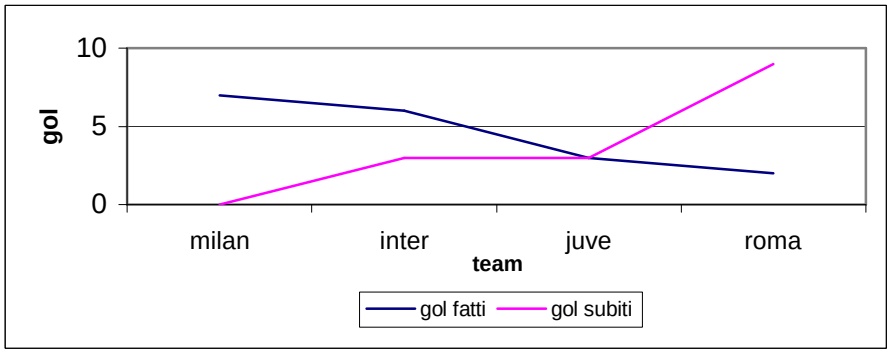
partita	gol team 1	gol team 2
milan-inter	2	0
milan-juve	3	0
milan-roma	2	0
inter-juve	1	1
inter-roma	5	0
juve-roma	2	2

La classifica è:

Tabella 10: classifica del quadrangolare

team	punti	gol fatti	gol subiti
milan	9	7	0
inter	4	6	3
juve	1	3	3
roma	1	2	9

Grafico 14: gol segnati e gol subiti da ciascuna squadra nel quadrangolare



Fin qui abbiamo simulato degli incontri tramite alcune determinazioni casuali della variabile Z, ottenuta come è stato descritto precedentemente.
Concentriamoci invece su una singola partita: milan-roma.

Abbiamo già descritto che il milan aveva, sulla base delle nostre ipotesi, molte più chance di vincere proprio perché i suoi parametri erano vantaggiosi sia in attacco che in difesa.
Andiamo a studiare le probabilità esatte per ciascun risultato.

Tabella 11: probabilità percentuale per ciascun risultato dell'incontro milan-roma

gol milan		**0**	**1**	**2**	**3**	**4**	**5**	**6**
gol roma	totale	16,2%	29,5%	26,8%	16,3%	7,42%	2,7%	0,82%
0	60,9%	9,85%	17,9%	16,3%	9,92%	4,52%	1,65%	0,5%
1	30,2%	4,89%	8,9%	8,1%	4,92%	2,24%	0,82%	0,25%
2	7,50%	1,21%	2,2%	2,01%	1,22%	0,56%	0,2%	0,06%
3	1,2%	0,2%	0,3%	0,33%	0,2%	0,09%	0,03%	0,01%
4	0,15%	0,02%	0,04%	0,04%	0,02%	0,01%	0%	0%
5	0,02%	0%	0%	0%	0%	0%	0%	0%
6	0%	0%	0%	0%	0%	0%	0%	0%

Vediamo come la probabilità che la partita finisse 2-0, risultato ottenuto simulando l'incontro, era del 16,3%.
Il risultato esatto più probabile, al contrario, era 1 a 0 per i rossoneri.
Calcoliamo invece le probabilità, per l'esito di questo incontro, ottenute come somma delle probabilità per i risultati esatti:

Tabella 12: probabilità percentuale per ciascun esito dell'incontro milan-roma

esito	**probabilità**
vittoria milan	69,29%
pareggio	20,97%
vittoria roma	9,74%

A questo punto è utile chiedersi se il nostro modello possa essere migliorato.
Qualunque persona appassionata di sport sa che la squadra che gioca sul proprio campo ha un indubbio vantaggio sulla squadra ospite.

Per questo motivo un modello che vuole prevedere i risultati di una partita di calcio dovrà tenere conto di questo aspetto.
Ho deciso quindi di introdurre un modello che tenga conto del fattore campo.
Per questo motivo va inserito un ulteriore parametro che tenga conto di quale delle due squadre giochi sul proprio campo. Il parametro utilizzato è γ (maggiore di 0). In realtà γ avrà quasi certamente un valore maggiore di 1 poiché giocare in casa, come detto, costituisce un vantaggio.
E' importante sottolineare che nella restante parte della tesi le squadre vengono distinte a seconda che giochino in casa oppure fuori casa. In questo caso, invece, teniamo conto della diversa capacità di ciascuna squadra di sfruttare il fattore campo.
Da questo momento in poi, quindi, la squadra *i* può essere correttamente chiamata squadra di casa e la squadra *j* squadra in trasferta.
I nuovi parametri per le due distribuzioni Poissoniane diventano:

$$\lambda_x = \alpha_i \times \beta_j \times \gamma$$

$$\lambda_y = \alpha_j \times \beta_i$$

Andiamo a simulare di nuovo il campionato introducendo quindi il parametro γ.

Tabella 13: valori assunti dai parametri α, β e γ per ciascuna squadra

team	α	β	γ
milan	1,38	0,68	2,00
inter	1,07	0,86	1,00
juve	0,83	1,14	0,50
roma	0,72	1,32	0,50

Vediamo come nell'esempio il Milan abbia un vantaggio nel giocare in casa, mentre per Roma e Juve giocare tra le mura amiche costituisce un fattore sconveniente.
Il fatto che γ assuma valore minore di uno, come detto, è possibile da un punto di vista teorico, ma è difficilmente riscontrabile empiricamente.

Anche per questo esempio c'è l'ipotesi di indipendenza tra i fenomeni "gol casa andata", "gol fuori andata", gol casa ritorno" e "gol fuori ritorno"

Tabella 14: gol segnati dalla squadra 1 e gol segnati dalla squadra 2 per ciascuna partita del quadrangolare

	ANDATA		**RITORNO**	
partita	**gol casa (team 1)**	**gol fuori (team 2)**	**gol casa (team 2)**	**gol fuori (team 1)**
milan-inter	3	1	1	0
milan-juve	2	0	0	0
milan-roma	3	1	1	1
inter-juve	5	2	0	1
inter-roma	1	2	1	0
juve-roma	2	0	1	1

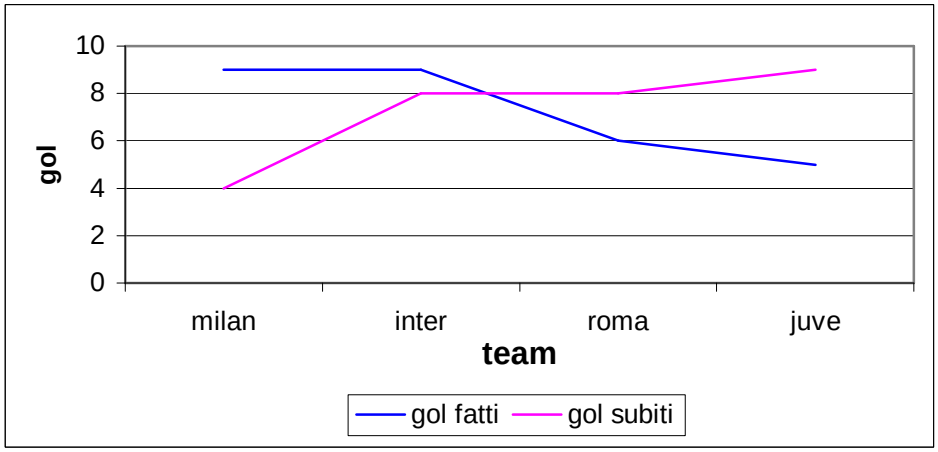
Come ci aspettavamo, l'andamento del milan è certamente più positivo tra le mura amiche.

I rossoneri ottengono 9 punti su 11 in casa.

Tabella 15: classifica del quadrangolare

classifica	**punti**	**gol fatti**	**gol subiti**
milan	11	9	4
inter	9	9	8
roma	8	6	8
juve	6	5	9

Grafico 15: gol segnati e gol subiti da ciascuna squadra nel quadrangolare



Questo grafico mostra che, come ci attendevamo, l'andamento dei gol fatti ripropone l'andamento dei parametri di attacco, mentre l'andamento dei gol subiti ripropone l'andamento dei parametri difensivi.

Studiamo le probabilità esatte per la partita milan-roma.

$X_{milan-roma} \sim$ Poisson(λ_x) con $\lambda_x = \alpha_{milan} \cdot \beta_{roma} \cdot \gamma = 3.64$

$Y_{milan-roma} \sim$ Poisson (λ_y) con $\lambda_y = \alpha_{roma} \cdot \beta_{milan} = 0.49$

Tabella 16: probabilità percentuale per ciascun risultato dell'incontro milan-roma

gol milan		0	1	2	3	4	5
gol roma	Totale	2,61%	9,53%	17,4%	21,1%	19,2%	14,0%
0	60,9%	1,59%	5,80%	10,6%	12,8%	11,7%	8,52%
1	30,2%	0,791%	2,88%	5,25%	6,37%	5,80%	4,23%
2	7,5%	0,196%	0,715%	1,302%	1,58%	1,441%	1,05%
3	1,2%	0,032%	0,118%	0,216%	0,262%	0,238%	0,176%
4	0,15%	0%	0,015%	0,027%	0,032%	0,030%	0,022%

gol milan	6	7	8	9	10	11	12
gol roma	8,50%	4,42%	2,01%	0,815%	0,287%	0,098%	0,030%
0	5,17%	2,69%	1,23%	0,496%	0,181%	0,060%	0,018%
1	2,57%	1,34%	0,608%	0,246%	0,090%	0,030%	0,009%
2	0,637%	0,332%	0,151%	0,061%	0,022%	0,007%	0,002%
3	0,105%	0,055%	0,025%	0,010%	0,004%	0,001%	0%
4	0,013%	0,007%	0,003%	0,001%	0%	0%	0%

La partita simulata finisce 3 a 1. La probabilità per questo risultato è del 6,37%: anche in questo caso il risultato che si è verificato virtualmente non aveva una probabilità molto elevata di verificarsi. Il risultato più probabile era 3 a 0. Rispetto a prima la probabilità di una goleada da parte del milan non è così remota. Addirittura in 1,23 casi su 100 esiste la possibilità che la partita finisca 8 a 0: per questa ragione si è ritenuto opportuno riportare la tabella fino al risultato di 12 a 4.

Tabella 17: probabilità percentuale per ciascun esito dell'incontro milan-roma

esito	probabilità
vittoria milan	91,74%
pareggio	6,07%
vittoria roma	2,14%

Oltre che ad abilità decisamente superiori, il milan può contare anche sul fattore campo. Questo fa sì che le sue chance di vittoria aumentino in maniera fortissima.

Studiamo che cosa accade al ritorno, quando cioè si gioca sul campo della Roma. Abbiamo già precisato che, contrariamente a quanto accede in realtà, il fattore campo influisce negativamente sulle prestazioni della squadra avendo un valore inferiore all'unità.

Tabella 18: probabilità percentuale per ciascun risultato dell'incontro milan-roma

gol milan		**0**	**1**	**2**	**3**	**4**	**5**	**6**
gol roma		16,2%	29,5%	26,8%	16,3%	7,42%	2,7%	0,82%
0	78,2%	12,6%	22,99%	20,94%	12,7%	5,79%	2,11%	0,640%
1	19,4%	3,13%	5,71%	5,20%	3,16%	1,44%	0,524%	0,159%
2	2,403%	0,389%	0,708%	0,645%	0,392%	0,178%	0,065%	0,020%
3	0,199%	0,032%	0,059%	0,053%	0,032%	0,015%	0,005%	0,002%
4	0,012%	0,002%	0,004%	0,003%	0,002%	0,001%	0%	0%

$X_{milan,roma} \sim$ Poisson(λ_x) con $\lambda_x = \alpha_{milan} \cdot \beta_{roma} = 1,82$

$Y_{milan,roma} \sim$ Poisson (λ_y) con $\lambda_y = \alpha_{roma} \cdot \beta_{milan} \cdot \gamma = 0,24$

L'incontro finisce 1 a 1. La probabilità che l'incontro finisse con questo punteggio era del 5,71%. Notiamo immediatamente che, nonostante giochi sul terreno avversario, la probabilità che vinca il milan è più alta della probabilità della vittoria romanista.

Tabella 19: probabilità percentuale per ciascun esito dell'incontro roma-milan

esito	**probabilità**
vittoria milan	76,62%
pareggio	19,00%
vittoria roma	4,38%

Questo è causato dal valore inferiore a 1 assegnato a γ della roma.

Una differenza rispetto a quanto accaduto con la prima simulazione, che non teneva conto del fattore campo, la roma in questo campionato, nonostante i parametri assegnati fossero peggiori, ottiene più punti della juve. Questo aspetto è dovuto esclusivamente al caso: una ulteriore simulazione di questo quadrangolare vede i bianconeri sopravanzare i giallorossi.
La differenza più interessante tra i due modelli è che senza tenere conto del fattore campo la roma aveva più probabilità di vincere la partita: 9,74%. Introducendo il fattore campo, al contrario, si ha il 2,15% quando la roma gioca sul terreno avversario e il 4,38% quando gioca in casa. Questo cambiamento è dovuto alla diminuzione del valore del parametro λ del milan piuttosto che all'aumento di quello della roma.

Conclusioni

Per concludere, possiamo sottolineare come il modello di Poisson sia caratterizzato da un unico parametro: questo se certamente costituisce un vantaggio dal punto di vista della semplicità nell'utilizzo del modello, per un altro verso non permette di considerare distintamente posizione e variabilità della distribuzione. Come detto precedentemente, il fatto che valore atteso e varianza attesa coincidano impedisce alla varianza di essere più o meno ampia della media.
Un ulteriore vantaggio è rappresentato dalla immediatezza e dalla facilità con cui possa essere stimato il parametro. Il fatto che la stima esatta di lambda sia pari alla media aritmetica degli eventi costituisce un notevole risparmio dal punto di vista del calcolo rispetto ad altri modelli caratterizzati da una notevole complessità nei metodi di stima.
Possiamo certamente affermare che il modello sembra ben adattarsi al fenomeno studiato. Tuttavia spesso vi sono delle distorsioni per gli eventi vicini all'origine, quelli cioè in cui la squadra segna pochi gol. Per questa ragione, un possibile perfezionamento per il modello prevede una funzione che corregge le probabilità per le partite con pochi gol.
L'analisi prosegue con la spiegazione di come sia possibile tenere in considerazione le reali capacità delle squadre in campo. Se nel resto della tesi si considerano le

squadre solo in base al fatto di giocare in casa o in trasferta, in questo capitolo vi è una parte che spiega come il modello di Poisson permetta di ottenere in modo intuitivo le probabilità a seconda delle maggiori o minori capacità difensive o offensive delle squadre in campo. Inoltre è possibile migliorare ulteriormente il modello prevedendo differenze tra squadre nel riuscire a sfruttare il fattore campo.
Non è stato trattato il problema della dipendenza tra eventi. L'analisi considera in modo marginale i gol fatti dalla squadra di casa e i gol fatti dalla squadra in trasferta sebbene i primi dipendano dalla capacità difensiva della squadra in trasferta e i secondi dipendano dalle capacità difensive della squadra di casa. Il problema della dipendenza per questo modello, nell'articolo da cui prende spunto la ricerca, viene risolto con l'introduzione di una funzione che corregga le probabilità per gli eventi 0-0, 1-0, 0-1, 1-1. Questa correzione nasce dalla considerazione che se una squadra tende a non segnare, anche l'altra avrà difficoltà a farlo a causa di un sostanziale equilibrio in campo. Come si vede nel confronto tra distribuzione empirica e distribuzione simulata, è necessaria una correzione delle probabilità per gli eventi con un basso numero di reti. La distorsione osservata è risolvibile quindi tramite l'introduzione di una funzione che permetta di tenere in considerazione la dipendenza tra i due fenomeni.

Il modello normale

Introduzione

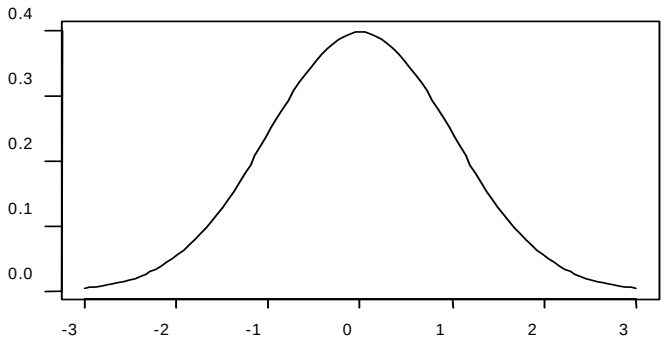
Il modello con distribuzione di Poisson è quello che tradizionalmente viene utilizzato per prevedere il numero di gol segnati da una squadra in un incontro di calcio.
Il problema di questo modello è di non potere considerare in modo distinto posizione e variabilità della distribuzione.
La tesi utilizza altri modelli basati su distribuzioni diverse per valutare se essi siano in grado di descrivere accuratamente il fenomeno di interesse.
Il primo modello statistico che verrà utilizzato per descrivere il numero di gol segnati da una squadra di calcio trae origine da una variabile X distribuita come una normale. I parametri di questo modello pertanto sono:

- μ cioè la media;
- σ^2 cioè la varianza.

$$X \sim Norm(\mu; \sigma^2)$$

Di seguito è riportata la distribuzione per una normale standard (media uguale a 0 e varianza uguale a 1).

Grafico 16: distribuzione di probabilità di una normale standard



Poiché ci interessa descrivere il numero di gol segnati da una squadra di calcio è necessario che la mia variabile assuma solamente valori positivi.
Per questa ragione creo, tramite una prima trasformazione, una nuova variabile.

$$Y = e^{X}$$

I parametri della nuova variabile Y sono:

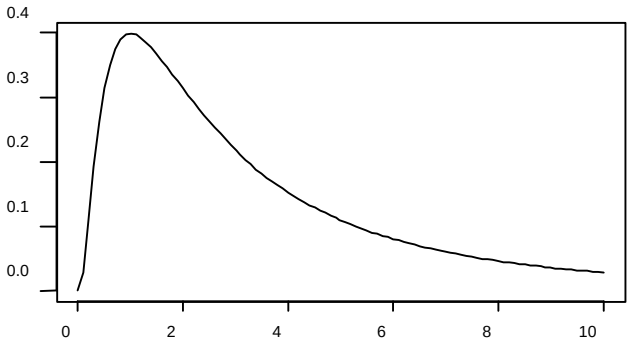
$$\mu_Y = e^{\mu+\frac{1}{2}\sigma^2}$$

$$\sigma_Y{}^2 = (e^{\sigma^2} - 1) * e^{2\mu+\sigma^2}$$

$$P(Y = y) = \phi\left(\frac{\log(y) - \mu}{\sigma}\right)$$

Assumendo solo valori positivi, la nuova variabile avrà una distribuzione asimettrica.

Grafico 17: distribuzione di probabilità della trasformazione esponenziale di una normale standard

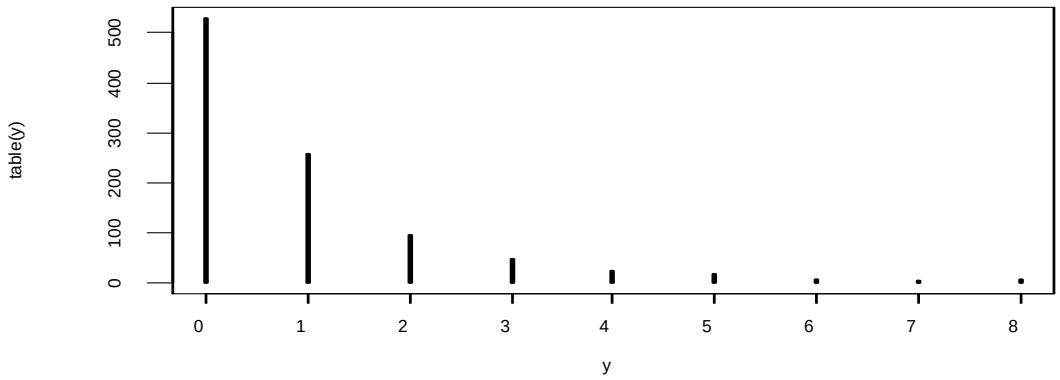


Ora però abbiamo un altro problema: la nostra variabile è continua. Al contrario, i gol assumono solamente valori interi: abbiamo bisogno quindi di una variabile discreta.
Per questo motivo abbiamo reso discreta la variabile Y.
Abbiamo pertanto una nuova variabile Z così definita:

$Z = j$ se e solo se Y: $j \le Y < j+1$

Per come abbiamo definito la nuova variabile Z, essa avrà come funzione di probabilità:

Grafico 18: distribuzione di probabilità della variabile Z ottenuta da una normale standard



$$P(Z=0)=P(0\le Y<1)=\phi\left(\frac{-\mu}{\sigma}\right)$$

$$P(Z=1)=P(1\le Y<2)=\phi\left(\frac{\log(2)-\mu}{\sigma}\right)-\phi\left(\frac{-\mu}{\sigma}\right)$$

$$P(Z=2)=P(2\le Y<3)=\phi\left(\frac{\log(3)-\mu}{\sigma}\right)-\phi\left(\frac{\log(2)-\mu}{\sigma}\right)$$

.

.

.

$$P(Z=100)=P(100\le Y<101)=\phi\left(\frac{\log(101)-\mu}{\sigma}\right)-\phi\left(\frac{\log(100)-\mu}{\sigma}\right)$$

.

.

.

In generale:

$$P(Z=j)=P(j\le Y<j+1)=\phi\left(\frac{\log(j+1)-\mu}{\sigma}\right)-\phi\left(\frac{\log(j)-\mu}{\sigma}\right)$$

Simulazione di dati

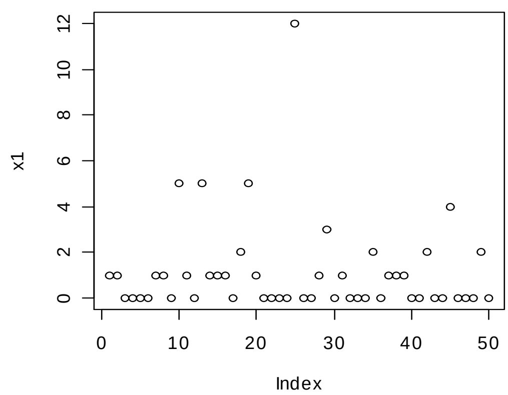
Osserviamo che cosa accade alla mia variabile Z facendo variare i parametri di partenza della normale.

$X \sim Norm(0;1)$

Le determinazioni di Z sono:

1 1 0 0 0 0 1 1 0 5 1 0 5 1 1 1 0 2 5 1 0 0 0 0 12 0 0 1 3 0
1 0 0 0 2 0 1 1 1 0 0 2 0 0 4 0 0 0 2 0

Grafico 19: distribuzione delle determinazione della variabile Z ottenuta da una normale standard



Vediamo che i valori per buona parte sono pari a 0. Questo perché con una distribuzione normale standard la probabilità che X assuma un valore inferiore a 1 è pari a 0,758.

$X \sim Norm(-5;1)$

Le determinazioni di Z sono:

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0

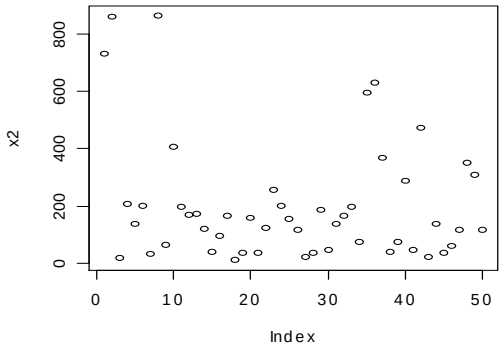
Tutti i valori sono pari a 0. Questo poiché la probabilità che X assuma un valore inferiore a 1 è quasi il 100%.

$X \sim Norm(5;1)$

Le determinazioni di Z sono:

730 857 17 205 135 200 30 862 62 405 197 169 172 118 40 95 166 11 34 158 36 121 254 199 154 114 21 36 187 46 137 164 195 75 592 628 368 39 74 287 44 473 23 135 36 58 117 350 309 116

Grafico 20: distribuzione delle determinazioni della variabile Z ottenuta da una normale con media 5 e varianza 1



I numeri casuali variano tra 21 e 857.

Studiamo ora che cosa accade se a variare non è più la media ma la varianza.

$X \sim Norm(0;10)$

Le determinazioni di Z sono:

22477 36 103070 0 0 1008 259 0 0 0 0 11 105 0 52 0 90953 11861 0 10190 23 0 0 2 30880 0 0 1806085 3253 0 0 0 0 0 6 2505 140 528 0 8 0 0 0 70617 539114 581 0 80837 0 0

In questo caso, come prevedibile, è presente un buon numero di 0 e contemporaneamente i valori in buona parte sono particolarmente alti.
Se invece andiamo a diminuire la variabilità della distribuzione vediamo come i dati assumano solamente valori pari a 0 o a 1.

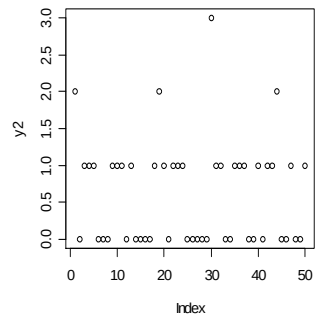
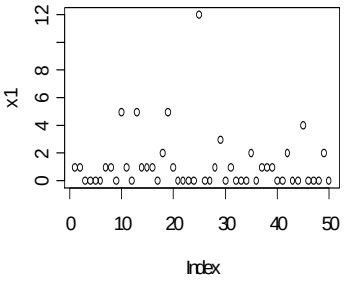
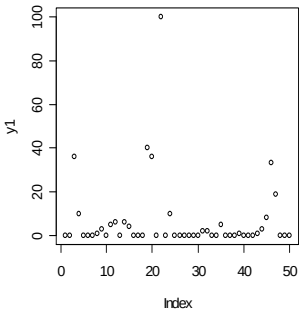
$X \sim Norm(0;0.1)$

Le determinazioni di Z sono:

0 1 1 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 0 1 0 1 1 1 0 1 0 1 0 0 0 1 1 0 0 1 1 0 1 1 0 1 1 0 0 0 0 1 1 1

A seconda del valore della varianza la distribuzioni dei numeri casuali sul piano cartesiano cambia notevolmente:

Grafico 21.a; 21.b; 21.c: distribuzione delle determinazione della variabile Z ottenuta da una normale con media 0 e varianza 0,5; 1; 2

$X \sim Norm(0;0.5)$	$X \sim Norm(0;1)$	$X \sim Norm(0;2)$
		

In tutti e tre i casi i valori tendono a concentrarsi sullo 0. Tuttavia una bassa variabilità, nel primo caso, obbliga i numeri casuali sull'unità, mentre una variabilità più alta consente ai numeri casuali di assumere valori molto alti: non sono rari valori superiori a 20, addirittura pari a 100,.

Valore atteso

Definiamo quindi le costanti caratteristiche della variabile aleatoria Z.

Il valore atteso della variabile X discreta è il numero definito da:

$$E(X) = \sum_{i=1}^{n} x_i \cdot f(x_i)$$

Il valore atteso della mia variabile Z sarà quindi:

$$E(Z)=\sum_{j=0}^{+\infty} j\cdot P(Z=j)=\sum_{j=0}^{+\infty} j\cdot\left(\phi\left(\frac{\log(j+1)-\mu}{\sigma}\right)-\phi\left(\frac{\log(j)-\mu}{\sigma}\right)\right)$$

Essendo definita da un numero infinito di addendi, il valore atteso può essere approssimato come:

$$E(Z)\simeq\sum_{j=0}^{100} j\cdot P(Z=j)=\sum_{j=0}^{100} j\cdot\left(\phi\left(\frac{\log(j+1)-\mu}{\sigma}\right)-\phi\left(\frac{\log(j)-\mu}{\sigma}\right)\right)=$$
$$=0+1\cdot\left(\phi\left(\frac{\log(2)-\mu}{\sigma}\right)-\phi\left(\frac{-\mu}{\sigma}\right)\right)+2\cdot\left(\phi\left(\frac{\log(3)-\mu}{\sigma}\right)-\phi\left(\frac{\log(2)-\mu}{\sigma}\right)\right)+\ldots$$
$$\ldots+100\cdot\left(\phi\left(\frac{\log(101)-\mu}{\sigma}\right)-\phi\left(\frac{\log(100)-\mu}{\sigma}\right)\right)$$

Vediamo che cosa accade facendo variare i parametri della normale di partenza: Proviamo a studiare tenendo fissa la varianza della normale di partenza pari a 1 e facendo variare il valore della media della normale di partenza.

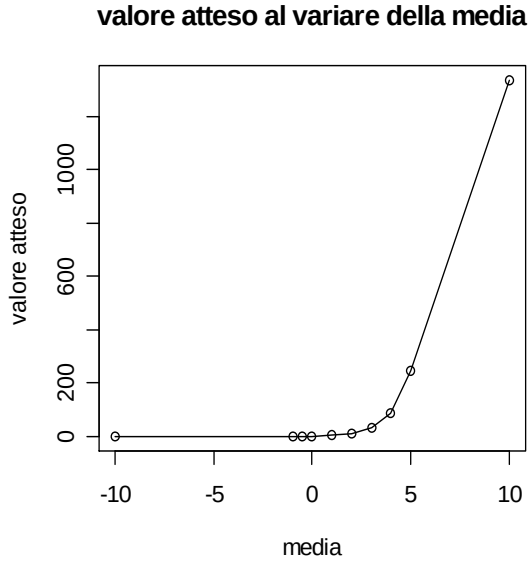
Tabella 20: valore atteso per la variabile Z al variare della media della normale di partenza

media	valore atteso
-10	0,0000
-1	0,2427
-0.5	0,5653
0	1,169
1	3,979
2	11,68
3	32,61
4	89,51
5	244,0
10	1335

Otteniamo quindi una funzione crescente.

Con la media che tende a -∞ il valore atteso tende a 0. Il valore atteso aumenta all'aumentare della media.

Grafico 22: valore atteso per la variabile Z al variare della media della normale di partenza



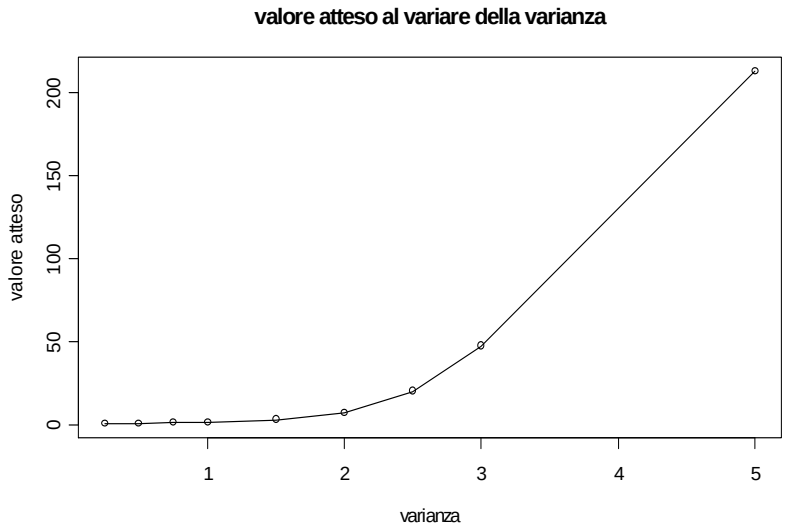
In questo caso la media della normale di partenza è fissata a 0 e facciamo variare il valore della varianza della normale di partenza.

Tabella 21: valore atteso per la variabile Z al variare della varianza della normale di partenza

varianza	valore atteso
0,005	0,5
0,05	0,5
0,25	0,5028
0,5	0,6005
0,75	0,8184
1	1,1698
1,5	2,6461
2	6,953
2,5	19,69
3	47,16
5	212,9

Il valore atteso in funzione della varianza della normale di partenza assume valori crescenti. Il valore atteso assume valore minimo con varianza della normale di partenza pari a 0,5. Il valore atteso ha come minimo pari a 0,5. Analogamente a quanto accade con la media, il valore atteso aumenta all'aumentare della varianza della normale di partenza.

Grafico 23: valore atteso per la variabile Z al variare della varianza della normale di partenza



Varianza attesa

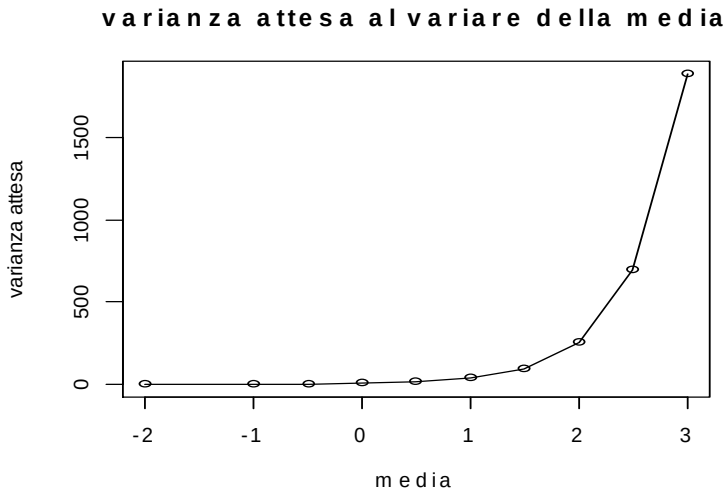
Andiamo a studiare cosa accade alla varianza attesa, facendo variare la media e tenendo fissa la varianza a 1.

Tabella 22: varianza attesa per la variabile Z al variare della media della normale di partenza

media	varianza attesa
-2	$\simeq 0$
-1	0,5540
-0,5	1,660
0	4,668
0,5	12,75
1	34,61
1,5	93,92
2	255,1
2,5	693,3
3	1884,4

È interessante notare che anche in questo caso otteniamo una funzione crescente. Analogamente a quanto accaduto per il valore atteso, la varianza attesa tende a 0 con la media che tende a $-\infty$. La varianza attesa cresce al crescere delle media.

Grafico 24: varianza attesa per la variabile Z al variare della media della normale di partenza



Infine analizziamo i valori assunti dalla varianza attesa al variare della varianza della normale di partenza.

Tabella 23: varianza attesa per la variabile Z al variare della varianza della normale di partenza

varianza	varianza attesa
0,005	0,25
0,05	0,25
0,25	0,2556
0,5	0,4861
0,75	1,3954
1	4,668
1,25	17,87
1,5	80,13
1,75	417,3
2	2113,1
2,25	8390,6
2,5	24847,4
3	10906,2
5	83810,8

Si scopre che la varianza attesa al minimo assume valore pari a 0,25. Ancora una volta notiamo che la varianza attesa aumenta col crescere della varianza della normale di partenza.

Grafico 24: varianza attesa per la variabile Z al variare della varianza della normale di partenza

Simulazione di dati

Partendo da dati reali abbiamo la necessità di stimare i due parametri del nostro modello.

Per fare questo è necessario creare dei vettori di numeri casuali che siano determinazioni della variabile Z con parametri mu e sigma scelti a priori

Nelle simulazioni posso far variare tre parametri: n, mu, sigma, dove n definisce la numerosità dei numeri casuali che vengono generati per testare i metodi di stima. Supponiamo che n sia pari a 100 oppure a 1000 per analizzare le differenze dovute alla quantità di dati a disposizione.

La variabile, dovendo rappresentare il numero di gol segnati da una squadra di calcio, deve assumere prevalentemente valori bassi: tra 0 e 2 nella maggior parte dei casi.

Per questo motivo abbiamo scelto un preciso dominio in cui far variare i valori dei nostri parametri: mu varia tra 0,6 e 1; sigma varia tra 0,5 e 0,3.

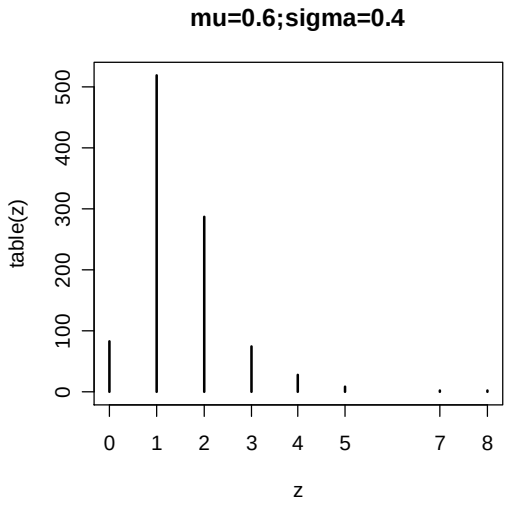
Tabella 24: valori assunti da n, mu e sigma nelle simulazioni

n	mu	sigma
100	0,6	0,3
1000	0,7	0,4
	0,8	0,5
	0,9	
	1	

Alcune di queste combinazioni sembrano poter descrivere i gol fatti.

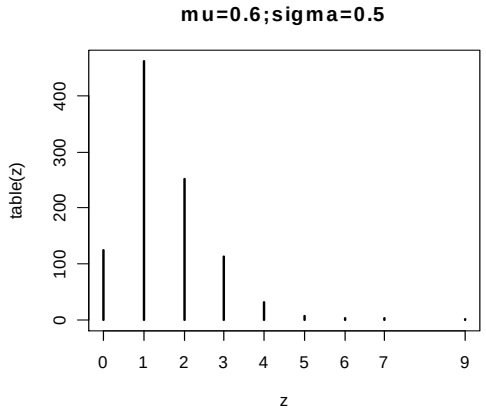
Con mu pari a 0,6 e sigma a 0,4 la moda è 1 gol: circa 500 determinazioni sulle 1000 generate assumono il valore uno.

Grafico 25: distribuzione di frequenza della variabile Z con mu e sigma fissati a priori



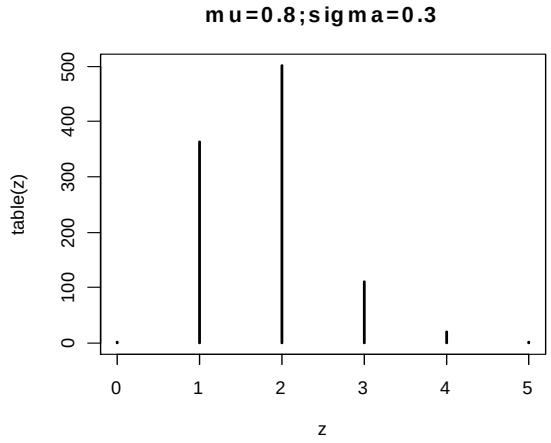
In questo secondo esempio con mu pari a 0,6 e sigma pari a 0,5 le squadre tendono a segnare di più, infatti è più probabile che si superino le quattro reti.

Grafico 26: distribuzione di frequenza della variabile Z con mu e sigma fissati a priori



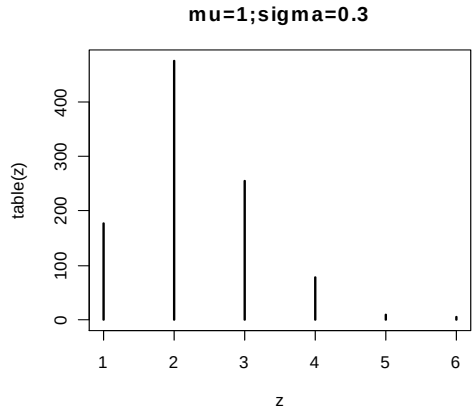
Nel terzo esempio, grafico 27, i gol segnati hanno moda in 2 mentre è molto raro che non si abbiano gol.

Grafico 27: distribuzione di frequenza della variabile Z con mu e sigma fissati a priori



Altre combinazioni possibili sono meno compatibili con il fenomeno studiato. Con la coppia 1 e 0,3 lo zero non appare mai. Questo è impensabile per una distribuzione che vuole descrivere il numero di reti segnate da una squadra di calcio.

Grafico 28: distribuzione di frequenza della variabile Z con mu e sigma fissati a priori



Con altre combinazioni, che non ho ritenuto necessario riportare, si ottengono valori troppo elevati per esprimere il numero di gol segnati da una squadra in un incontro di calcio.

Funzione di verosimiglianza

Per il modello di Poisson la stima di massima verosimiglianza è il valore che annulla la derivata prima della funzione di log-verosimiglianza.
Come abbiamo, visto la funzione di probabilità della variabile che stiamo analizzando è:

$$\Pr\left(X_i = x\right) = \left[\phi\left(\frac{\log\left(i+1\right)-\mu}{\sigma}\right) - \phi\left(\frac{\log\left(i\right)-\mu}{\sigma}\right)\right]$$

Quindi la funzione di verosimiglianza diviene:

$$L\left(\mu_i, \sigma_i; i = 1,..n\right) = \prod_{k=1}^{N}\left[\phi\left(\frac{\log\left(i_k+1\right)-\mu_{i(k)}}{\sigma_{i(k)}}\right) - \phi\left(\frac{\log\left(i_k\right)-\mu_{i(k)}}{\sigma_{i(k)}}\right)\right].$$

Dove l' i-esima squadra nel momento in cui gioca la k-esima partita.
La funzione di log-verosimiglianza:

$$L\left(\mu_i, \sigma_i; i = 1,..n\right) = \prod_{k=1}^{N}\left\{\left[\phi\left(\frac{\log\left(i_k+1\right)-\mu_{i(k)}}{\sigma_{i(k)}}\right) - \phi\left(\frac{\log\left(i_k\right)-\mu_{i(k)}}{\sigma_{i(k)}}\right)\right]\right\}.$$

Per la proprietà dei logaritmi:

$$l\left(\mu_i, \sigma_i; i = 1,..n\right) = \sum_{k=1}^{N}\log\left\{\left[\phi\left(\frac{\log\left(i_k+1\right)-\mu_{i(k)}}{\sigma_{i(k)}}\right) - \phi\left(\frac{\log\left(i_k\right)-\mu_{i(k)}}{\sigma_{i(k)}}\right)\right]\right\}.$$

A questo punto dovremmo agire analogamente a quanto fatto per il modello di Poisson e ricavare la stima di massima verosimiglianza in modo analitico. Tuttavia non esiste una derivata prima nota di questa funzione, perciò abbiamo dovuto utilizzare altri metodi di stima.

Nelle simulazioni le distribuzioni sono originate tramite la variabile Z avente mu e sigma fissati e poi sulla base di queste distribuzioni abbiamo stimato i parametri attraverso i diversi metodi.
Nelle seguenti simulazioni vengono creati 10 campioni di numeri casuali. Per ciascun campione vengono stimati i parametri e i loro intervalli di confidenza.

A questo punto abbiamo testato tre metodi di stima per valutare la loro precisione.

Metodo di stima tramite massima verosimiglianza

Questo metodo si basa sul metodo numerico di Newton-Raphson della funzione di log-verosimiglianza della variabile Z. Per questa ragione la precisione nelle stime dipende dalla numerosità del campione che viene generato.
Le stime dei parametri sono i valori in cui la funzione assume il massimo valore, cioè sono i punti di massimo. Vale a dire i valori per cui le derivate prime della funzione di log-verosimiglianza si annullano.
Le stime sono quindi:

$$\hat{Mu} = \tilde{Mu}$$

$$\hat{Sigma} = \tilde{Sigma}$$

Dove $\tilde{Mu}$ e $\tilde{Sigma}$ sono i valori che annullano la derivata prima della log-verosimiglianza.
Gli intervalli di confidenza sono determinati per Mu tramite:

$$\left(\tilde{Mu} - q_{0,975} \cdot \sqrt{\tau_{mu}} ; \tilde{Mu} + q_{0,975} \cdot \sqrt{\tau_{mu}} \right)$$

Per Sigma tramite:

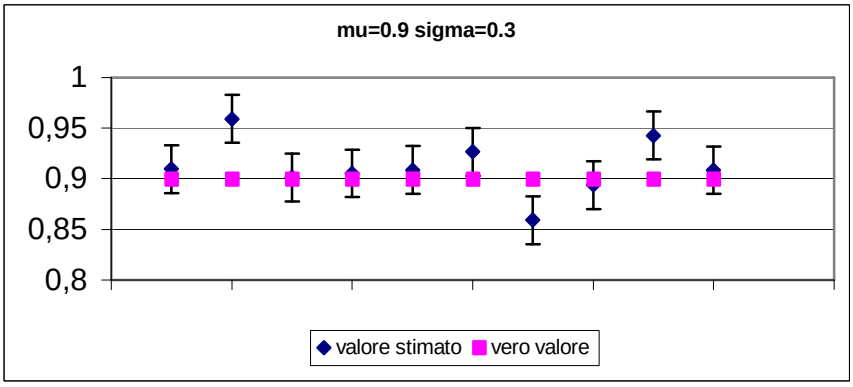
$$\left(\tilde{Sigma} - q_{0,975} \cdot \sqrt{\tau_{sigma}} ; \tilde{Sigma} + q_{0,975} \cdot \sqrt{\tau_{sigma}} \right)$$

Dove τ^2_{mu} e τ^2_{sigma} rappresentano gli elementi della diagonale principale della inversa della matrice di informazione della funzione di log-verosimiglianza.

Generazione di 100 numeri

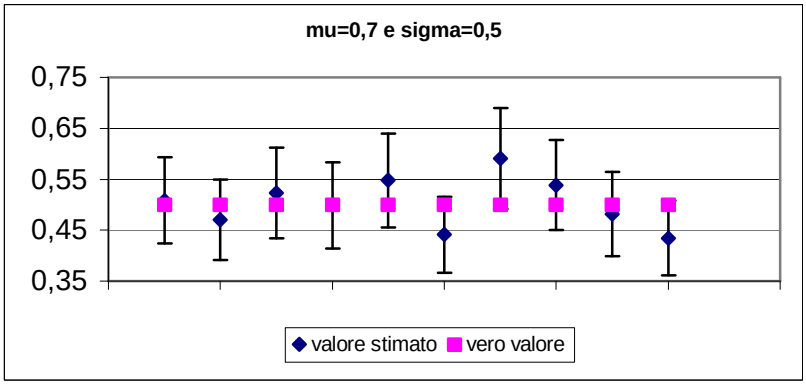
Per il parametro mu la precisione è buona.
Non ci sono fenomeni di sovrastima o sottostima.
Avendo creato intervalli di confidenza al 95%, in 95 casi su 100 ci si aspetta che il vero appartenga all'intervallo di confidenza della stima.

Grafico 29: stima di mu con metodo di stima tramite massima verosimiglianza



Gli stessi aspetti si riscontrano con sigma: per ciascuno dei dieci vettori creati il vero valore appartiene all'intervallo di confidenza.

Grafico 30: stima di Sigma con metodo di stima tramite massima verosimiglianza



Generazione di 1000 numeri

Le ampiezze degli intervalli si riducono: avendo a disposizione più dati, l'approssimazione ci porta ad una maggiore precisione delle stime. Tuttavia abbiamo un aumento dei tempi necessari per ottenere le stime. Un metodo iterativo, infatti, procedendo per approssimazioni successive con un elevato numero di dati, necessita di tempi maggiori.

Grafico 31: stima di Mu con metodo di stima tramite massima verosimiglianza.

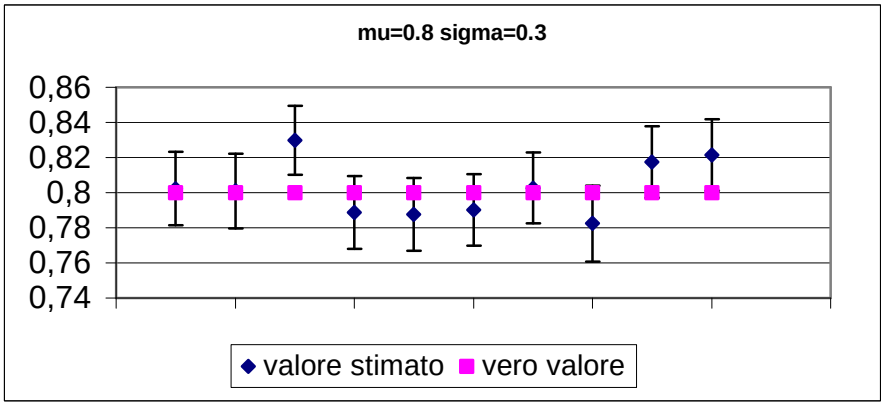
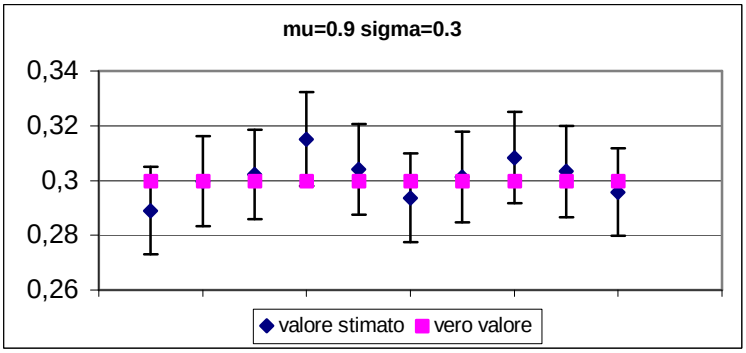


Grafico 32: stima di Sigma con metodo di stima tramite massima verosimiglianza



Il metodo numerico di Newton-Raphson si dimostra molto preciso: per questa ragione è un "metodo esatto". Tuttavia basandosi sull'iteratività comporta una lunga serie di approssimazioni successive. Per questa ragione risulta oneroso se per la stima si ha a disposizione un alto numero di osservazioni.

Metodo di stima approssimato

Questo metodo sfrutta le caratteristiche con le quali è stata costruita la nostra variabile. In altre parole, considerando che la variabile Gaussiana ha delle caratteristiche molto particolari, vogliamo provare a sfruttarle per ottenere un metodo di stima preciso tanto quanto il precedente, senza però l'onerosità dal punto di vista dei calcolo dello stesso.

Poiché Z vale 0 se Y è compresa tra 0 e 1, la prima operazione da fare è riscalare la nostra Z aggiungendo 0,5. Poiché Y è la trasformazione esponenziale di X, bisogna effettuare una trasformazione logaritmica delle determinazioni di Y.

Quindi partendo dalle n determinazioni z di Z:

y=z+0,5

x=log(y)

A questo punto i parametri mu e sigma vengono stimati come media e deviazione standard delle n determinazioni della variabile aleatoria X.

$$\hat{mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

$$\hat{sigma} = \sqrt{\frac{1}{n-1} \bullet \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

L'intervallo di confidenza per mu al livello 95%è:

$$\left(\overline{x} - q_{0,975} \cdot \sqrt{\frac{\sigma^2}{n}}; \overline{x} + q_{0,975} \cdot \sqrt{\frac{\sigma^2}{n}} \right)$$

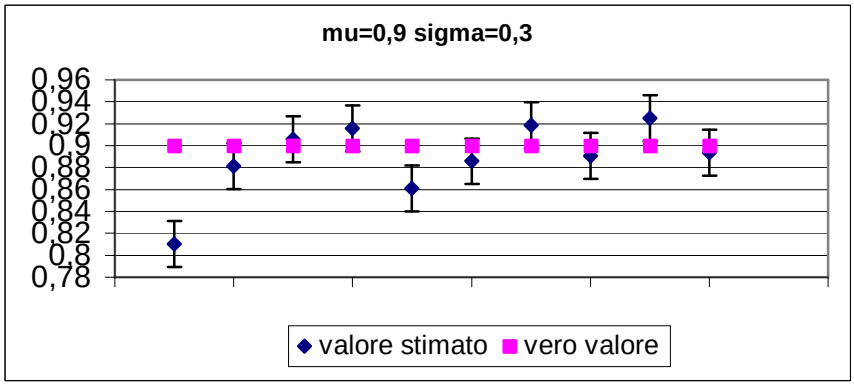
L'intervallo di confidenza per sigma al livello 95% è:

$$\left(\sqrt{\frac{1}{n-1} \bullet \sum_{i=1}^{n} (x_i - \overline{x})^2} \pm q_{0,975} \cdot \frac{(n_{camp} - 1)\left[(n_{camp} - 1) \cdot 3\sigma^4 - (n_{camp} - 3) \cdot \sigma^4 \right]}{n^3} \right)$$

Generazione di 100 numeri

Per il parametro mu non si riscontrano errori sistematici. Il vero valore del parametro è compreso nell'intervallo di stima in poco più del 50% dei casi.
Nei casi in cui il vero valore non è compreso la stima a volte sovrastima e a volte sottostima il valore del parametro.
Esaminando il caso con mu pari a 0,9 e sigma pari a 0,3, il metodo approssimato sembra garantire una buona precisione. Per questo esempio, infatti, 8 volte su 10 il vero valore del parametro mu appartiene all'intervallo di confidenza che si stima.

Grafico 33: Stima di mu con metodo approssimato



Per il parametro sigma le cose cambiano: vi è una sovrastima del vero valore del parametro. Questo accade sia se il vero valore non viene mai centrato come nel primo esempio, sia se viene centrato un discreto numero di volte come nel secondo esempio.

Grafico 34: Stima di sigma con metodo approssimato

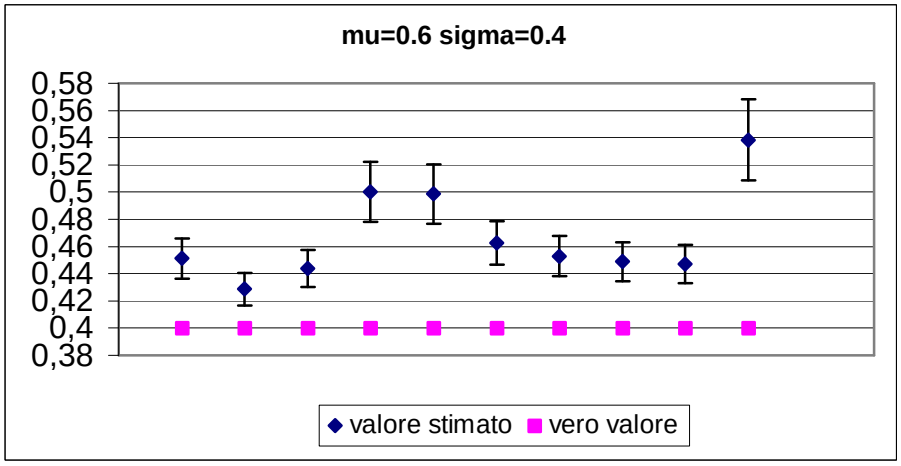
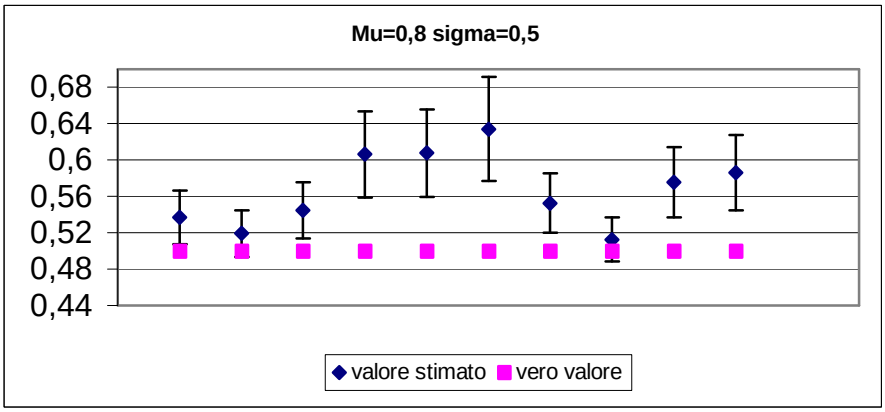


Grafico 35: Stima di sigma con metodo approssimato

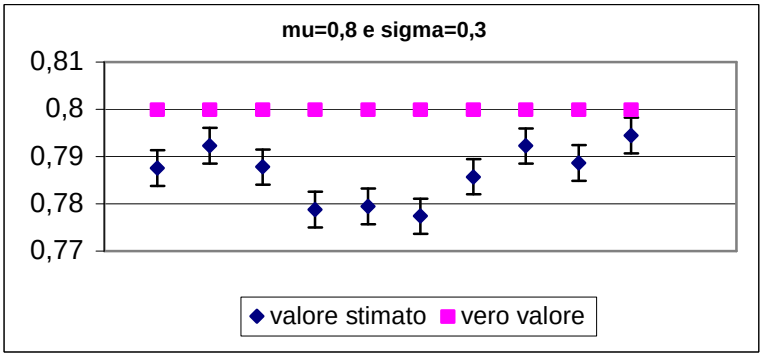


Generazione di 1000 numeri

La prima conseguenza nel generare vettori più numerosi è che l'ampiezza degli intervalli di confidenza si restringe. Questo metodo tuttavia non ha lo svantaggio del metodo esatto in quanto il calcolo di media e varianza di un vettore di dati è quasi immediato per ogni dimensione vettoriale.

Per il parametro mu abbiamo che il vero valore viene centrato circa il 30% dei casi. In questo caso è da sottolineare che il parametro è sistematicamente sottostimato.

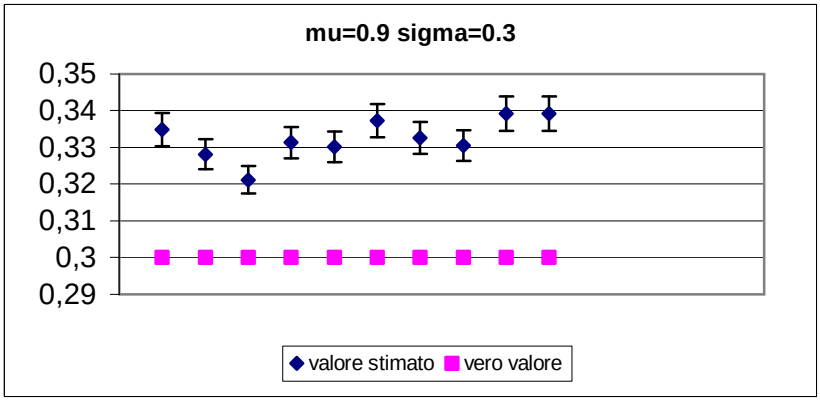
Grafico 36: Stima di mu con metodo approssimato



Questo esempio mostra in modo evidente la distorsione che caratterizza questo metodo di stima del parametro mu. Naturalmente la stessa distorsione si riscontra per ogni coppia di valori assegnata ai due parametri.

Al contrario, il parametro sigma viene costantemente sovrastimato. Addirittura per ogni combinazione analizzata il vero valore non appartiene all'intervallo di confidenza calcolato in nessuno dei dieci campioni.

Grafico 37: Stima di sigma con metodo approssimato



Confronto fra i due metodi

Eravamo consapevoli fin dalla partenza che un metodo basato sul metodo numerico di Newton-Raphson ci avrebbe dato risultati soddisfacenti. Tuttavia il difetto di questo metodo è la dispendiosità delle stime in termini di tempo e costi nel momento in cui si hanno a disposizione grandi quantità di dati.

Per questa ragione si è tentato di sfruttare le proprietà della normale con cui viene costruita la variabile Z, inventando il secondo metodo di stima: questo metodo tuttavia non è preciso.

Il confronto tra i due metodi evidenzia la costante sottostima del parametro mu e la costante sovrastima di sigma.

Grafico 38: Stima di mu con metodo esatto e con metodo approssimato

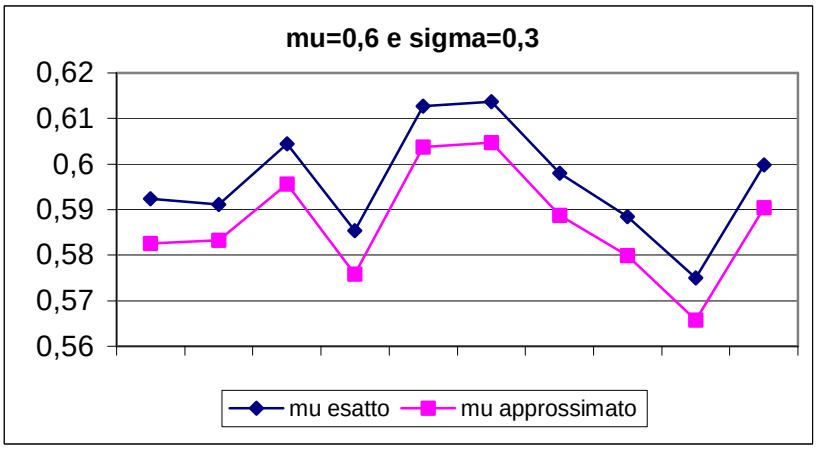
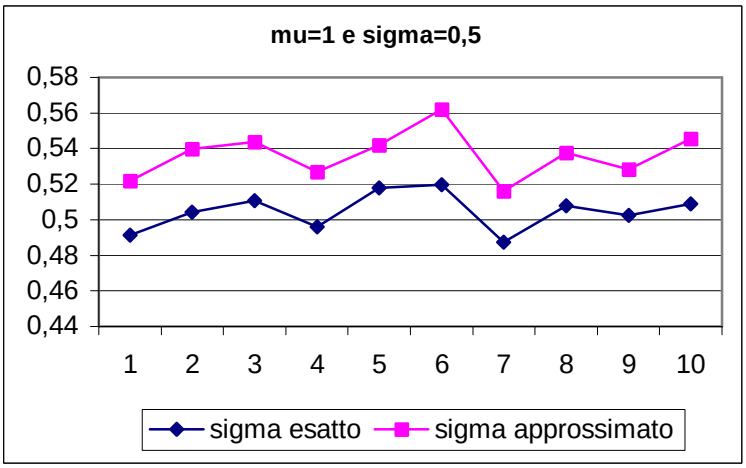


Grafico 39: Stima di sigma con metodo esatto ed approssimato.



Secondo metodo di stima approssimato

Una volta riscontrata la distorsione per il metodo approssimato, si è provato a modificare il metodo di stima approssimato in modo tale da potere sfruttare le proprietà della normale da cui Z trae origine.

Il problema negli errori di stima poteva essere causato dal fatto che il metodo forzava tutti i valori al valore intermedio 0,5. Questa potrebbe essere la causa della distorsione nelle stime.

Un metodo ulteriore è quindi, partendo dalle n determinazioni z di Z, il trasformare

y=z+k

dove k è il rapporto tra le numerosità di classi successive di gol segnati.

Successivamente si procede analogamente a quanto fatto per il primo metodo di stima approssimato.

Poiché Y è la trasformazione esponenziale di X, operiamo la trasformazione inversa:

x=log(y).

A questo punto i parametri mu e sigma vengono stimati come media e deviazione standard delle x.

$$\hat{Mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

$$\hat{Sigma} = \sqrt{\frac{1}{n-1} \bullet \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

Con questo nuovo metodo si ha una sovrastima per il parametro mu e, viceversa, una sottostima di sigma.

Grafico 40: Stima di mu con il secondo metodo approssimato

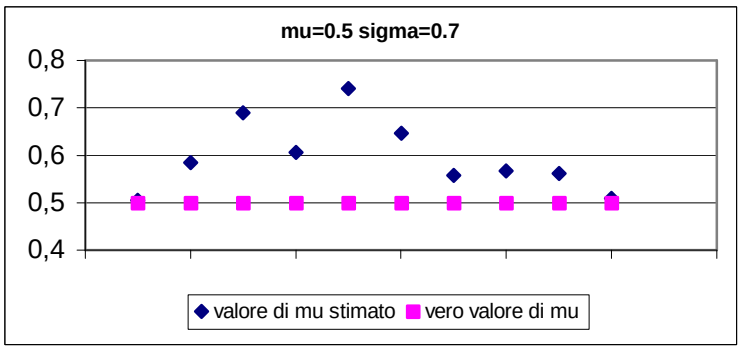
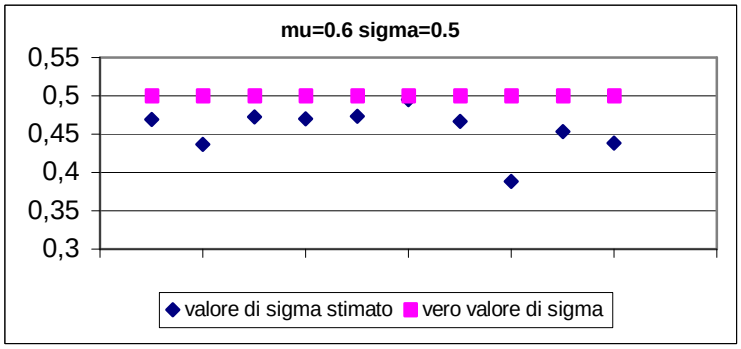


Grafico 41: Stima di sigma con il secondo metodo approssimato



Anche questo secondo metodo di approssimazione non è quindi adeguato per ottenere delle stime soddisfacenti dei parametri.

Stime

Nella parte precedente siamo andati a valutare il comportamento della variabile Z al variare dei nostri parametri.
Abbiamo deciso utilizzare il metodo esatto per la stima dei nostri parametri.
Per questa ragione abbiamo provato a stimare il valore dei parametri mu e sigma sulla base di un campione composto da un numero variabile di partite disputate nelle varie nazioni.
Ecco le stime ottenute:

Tabella 25: stime con metodo esatto dei parametri Mu e Sigma e dei rispettivi standard error per squadra 1 in ciascuna nazione

nazione	mu1	s.e. mu1	sigma1	s.e.sigma 1
Argentina	0,521	0,087	0,595	0,064
Australia	0,485	0,391	0,649	0,296
Austria	0,605	0,076	0,645	0,061
Belgio	0,581	0,070	0,615	0,054
Brasile	0,637	0,055	0,583	0,040
Cile	0,579	0,113	0,581	0,082
Croazia	0,595	0,258	0,729	0,238
Finlandia	0,540	0,101	0,652	0,081
Francia	0,437	0,052	0,627	0,039
Germania	0,534	0,046	0,670	0,037
Giappone	0,514	0,118	0,590	0,087
Grecia	0,499	0,111	0,658	0,089
Inghilterra	0,513	0,042	0,618	0,032
Irlanda	0,438	0,087	0,651	0,068
Islanda	0,543	0,129	0,634	0,101
Italia	0,533	0,078	0,603	0,058
Messico	0,570	0,131	0,591	0,098
Norvegia	0,708	0,098	0,657	0,080
Olanda	0,671	0,060	0,632	0,048
Polonia	0,496	0,107	0,666	0,086
Portogallo	0,437	0,076	0,639	0,058
Rep.Ceka	0,466	0,159	0,586	0,114
Romania	0,485	0,103	0,580	0,073
Russia	0,512	0,073	0,653	0,058
Scozia	0,595	0,113	0,643	0,090
Spagna	0,449	0,058	0,647	0,045
Svezia	0,502	0,067	0,652	0,053
Svizzera	0,613	0,115	0,644	0,093
Turchia	0,534	0,120	0,653	0,095
Usa	0,449	0,210	0,638	0,162
generale	0,537	0,014	0,633	0,011

Nota x: nel campionato statunitense e nel campionato australiano la squadra 1 è la squadra in trasferta, la squadra 2 è quella in casa

Tabella 26: stime con metodo esatto dei parametri Mu e Sigma e dei rispettivi standard error per squadra 2 in ciascuna nazione

nazione	mu2	s.e. mu2	sigma2	s.e. sigma 2
Argentina	0,242	0,079	0,640	0,057
Australia	0,594	0,475	0,534	0,328
Austria	0,328	0,074	0,650	0,056
Belgio	0,305	0,066	0,636	0,049
Brasile	0,258	0,049	0,632	0,035
Cile	0,513	0,116	0,560	0,082
Croazia	0,328	0,281	0,652	0,218
Finlandia	0,270	0,091	0,701	0,072
Francia	0,204	0,050	0,641	0,036
Germania	0,329	0,045	0,658	0,035
Giappone	0,397	0,109	0,636	0,083
Grecia	0,164	0,106	0,662	0,079
Inghilterra	0,338	0,042	0,609	0,030
Irlanda	0,247	0,090	0,612	0,063
Islanda	0,499	0,135	0,605	0,100
Italia	0,316	0,079	0,582	0,055
Messico	0,345	0,124	0,612	0,089
Norvegia	0,417	0,098	0,641	0,076
Olanda	0,370	0,057	0,652	0,044
Polonia	0,189	0,095	0,722	0,076
Portogallo	0,191	0,079	0,595	0,054
Rep.Ceka	0,200	0,155	0,581	0,103
Romania	0,245	0,099	0,582	0,067
Russia	0,297	0,073	0,635	0,054
Scozia	0,321	0,109	0,649	0,082
Spagna	0,279	0,061	0,609	0,043
Svezia	0,353	0,064	0,669	0,051
Svizzera	0,316	0,099	0,728	0,083
Turchia	0,269	0,122	0,624	0,088
Usa	0,429	0,244	0,545	0,159
generale	0,305	0,014	0,636	0,010

Grafico 42: stime con metodo esatto dei parametri Mu per squadra 1 e squadra 2 in ciascuna nazione

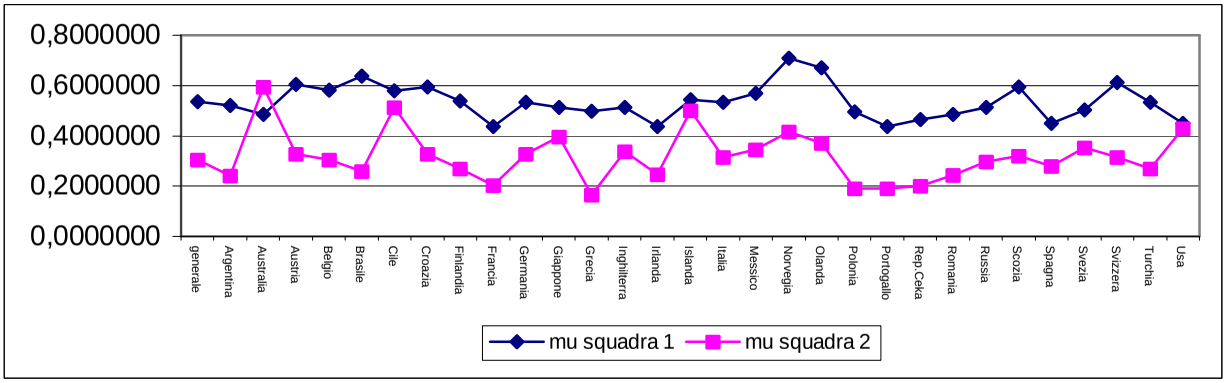
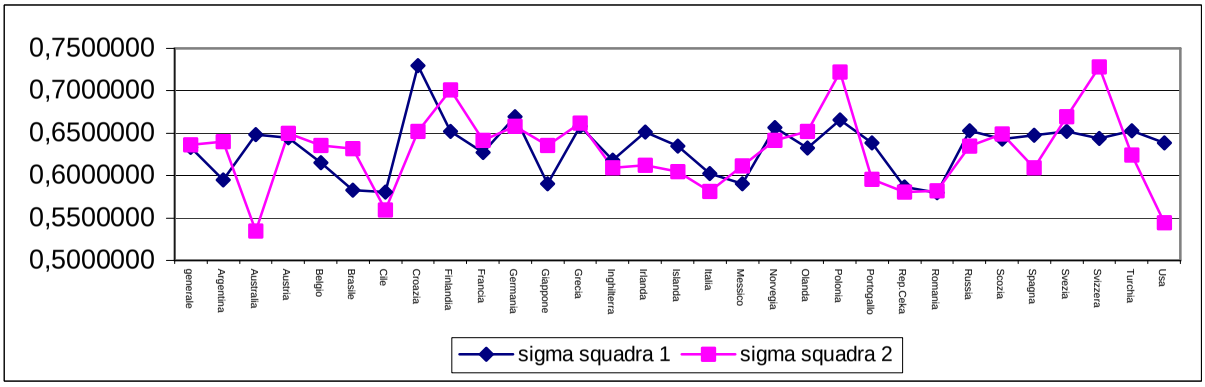


Grafico 43: stime con metodo esatto dei parametri Sigma per squadra 1 e squadra 2 in ciascuna nazione



Con il metodo approssimato le stime invece sono:

Tabella 27: stime con metodo approssimato dei parametri Mu e Sigma per squadra 1 e squadra 2 in ciascuna nazione

nazione	mu1	sigma1	mu2	sigma2
Argentina	0,470	0,698	0,181	0,722
Australia	0,419	0,757	0,557	0,628
Austria	0,555	0,741	0,272	0,733
Belgio	0,533	0,713	0,244	0,725
Brasile	0,593	0,681	0,196	0,719
Cile	0,530	0,686	0,468	0,660
Croazia	0,565	0,784	0,276	0,730
Finlandia	0,492	0,740	0,217	0,762
Francia	0,379	0,727	0,141	0,719
Germania	0,485	0,756	0,271	0,741
Giappone	0,467	0,688	0,343	0,727
Grecia	0,452	0,742	0,108	0,723
Inghilterra	0,460	0,718	0,278	0,708
Irlanda	0,384	0,740	0,181	0,707
Islanda	0,492	0,730	0,446	0,708
Italia	0,481	0,705	0,255	0,687
Messico	0,525	0,687	0,281	0,714
Norvegia	0,662	0,748	0,365	0,730
Olanda	0,626	0,726	0,315	0,737
Polonia	0,444	0,754	0,143	0,762
Portogallo	0,381	0,733	0,124	0,689
Rep.Ceka	0,411	0,692	0,132	0,681
Romania	0,431	0,689	0,178	0,686
Russia	0,458	0,748	0,236	0,724
Scozia	0,548	0,734	0,261	0,734
Spagna	0,394	0,740	0,214	0,706
Svezia	0,449	0,745	0,300	0,748
Svizzera	0,569	0,732	0,269	0,782
Turchia	0,480	0,749	0,208	0,713
Usa	0,395	0,731	0,361	0,677
generale	0,486	0,729	0,245	0,724

Grafico 44: stime con metodo approssimato del parametro Mu per squadra 1 e squadra 2 in ciascuna nazione

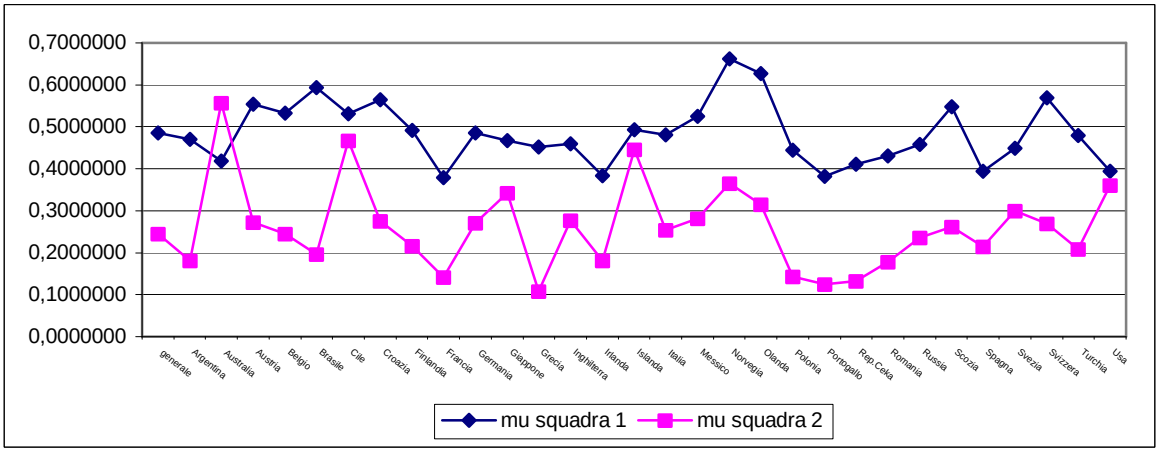
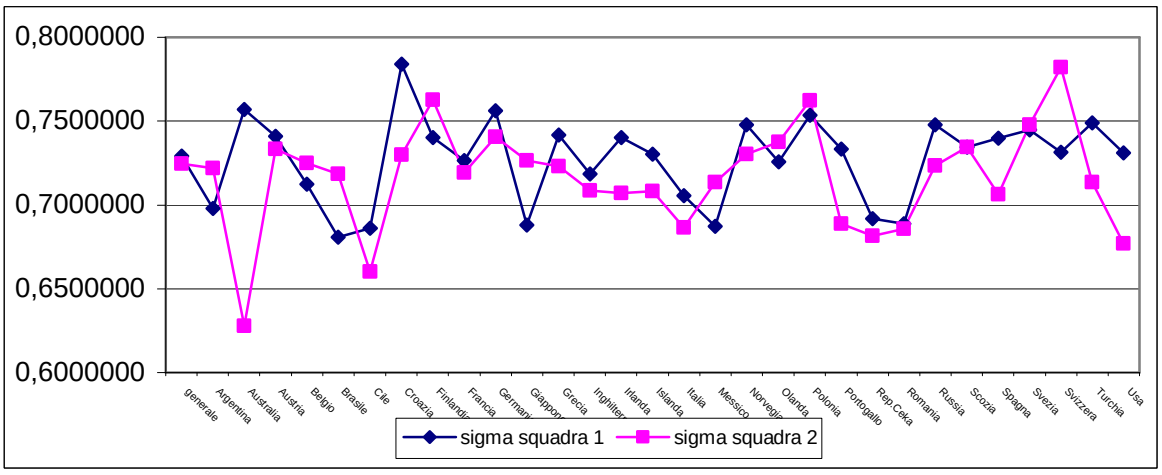


Grafico 45: stime con metodo approssimato del parametro Sigma per squadra 1 e squadra 2 in ciascuna nazione



Una volta stimati i parametri, abbiamo provato ad utilizzare la nostra variabile Z avente come parametri le stime appena calcolate con il metodo di stima tramite massima verosimiglianza.

Confronto fra la distribuzione empirica e la distribuzione ottenuta tramite modello

A questo punto è possibile confrontare la distribuzione dei gol fatti negli incontri presi in considerazione e la distribuzione di probabilità della nostra variabile Z avente parametri le stime ottenute con il metodo esatto.

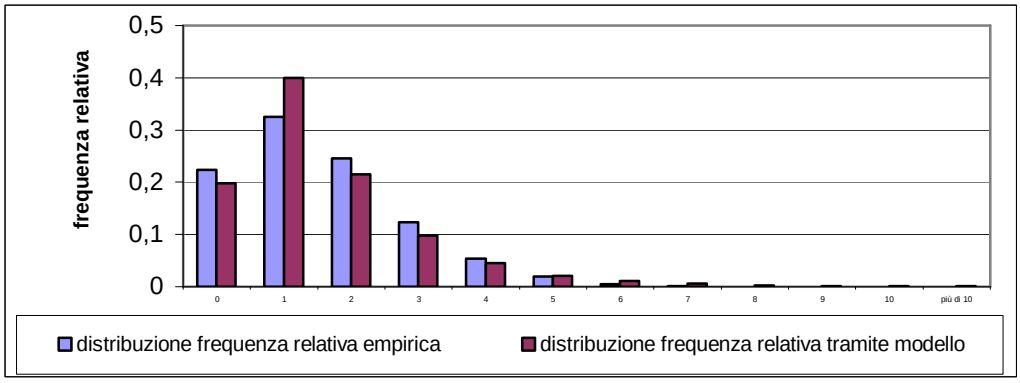
Vediamo che cosa accade alla distribuzione della squadra 1 prendendo in considerazione tutte le partite a disposizione.

Tabella 28: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione simulata

numero di gol segnati	Distribuzione empirica	Distribuzione modello
0	22,40%	19,95%
1	32,54%	39,67%
2	24,60%	21,47%
3	12,39%	10,05%
4	5,43%	4,46%
Più di 4 goal	2,63%	4,35%

Appare evidente che vi è una forte differenza nelle due distribuzioni di probabilità. La probabilità che la squadra segni almeno 5 goal è del 4,35% nella simulazione tramite la variabile Z. Al contrario ciò avviene il 2,63% delle volte. Effettivamente la coda è troppo pesante. Il rapporto tra le due probabilità è pari a 1,65.
Addirittura,facendo una simulazione degli incontri, in 2 casi su 11mila la squadra segna 21 goal.

Grafico 46: frequenza relativa del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione tramite modello



Il grafico mostra che esiste una differenza marcata anche per l'evento "1 gol" che è troppo probabile rispetto alla probabilità empirica.
Proviamo a studiare che cosa accade per la squadra in trasferta.

Tabella 29: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione tramite modello

numero di gol segnati	Distribuzione empirica	Distribuzione modello
0	33,08%	31,58%
1	36,49%	41,34%
2	19,54%	16,48%
3	7,62%	6,15%
4	2,39%	2,44%
Più di 4 gol	0,88%	1,96%

Gli stessi problemi si riscontrano anche per la squadra 2: la probabilità di una goleada è doppia e l'evento "1 gol" è troppo probabile. Questo secondo aspetto viene bilanciato da probabilità troppo basse per gli eventi "0 gol", "2 gol" e "3 gol".
Tuttavia queste problematiche potrebbero dipendere dai valori assegnati ai parametri.
Studiamo che cosa accade scegliendo la distribuzione con stima di mu maggiore. Confrontiamo la distribuzione empirica dei goal segnati dalla squadra di casa nel campionato olandese con quella simulata tramite il nostro modello avente come parametri i valori stimati con metodo esatto.

Tabella 30: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato olandese per la distribuzione empirica e per la distribuzione tramite modello

numero di gol	Distribuzione empirica	Distribuzione modello
0	17,54%	14,42%
1	28,57%	36,98%
2	26,86%	23,67%
3	13,51%	12,05%
4	8,07%	6,01%
Più di 4 gol	5,43%	6,56%

Anche in questo esempio l'evento "1 gol" ha una probabilità molto alta, bilanciata da probabilità basse per gli eventi contigui.
Il problema rimane anche se mu ha una stima bassa.
Consideriamo i gol segnati dalla squadra in trasferta nel campionato polacco.

Tabella 31: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato polacco per la distribuzione empirica e per la distribuzione tramite modello

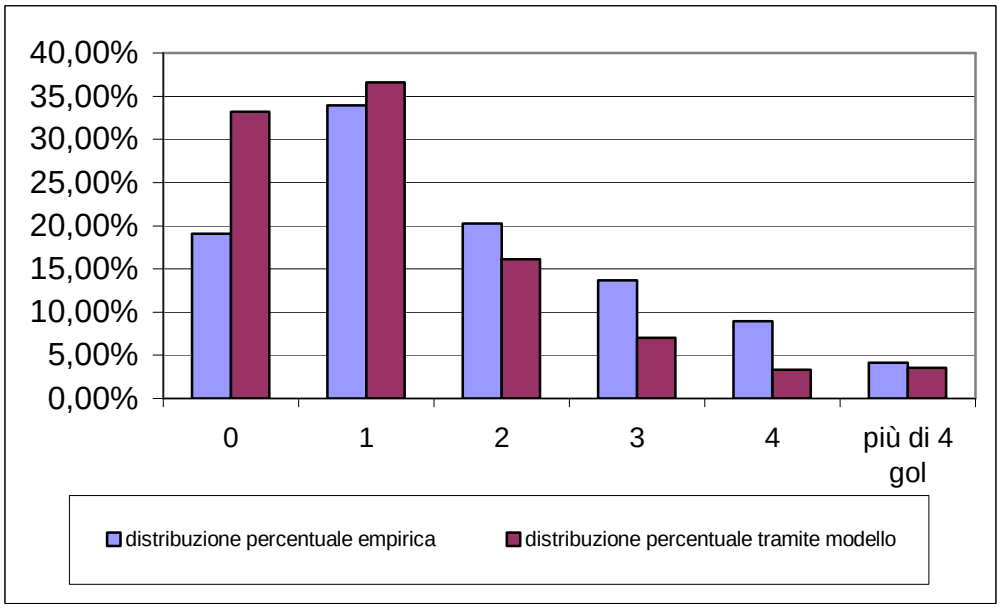
numero di gol	distribuzione empirica	distribuzione modello
0	42,13%	39,67%
1	26,40%	36,07%
2	21,91%	13,86%
3	7,87%	5,52%
4	1,12%	2,41%
più di 5 gol	0,56%	2,19%

Per questo esempio si ha una probabilità per l'evento "1 gol" più alta addirittura del 10% rispetto a quella empirica. La coda della distribuzione simulata tramite il modello è troppo pesante rispetto alla coda della distribuzione empirica.
Vediamo che cosa accade con un alto valore di sigma: gol segnati dalla squadra 2 nel campionato svizzero.

Tabella 32: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato svizzero per la distribuzione empirica e per la distribuzione tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0	19,05%	33,21%
1	33,93%	36,57%
2	20,24%	16,10%
3	13,69%	7,04%
4	8,93%	3,29%
più di 4 gol	4,17%	3,57%

Grafico 47: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato svizzero per la distribuzione empirica e per la distribuzione tramite modello



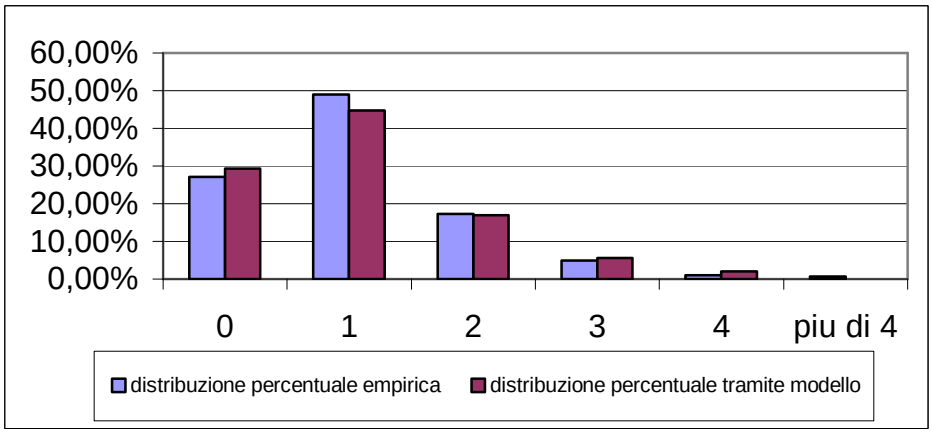
In questo caso il problema più evidente è nella parte sinistra cioè quando ci sono pochi gol. Il grafico 47 mostra in modo chiaro il sovradimensionamento per le probabilità degli eventi "0 gol" e "1 gol" ed il sottodimensionamento di quelle per gli eventi della coda destra.

Studiamo cosa accade se invece sigma è particolarmente basso come nel caso dei gol segnati in Italia dalla squadra in trasferta.

Tabella 33: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato italiano per la distribuzione empirica e per la distribuzione tramite modello

numero di gol	Distribuzione empirica	Distribuzione modello
0	27,1%	29,36%
1	49,0%	44,79%
2	17,3%	16,91%
3	4,88%	5,64%
4	0,98%	1,98%
più di 4 gol	0,73%	1,28%

Grafico 48: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato italiano per la distribuzione empirica e per la distribuzione tramite modello



In questo caso il problema principale torna ad essere lo sbilanciamento della distribuzione ottenuta con la nostra variabile nella coda destra. Inoltre si nota che vi è una sovrastima per la probabilità dell' evento "0 gol" mentre risulta sottodimensionata la probabilità per l'evento "1 gol".

Conclusioni

Riassumendo, possiamo affermare che sono presenti due problemi.

Il primo problema riguarda la previsione nel caso di un numero elevato di gol: la coda destra in quasi tutti i casi studiati risulta troppo pesante rispetto a quanto si osserva nel campione analizzato: non si riscontra una precisione soddisfacente.

Il secondo problema riguarda gli eventi vicini all'origine: le differenze nelle probabilità per gli eventi "0 gol" ed "1 gol" sono marcate a prescindere dal valore fissato ai parametri.

Tornando al caso generale, è immediato osservare che per ogni evento vi è una forte differenza tra le due distribuzioni di probabilità.

Pertanto, possiamo concludere che la forma del nostro modello si adatta male alla descrizione del fenomeno studiato.

L'ultima parte della tesi riprende questo modello e analizza congiuntamente, tramite il modello bivariato, la distribuzione dei gol segnati dalla squadra 1 e i gol

segnati dalla squadra 2, tenendo in considerazione anche la dipendenza tra i fenomeni.

È bene sottolineare che, sebbene il modello normale non sembri essere efficace, potremmo riscontrare nel modello normale bivariato una buona descrizione della dipendenza. A questo punto sarebbe opportuno utilizzare il modello bivariato con una distribuzione marginale diversa da quella normale che meglio si adatti a descrivere il fenomeno marginalmente.

Il modello Gamma

Introduzione

Il modello Poisson, sebbene descriva in modo opportuno il nostro fenomeno di interesse, ha il difetto di avere un unico parametro che non permette di analizzare distintamente posizione e variabilità della distribuzione: la varianza non può essere più o meno ampia della media.
Il modello Normale, viceversa, consente di distinguere tra il parametro relativo alla posizione e quello relativo alla variabilità della distribuzione. Tuttavia questo modello non si adatta bene ai dati a disposizione.
Il seguente capitolo introduce un nuovo modello basato sulla distribuzione Gamma.
La variabile casuale Gamma è una variabile continua che assume valori positivi ed è definita da due parametri strettamente positivi.
Questa distribuzione di probabilità è utilizzata nell'ambito delle file di attesa e delle telecomunicazioni; ha un ruolo importante anche nella statistica bayesiana.
Definiamo quindi:

$$X \sim Gamma(shape, rate)$$ con shape>0 e rate>0.

La funzione di probabilità è:

$P(X=x)=\frac{a^p \cdot e^{-ax} \cdot x^{p-1}}{\Gamma(p)}$ dove shape viene indicato con p e rate con a.

Il valore atteso della variabile è $E(X)=\frac{shape}{rate}$; la varianza attesa è $Var(X)=\frac{shape}{rate^2}$.

Poiché i gol segnati sono eventi caratterizzati dall'assumere valori discreti operiamo una trasformazione e rendiamo discreta la variabile X.
La nuova variabile Y è tale che

$Y = y$ se e solo se X: $x \le X < x+1$

La nuova funzione di probabilità è:

$$P(Y=0)=\frac{a^p\cdot e^{-a}}{\Gamma(p)}$$

$$P(Y=1)=\frac{a^p\cdot e^{-2a}\cdot 2^{p-1}}{\Gamma(p)}-\frac{a^p\cdot e^{-a}}{\Gamma(p)}$$

$$P(Y=2)=\frac{a^p\cdot e^{-3a}\cdot 3^{p-1}}{\Gamma(p)}-\frac{a^p\cdot e^{-2a}\cdot 2^{p-1}}{\Gamma(p)}$$

.

.

.

In generale:

$$P(Y=y)=\frac{a^p\cdot e^{-(y+1)a}\cdot(y+1)^{p-1}}{\Gamma(p)}-\frac{a^p\cdot e^{-ya}\cdot y^{p-1}}{\Gamma(p)}$$

Simulazione di dati

Analogamente a quanto fatto per la variabile normale, per descrivere i gol segnati da una squadra di calcio i parametri apparterranno ad un ben preciso dominio.
Il parametro shape appartiene all'intervallo tra 2 e 3, il parametro rate appartiene all'intervallo tra 1 e 2.

Tabella 34: valori assunti da n, shape e rate nelle simulazioni

n	shape	rate
100	2	1
1000	2,25	1,5
	2,5	2
	2,75	
	3	

In questo modo si possono ottenere dei vettori composti da 1000 numeri casuali che sembrano poter descrivere l'andamento dei gol segnati da una squadra di calcio.

Grafico 49: distribuzione di frequenza della variabile Y con shape e rate fissati a priori

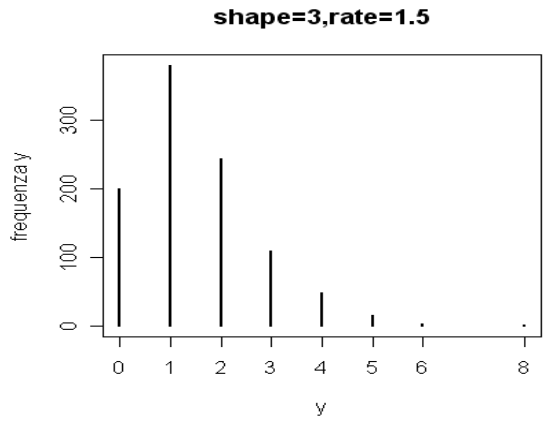
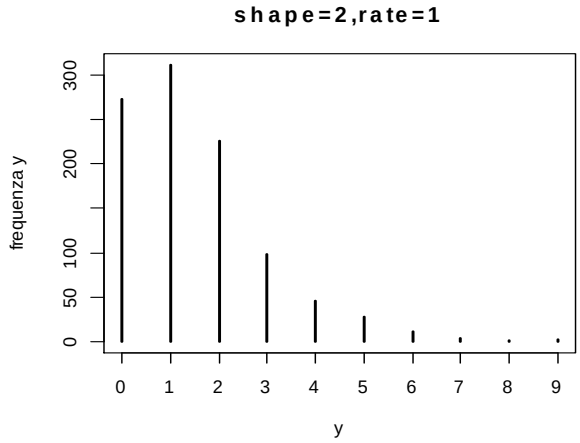
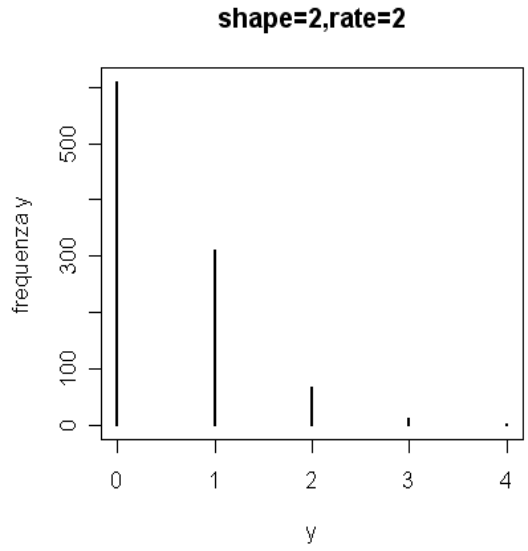


Grafico 50: distribuzione di frequenza della variabile Y con shape e rate fissati a priori



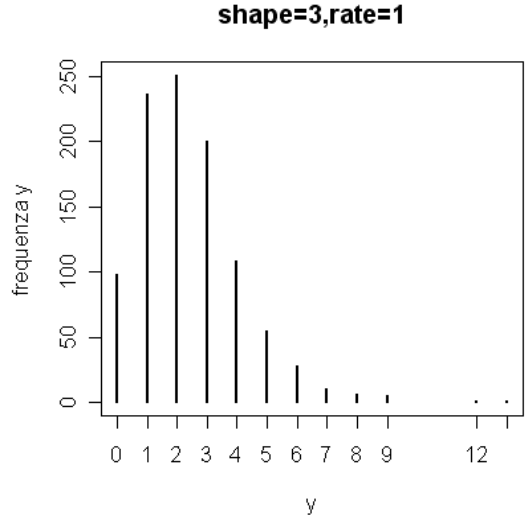
In alcuni casi, come mostra il grafico 51, la combinazione dei valori non è utile a descrivere il fenomeno in questione, poichè l'evento "0 gol" è troppo probabile.

Grafico 51: distribuzione di frequenza della variabile Y con shape e rate fissati a priori



In altri casi, grafico 52, la coda destra ha una probabilità troppo alta.

Grafico 52: distribuzione di frequenza della variabile Y con shape e rate fissati a priori



A questo punto abbiamo testato vari metodi di stima

Metodo di stima tramite massima verosimiglianza

Come fatto per la variabile costruita tramite trasformazioni della variabile normale, poiché non esiste una derivata prima esatta della funzione di log-verosimiglianza,

la prima stima si basa sul metodo numerico di Newton-Raphson. Tramite questo metodo vengono determinate le soluzioni per le equazioni ottenute dalla funzione di log-verosimiglianza della variabile. La precisione nelle stime dipende pertanto dalla numerosità del campione che viene generato.
Le stime sono quindi:

$$\hat{shape} = \tilde{shape}$$

$$\hat{rate} = \tilde{rate}$$

.
Gli intervalli di confidenza con alfa fissato al 95% sono determinati per shape tramite:

$$\left(\tilde{shape} - q_{0,975} \cdot \sqrt{\sigma^2_{shape}} ; \tilde{shape} + q_{0,975} \cdot \sqrt{\sigma^2_{shape}} \right)$$

Per rate tramite:

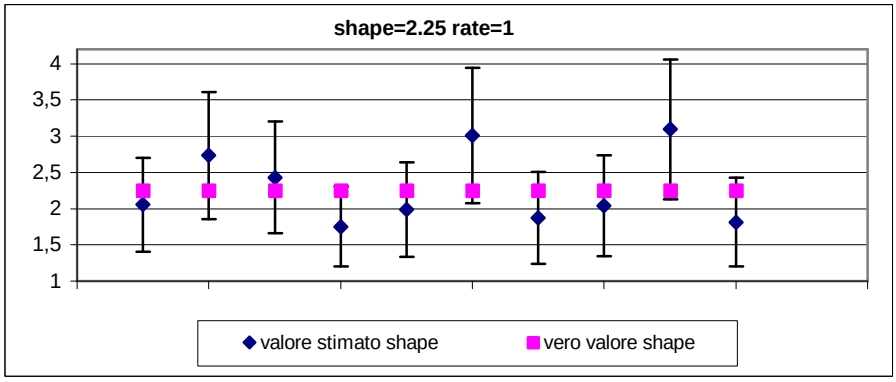
$$\left(\tilde{rate} - q_{0,975} \cdot \sqrt{\sigma^2_{rate}} ; \tilde{rate} + q_{0,975} \cdot \sqrt{\sigma^2_{rate}} \right)$$

Dove σ^2_{shape} e σ^2_{rate} rappresentano gli elementi della diagonale principale della inversa della matrice di informazione della funzione di log-verosimiglianza.

Generazione di 100 numeri

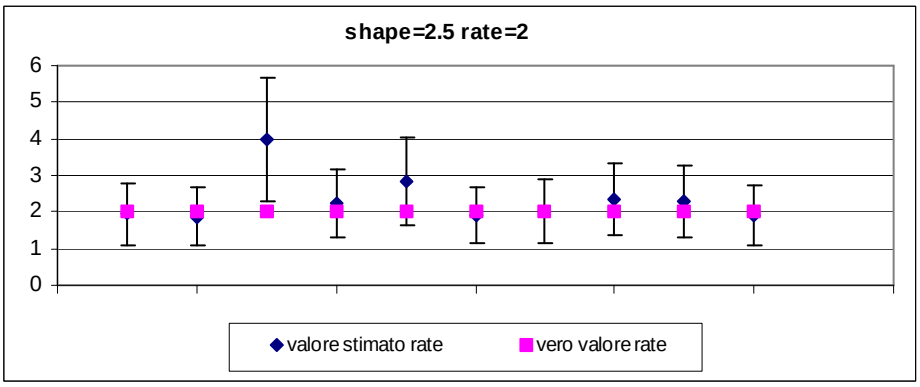
La generazione di un numero limitato di dati non compromette l'affidabilità delle stime. Con questo metodo infatti otteniamo stime precise in quanto gli intervalli di confidenza comprendono in quasi tutti i casi studiati il vero valore assegnato a priori al parametro, caratterizzandosi però per una ampiezza molto elevata.
Andando ad analizzare che cosa accade per il parametro shape, otteniamo per 10 campioni di 100 numeri casuali degli intervalli molto ampi, seppure comprendenti il valore assegnato.

Grafico 53: Stima di shape con metodo tramite massima verosimiglianza



La stessa cosa accade per l'altro parametro caratteristico della distribuzione gamma.

Grafico 54: Stima di rate con metodo tramite massima verosimiglianza



In questo esempio osserviamo che in un campione su dieci il vero valore non appartiene all'intervallo stimato. Questo è dovuto al fatto che la creazione di intervalli di confidenza di livello 95% porta ad includere in 95 casi su 100 il vero valore del parametro.

Generazione di 1000 numeri

La precisione delle stime con metodo esatto era osservabile anche avendo a disposizione un numero limitato di dati. Generando più numeri casuali, otteniamo inoltre degli intervalli di confidenza con una ampiezza minore: gli estremi

dell'intervallo sono più vicini al valore assegnato al parametro per generare il vettore di numeri casuali.

Anche in questo caso per ogni coppia di valori assegnata ai parametri sono stati originati 10 campioni.

Grafico 55: Stima di shape con metodo tramite massima verosimiglianza

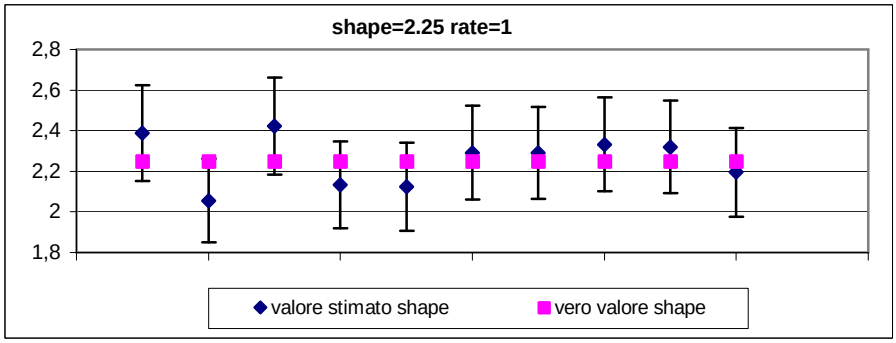
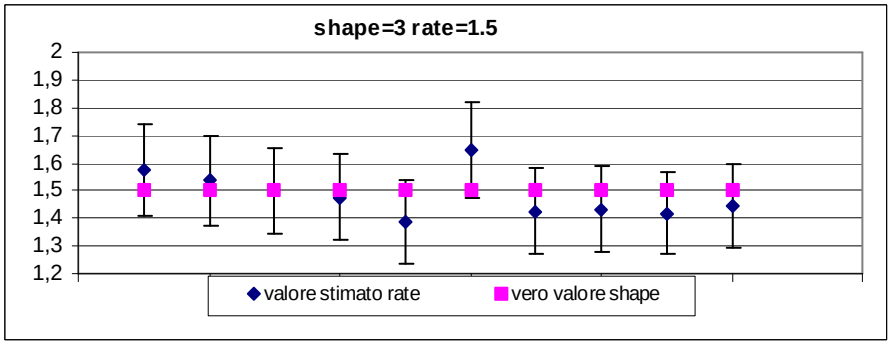


Grafico 56: Stima di rate con metodo tramite massima verosimiglianza



Metodo di stima approssimato

Questo secondo metodo si basa sulle caratteristiche della variabile X di partenza avente distribuzione Gamma.

Poiché Y vale 0, se X è compresa tra 0 e 1, la prima operazione da fare è riscalare la nostra Y aggiungendo 0,5.

Quindi partendo dalle n determinazioni y di Y:

y=y+0,5

A questo punto i parametri mu e sigma vengono stimati come media e deviazione standard delle y.

Poiché $E(X)=\frac{shape}{rate}$ e $Var(X)=\frac{shape}{rate^2}$, abbiamo supposto che $\bar{y}=\frac{\hat{shape}}{\hat{rate}}$ e che $\sigma_y^2=\frac{\hat{shape}}{\hat{rate}^2}$.

In questo modo è immediato ottenere le stime dei parametri:

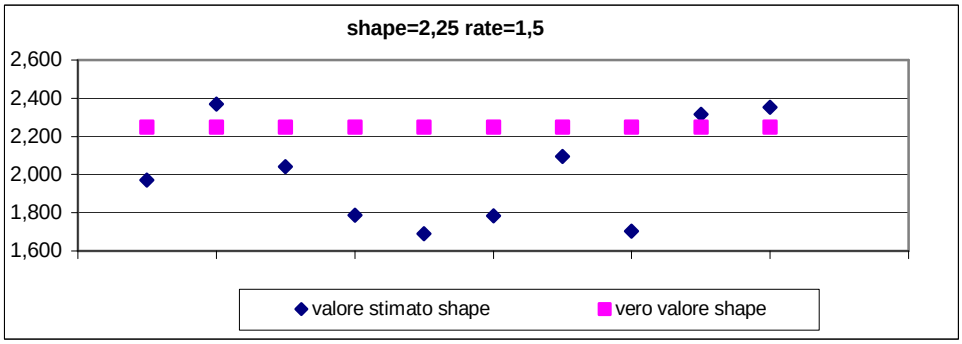
$$\hat{shape}=\frac{\bar{y}^2}{\sigma_y^2}$$

$$\hat{rate}=\frac{\bar{y}}{\sigma_y^2}$$

Generazione di 100 numeri

Fin dalle simulazioni ci si accorge che questo secondo metodo non riesce ad avere la stessa precisione del metodo esatto. Le stime dei due parametri infatti in tutti i casi si distanziano in modo rilevante dai veri valori con i quali le distribuzioni di numeri casuali sono state generate.

Grafico 57: Stima di shape con metodo approssimato



Da questa combinazione di valori non sembrano esserci delle distorsioni sistematiche nelle stime. La stima del parametro shape si colloca in modo asistematico al di sopra e al di sotto del vero valore del parametro.

I seguenti due grafici invece permettono di cogliere lo stretto legame esistente negli errori delle stime.

Grafico 58: Stima di shape con metodo approssimato

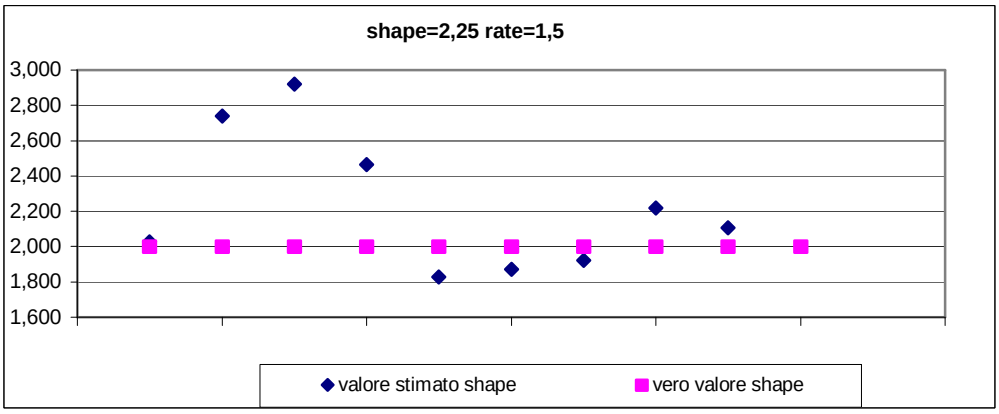
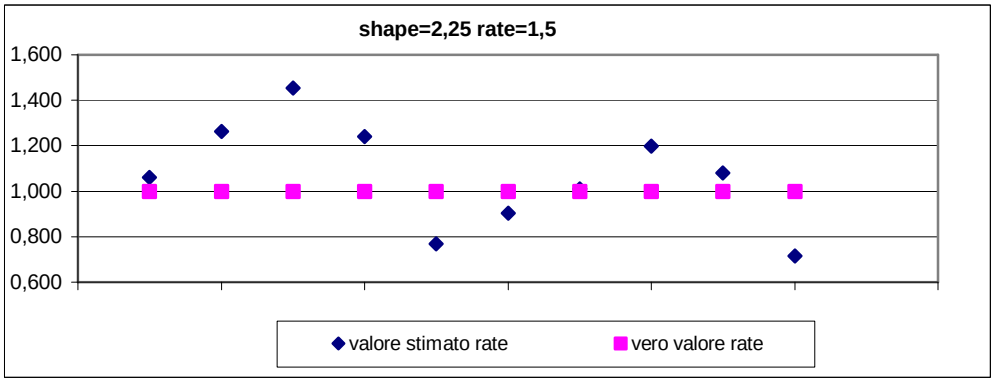


Grafico 59: Stima di rate con metodo approssimato



E' interessante osservare, confrontando i grafici 58 e 59 che sia la sovrastima che la sottostima del parametro shape sono accompagnate dall'analoga distorsione per il parametro rate. Questo è dovuto alla dipendenza delle stime ottenute con questo metodo dalla media e dalla varianza dei numeri casuali generati.
Tuttavia per nessuna combinazione di valori assegnata ai due parametri si osserva una distorsione sistematica delle stime.

Generazione di 1000 numeri

Generalmente un aumento di quantità di dati a disposizione delle stime dovrebbe aumentare la precisione delle stime.

In questo caso ciò non accade: generando un maggior numero di determinazioni casuali si osserva una sistematicità nella distorsione delle stime; inoltre andando a generare mille numeri anziché cento ci si accorge che la stima tramite metodo approssimato è inferiore al vero valore del parametro.

Grafico 60: Stima di shape con metodo approssimato

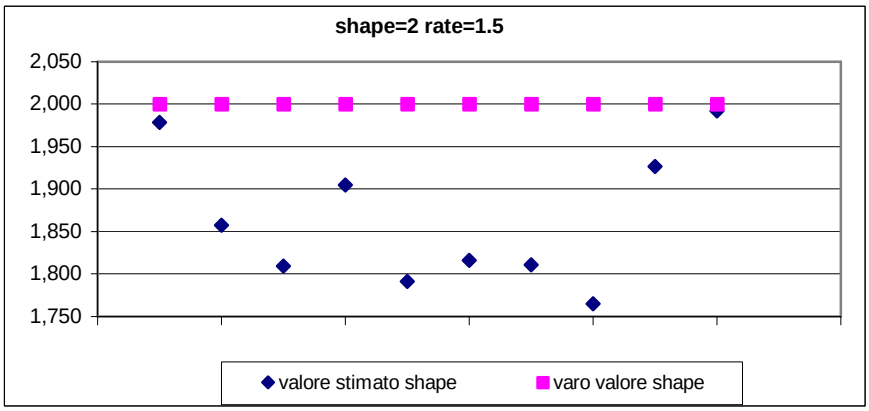
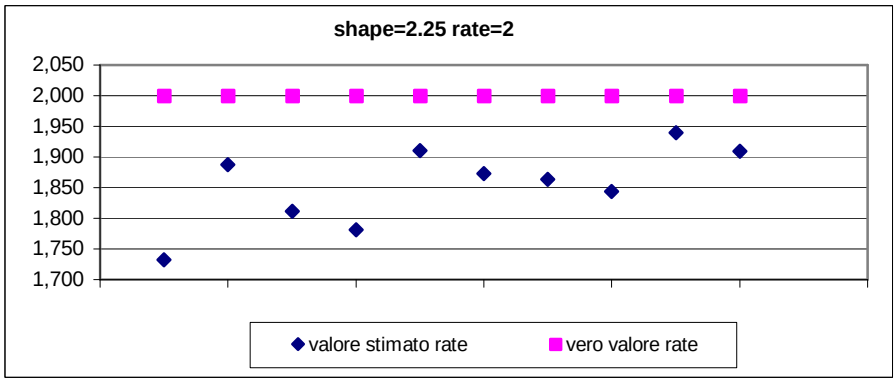


Grafico 61: Stima di rate con metodo approssimato



I precedenti grafici evidenziano la sottostima che contraddistingue il metodo approssimato. È bene sottolineare che tale distorsione nelle stime approssimate si riscontra per ogni coppia di valori utilizzata per generare il vettore di numeri casuali.

Stime

Andiamo ad analizzare la distribuzione della nostra variabile sulla base delle stime ottenute tramite il metodo esatto.

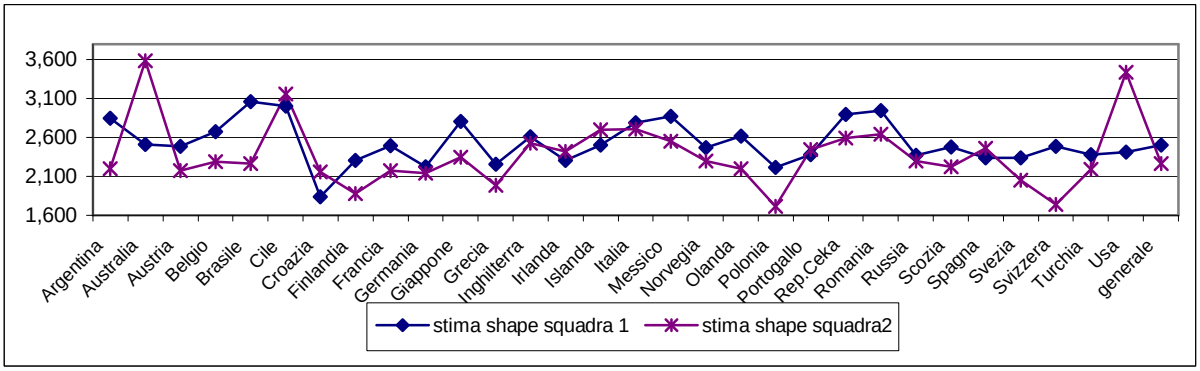
Tabella 35: stime con metodo esatto dei parametri shape e rate e dei rispettivi standard error per squadra 1 in ciascuna nazione

nazione	Shape team 1	s.e. shape 1	Rate team 1	s.e. rate 1
Argentina	2,844	0,6030	1,437	1,142
Australia	2,508	0,6407	1,295	1,192
Austria	2,489	0,6569	1,130	1,369
Belgio	2,679	0,6288	1,260	1,270
Brasile	3,060	0,5858	1,386	1,232
Cile	3,000	0,5889	1,439	1,172
Croazia	1,838	0,7923	0,786	1,703
Finlandia	2,308	0,6830	1,105	1,347
Francia	2,496	0,6419	1,352	1,137
Germania	2,228	0,6956	1,067	1,372
Giappone	2,810	0,6078	1,426	1,143
Grecia	2,255	0,6893	1,121	1,311
Inghilterra	2,613	0,6323	1,315	1,199
Irlanda	2,307	0,6727	1,230	1,204
Islanda	2,502	0,6508	1,211	1,277
Italia	2,789	0,6105	1,388	1,172
Messico	2,873	0,6040	1,380	1,198
Norvegia	2,471	0,6664	1,007	1,537
Olanda	2,616	0,6431	1,117	1,421
Polonia	2,216	0,6941	1,105	1,319
Portogallo	2,383	0,6598	1,280	1,175
Rep.Ceka	2,898	0,5923	1,555	1,061
Romania	2,949	0,5879	1,555	1,071
Russia	2,371	0,6677	1,176	1,280
Scozia	2,479	0,6586	1,135	1,360
Spagna	2,341	0,6675	1,238	1,205
Svezia	2,337	0,6732	1,168	1,279
Svizzera	2,483	0,6598	1,114	1,387
Turchia	2,380	0,6680	1,155	1,307
Usa	2,415	0,6562	1,283	1,180
generale	2,504	0,6502	1,220	1,267

Tabella 36: stime con metodo esatto dei parametri shape e rate e dei rispettivi standard error per squadra 2 in ciascuna nazione

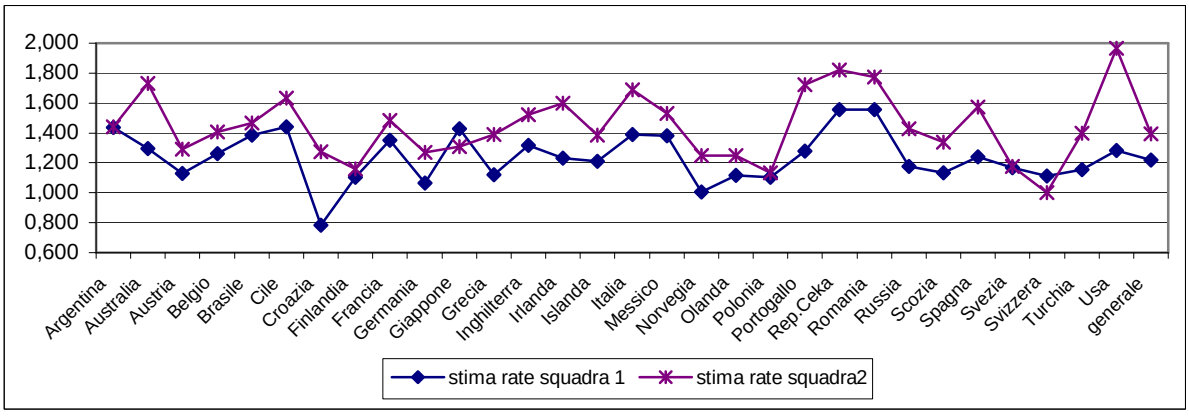
nazione	Shape team 1	s.e. shape 1	Rate team 1	s.e. rate 1
Argentina	2,202	0,6664	1,441	0,9942
Australia	3,590	0,5325	1,731	1,060
Austria	2,176	0,6840	1,293	1,108
Belgio	2,286	0,6609	1,405	1,042
Brasile	2,268	0,6576	1,466	0,992
Cile	3,159	0,5681	1,631	1,056
Croazia	2,155	0,6896	1,275	1,118
Finlandia	1,877	0,7342	1,158	1,151
Francia	2,178	0,6644	1,484	0,9579
Germania	2,143	0,6891	1,270	1,120
Giappone	2,346	0,6628	1,310	1,138
Grecia	1,988	0,6953	1,392	0,9763
Inghilterra	2,527	0,6281	1,521	1,010
Irlanda	2,421	0,6314	1,601	0,9349
Islanda	2,700	0,6196	1,386	1,154
Italia	2,712	0,6015	1,689	0,9379
Messico	2,554	0,6239	1,530	1,009
Norvegia	2,296	0,6736	1,251	1,181
Olanda	2,198	0,6850	1,250	1,154
Polonia	1,711	0,7610	1,133	1,121
Portogallo	2,447	0,6210	1,721	0,8692
Rep.Ceka	2,592	0,6022	1,819	0,8449
Romania	2,646	0,6009	1,774	0,8781
Russia	2,298	0,6578	1,426	1,028
Scozia	2,228	0,6725	1,340	1,080
Spagna	2,459	0,6305	1,574	0,9599
Svezia	2,055	0,7104	1,176	1,188
Svizzera	1,735	0,7777	1,001	1,287
Turchia	2,195	0,6736	1,397	1,026
Usa	3,441	0,5324	1,967	0,9054
generale	2,269	0,6641	1,393	1,047

Grafico 62: stime con metodo esatto del parametro shape per squadra 1 e squadra 2 in ciascuna nazione



Vediamo che il parametro shape è maggiore per la squadra 1 in quasi tutti i campionati ad eccezione di quello statunitense e di quello australiano. Questo era prevedibile in quanto il parametro a rappresenta il parametro di posizione della distribuzione. Un valore alto equivale ad un elevato numero di gol segnati.

Grafico 63: stime con metodo esatto dei parametri rate per squadra 1 e squadra 2 in ciascuna nazione



Viceversa il parametro rate tende ad essere più alto per la squadra 2 anche se in questo caso l'andamento è meno evidente.

Una volta stimati i parametri, è possibile utilizzare la nostra variabile Z per valutare se la distribuzione simulata tramite modello rispecchia la distribuzione empirica.

Confronto fra la distribuzione empirica e la distribuzione ottenuta tramite modello

La probabilità di "k gol segnati dalla squadra 1" per la distribuzione empirica è calcolato come rapporto tra il numero di partite appartenenti al campione in cui la squadra 1 ha segnato k gol e il numero totale di partite appartenenti al campione.
La probabilità di "k gol segnati dalla squadra 1" per la distribuzione tramite modello è ottenuta tramite il calcolo della probabilità esatta. Per questo modello

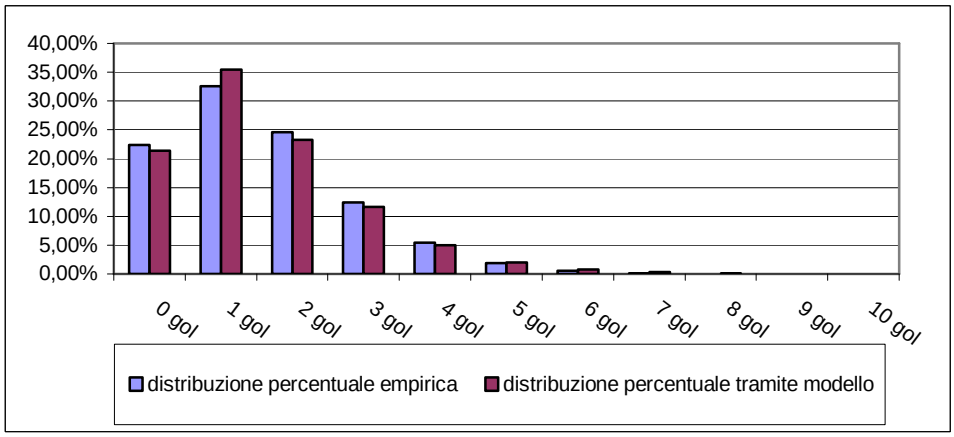
$$P(Y=k)=\frac{a^{p}\cdot e^{-(k+1)a}\cdot(k+1)^{p-1}}{\Gamma(p)}-\frac{a^{p}\cdot e^{-ka}\cdot k^{p-1}}{\Gamma(p)}$$

Tabella 37: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	22,40%	21,36%
1 gol	32,54%	35,46%
2 gol	24,60%	23,31%
3 gol	12,39%	11,59%
4 gol	5,43%	5,04%
5 gol	1,90%	2,02%
6 gol	0,55%	0,77%
7 gol	0,11%	0,28%
8 gol	0,05%	0,10%
9 gol	0,02%	0,04%
10 gol	0%	0,01%

La distribuzione simulata tramite il modello si adatta bene ai dati. L'errore più significativo è la sovrastima della probabilità dell'evento "1 gol".

Grafico 64: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione tramite modello



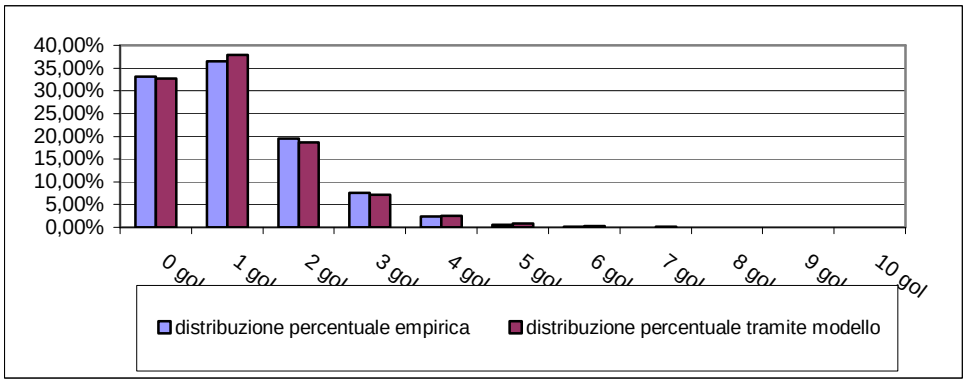
E' evidente che tra le due distribuzioni sono presenti delle differenze, anche se non sembra esserci una distorsione sistematica nella distribuzione.
Per approfondire questo problema è utile fare degli altri tentativi.

Tabella 38: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	33,08%	32,67%
1 gol	36,49%	37,85%
2 gol	19,54%	18,64%
3 gol	7,62%	7,21%
4 gol	2,39%	2,48%
5 gol	0,61%	0,80%
6 gol	0,17%	0,25%
7 gol	0,05%	0,07%
8 gol	0,02%	0,02%
9 gol	0,02%	0,00%
10 gol	0,01%	0,00%

In questo caso gli errori sono meno evidenti.

Grafico 65: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione tramite modello



Nel grafico 65 le due distribuzioni sembrano quasi coincidere. Le differenze tra le due distribuzioni sembrano annullarsi: la differenza è inferiore al punto percentuale per ogni classe di gol.

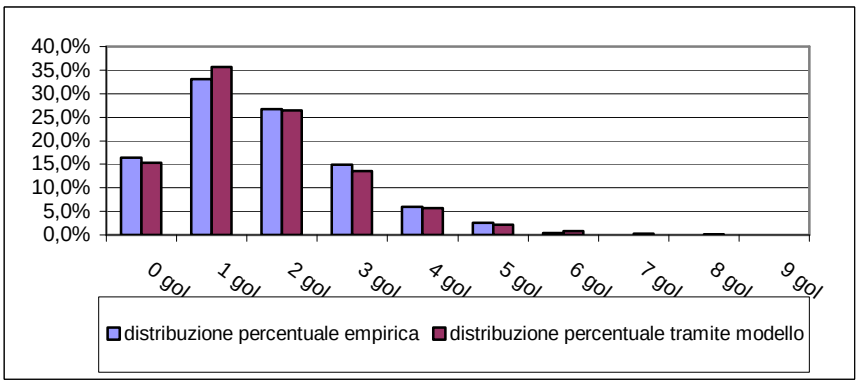
Per continuare la nostra analisi, scegliamo la distribuzione empirica dei gol segnati dalla squadra in casa nel campionato brasiliano, caratterizzata dalla stima del parametro shape, che risulta essere tra le maggiori.

Tabella 39: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato brasiliano per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	16,37%	15,33%
1 gol	33,08%	35,60%
2 gol	26,66%	26,48%
3 gol	14,93%	13,51%
4 gol	5,97%	5,74%
5 gol	2,54%	2,19%
6 gol	0,44%	0,78%
7 gol	0,00%	0,26%
8 gol	0,00%	0,09%
9 gol	0,00%	0,03%

Vediamo che le due distribuzioni sono simili, anche se fino ai tre gol esiste una certa differenza. Il nostro modello tende a dare troppo peso all'evento "1 gol" rispetto agli altri.
Questo è certamente dovuto al fatto che, rispetto all'analisi dell'intero campione , abbiamo meno dati a disposizione.

Grafico 66: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato brasiliano per la distribuzione empirica e per la distribuzione tramite modello



Il grafico 66 mostra che l'evento "1 gol" soffre di una sovrastima della probabilità del modello rispetto a quella empirica bilanciata da una sottostima per gli eventi "0 gol" e "2 gol". La probabilità dell'evento "2 gol" coincide per le due distribuzioni.

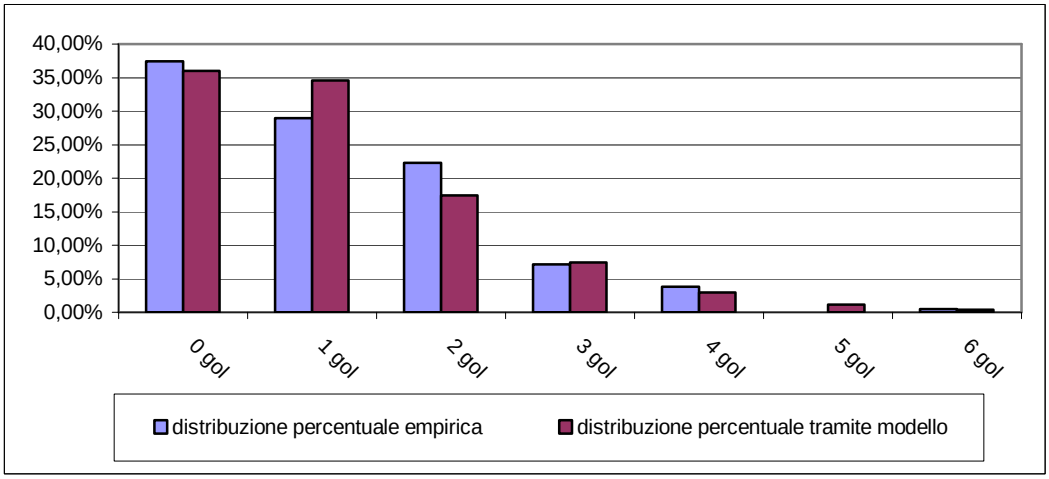
Se invece scegliamo la distribuzione per i gol della squadra fuori casa in Finlandia, caratterizzata da una stima del parametro shape particolarmente bassa, ci accorgiamo che anche in questo caso le cose non cambiano e vi è una sovrastima dell'evento "1 gol".

Tabella 40: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato finlandese per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	37,44%	35,99%
1 gol	28,91%	34,54%
2 gol	22,27%	17,41%
3 gol	7,11%	7,42%
4 gol	3,79%	2,92%
5 gol	0,00%	1,10%
6 gol	0,47%	0,40%
7 gol	0,00%	0,14%
8 gol	0,00%	0,05%
9 gol	0,00%	0,02%

Il fatto di avere un basso numero di partite a disposizione (211) fa sì che vi sia una distorsione abbastanza evidente.

Grafico 67: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato finlandese per la distribuzione empirica e per la distribuzione tramite modello



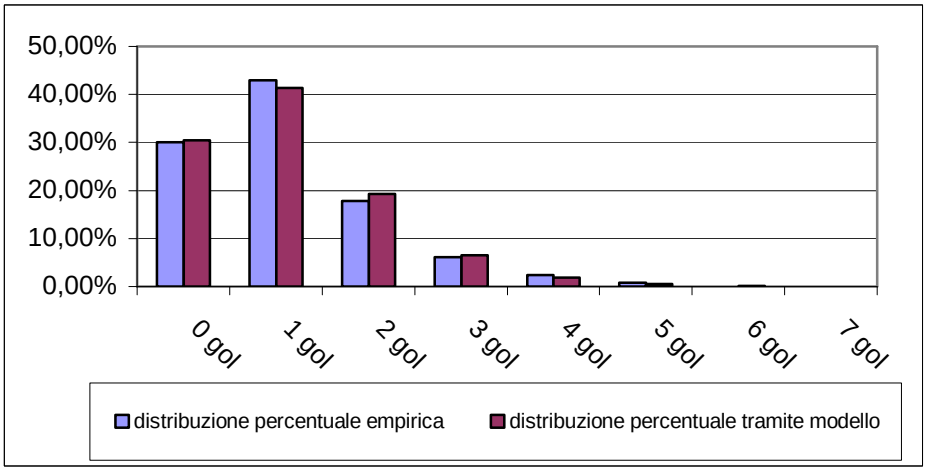
Il grafico 67 mostra la distorsione del modello dovuta allo scarso numero di dati a disposizione.

Andiamo ora a verificare che cosa accade al variare del parametro rate.
Prendiamo in considerazione i gol segnati dalla squadra in trasferta nel campionato italiano.

Tabella 41: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato italiano per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	30,00%	30,43%
1 gol	42,93%	41,33%
2 gol	17,80%	19,25%
3 gol	6,10%	6,48%
4 gol	2,44%	1,86%
5 gol	0,73%	0,49%
6 gol	0,00%	0,12%
7 gol	0,00%	0,03%

Grafico 68: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato italiano per la distribuzione empirica e per la distribuzione tramite modello



Anche per questo esempio il problema è relativo alle prime classi di eventi. Quello che si verifica, contrariamente al caso finlandese, è una sottostima della probabilità per l'evento "1 gol" compensata da una sovrastima per l'evento "0 gol" e "2 gol".

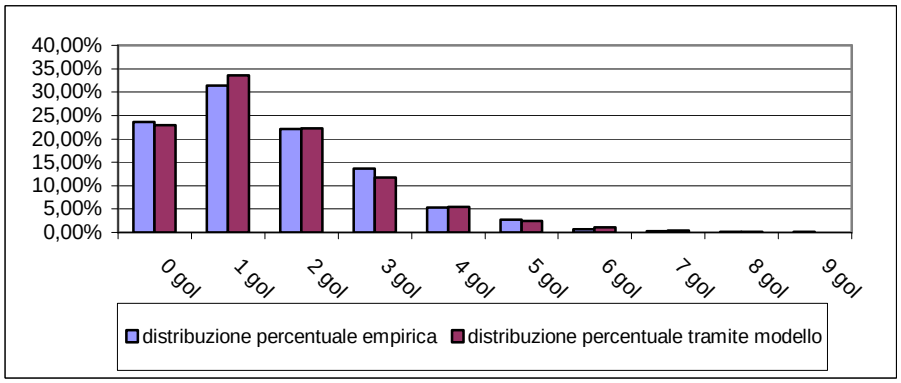
Studiamo invece il caso della distribuzione dei gol segnati dalla squadra di casa nel campionato tedesco, distribuzione caratterizzata da un valore stimato per il parametro rate piuttosto basso.

Tabella 42: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato tedesco per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	23,67%	22,87%
1 gol	31,45%	33,57%
2 gol	22,13%	22,20%
3 gol	13,63%	11,68%
4 gol	5,33%	5,51%
5 gol	2,66%	2,43%
6 gol	0,61%	1,03%
7 gol	0,31%	0,42%
8 gol	0,10%	0,17%
9 gol	0,10%	0,07%
10 gol	0,00%	0,03%
11 gol	0,00%	0,01%

Vediamo che nella distribuzione tramite modello è presente una sovrastima nelle frequenze percentuali degli eventi "1 gol" e "2 gol" compensati da una sottostima per i contigui eventi "0 gol" e "3 gol".

Grafico 69: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato tedesco per la distribuzione empirica e per la distribuzione tramite modello



Anche per l'analisi dei gol segnati dalla squadra 1 nel campionato tedesco si osservano delle differenze tra le due distribuzioni. In questo esempio la distorsione presenta un andamento ancora diverso da quello registrato negli altri esempi.

Conclusioni

Possiamo quindi concludere che la distribuzione ottenuta tramite modello Gamma sembra riuscire a descrivere in maniera opportuna il fenomeno studiato nel caso dell'intero campione.

Quando invece si hanno meno dati a disposizione si ha una distorsione nelle probabilità per gli eventi vicini all'origine. Questa distorsione porta in quasi tutti i casi ad una sovrastima dell'evento "1 gol" che viene compensato dagli eventi contigui.

Il risultato è comunque molto soddisfacente in quanto per questo modello abbiamo da un lato un buon adattamento ai dati, analogamente a quanto riscontrato per il modello di Poisson, dall'altro lato la possibilità di separare il valore fissato per il parametro di posizione e per il parametro di variabilità, analogamente a quanto accade nel modello normale.

Il modello Weibull

Introduzione

Dopo aver effettuato l'analisi tramite l'utilizzo dei modelli Poisson, Normale e Gamma, ho deciso di valutare l'adattamento ai dati calcistici di una variabile avente come distribuzione di probabilità quella originata dalla distribuzione di Weibull.
Il modello Weibull è originato da una variabile casuale continua che assume valori positivi. La distribuzione di Weibull è utilizzata, nell'ambito dei controlli di qualità in campo industriale, per misurare la vita media e l'affaticamento delle componenti materiali dei prodotti; viene utilizzata inoltre per studi sulla velocità del vento.
La distribuzione viene definita da due parametri chiamati shape e scale.

$$X \sim Weibull(a,b)$$

dove con a indichiamo shape e con b scale.
La funzione di probabilità è data da:

$$P(X=x)=\frac{a}{b}\cdot\left(\frac{x}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{x}{b}\right)^{a}\right)$$

Dal momento che i gol segnati sono un fenomeno discreto è necessario che la nostra variabile assuma solamente valori discreti. Per questa ragione creiamo una nuova variabile discreta. La nuova variabile Y è tale che:

$$Y=y \qquad \text{se solo se } x\le X<x+1.$$

Pertanto possiamo descrivere la funzione di probabilità della variabile Y:

$$P(Y=0)=\frac{a}{b}\cdot\left(\frac{1}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{1}{b}\right)^{a}\right)$$

$$P(Y=1)=\frac{a}{b}\cdot\left(\frac{2}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{2}{b}\right)^{a}\right)-\frac{a}{b}\cdot\left(\frac{1}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{1}{b}\right)^{a}\right)$$

$$P(Y=2)=\frac{a}{b}\cdot\left(\frac{3}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{3}{b}\right)^{a}\right)-\frac{a}{b}\cdot\left(\frac{2}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{2}{b}\right)^{a}\right)$$

.

.

.

In generale:

$$P(Y=y)=\frac{a}{b}\cdot\left(\frac{y+1}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{y+1}{b}\right)^{a}\right)-\frac{a}{b}\cdot\left(\frac{y}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{y}{b}\right)^{a}\right)$$

Simulazione di dati

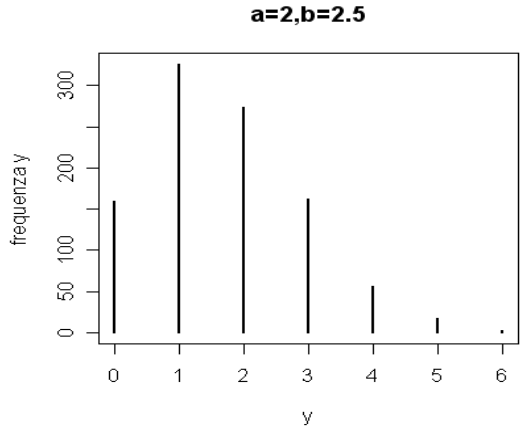
A questo punto cerchiamo di capire quali valori di a e b possono essere tali da descrivere il numero di gol segnati da una squadra di calcio.

Tabella 43: valori assunti da n, a e b nelle simulazioni

n	a	b
100	1	1,5
1000	1,25	2
	1,5	2,5
	1,75	
	2	

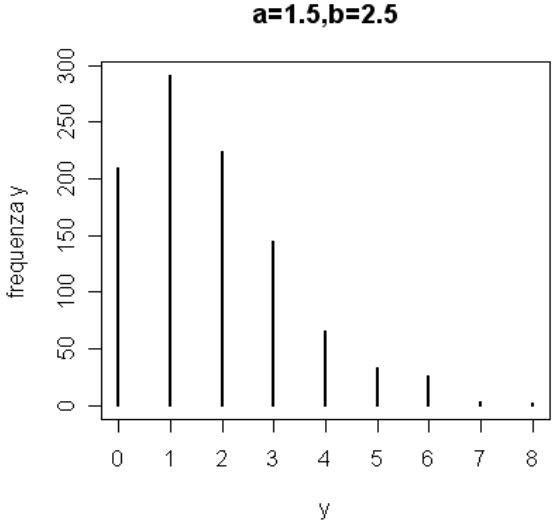
Nelle seguenti simulazioni sono stati generati mille numeri casuali, determinazioni di questa variabile casuale.

Grafico 70: distribuzione di frequenza della variabile Y con a e b fissati a priori



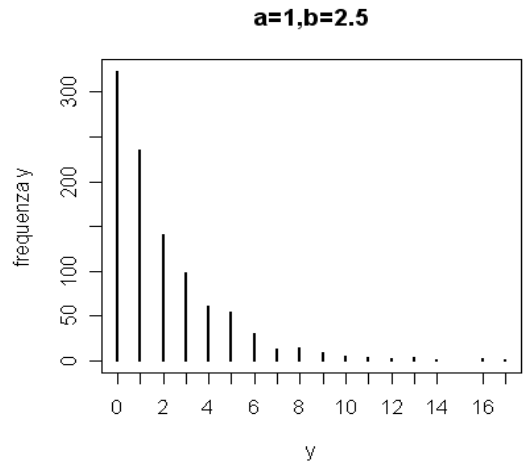
La coppia di valori 2 e 2,5 per i parametri a e b, da quanto emerge nel grafico 70, sembra poter descrivere in maniera soddisfacente la distribuzione di gol segnati da una squadra di calcio.

Grafico 71: distribuzione di frequenza della variabile Y con a e b fissati a priori



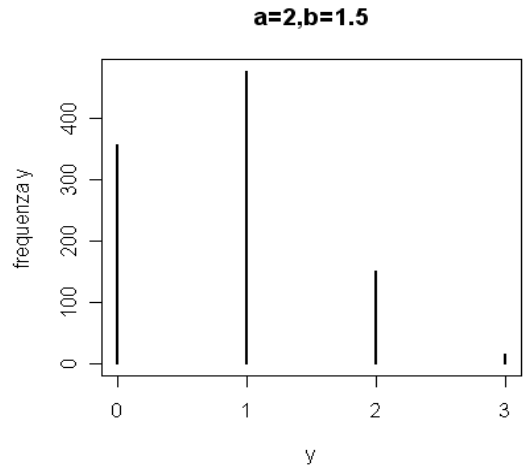
Anche se fissiamo a pari a 1,5 e b pari a 2,5, osserviamo una distribuzione di frequenza simile a quella dei gol segnati da una squadra di calcio.
Tuttavia non ogni combinazione tra quelle ottenibili dalla tabella 43 genera una distribuzione adatta a descrivere i gol segnati da una squadra di calcio.

Grafico 72: distribuzione di frequenza della variabile Y con a e b fissati a priori



Nel caso in cui il parametro a sia uguale ad 1 e il parametro b sia uguale a 2,5, gli eventi vicini all'origine pesano troppo perché il modello possa essere esplicativo del nostro fenomeno di interesse. Inoltre, come emerso dal grafico 72, si hanno alcune determinazioni di Y con valori troppo elevati per rappresentare i gol segnati da una squadra di calcio in una partita.

Grafico 73: distribuzione di frequenza della variabile Y con a e b fissati a priori



Per questa ultima coppia di valori,a=2 e b=1,5, si riscontra un range per le determinazioni di Y troppo ristretto rispetto a quello che si osserva empiricamente per il fenomeno studiato.

Metodo di stima tramite massima verosimiglianza

Le stime dei parametri sono i valori in base ai quali le derivate prime della funzione di log-verosimiglianza si annullano.

Poiché anche in questo caso la funzione di log-verosimiglianza non ha una derivata prima nota, i valori vengono determinati tramite il metodo numerico Newton-Raphson.

Le stime sono quindi:

$\hat{a} = \tilde{a}$

$\hat{b} = \tilde{b}$

Dove $\tilde{a}$ e $\tilde{b}$ sono i valori che annullano la derivata prima della log-verosimiglianza, rispettivamente rispetto ad a e b, determinati con metodo numerico.

L'intervallo di confidenza per il parametro shape è:

$$\left(\tilde{shape} - q_{0,975} \cdot \sqrt{\sigma^2_{shape}} ; \tilde{shape} + q_{0,975} \cdot \sqrt{\sigma^2_{shape}} \right)$$

L'intervallo di confidenza per il parametro rate è:

$$\left(\tilde{rate} - q_{0,975} \cdot \sqrt{\sigma^2_{rate}} ; \tilde{rate} + q_{0,975} \cdot \sqrt{\sigma^2_{rate}} \right)$$

Dove σ^2_{shape} e σ^2_{rate} rappresentano gli elementi della diagonale principale della inversa della matrice di informazione della funzione di log-verosimiglianza.

Generazione di 100 numeri

Per testare questo metodo di stima ho generato dei vettori composti da numeri casuali tramite la distribuzione di probabilità definita nella parte introduttiva di questo capitolo.
È interessante notare che la creazione di 100 numeri casuali porta a stime con intervalli di confidenza particolarmente ampi.
Il metodo numerico di Newton-Raphson, nonostante l'ampiezza degli intervalli, si dimostra molto preciso in quanto gli intervalli di confidenza comprendono per tutti i casi studiati la vera coppia di valori mediante la quale è stato generato il vettore casuale.
Di seguito è riportato un esempio per la stima di a e un esempio per la stima di b: in ciascuno di questi esempi sono stati creati dieci vettori composti da 100 determinazioni casuali della variabile con distribuzione di probabilità di cui sopra che ha avente come parametri fissati una combinazione della tabella 43.

Grafico 74: Stima di a con metodo esatto

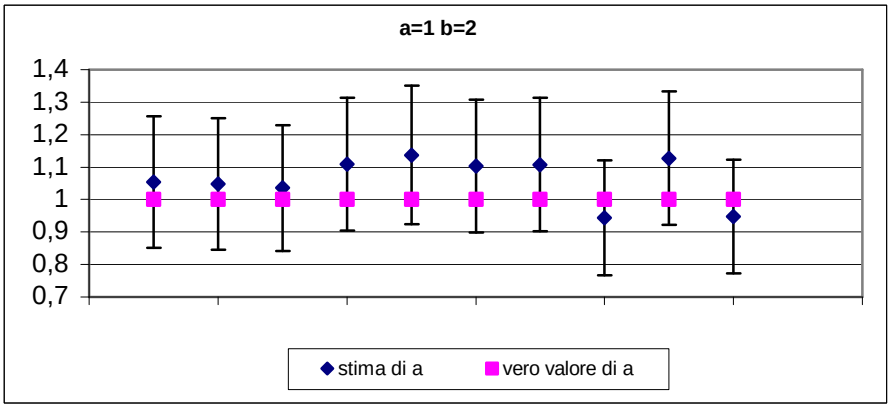
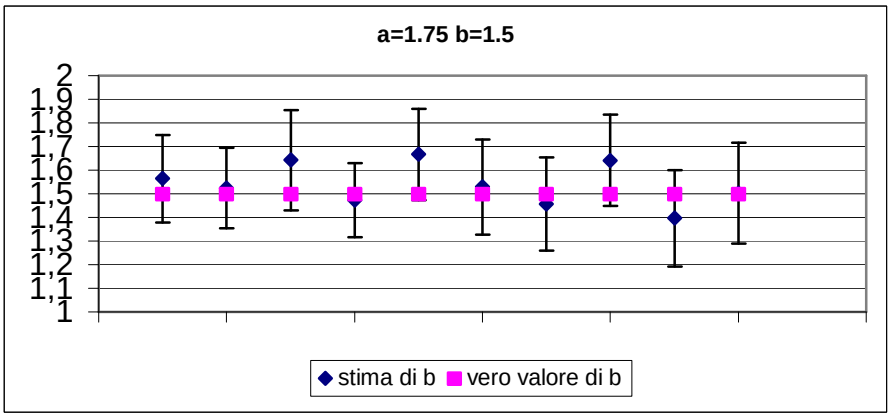


Grafico 75: Stima di b con metodo esatto

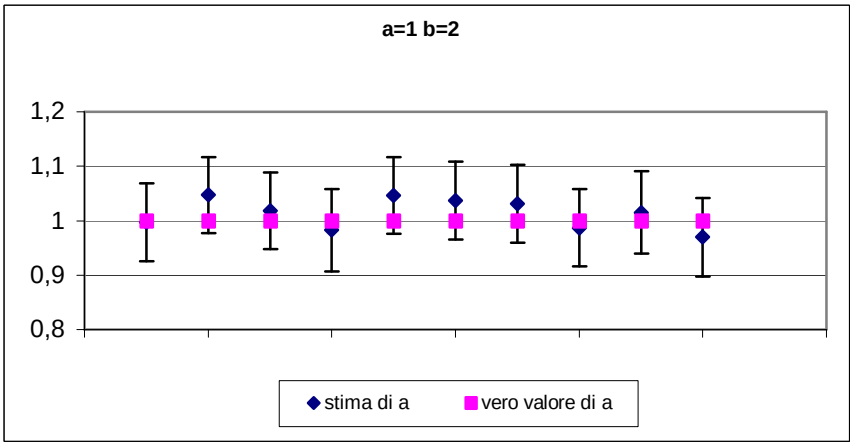


Dai due grafici è possibile osservare da un lato l'ampiezza degli intervalli di confidenza e dall'altro la correttezza del metodo di stima utilizzato. Vale la pena sottolineare infine che la stima puntuale non ha distorsioni sistematiche per nessuno dei due parametri.

Generazione di 1000 numeri

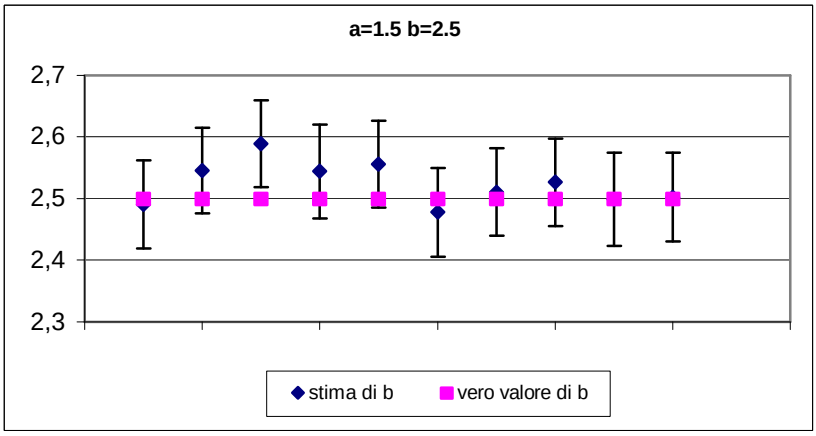
Generando un vettore casuale composto da un numero maggiore di determinazioni, si ha una riduzione nell'ampiezza degli intervalli.
Questo aspetto si accompagna ad una stima puntuale corretta.

Grafico 76: Stima di a con metodo esatto



Questo grafico, confrontato con il grafico 74, ci permette di cogliere la maggiore precisione ottenibile con una maggiore quantità di dati a disposizione.
Nel caso in cui il parametro a è fissato a 1, tutti gli intervalli si collocano nella fascia compresa tra 0,9 e 1,1.

Grafico 77: Stima di b con metodo esatto



Anche per il parametro b valgono i ragionamenti che sono stati fatti in precedenza. Per questo specifico esempio constatiamo come cinque stime puntuali su dieci corrispondano al valore fissato per il parametro b.
È bene sottolineare però che per ogni combinazione della tabella 43, sia generando cento numeri che generandone mille, le stime sono precise.
Addirittura in tutti i casi studiati ogni intervallo di confidenza, per ogni vettore creato, contiene la vera coppia di valori mediante la quale la distribuzione è stata creata.

Stime

I valori delle stime ottenute con il metodo descritto nella parte precedente sono:

Tabella 44: stime con metodo esatto dei parametri a e b e dei rispettivi standard error per la squadra 1 in ciascuna nazione

nazione	a team 1	s.e. a team 1	b team1	s.e. b team 1
Argentina	1,78	0,676	2,22	0,783
Australia	1,82	0,590	2,17	0,817
Austria	1,69	0,715	2,46	0,673
Belgio	1,71	0,726	2,38	0,704
Brasile	1,86	0,662	2,48	0,734
Cile	1,80	0,696	2,34	0,752
Croazia	1,36	0,934	2,56	0,526
Finlandia	1,54	0,808	2,31	0,656
Francia	1,66	0,709	2,06	0,786
Germania	1,55	0,784	2,31	0,658
Giappone	1,70	0,734	2,20	0,757
Grecia	1,54	0,791	2,23	0,679
Inghilterra	1,69	0,718	2,22	0,745
Irlanda	1,59	0,739	2,08	0,746
Islanda	1,66	0,728	2,30	0,707
Italia	1,77	0,679	2,25	0,767
Messico	1,75	0,715	2,33	0,734
Norvegia	1,69	0,722	2,74	0,608
Olanda	1,72	0,715	2,62	0,646
Polonia	1,54	0,786	2,22	0,680
Portogallo	1,60	0,739	2,07	0,757
Rep.Ceka	1,81	0,649	2,09	0,840
Romania	1,79	0,672	2,12	0,822
Russia	1,63	0,727	2,24	0,713
Scozia	1,66	0,736	2,44	0,669
Spagna	1,60	0,737	2,10	0,745
Svezia	1,59	0,758	2,22	0,703
Svizzera	1,67	0,730	2,49	0,659
Turchia	1,64	0,726	2,29	0,700
Usa	1,63	0,717	2,10	0,762
generale	1,65	0,734	2,29	0,709

La stima relativa alla squadra di casa per il parametro a varia tra 1,3 e 1,9; la stima relativa alla squadra di casa per il parametro b invece varia tra 2 e 2,8.
Balza agli occhi che gli errori standard sono più alti di quelli ottenuti per le altre distribuzioni.

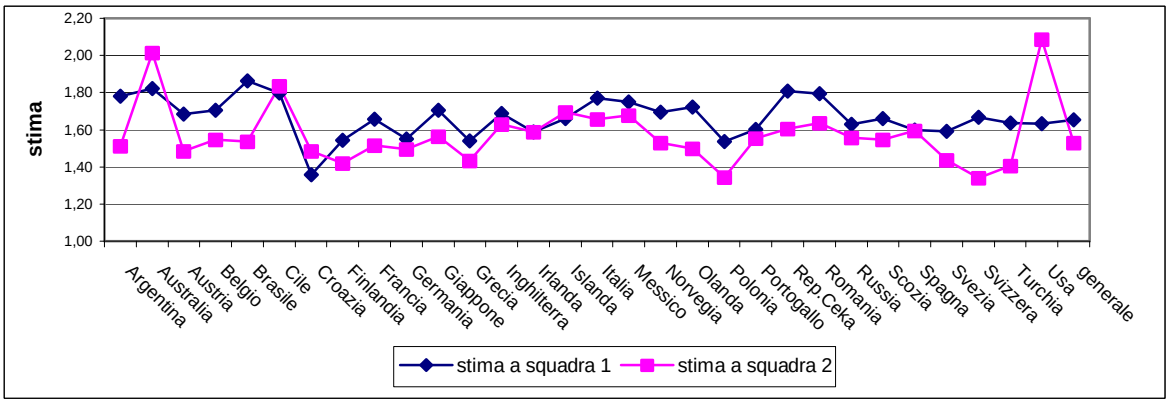
Tabella 45: stime con metodo esatto dei parametri a e b e dei rispettivi standard error per la squadra 2 in ciascuna nazione

nazione	a team 2	s.e. a team 2	b team2	s.e. b team 2
Argentina	1,51	0,744	1,68	0,872
Australia	2,01	0,610	2,34	0,836
Austria	1,49	0,793	1,85	0,784
Belgio	1,55	0,744	1,80	0,837
Brasile	1,53	0,736	1,71	0,873
Cile	1,84	0,672	2,18	0,820
Croazia	1,48	0,791	1,86	0,779
Finlandia	1,42	0,781	1,77	0,782
Francia	1,52	0,721	1,62	0,908
Germania	1,50	0,776	1,86	0,786
Giappone	1,56	0,761	1,98	0,770
Grecia	1,43	0,760	1,56	0,890
Inghilterra	1,63	0,713	1,85	0,858
Irlanda	1,59	0,708	1,67	0,919
Islanda	1,70	0,722	2,17	0,762
Italia	1,66	0,713	1,79	0,898
Messico	1,68	0,675	1,86	0,875
Norvegia	1,53	0,794	2,03	0,738
Olanda	1,50	0,794	1,94	0,757
Polonia	1,34	0,807	1,64	0,803
Portogallo	1,55	0,727	1,57	0,957
Rep.Ceka	1,60	0,704	1,58	0,980
Romania	1,64	0,699	1,65	0,953
Russia	1,56	0,729	1,78	0,851
Scozia	1,55	0,732	1,84	0,821
Spagna	1,59	0,720	1,73	0,893
Svezia	1,44	0,835	1,91	0,735
Svizzera	1,34	0,857	1,88	0,700
Turchia	1,40	0,888	1,71	0,802
Usa	2,09	0,525	1,97	1,01
generale	1,53	0,758	1,80	0,829

Per quanto riguarda la squadra in trasferta, le stime del parametro a sono comprese tra 1,3 e 2,1, mentre quelle del parametro b tra 1,5 e 2,3. Le stime

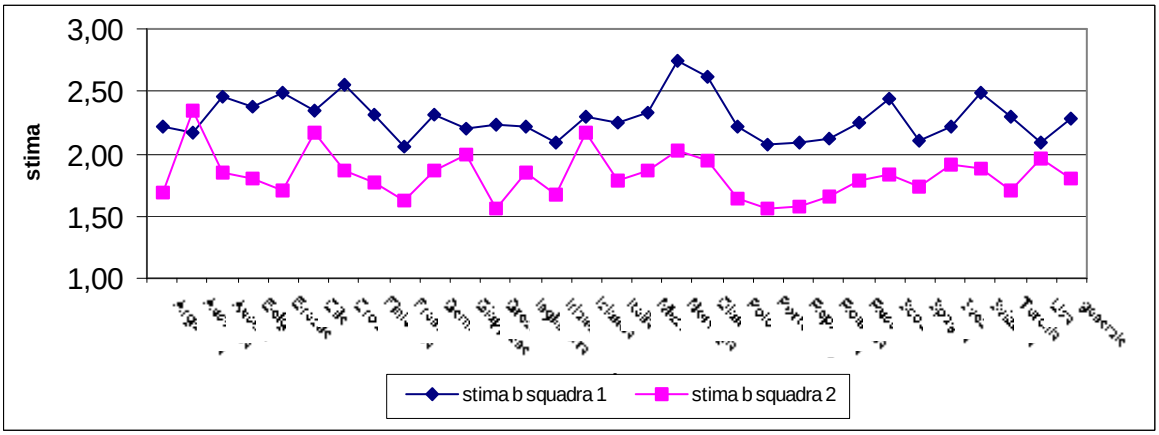
relative alla seconda squadra, quindi, si collocano in un range più ampio di possibili valori.

Grafico 78: stime con metodo esatto del parametro a per la squadra 1 e la squadra 2 in ciascuna nazione



Pur essendo un parametro di posizione, le stime di a per la squadra di casa e le stime di a per la squadra in trasferta non sono così diverse. Si osservi, però che in quasi tutti i casi, ad eccezione di Australia e Stati Uniti, che il valore stimato per la squadra in casa è più elevato di quello relativo alla squadra che gioca fuori casa.

Grafico 79: stime con metodo esatto del parametro b per la squadra 1 e la squadra 2 in ciascuna nazione



Le stime del parametro b ci permettono di osservare le differenze tra la squadra di casa e quella in trasferta.

Il valore di b per la squadra 1 è in quasi tutti i casi, esclusa l'Australia, maggiore rispetto al valore della squadra 2.

Confronto fra la distribuzione empirica e la distribuzione ottenuta tramite modello

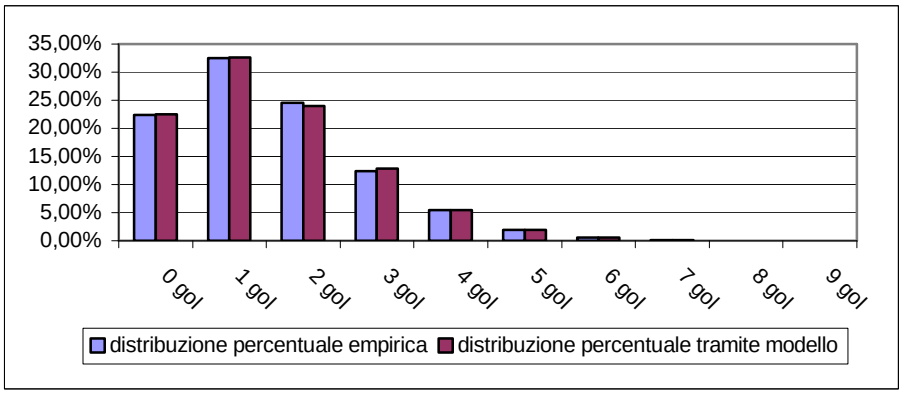
Andiamo ad analizzare ora la capacità del modello, ottenuto tramite la distribuzione Weibull, di descrivere il fenomeno studiato. Vogliamo cioè confrontare la distribuzione empirica dei gol segnati con la distribuzione simulata tramite il nostro modello avente come parametri le stime esatte.

Tabella 46: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	22,40%	22,50%
1 gol	32,54%	32,56%
2 gol	24,60%	23,96%
3 gol	12,39%	12,86%
4 gol	5,43%	5,47%
5 gol	1,90%	1,91%
6 gol	0,55%	0,56%
7 gol	0,11%	0,14%
8 gol	0,05%	0,03%
9 gol	0,02%	0,00%

La distribuzione di probabilità ottenuta tramite il nostro modello è molto precisa e si avvicina in modo impressionante a quella empirica. È evidente come il modello Weibull sembri riuscire a descrivere il fenomeno studiato in modo molto più soddisfacente rispetto ai modelli precedentemente utilizzati.

Grafico 80: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per la distribuzione simulata tramite modello

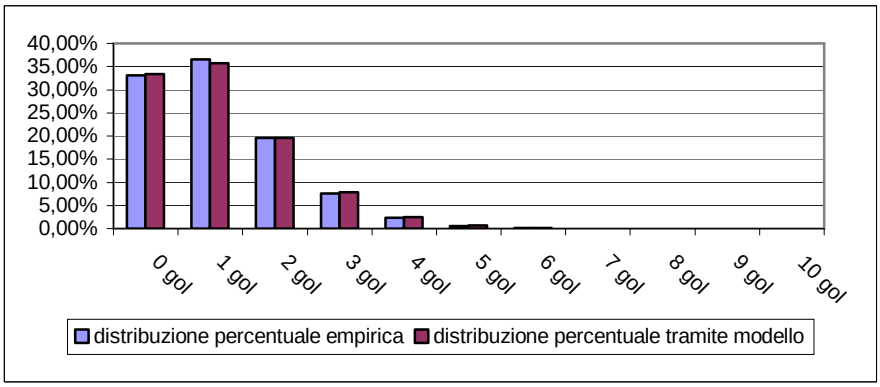


Il grafico 80 permette di apprezzare la quasi assoluta coincidenza tra le due distribuzioni. Per ciascun evento le due colonne dell'istogramma sembrano addirittura coincidere. Questo risultato è molto importante poiché possiamo ipotizzare che i gol segnati da una squadra di calcio possano essere descritti tramite una variabile discreta originata da una distribuzione di Weibull.

Tabella 47: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	33,08%	33,43%
1 gol	36,49%	35,69%
2 gol	19,54%	19,64%
3 gol	7,62%	7,89%
4 gol	2,39%	2,52%
5 gol	0,61%	0,66%
6 gol	0,17%	0,15%
7 gol	0,05%	0,03%
8 gol	0,02%	0,00%
9 gol	0,02%	0,00%
10 gol	0,01%	0,00%

Grafico 81: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per la distribuzione simulata tramite modello



Anche per la squadra 2 la somiglianza è molto alta: non si riscontrano, infatti, significative differenze tra le due distribuzioni. Possiamo quindi affermare che, avendo a disposizione una grande mole di dati, è possibile prevedere in modo esatto il numero di gol segnati tramite una modello basato sulla discretivizzazione della distribuzione di Weibull.

Vediamo che cosa accade per distribuzioni con minori dati a disposizione e facendo assumere ai parametri i valori estremi delle stime.

La distribuzione dei gol segnati da una squadra in casa nel campionato italiano è caratterizzata da una stima per il parametro a particolarmente alta rispetto a quelle di altre distribuzioni: vediamo il confronto tra questa distribuzione simulata e quella empirica.

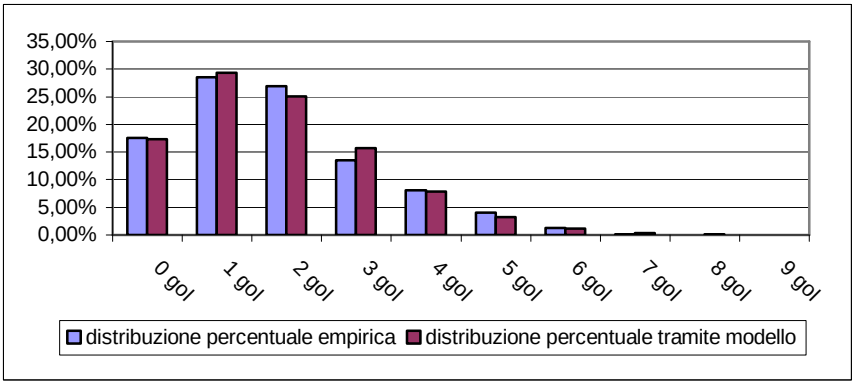
Tabella 48: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato italiano per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	21,46%	21,18%
1 gol	34,15%	34,41%
2 gol	24,88%	25,47%
3 gol	13,66%	12,66%
4 gol	3,66%	4,63%
5 gol	2,20%	1,30%
6 gol	0%	0,29%
7 gol	0%	0,05%

Le due distribuzioni non sono molto diverse tra loro.

La massima distanza tra le due distribuzioni è inferiore all' 1%, questo potrebbe dipendere dal fatto che i dati a disposizione sono meno numerosi rispetto ai casi presentati nelle tabelle 47 e 48.

Grafico 82: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato italiano per la distribuzione empirica e per la distribuzione simulata tramite modello



Il grafico 82 evidenzia le differenze esistenti tra le due distribuzioni. Tuttavia, come è possibile verificare dalla tabella 48, per nessun evento la differenza tra le due distribuzioni di probabilità supera il punto percentuale.

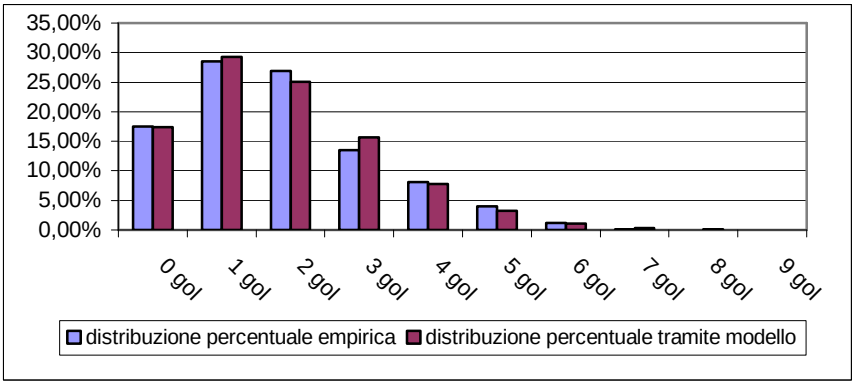
La stima del parametro b per la squadra di casa nel campionato olandese è una delle più alte. Ci chiediamo se questo possa influenzare la precisione della distribuzione simulata tramite questo modello.

Tabella 49: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato olandese per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	17,55%	17,37%
1 gol	28,57%	29,29%
2 gol	26,86%	25,04%
3 gol	13,51%	15,69%
4 gol	8,07%	7,83%
5 gol	4,04%	3,22%
6 gol	1,24%	1,12%
7 gol	0,16%	0,33%
8 gol	0,00%	0,09%
9 gol	0,00%	0,02%

In questo esempio sono presenti maggiori differenze tra le due distribuzioni. Tuttavia, non ci sono degli errori sistematici volti a sovradimensionare un determinato gruppo di eventi rispetto ad un altro gruppo.

Grafico 83: probabilità percentuale del numero di gol segnati dalla squadra 1 nel campionato olandese per la distribuzione empirica e per la distribuzione simulata tramite modello



La non sistematicità dell'errore si evidenzia nel grafico. Gli eventi "1 gol" e "3 gol" hanno una probabilità simulata troppo alta rispetto a quanto accade nella distribuzione empirica. Al contrario, gli eventi "0 gol" e "2 gol" sono poco probabili nella distribuzione simulata rispetto a ciò che accade nel campione di partite analizzato.

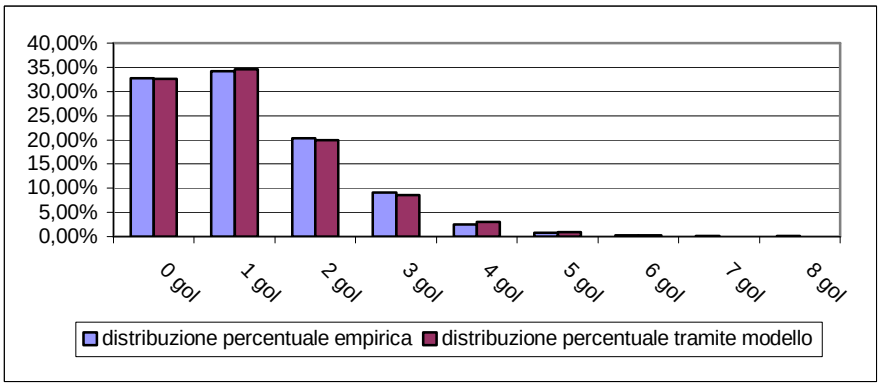
Osserviamo che cosa accade alla distribuzione simulata per i gol segnati in trasferta nel campionato tedesco.

Tabella 50: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato tedesco per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	32,68%	32,58%
1 gol	34,22%	34,63%
2 gol	20,29%	19,90%
3 gol	9,12%	8,62%
4 gol	2,46%	3,05%
5 gol	0,82%	0,91%
6 gol	0,20%	0,24%
7 gol	0,10%	0,05%
8 gol	0,10%	0,01%

La tabella mostra come le due distribuzioni abbiano delle probabilità molto simili tra loro; le differenze sono talmente ridotte da non superare il mezzo punto percentuale.

Grafico 84: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato tedesco per la distribuzione empirica e per la distribuzione simulata tramite modello



Il grafico 84 permette di apprezzare la quasi coincidenza tra le due diverse distribuzioni di probabilità: il modello Weibull, infatti, si adatta perfettamente alla descrizione dei gol segnati dalla squadra fuori casa nel campionato tedesco.

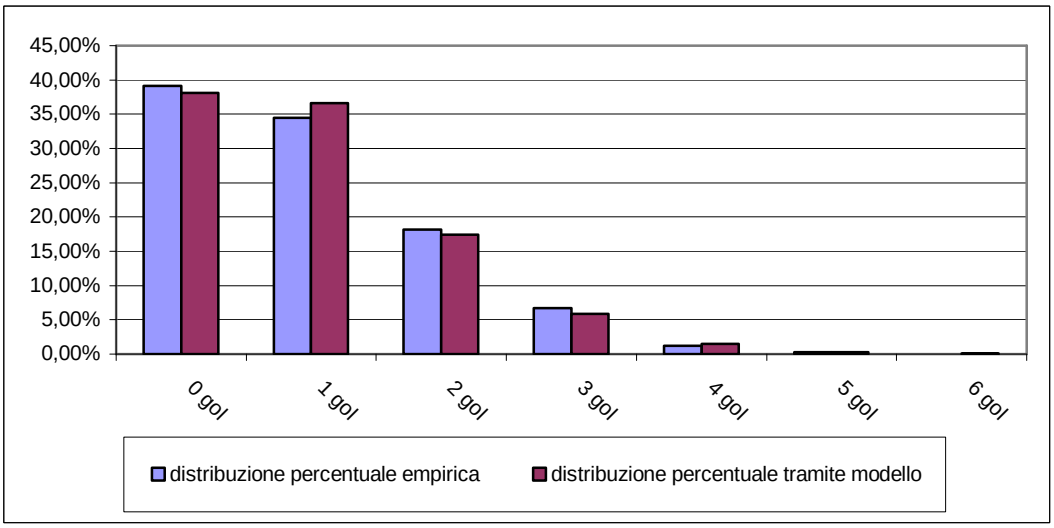
L'ultimo esempio riguarda la squadra in trasferta nel campionato francese.

Tabella 52: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato francese per la distribuzione empirica e per la distribuzione simulata tramite modello

numero di gol	distribuzione empirica	distribuzione modello
0 gol	39,16%	38,14%
1 gol	34,49%	36,64%
2 gol	18,20%	17,42%
3 gol	6,71%	5,87%
4 gol	1,20%	1,53%
5 gol	0,24%	0,32%
6 gol	0%	0,06%

In questo esempio ci sono delle differenze tra le due distribuzioni. Possiamo apprezzare un buon adattamento ai dati anche nel caso dei gol segnati dalla squadra 2 nel campionato francese, sebbene la precisione non sia paragonabile a quella riscontrata nel caso precedente.

Grafico 85: probabilità percentuale del numero di gol segnati dalla squadra 2 nel campionato francese per la distribuzione empirica e per la distribuzione simulata tramite modello



Per le squadre francesi che giocano in trasferta la previsione dei gol segnati è meno precisa: per questo esempio si hanno degli errori superiori ai due punti percentuali.

Conclusioni

Il modello Weibull prevede il fenomeno di interesse in modo molto preciso.
Nei due casi in cui si considera l'intero campione a disposizione, l'andamento della distribuzione simulata è quasi identico alla distribuzione osservata empiricamente. Poiché questo modello non è mai stato utilizzato per descrivere i gol fatti da una squadra di calcio questo aspetto è molto interessante. Inoltre, il modello potrebbe essere ulteriormente migliorato tramite l'introduzione di una correzione che permetta di tenere conto della dipendenza tra i gol segnati dalla squadra di casa e i gol segnati dalla squadra in trasferta.
Riducendo i dati a disposizione, si osserva in generale una diminuzione nella precisione delle stime: per le squadre francesi che giocano fuori casa, tale imprecisione è molto evidente. In altri casi, invece, come per le squadre tedesche che giocano in casa, si osserva una precisione paragonabile a quella ottenuta tenendo conto dell'intero campione a disposizione.

Confronto tra modelli

Introduzione

Abbiamo utilizzato diversi modelli per tentare di descrivere il numero di reti segnati da una squadra di calcio, ottenendo risultati più o meno soddisfacenti per ciascuna di essi.
La nostra analisi pertanto termina con il confronto tra le varie distribuzioni utilizzate per determinare quella più adatta a prevedere il fenomeno di interesse.
Il primo modello utilizzato ha una distribuzione poissoniana. Questo modello è quello che viene utilizzato tradizionalmente per descrivere il fenomeno studiato. La variabile introdotta assume valori interi positivi, pertanto non sono necessarie trasformazioni.
La seconda variabile è originata da una distribuzione normale. Poiché la normale è una variabile continua, abbiamo effettuato una trasformazione esponenziale per far sì che la variabile ammettesse solo valori positivi ed infine abbiamo discretivizzato la variabile affinché assumesse solo valori interi.
Successivamente abbiamo testato una variabile ottenuta da una distribuzione Gamma. La distribuzione Gamma è caratterizzata da determinazioni positive, quindi è stato necessario solamente la discretivizzazione dei valori assunti da tale variabile.
Infine, la quarta variabile ha distribuzione di Weibull. Anche questo modello è continuo e ammette valori positivi, per questa ragione quindi abbiamo reso discrete le determinazioni della variabile analogamente a quanto fatto per la variabile con distribuzione Gamma.
Detto ciò, in questa parte abbiamo confrontato le varie distribuzioni utilizzate per stabilire quale è la più adatta a descrivere il numero di gol segnati da una squadra di calcio.

Confronto fra la distribuzione empirica e la distribuzione ottenuta tramite modello

Iniziamo il confronto con lo studio della distribuzione dei gol segnati dalla squadra 1, prendendo in considerazione tutte le partite appartenenti al campione.
Tutte le distribuzioni simulate hanno come parametri le stime ottenute con il metodo Newton-Raphson.

Tabella 53: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per le distribuzioni simulate

numero di gol	Distribuzione empirica	Distribuzione Poisson	Distribuzione normale	Distribuzione Gamma	Distribuzione Weibull
0 gol	22,40%	21,27%	19,81%	21,36%	22,50%
1 gol	32,54%	32,92%	39,93%	35,46%	32,56%
2 gol	24,60%	25,48%	21,51%	23,31%	23,96%
3 gol	12,39%	13,15%	9,76%	11,59%	12,86%
4 gol	5,43%	5,09%	4,47%	5,04%	5,47%
5 gol	1,90%	1,58%	2,14%	2,02%	1,91%
6 gol	0,55%	0,41%	1,07%	0,77%	0,56%
7 gol	0,11%	0,09%	0,56%	0,28%	0,14%
8 gol	0,05%	0,02%	0,30%	0,10%	0,03%
9 gol	0,02%	0,00%	0,17%	0,04%	0,00%
10 gol	0,00%	0,00%	0,10%	0,01%	0,00%
più di 10 gol	0,00%	0,00%	0,13%	0,00%	0,00%

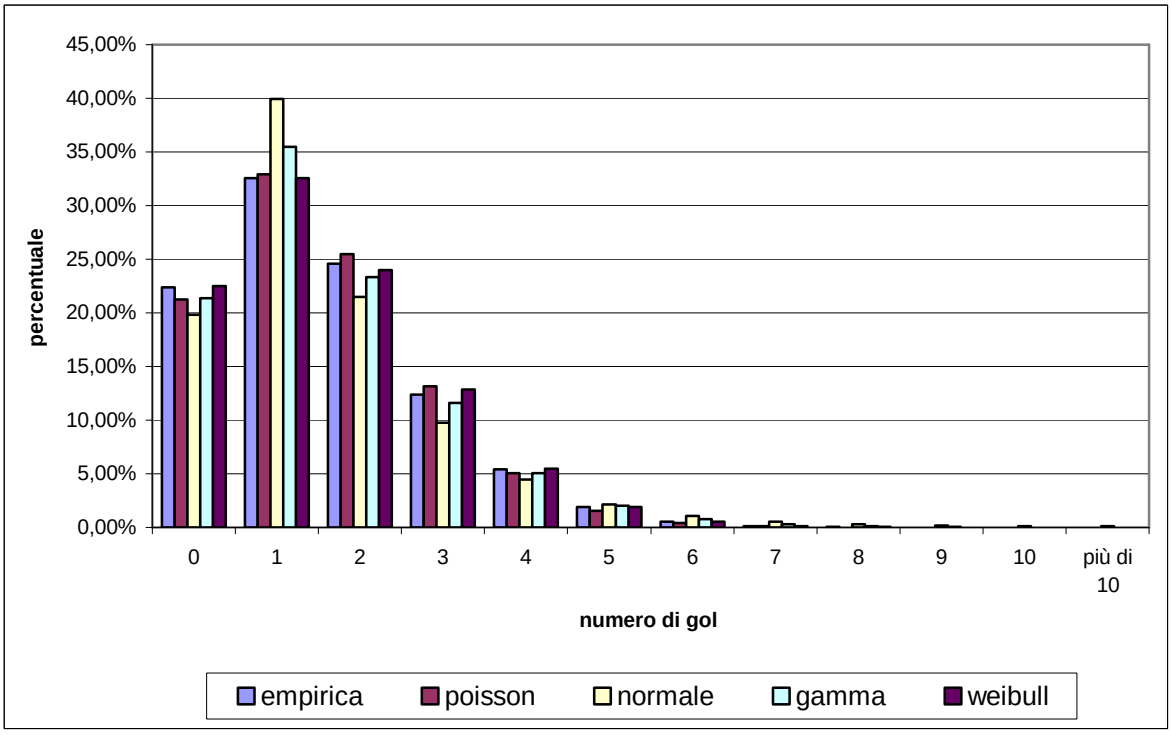
Da una prima lettura della tabella, il modello che più si avvicina alla distribuzione empirica osservata nel campione è il modello basato sulla distribuzione di Weibull. Le differenze sono minime per ogni evento e non si riscontrano particolari errori esclusa forse una sottostima per gli eventi della coda.
La simulazione ottenuta dal modello di Poisson si adatta bene ai dati e ripropone una forma analoga a quella che si osserva empiricamente dai dati appartenenti al campione: si distanzia dalla distribuzione empirica, per ciascun evento, meno di un punto percentuale. Questo aspetto era abbastanza prevedibile dal momento che la

distribuzione di Poisson viene solitamente utilizzata per descrivere questo fenomeno.

Anche la variabile Gamma ha una precisione simile a quella della variabile di Poisson.

La variabile originata da una variabile normale è caratterizzata da una sovrastima per l'evento più probabile compensata da frequenze percentuali troppo basse per gli eventi attigui. Inoltre abbiamo il problema della coda destra avente un peso troppo alto per descrivere opportunamente i gol segnati da una squadra di calcio.

Grafico 86: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per le distribuzioni simulate

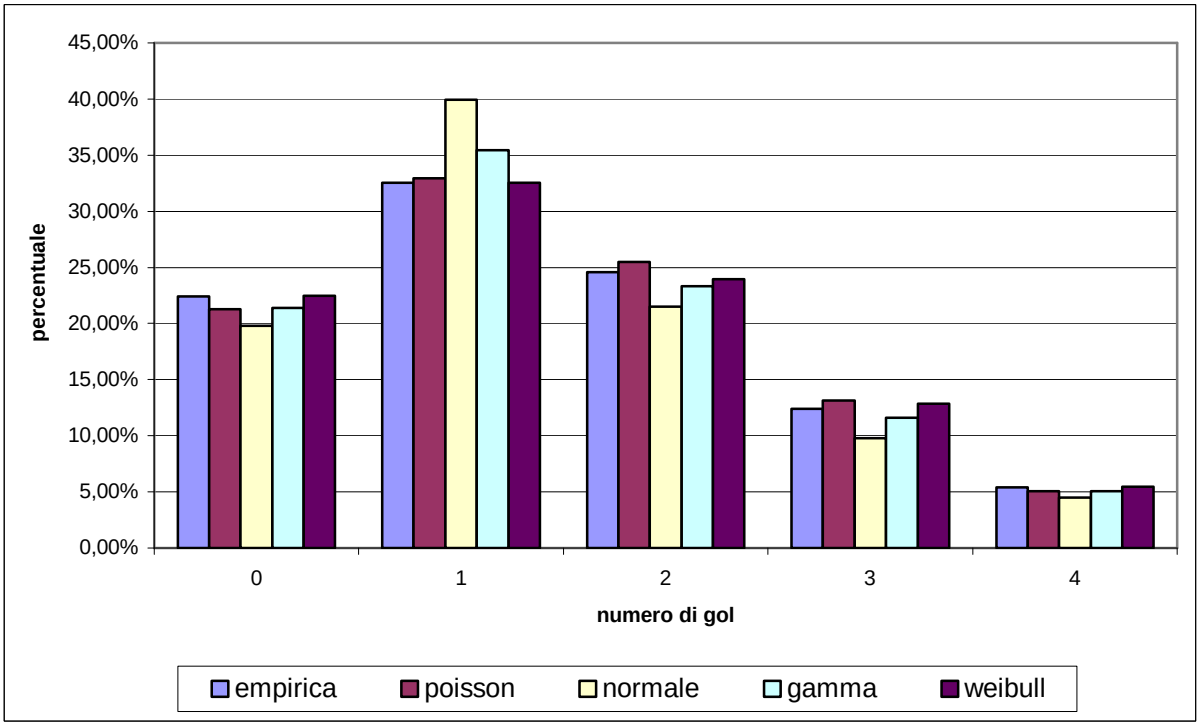


Il grafico ci permette di vedere con chiarezza il buon adattamento delle variabili basate sul modello di Poisson, di Gamma e di Weibull, ma non ci permette di cogliere le significative differenze esistenti tra queste tre distribuzioni.

Per meglio cogliere la bontà di adattamento ai dati della variabile ottenuta da una distribuzione di Weibull rispetto alle altre due è utile soffermarsi su un determinato sottoinsieme di eventi.

Vediamo che cosa accade per gli eventi più probabili, quelli, cioè, vicini all'origine.

Grafico 87: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per le distribuzioni simulate

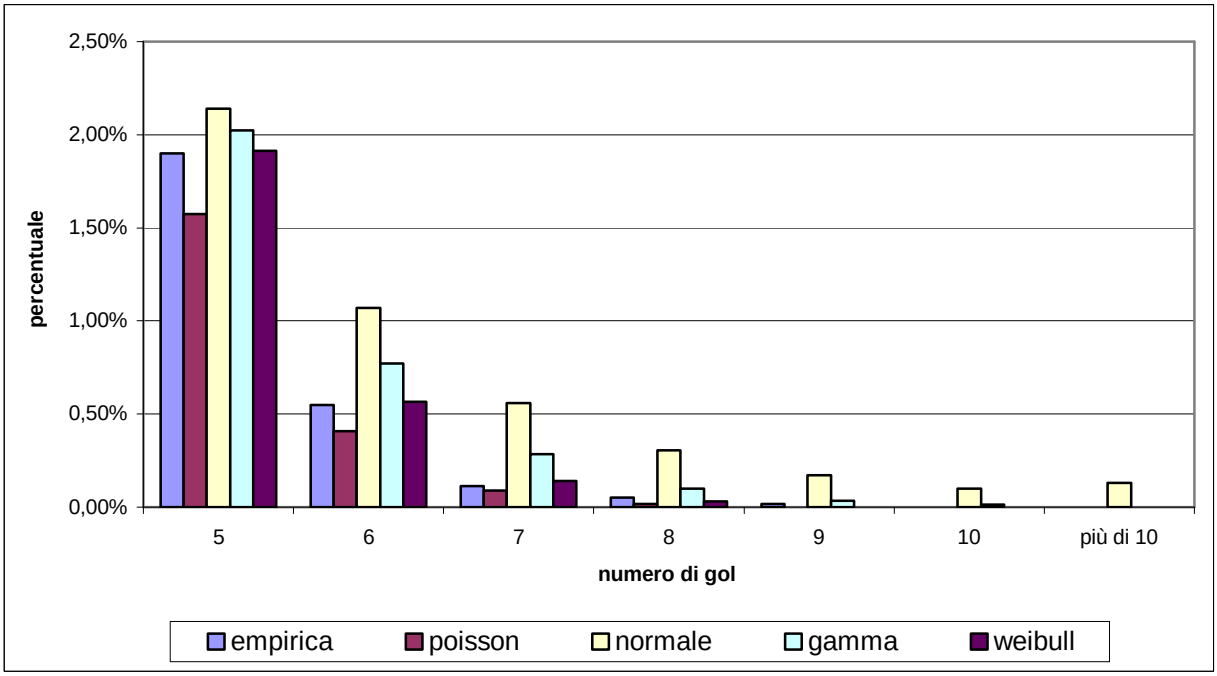


Il grafico precedente permette di apprezzare come la distribuzione di Poisson e quella di Weibull abbiano andamenti molto vicini a quanto osservato empiricamente.

Il modello normale prevede una probabilità troppo alta per l'evento "1 gol" compensata da probabilità basse per gli altri quattro eventi rappresentati.

La variabile costruita partendo da una distribuzione Gamma sembra avere un problema analogo, seppure di entità minore rispetto al modello normale.

Grafico 88: probabilità percentuale del numero di gol segnati dalla squadra 1 per la distribuzione empirica e per le distribuzioni simulate



Per la coda destra la distribuzione nata dalla Weibull conserva la sua vicinanza alla distribuzione empirica, mentre la distribuzione di Poisson perde precisione.
La distorsione della variabile originata da una normale in questo grafico appare in tutta la sua evidenza: la probabilità che una squadra segni un alto numero di gol è troppo alta rispetto a quanto accade empiricamente.
La variabile Gamma anche per la coda destra ha lo stesso problema della precedente distribuzione, anche se l'inesattezza è minore.
Infine, è bene sottolineare che questi eventi sono caratterizzati da probabilità basse: lo scarto tra probabilità empirica e quella simulata non supera lo 0,5%.

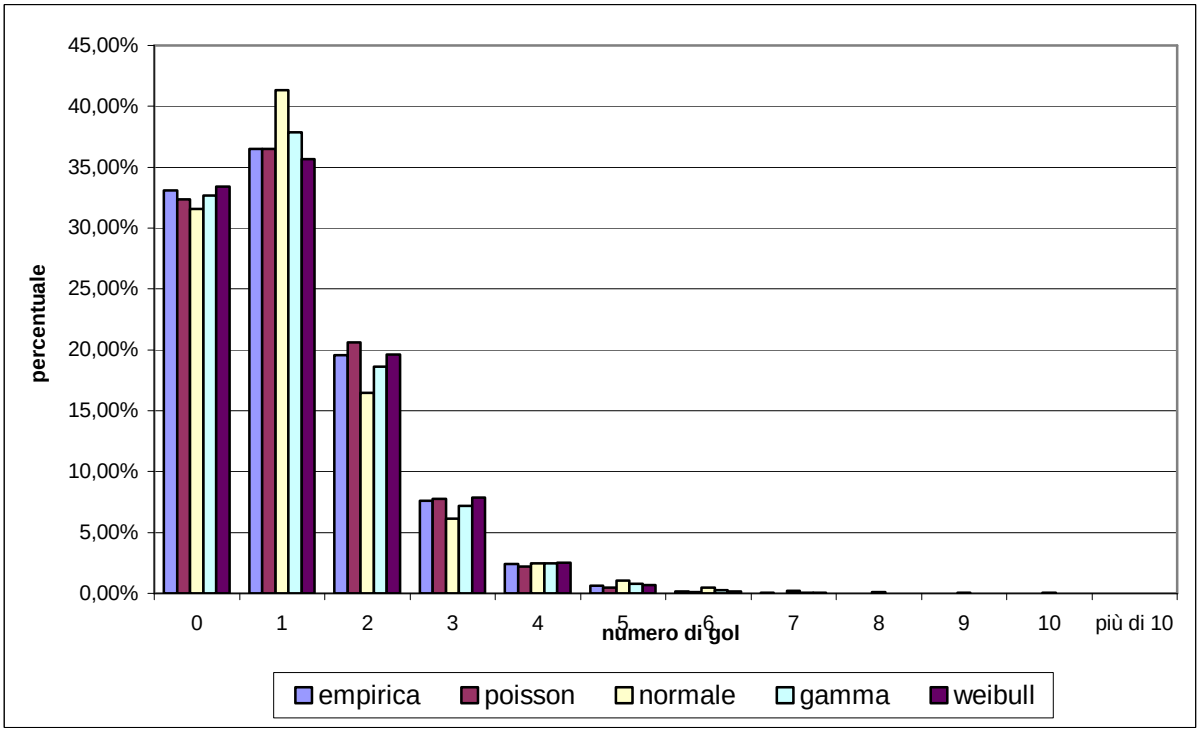
Studiamo la distribuzione dei gol segnati dalla squadra fuori casa.

Tabella 54: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per le distribuzioni simulate

numero di gol	Distribuzione empirica	Distribuzione Poisson	Distribuzione normale	Distribuzione Gamma	Distribuzione Weibull
0 gol	33,08%	32,37%	31,58%	32,67%	33,43%
1 gol	36,49%	36,51%	41,34%	37,85%	35,69%
2 gol	19,54%	20,59%	16,48%	18,64%	19,64%
3 gol	7,62%	7,74%	6,15%	7,21%	7,89%
4 gol	2,39%	2,18%	2,44%	2,48%	2,52%
5 gol	0,61%	0,49%	1,04%	0,80%	0,66%
6 gol	0,17%	0,09%	0,48%	0,25%	0,15%
7 gol	0,05%	0,01%	0,23%	0,07%	0,03%
8 gol	0,02%	0,00%	0,12%	0,02%	0,00%
9 gol	0,02%	0,00%	0,06%	0,00%	0,00%
10 gol	0,01%	0,00%	0,03%	0,00%	0,00%

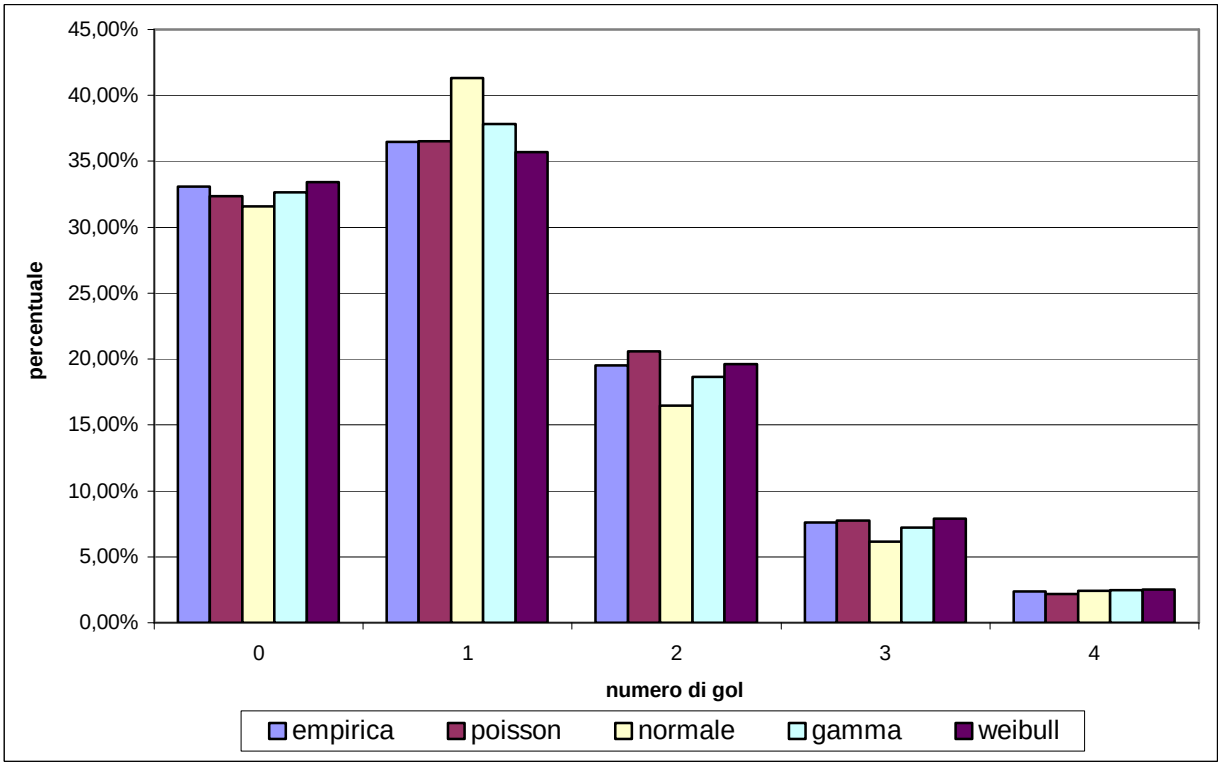
Dalla analisi sulla distribuzione dei gol per le squadre in trasferta sembrano ripresentarsi gli stessi aspetti analizzati per le squadre di casa. Notiamo che le variabili originate da distribuzioni poissoniane, Gamma e Weibull hanno un andamento molto vicino alla distribuzione empirica. La distribuzione ottenuta tramite trasformazioni della Gaussiana invece è incapace di avere un andamento comparabile a quello riscontrato empiricamente.

Grafico 89: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per le distribuzioni simulate



Il grafico 89 evidenzia l'inesattezza della variabile normale, ma non permette di apprezzare le differenze tra gli andamenti delle altre distribuzioni.

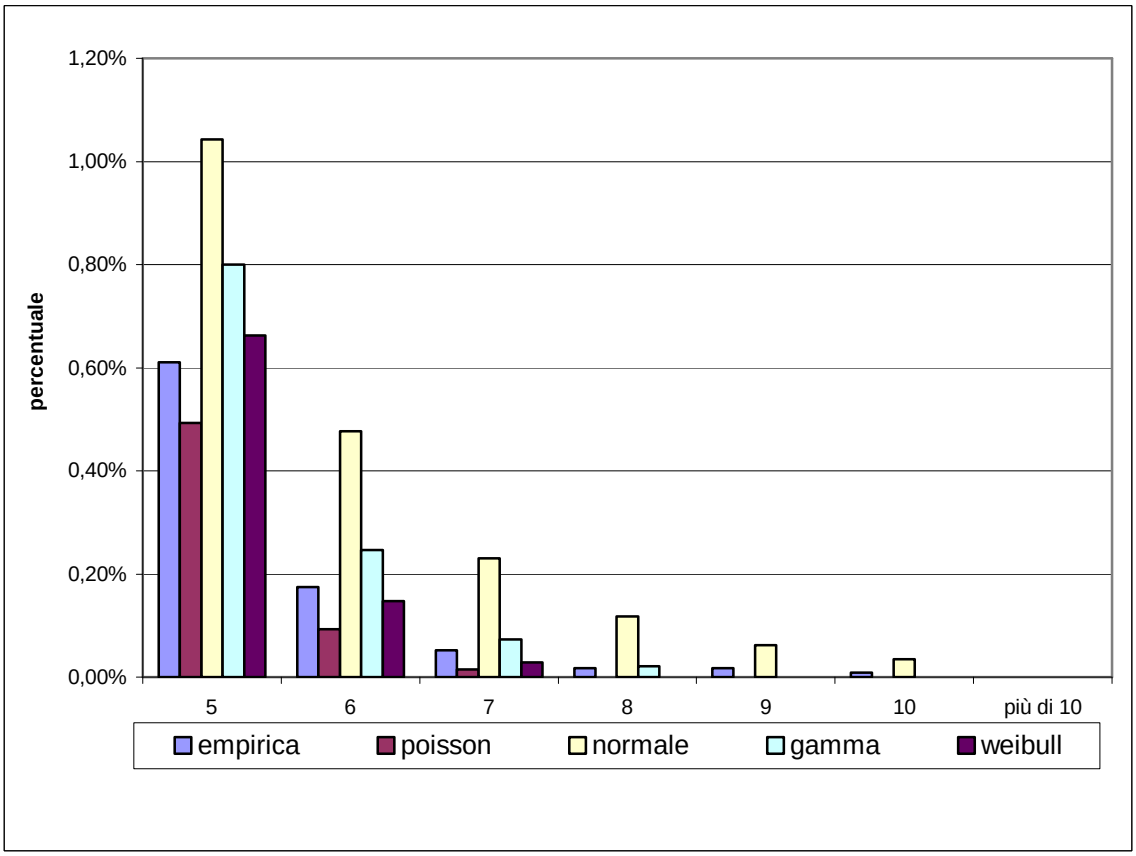
Grafico 90: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per le distribuzioni simulate



Soffermandoci sulla parte sinistra delle distribuzioni, si può osservare come le variabili basata sulla distribuzione di Poisson e quelle originate dalla distribuzione Gamma e Weibull abbiano andamenti molto simili tra loro e siano altresì vicine alla distribuzione empirica.

Anche per il fenomeno dei gol segnati dalla squadra in trasferta la variabile originata dalle trasformazioni della normale è caratterizzata da probabilità troppo basse per gli eventi "0 gol", "2 gol", "3 gol" e "4 gol" contrapposti ad una probabilità eccessivamente alta per l'evento "1 gol".

Grafico 91: probabilità percentuale del numero di gol segnati dalla squadra 2 per la distribuzione empirica e per le distribuzioni simulate



Le probabilità nella coda destra per la distribuzione empirica sono molto basse, di conseguenza anche le differenze per ciascuna distribuzione rispetto a quella empirica sono minime.

Detto questo, appare fin da subito evidente come la trasformata della Gaussiana dia troppo peso agli eventi estremi e quindi a tutti gli eventi appartenenti alla coda.

Le altre tre variabili mantengono la loro vicinanza alla distribuzione empirica, anche se il modello Weibull ha una precisione maggiore rispetto alle altre due distribuzioni.

La variabile con distribuzione di Poisson è caratterizzata da probabilità più basse rispetto a quelle che si verificano empiricamente, mentre la variabile originata da una distribuzione Gamma tende ad avere probabilità troppo alte.

Possiamo concludere dicendo che la trasformazione della variabile normale non è adatta a prevedere con sufficiente precisione il numero di gol segnati da una squadra di calcio.

Anche la trasformata di una variabile Gamma, sebbene sia più positiva della precedente, non ha sufficiente precisione nel prevedere i risultati di una partita di calcio.

Tra il modello di Poisson e il modello originato da una distribuzione di Weibull, il secondo sembra essere più preciso. Questo risultato è certamente molto importante se pensiamo che il primo modello è quello che viene utilizzato tradizionalmente per la previsione del fenomeno che abbiamo studiato.

Il modello bivariato

Introduzione

Nei modelli precedenti abbiamo analizzato una variabile allo scopo di descrivere il numero di gol segnati da una singola squadra. Abbiamo cioè analizzato separatamente, in un incontro di calcio, le reti segnate dalla squadra 1 e dalla squadra 2. Tuttavia è impensabile che i gol segnati da due squadre che si fronteggiano possano essere assunti come indipendenti.

Per questa ragione è necessario descrivere con una unica variabile il punteggio di una partita di calcio.

Nonostante il modello normale abbia evidenziato dei difetti nel prevedere i gol segnati, possiamo provare a prevedere il punteggio con una variabile normale bivariata. Questa variabile avrà come distribuzioni marginali le due distribuzioni normali che descrivono i gol segnati da ciascuna squadra; sebbene le marginali sono difettose, questo approccio potrebbe funzionare nel prevedere la dipendenza. In altri termini, potremmo usare distribuzioni più precise nelle marginali e descrivere la dipendenza tramite il modello bivariato.

Analogamente a quanto fatto per la variabile utilizzata per descrivere i gol segnati da una singola squadra, consideriamo la variabile X:

$$X \sim N_2(\mu;\Sigma).$$

dove $\mu=[\mu_1;\mu_2]^T$ con μ_1 media della marginale 1 e μ_2 media della marginale 2.

E dove $\Sigma=\begin{bmatrix}\sigma_1^2 & \rho\cdot\sigma_1\cdot\sigma_2\\ \rho\cdot\sigma_1\cdot\sigma_2 & \sigma_2^2\end{bmatrix}$ con σ_1 deviazione standard della marginale 1, σ_2 deviazione standard della marginale 2 e ρ che esprime la correlazione tra la variabile marginale 1 e la variabile marginale 2.

La funzione di probabilità è:

$$P\left(X=(x_1;x_2)\right)=\frac{1}{2\pi\cdot\sigma_1\cdot\sigma_2\cdot\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2\left(1-\rho^2\right)}\cdot\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2\rho\cdot(x_1-\mu_1)\cdot(x_2-\mu_2)}{\sigma_1\cdot\sigma_2}\right)\right)$$

Anche in questo caso però il problema è che il dominio va da $-\infty$ a $+\infty$, noi invece abbiamo bisogno che la variabile assuma solo valori positivi.
Facciamo una prima trasformazione esponenziale:

$$Y = e^X$$

Rimane da risolvere il problema della continuità. Introduciamo tramite una ulteriore trasformazione la terza variabile Z, che è una variabile discreta.

$$Z = (j_1; j_2) \text{ se e solo se}$$

$$Y_1 : j_1 \le Y_1 < j_1 + 1$$

e

$$Y_2 : j_2 \le Y_2 < j_2 + 1$$

Questa nuova variabile, trasformazione di una bivariata, sarà caratterizzata da 5 parametri a loro volta trasformati dalla variabile bivariata di partenza. Chiameremo i 5 parametri mu1, mu2, sigma1, sigma2 e Ro, trasformazioni rispettivamente di μ_1, μ_2, σ_1, σ_2 e ρ.

Simulazione di dati

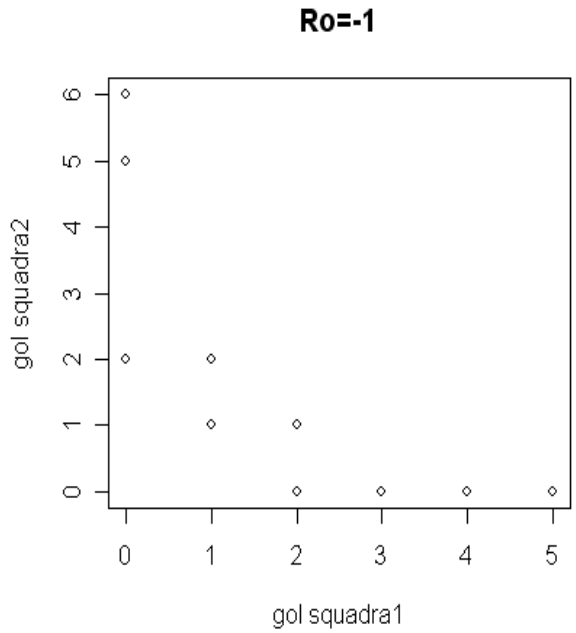
Abbiamo già analizzato i possibili valori assunti dai parametri mu e sigma per una variabile che, originata da una Normale, voglia descrivere il numero di gol segnati da una squadra di calcio.
Per questa ragione nelle simulazioni che seguiranno, fisseremo i valori dei parametri mu1, mu2, sigma1 e sigma2 e faremo variare il valore di Ro.
I valori assunti dai primi quattro parametri sono quelli stimati per la variabile normale tramite metodo esatto sia per la squadra 1 sia per la squadra 2 e prendendo in esame l'intero campione a disposizione.

Tabella 55:valori assunti ai parametri Mu1, Mu2, Sigma1 e Sigma2 nelle simulazioni

Parametro	Valore fissato
Mu1	0,537
Mu2	0,305
Sigma1	0,633
Sigma2	0,636

Il parametro Ro invece assume invece i valori : –1; -0,5; 0; 0,5; 1.
Per ogni simulazione vengono creati 50 determinazioni casuali.

Grafico 92: grafico di dispersione dei gol segnati dalla squadra 1 e dei gol segnati dalla squadra 2 tramite simulazione della variabile Z con i parametri fissati a priori



Come ci aspettavamo, se il parametro Ro, trasformazione del parametro di correlazione, vale –1, la distribuzione dei gol per la squadra 1 è inversamente proporzionale alla distribuzione dei gol per la squadra 2. In altri termini, se la squadra di casa segna un alto numero di gol, la squadra in trasferta invece tenderà a non segnare. Viceversa, se la squadra di casa totalizza zero reti, quella fuori casa è caratterizzata da un numero elevato di segnature.

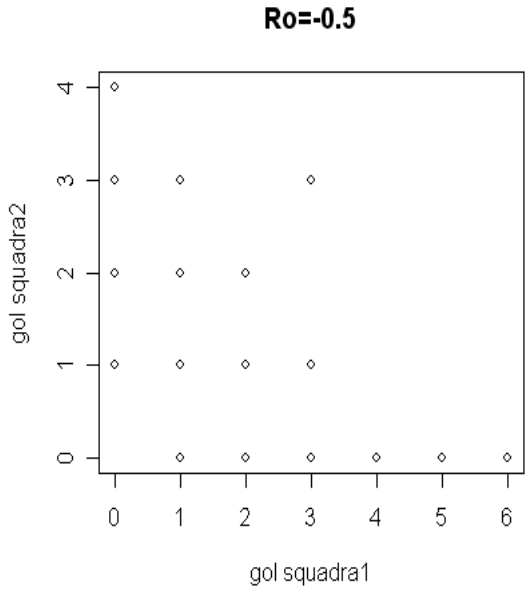
Questo fenomeno si può apprezzare meglio dalla seguente tabella che riporta i risultati di 50 partite generati tramite distribuzione casuale avente Mu1, Mu2, Sigma1 e Sigma2 come sopra fissati e Ro pari a –1.

Tabella 56: distribuzione di frequenza per gol di scarto e per esito tramite simulazione della variabile Z con i parametri fissati a priori

numero gol di scarto	esito	frequenza
5 gol	vittoria squadra 1	2
4 gol	vittoria squadra 1	2
3 gol	vittoria squadra 1	5
2 gol	vittoria squadra 1	9
1 gol	vittoria squadra 1	3
0 gol	pareggio	20
1 gol	vittoria squadra 2	4
2 gol	vittoria squadra 2	3
3 gol	vittoria squadra 2	0
4 gol	vittoria squadra 2	0
5 gol	vittoria squadra 2	1
6 gol	vittoria squadra 2	1

In undici casi su 50 lo scarto è superiore ai 2 gol, i pareggi sono 20. Nonostante il valore di mu1 sia superiore a quello di mu2, il fatto che Ro valga –1 fa sì che si registrino delle vittorie in trasferta.
Vediamo che cosa accade con Ro pari a –0,5.

Grafico 93: grafico di dispersione dei gol segnati dalla squadra 1 e dei gol segnati dalla squadra 2 tramite simulazione della variabile Z con i parametri fissati a priori



La forma è simile a quella riscontrata per Ro pari a –1. In questo caso però tende ad attenuarsi la delineatezza della curva, tipica della proporzionalità inversa. I punti nel grafico sono più equamente distribuiti.

Tabella 57: distribuzione di frequenza per gol di scarto e per esito tramite simulazione della variabile Z con i parametri fissati a priori

numero gol di scarto	esito	frequenza
6 gol	vittoria squadra 1	1
5 gol	vittoria squadra 1	1
4 gol	vittoria squadra 1	2
3 gol	vittoria squadra 1	3
2 gol	vittoria squadra 1	5
1 gol	vittoria squadra 1	7
0 gol	pareggio	16
1 gol	vittoria squadra 2	6
2 gol	vittoria squadra 2	6
3 gol	vittoria squadra 2	2
4 gol	vittoria squadra 2	1

Lo scarto supera i 2 gol in 10 casi, i pareggi sono 16. Anche in questo caso, il segno negativo di Ro comporta la presenza di vittorie fuori casa.

Se Ro vale 0, il numero di reti segnate dalla squadra 1 e quello segnato dalla squadra 2 sono indipendenti. In altre parole, è come se la squadra di casa e la squadra in trasferta non giocassero contemporaneamente e le prestazioni di una squadra non dipendessero dalla squadra avversaria. Questa ipotesi è difficilmente accettabile nella realtà: il numero di gol segnati da una squadra dipende infatti anche dall'abilità difensiva della squadra avversaria. Inoltre un sostanziale equilibrio in campo farà sì che entrambe le squadre avranno difficoltà a segnare.

Grafico 94: grafico di dispersione dei gol segnati dalla squadra 1 e dei gol segnati dalla squadra 2 tramite simulazione della variabile Z con i parametri fissati a priori

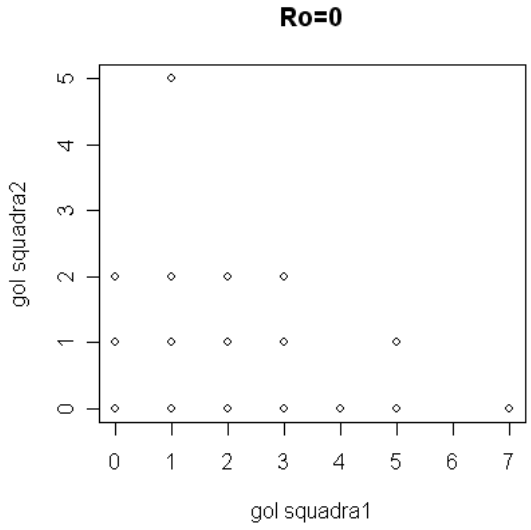


Tabella 58: distribuzione di frequenza per gol di scarto e per esito tramite simulazione della variabile Z con i parametri fissati a priori

numero gol di scarto	esito	frequenza
7 gol	vittoria squadra 1	1
6 gol	vittoria squadra 1	0
5 gol	vittoria squadra 1	1
4 gol	vittoria squadra 1	2
3 gol	vittoria squadra 1	1
2 gol	vittoria squadra 1	3
1 gol	vittoria squadra 1	17
0 gol	pareggio	13
1 gol	vittoria squadra 2	8
2 gol	vittoria squadra 2	2
3 gol	vittoria squadra 2	0
4 gol	vittoria squadra 2	2

I pareggi 13 e sono causati dalla scarsa differenza per i valori assunti dai parametri; se assumessimo una differenza marcata tra mu1 e mu2 allora pareggi diminuirebbero in modo significativo. Le vittorie con tante reti di scarto si riducono a 7.

È interessante osservare che le vittorie della squadra di casa sono molto di più di quelle ottenute dalla squadra in trasferta. Questa asimmetricità deriva dal valore maggiore di mu1 rispetto a mu2.

Grafico 95: grafico di dispersione dei gol segnati dalla squadra 1 e dei gol segnati dalla squadra 2 tramite simulazione della variabile Z con i parametri fissati a priori

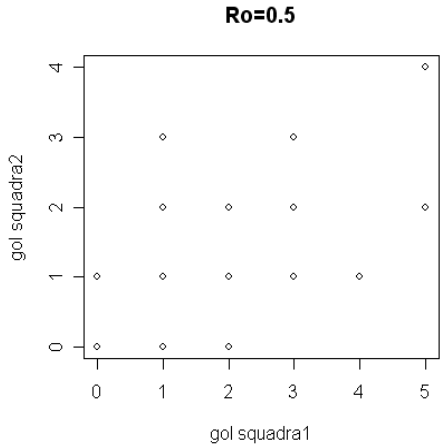


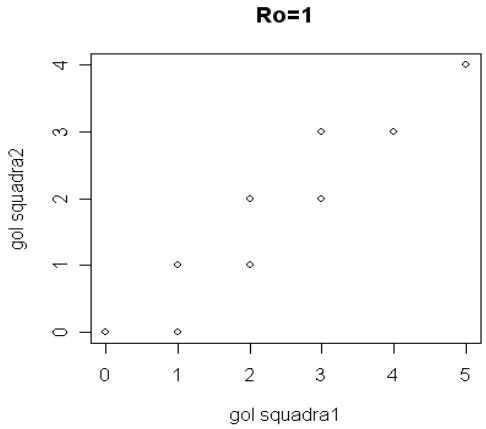
Tabella 59: distribuzione di frequenza per gol di scarto e per esito tramite simulazione della variabile Z con i parametri fissati a priori

numero gol di scarto	esito	frequenza
3 gol	vittoria squadra 1	3
2 gol	vittoria squadra 1	4
1 gol	vittoria squadra 1	10
0 gol	pareggio	25
1 gol	vittoria squadra 2	7
2 gol	vittoria squadra 2	1

È chiaro che fissando Ro a 0,5 otteniamo due distribuzioni simili, perché se la squadra 1 segna un numero alto di gol, anche la squadra 2 tenderà a fare lo stesso, proprio per la correlazione positiva. Ci sono solamente tre partite in cui lo scartò è di 3 gol. In nessuna lo scarto supera i 3 gol.
Un valore positivo di Ro fa sì che i pareggi aumentino: in questo caso sono addirittura 25, la metà. Osserviamo inoltre la diminuzione delle vittorie per la squadra in trasferta.

Se forziamo Ro a 1 ci aspettiamo un andamento ancora più simile per la distribuzione delle reti della squadra 1 e della squadra 2.

Grafico 96: grafico di dispersione dei gol segnati dalla squadra 1 e dei gol segnati dalla squadra 2 tramite simulazione della variabile Z con i parametri fissati a priori



I punti del grafico confermano questa nostra ipotesi di proporzionalità diretta tra i due fenomeni studiati.

Tabella 60: distribuzione di frequenza per gol di scarto e per esito tramite simulazione della variabile Z con i parametri fissati a priori

gol di scarto	esito	frequenza
1 gol	vittoria squadra 1	20
0 gol	pareggio	30

Per questa matrice di numeri casuali i pareggi sono addirittura 30. Il massimo scarto registrato è di una rete; è interessante osservare che alcuni incontri sono caratterizzate da un alto numero di reti complessive equamente distribuite tra le due squadre in campo. Dal momento che il valore di mu 1 è maggiore del valore di mu2 e il parametro Ro assume valore 1, non si registrano vittorie per la squadra in trasferta.

Riassumendo:

Tabella 61: distribuzione di frequenza per i risultati di partite ottenute tramite simulazione della variabile Z con i parametri fissati a priori

Ro	vittorie casa	pareggi	vittorie trasferta
-1	21	20	9
-0,5	19	16	15
0	25	13	12
0,5	17	25	8
1	20	30	0

Il primo aspetto che si nota per primo è che la squadra di casa fa registrare un più alto numero di vittorie rispetto alla squadra in trasferta a prescindere dal valore di Ro: questo è dovuto al valore maggiore fissato a Mu1 rispetto a quello fissato a Mu2.

Il numero di pareggi aumenta in modo significativo all'aumentare del valore di Ro. Quest'ultimo parametro rappresenta, infatti, il legame esistente tra le due distribuzioni. Le due distribuzioni delle determinazioni sono tanto simili quanto più il valore di Ro è alto.

I risultati tendono ad essere più equilibrati con Ro pari a –0,5 e a -1. Questo aspetto è dovuto da un lato al valore negativo di questo parametro, dall'altro alla differenza non particolarmente alta tra le due coppie degli altri parametri.

Metodo di stima tramite massima verosimiglianza

Dopo avere effettuato le varie simulazioni, proviamo a stimare il valore del parametro Ro in ciascun campionato appartenente al campione.

Le stime sono ottenute tramite il metodo numerico di Newton-Raphson della funzione di log-verosimiglianza, avente cinque parametri. Nella successiva tabella viene riportato solamente il valore della stima di Ro, dal momento che i valori stimati per gli altri quattro parametri coincidono con quelli ottenuti stimando tramite metodo esatto i parametri per la distribuzione dei gol segnati in casa e i parametri per la distribuzione dei gol segnati in trasferta, separatamente considerati.

Stime

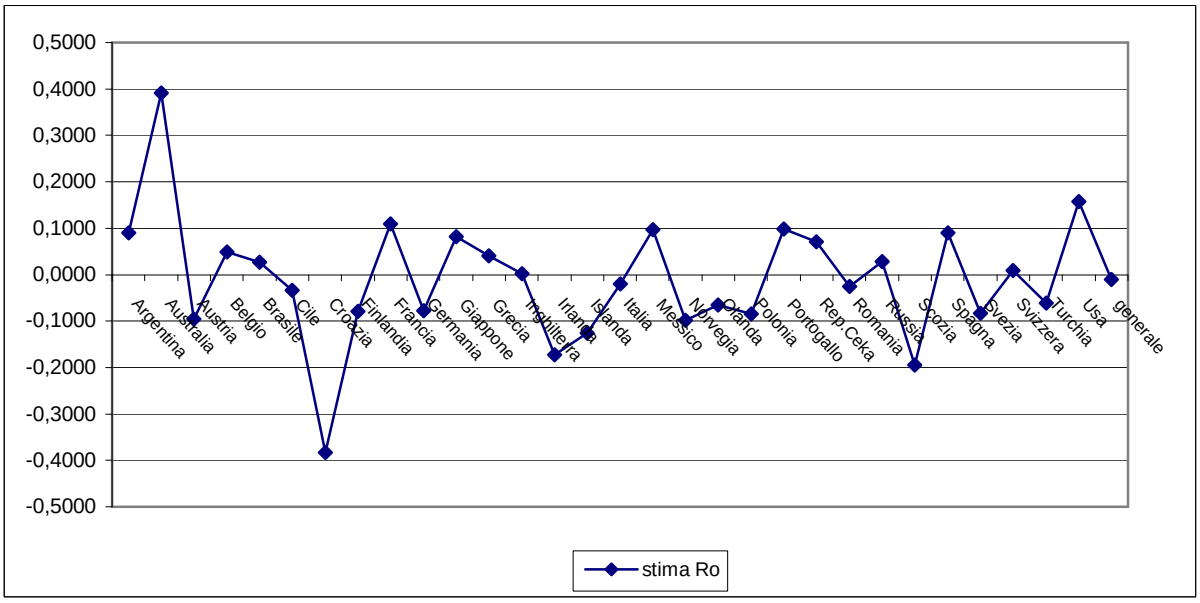
Tabella 62: stima del parametro Ro e del suo standard error tramite metodo esatto per ciascun campionato

nazione	Ro	s.e. Ro
Argentina	0,0904	0,0038
Australia	0,3917	0,0651
Austria	-0,0959	0,0033
Belgio	0,0488	0,0026
Brasile	0,0266	0,0014
Cile	-0,0334	0,0058
Croazia	-0,3826	0,0354
Finlandia	-0,0790	0,0060
Francia	0,1094	0,0016
Germania	-0,0776	0,0013
Giappone	0,0823	0,0068
Grecia	0,0401	0,0075
Inghilterra	0,0019	0,0009
Irlanda	-0,1728	0,0045
Islanda	-0,1255	0,0084
Italia	-0,0198	0,0031
Messico	0,0967	0,0083
Norvegia	-0,0981	0,0052
Olanda	-0,0659	0,0019
Polonia	-0,0846	0,0071
Portogallo	0,0986	0,0035
Rep.Ceka	0,0715	0,0052
Romania	-0,0259	0,0129
Russia	0,0281	0,0032
Scozia	-0,1941	0,0065
Spagna	0,0905	0,0020
Svezia	-0,0835	0,0026
Svizzera	0,0093	0,0075
Turchia	-0,0612	0,0085
Usa	0,1578	0,0242
generale	-0,0098	0,0001

È interessante notare come la stima di Ro assuma sia valori positivi che valori negativi.

I valori degli errori standard sono sufficientemente bassi. Il basso valore assoluto stimato per Ro ci fa sospettare che il parametro non abbia la capacità di riuscire ad esprimere la dipendenza dei due fenomeni analizzati congiuntamente. Questo aspetto ci viene confermato dal valore vicino allo zero che la stima di Ro assume considerando l'intero campione a nostra disposizione. Se Ro è pari a zero i due fenomeni sono indipendenti linearmente. Con Ro pari a zero potrebbe esistere una dipendenza di tipo non lineare.

Grafico 97: stima del parametro Ro tramite metodo esatto per ciascun campionato



Dal grafico è possibile apprezzare l'ampio dominio dei valori assunti dalla stima di Ro. Al di là dei campionati caratterizzati da un basso numero di partite campionarie (ad esempio Australia e Croazia), si può constatare che i campionati francese e spagnolo hanno una stima positiva, mentre quello tedesco ha una stima di Ro con valore negativo. La stima di Ro per il campionato italiano, quello inglese e per l'intero campione è molto vicina allo zero.

Confronto fra la distribuzione empirica e la distribuzione ottenuta tramite modello

Andiamo a confrontare la distribuzione empirica dei risultati delle partite appartenenti al campione con una distribuzione simulata tramite la nostra variabile Z.

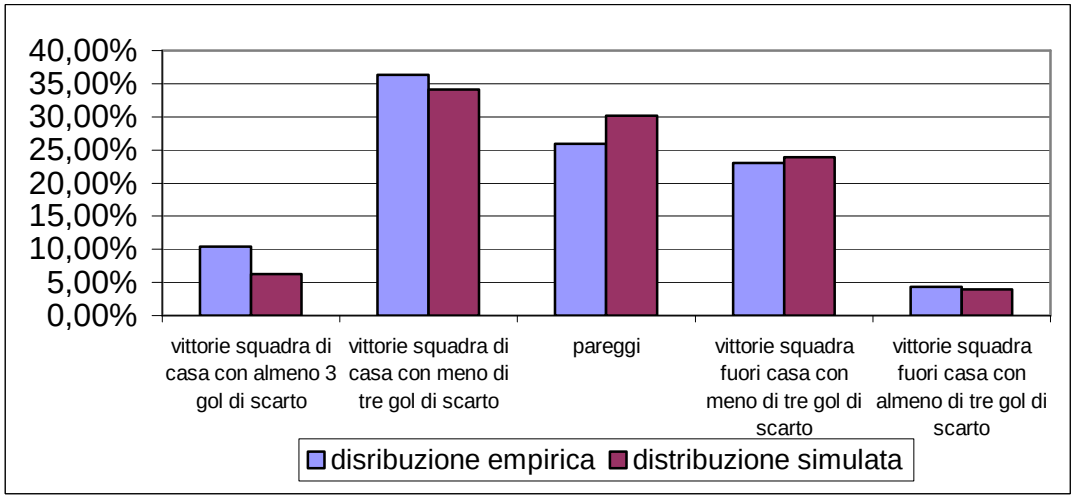
È giusto sottolineare che, per questo confronto, le probabilità della distribuzione simulata non sommano cento. Questo perché il procedimento in primo luogo prevedeva il calcolo della probabilità per ciascun risultato esatto ed in secondo luogo la somma di essi nei cinque casi riportati nelle tabelle 63, 64, 65. Essendo infiniti i possibili risultati esatti, la somma delle probabilità non è cento ma un valore leggermente inferiore, tendente a cento.

Tabella 63: probabilità percentuale dei risultati per la distribuzione empirica e per la distribuzione simulata tramite modello

esito	Gol di scarto	Distribuzione empirica	Distribuzione simulata
Vittoria squadra di casa	Tre o più	10,41%	6,27%
Vittoria squadra di casa	Meno di tre	36,31%	34,09%
pareggi	Nessuno	25,94%	30,15%
Vittoria squadra fuori casa	Meno di tre	23,04%	23,87%
Vittoria squadra fuori casa	Tre o più	4,30%	3,97%

Vediamo che la distribuzione dei risultati aggregati, simulata tramite un modello originato da una distribuzione bivariata, perde la distorsione che invece si riscontrava nella analisi dei gol segnati da ciascuna squadra attraverso un modello originato da una variabile normale; tuttavia persistono delle differenze abbastanza nette tra le due distribuzioni.

Grafico 98: probabilità percentuale dei risultati per la distribuzione empirica e per la distribuzione simulata tramite modello



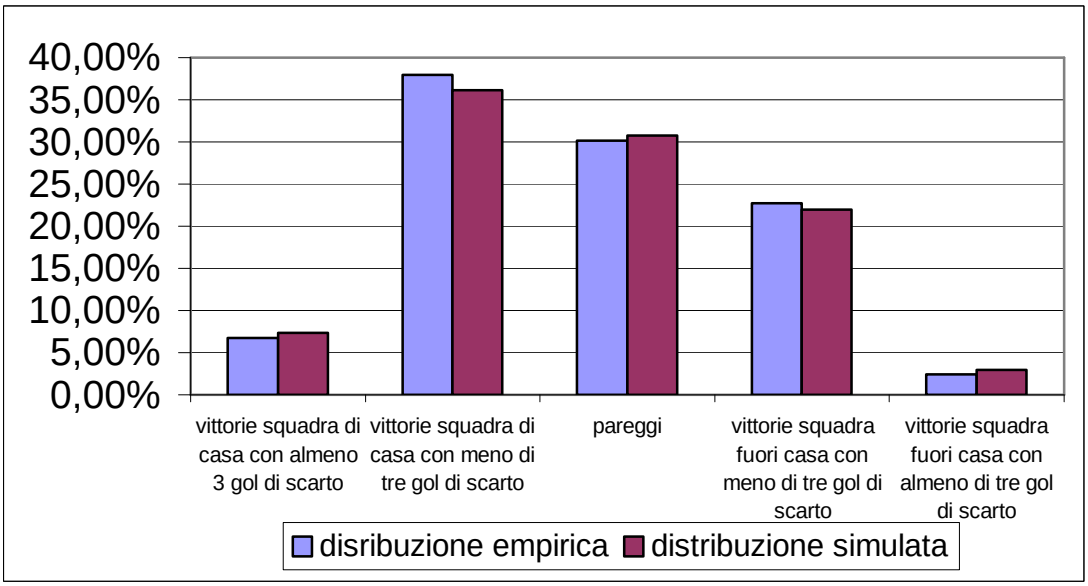
Il problema differenza principale della distribuzione simulata è di sovrastimare il pareggio nonostante la stima di Ro sia negativa, seppure di poco. Ciò significa che la nostra variabile non riesce a prevedere in modo soddisfacente l'andamento delle due distribuzioni, congiuntamente analizzato. Se così fosse, il valore negativo di Ro, seppure prossimo allo zero, avrebbe dovuto portare ad una probabilità maggiore di riscontrare alti scarti di reti segnate in una partita.
Simuliamo i risultati del campionato francese mediante una distribuzione bivariata con valore di Ro positivo.

Tabella 64: probabilità percentuale dei risultati nel campionato francese per la distribuzione empirica e per la distribuzione simulata tramite modello

esito	Gol di scarto	Distribuzione empirica	Distribuzione simulata
Vittoria squadra di casa	Tre o più	6,71%	7,37%
Vittoria squadra di casa	Meno di tre	37,96%	36,16%
pareggi	Nessuno	30,18%	30,74%
Vittoria squadra fuori casa	Meno di tre	22,75%	21,95%
Vittoria squadra fuori casa	Tre o più	2,40%	2,97%

Per il campionato francese si ottiene una distribuzione simulata non molto diversa da quella empirica. Il grafico 99 ci permette di osservare la buona precisione ottenuta tramite il modello bivariato in modo chiaro.

Grafico 99: probabilità percentuale dei risultati nel campionato francese per la distribuzione empirica e per la distribuzione simulata tramite modello



L'andamento della distribuzione simulata approssima correttamente l'andamento della distribuzione empirica. In particolare si osservi che la sovrastima per le vittorie con un alto numero di gol di scarto è bilanciata da una sottostima per le vittorie con pochi gol di scarto.

Andando a simulare le distribuzioni per altri campionati, si trovano degli errori.

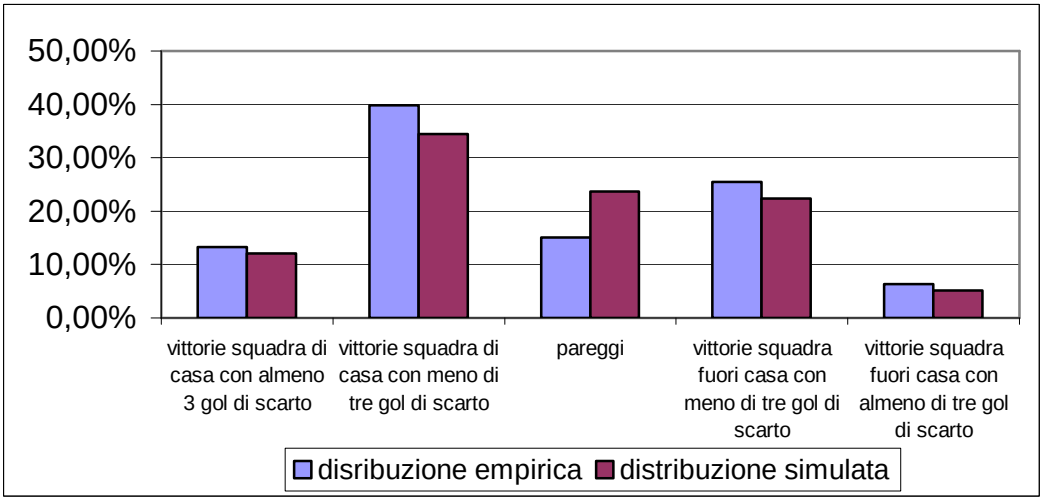
Studiamo infine il caso della Scozia, in cui Ro ha valore negativo.

Tabella 65: probabilità percentuale dei risultati nel campionato scozzese per la distribuzione empirica e per la distribuzione simulata tramite modello

esito	Gol di scarto	Distribuzione empirica	Distribuzione simulata
Vittoria squadra di casa	Tre o più	13,29%	12,03%
Vittoria squadra di casa	Meno di tre	39,88%	34,40%
pareggi	Nessuno	15,03%	23,67%
Vittoria squadra fuori casa	Meno di tre	25,43%	22,35%
Vittoria squadra fuori casa	Tre o più	6,36%	5,09%

In questo ultimo caso le probabilità simulate tramite il modello bivariato sono diverse da quelle registrate empiricamente. Si osservi in particolare che la probabilità di pareggio è troppo alta rispetto a quella empirica e porta ad una sottostima per gli altri quattro eventi.

Grafico 100: probabilità percentuale dei risultati nel campionato scozzese per la distribuzione empirica e per la distribuzione simulata tramite modello



La sovrastima per la probabilità pareggio non si concilia con il valore negativo assunto da Ro in questa simulazione. Questo aspetto avrebbe dovuto portare viceversa a un alto scarto nel numero di gol segnati dalla due squadre. Le differenze tra la distribuzione empirica a e quella simulata sono molto evidenti.

Conclusioni

Nella prima parte della tesi abbiamo sperimentato che, trasformando una variabile Gaussiana, si ottiene un modello caratterizzato da una sovrastima nella frequenza percentuale per l'evento modale e da una coda troppo pesante. Per questa ragione abbiamo scartato il modello Normale per la previsione del numero di gol segnato da una squadra di calcio.

Nonostante questo, si è voluto testare il modello originato da una trasformazione di una Normale Bivariata per la previsione del risultato di una partita di calcio.

Introducendo il modello bivariato non si riescono sempre ad ottenere i risultati sperati: il modello non riesce a prevedere con sufficiente precisione il risultato di una partita di calcio.

Questi errori previsivi tuttavia, presenti in due esempi su tre, potrebbero essere causati dalla errata forma assunta dalle distribuzioni a livello marginale piuttosto che dalla incapacità del modello di descrivere la dipendenza tra i due fenomeni. Non possiamo, quindi, escludere l'ipotesi che il modello bivariato possa descrivere la dipendenza tra i gol segnati da due squadre in una singola partita di calcio.

Per questa ragione un ulteriore sviluppo della ricerca porterebbe all'utilizzo di questo modello per l'analisi congiunta e l'utilizzo di altri modelli, in particolare il modello Weibull, per le distribuzioni marginali.

Conclusioni

La tesi nasce dalla lettura dell'articolo scientifico Modelling Association Football Scorse and Inefficiencies in the Football Betting Market".
Per questa ragione, il primo modello utilizzato è il modello di Poisson. Esso si dimostra abbastanza affidabile nel prevedere i gol segnati da una squadra di calcio; inoltre, ha il vantaggio che può essere utilizzato senza necessità di trasformazioni. Altro aspetto positivo è costituito dal fatto che, essendo stato già utilizzato per gli scopi esplorati nella tesi, esistono dei metodi per adattare la distribuzione dei gol di una squadra non solo rispetto alle proprie capacità offensive, ma anche rispetto alle capacità difensive della squadra avversaria e rispetto al fattore campo, uno di questi metodi viene presentato nel capitolo relativo al modello Poisson. Inoltre, sono state create delle funzioni correttive in grado di tenere in considerazione la dipendenza nelle reti segnate dalle due squadre in campo.
Un ulteriore aspetto positivo di questo modello è il metodo di stima: la stima del parametro Lambda è pari, infatti, alla media aritmetica delle determinazioni della variabile di partenza.
Lo svantaggio principale di questo modello viene dall'unico parametro caratteristico che impedisce alla varianza di essere più o meno ampia del valore atteso.
Il secondo modello nasce dalla speranza che le caratteristiche della distribuzione statistica più utilizzata al mondo, la Normale, potessero permettere una buona precisione nella previsione del fenomeno studiato. Prima di utilizzare questa distribuzione, si sono rese necessarie delle trasformazioni; inoltre ci si augurava che un buon adattamento marginale fosse accompagnato da un buon adattamento a livello bivariato in modo tale da potere descrivere il risultato di una partita di calcio tramite il modello bivariato.
Queste ipotesi non trovano conferma negli esperimenti: il modello Normale non si adatta bene al fenomeno studiato.

Tuttavia rimane in piedi l'ipotesi legata al modello bivariato in quanto si potrebbe utilizzare questo modello per la descrizione della dipendenza tra i gol della squadra 1 e i gol della squadra 2 ed un altro modello per le distribuzioni marginali.
Il modello bivariato, introdotto nell'ultima parte della tesi, completa l'analisi delle distribuzioni marginali tramite un modello che permetta la descrizione del punteggio di una partita di calcio. L'adattamento ai dati non è buono. Questo cattivo adattamento potrebbe essere però dovuto alle distribuzioni marginali.
Il terzo modello è il modello Gamma. In questo caso è necessaria una semplice discretivizzazione della distribuzione. I risultati in questo caso non sono soddisfacenti e suggeriscono di non utilizzare questo modello per la descrizione del fenomeno studiato.
Il modello Weibull, discretivizzazione di una variabile avente distribuzione di Weibull, porta a dei risultati sbalorditivi in quanto riesce a descrivere in modo molto preciso i gol segnati da una squadra di calcio. Tale precisione è addirittura sorprendente nel caso si abbia a disposizione un numero elevato di dati.
Prima di proseguire l'analisi, è opportuno sottolineare però che, sebbene il modello Weibull risulti più preciso di quello di Poisson, i vantaggi di utilizzare un modello di Poisson sono molti: innanzitutto, trattandosi di una variabile che assume valori interi positivi, non necessita di alcuna trasformazione per descrivere il fenomeno di interesse, contrariamente alla variabile ottenuta dalla variabile weibulliana; in secondo luogo, il modello di Poisson ha un solo parametro mentre quello di Weibull necessita di due parametri; infine, le stime per il primo modello sono stime di massima verosimiglianza ottenute senza l'ausilio di metodi numerici. Al contrario, le stime di massima verosimiglianza per il secondo modello sono ottenute tramite il metodo numerico di Newton-Raphson. Questo aspetto fa sì che all'aumentare dei dati a disposizione aumenti l'onerosità dei calcoli per determinare le stime. La precisione di tale modello dipende dalla numerosità campionaria, il che costituisce un indubbio svantaggio nell'utilizzo del modello di Weibull.
Gli sviluppi di ricerca successivi si baseranno su due passi successivi. In primo luogo, è necessario un metodo che permetta di tenere conto nella distribuzione del modello Weibull delle capacità in campo. Per quanto riguarda il parametro di posizione, l'idea è di sviluppare un metodo simile a quello utilizzato dal modello Poisson. In altre parole, il parametro di posizione per la distribuzione dei gol

segnati da una squadra dovrà necessariamente tenere in considerazione le capacità offensive della squadra stessa, le capacità difensive della squadra avversaria ed il fattore campo. Per quanto riguarda il parametro di variabilità non ci sono modelli dai quali prendere spunto. Tuttavia, il ragionamento potrebbe essere analogo a quello fatto per il parametro di posizione. In altri termini, il parametro di variabilità per la distribuzione dei gol segnati da una squadra dipenderà dalla variabilità delle capacità offensive della squadra stessa, dalla variabilità delle capacità difensive della squadra avversaria e della variabilità nel riuscire a sfruttare il fattore campo da parte della squadra di casa.

In secondo luogo, potrebbe rendersi necessaria una correzione per la dipendenza tra i gol segnati dalla squadra di casa e i gol segnati dalla squadra in trasferta, analoga a quella introdotta nel modello Poisson.

Bibliografia

Il riferimento principale è stato l'articolo:
[1] "Modelling Association Football Scores and Inefficiencies in Football Betting Market" di Mark J. Dixon e Stuart G. Coles – Lanchester University, 1997.

È stato utilizzato il testo didattico:
[2] "Laboratorio di Statistica con R", di Stefano Iacus e Guido Masarotto, McGraw Hill, 2003.

Sono state utilizzate le dispense didattiche:
[3] "Lucidi delle lezioni di inferenza statistica I (a.a. 2006/07)" di Guido Masarotto, 2007;

[4] "Materiale didattico per il corso di Inferenza Statistica I" di Alessandra Dalla Valle e Carlo Gaetan, 2001;

[5] "Materiale didattico per i laboratori di Modelli Statistici I" di Monica Chiogna, Alessandra Salvan e Nicola Sartori, 2006.

Sono stati consultati i siti web:
[6] wikipedia.org;

[7] mathworld.wolfram.com.

Le elaborazioni sono state ottenute tramite i software:
R 2.6.2.;

Microsoft Excel.

Ringraziamenti

Dopo questa lunga tesi, rimane da dire solamente che senza l'aiuto, consapevole o no, di tante persone non avrei mai potuto portare a termine questo progetto.
Il primo e più grande grazie va alla mia famiglia: a mio padre, infaticabile correttore di bozze, a mia madre, che mi ha trasmesso l'amore per i numeri, e alla mia sorellina, alla quale nonostante tutto voglio bene.
Un enorme ringraziamento al professor Coles che sin dall'inizio, pur non conoscendomi, si è dimostrato interessato e disponibile a chiarire ogni mio dubbio.
Ringrazio tutti i docenti che in questi tre anni, chi più e chi meno, hanno saputo farmi apprezzare la Statistica; tra essi un sincero grazie al professor Scarpa che, per primo e più di ogni altro, ha saputo trasmettermi l'amore per questa materia.
Un grazie a tutti i miei parenti: alla nonna Maria e alla nonna Gemma. Un grazie anche a nonno Taddeo e nonno Mauro, certamente orgogliosi da lassù.
Grazie ai miei amici, i Popguys, sempre presenti anche se mai tutti contemporaneamente; grazie ai Povoboys, molto più che semplici compagni di squadra; grazie agli amici dell'UdS di ieri e della ReDS di oggi; grazie ai miei amici di Cavazzale con cui ho trascorso i più bei momenti della mia adolescenza; grazie ai colleghi di SPS, con i quali ho vissuto tutti i miei momenti padovani e infine grazie a tutte le persone con le quali ho condiviso qualcosa in questi tre anni.
Tra tutte le persone care un grazie di cuore a Bracco, fidato consigliere per tante questioni e per tutti i problemi, e a Gio, compagno di viaggio e di mille avventure.